Queen Margaret University

EDINBURGH

ENHANCING KINEMATIC SHOULDER FUNCTION EVALUATION THROUGH A VALID, SIMPLE AND CLINICALLY APPLICABLE SCORE

CLAUDE PICHONNAZ

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

QUEEN MARGARET UNIVERSITY

2019

Abstract

Introduction

Controversies surrounding Patient-Reported Outcome Measures (PROMs) and the cumbersome-nature of movement analysis-based (MAB) methods for shoulder function evaluation make the exploration of alternatives needed. Research aimed at the simplification of MAB outcome measures had demonstrated previously that the B-B Score, which relies on two movements only, was valid for out-of-laboratory evaluations of shoulder function. Nevertheless, further investigations were needed to optimise testing procedures, test the B-B Score's capability of acquisition using a user-friendly device, and critically evaluate its measurement properties in comparison to current methods.

Objective

The aim of this thesis was to develop and assess the simplest possible MAB shoulder function scoring procedure for clinical measurement.

Methods

The research included four steps: 1) Optimisation of the B-B Score testing procedure (Phase 1 study [data-driven]), 2) Comparison of measurements using a smartphone or an inertial sensor system (Phase 2 study [data-driven]), 3) Validation in frequently-occurring pathologies (rotator cuff conditions, instability, fracture, capsulitis) (Phase 3 study [data-driven]), 4) Benchmarking of the new approach with concurrent MAB outcome measures and PROMs (literature review).

Results

Amongst the tested methods, the B-B score was optimised by using the mean of three replicates in the computation of the range of accelerations by angular velocities. The comparison of easily-used smartphone and reference device showed non-significant differences and excellent relationships between measurements (Intraclass Correlation Coefficient [ICC=0.97]). The smartphone's B-B Score intra-rater and inter-rater reliability was excellent (ICC=0.92), but limits of agreement could reach up to ±19.4%. The score was responsive (area under the curve [AUC≥0.70]) and demonstrated excellent discriminative power between patients and controls (AUC≥0.90), except for shoulder instability (AUC=0.67). The correlations with PROMs were moderate to high. The benchmarking established that the measurement properties of the B-B Score

compared equivalently with those of PROMs and MAB outcome measures, except for shoulder instability.

Conclusion

Shoulder function can be efficiently evaluated using a simple scoring procedure performed with a smartphone, which facilitates its objective assessment. Further research is needed to understand how best to reduce the effects of variability associated with single measurements in order to optimise clinical applicability and to explore the B-B Score's properties in other situations requiring functional assessments of the shoulder.

.

Keywords: shoulder, shoulder function; outcome assessment; validation studies, reliability and validity; inertial sensors; smartphone sensors; body-worn sensors; kinematics; sensitivity and specificity.

Acknowledgements

I would like to express my deepest gratitude to all the participants who voluntarily participated in this research. Their willingness to contribute to the improvement of patients' care, beyond their own interest and for the good of all, was a strong motivation to invest in a research that would be useful to them, hopefully.

I would like to offer my especial thanks to Prof. Nigel Gleeson, my primary research supervisor. Prof. Gleeson is a professor of exercise and rehabilitation sciences in the School of Health Sciences, Queen Margaret University, Edinburgh. He has guided me on a very long path over the last nine years, from the beginning of the MSc dissertation to the end of the PhD thesis, and always with rigor and humanity. Beyond his indisputable competencies in the fields of research and rehabilitation, I have appreciated his friendly manner to incite me to give the best of myself, with sympathetic consideration, confidence and even humour. I am also thankful to my co-supervisor Dr. Fiona Coutts, Dean Faculty of Health Sciences, for her efficient and indispensable attention to the PhD progress and her contribution to the smooth running of all stages of this PhD thesis.

The completion of this thesis would not have been possible without the support of my two employers, HESAV (Haute Ecole de Santé Vaud (HESAV), HES-SO, (University of Applied Sciences and Arts Western Switzerland, Lausanne, Switzerland) and the DAL-CHUV (Service of Orthopaedics and Traumatology, Department of Musculoskeletal Medicine, University Hospital of Lausanne, Lausanne, Switzerland). I am grateful for the trust they have placed in me.

I am also grateful to the Swiss National Science Foundation (SNF) for funding the main research of this project, through the DORE found for the development of applied research in universities of applied sciences (DORE 135061).

The completion of this thesis would also not have been possible without the implementation of collaborations with several institutions, involving HESAV, Queen Margaret University, the DAL-CHUV and the Laboratory of Movement Analysis and Measurement of the Swiss Institute of Technology (LMAM-EPFL). There are many people working in these institutions that I want to thank personally for their active support:

Prof. Kamiar Aminian, Head of the Laboratory of Movement Analysis and Measurement of the Swiss Institute of Technology (LMAM-EPFL), for engaging his laboratory

resources on this project and providing his focused and experienced support whenever needed. Special thanks to Cyntia Duc, biomedical engineer, for her generous involvement, her willingness to share and her flawless dependability. Her contribution went far beyond what could be expected of her.

Prof. Alain Farron, orthopaedic surgeon, Head of the DAL-CHUV, and Prof. Brigitte Jolles, orthopaedic surgeon, Director of the operational unit of the DAL-CHUV, for their support, scientific counselling and contribution to the research implementation in the clinical setting.

Jean Lambert, Head of the Physiotherapy Service of the DAL-CHUV and Valérie Zoll, Head of the Physiotherapy Unit of the Orthopaedic Hospital-CHUV, for the trust they have placed in me and their willingness to promote the quality of physiotherapy through research. My warmest thanks also to my physiotherapist colleagues Anne Rothenbacher and Barbara Balmelli, for their dependability and their care in collecting good quality data.

Dr. Estelle Lécureux, statistician at the Medical Direction of the CHUV and at Stat'Elite, for her invaluable ability to understand clinical issues, and to make statistical approaches relevant and understandable for clinicians in return.

Last but not least, Céline Ancey, Pierre Balthazard, Jean-Philippe Bassin, Guillaume Christe, Hervé Jaccard and Noémie Sauvage Pasche, by alphabetical order, my estimated colleagues at HESAV who, each at a different stage of progress of the project, actively contributed to the study management, coordination, implementation and/or data collection. I cannot detail their respective contributions here, but they can be assured that I am grateful to each and every one of them. Beyond their impeccable professional involvement, it was a pleasure to share the friendly and constructive relationship that we all had together, by being serious without taking ourselves too seriously.

Dedication

The completion of this PhD was a long process, which was conducted in parallel to all my daily life commitments, above all as a husband and father. It was a mind-opening process, but I hope that it was not a process that distracted me from you, my family, who are the people I care about most, because you are more important than anything in the world to me.

That's why I want to dedicate this thesis to my wife Mireille who has been at my side for over 20 years, with all the love that this implies, and to Laure and Florine who were my little children at the beginning of the thesis and who are now young adults who make me proud.

Table of Contents

List of figures

List of tables

List of peer-reviewed articles and conference papers associated with and underpinning aspects of the work within this thesis

1. Articles in international peer-reviewed journals

PICHONNAZ, C., AMINIAN, K., ANCEY, C., JACCARD, H., LÉCUREUX, E., DUC, C., FARRON, A., JOLLES, B. M. & GLEESON, N. 2017. Heightened clinical utility of smartphone versus body-worn inertial system for shoulder function B-B Score. *PLOS ONE,* 12**,** e0174365.

PICHONNAZ, C., DUC, C., GLEESON, N., ANCEY, C., JACCARD, H., LECUREUX, E., et al. 2015. Measurement Properties of the Smartphone-Based B-B Score in Current Shoulder Pathologies. *Sensors (Basel),* 15**,** 26801-26817.

PICHONNAZ, C., LECUREUX, E., BASSIN, J. P., DUC, C., FARRON, A., AMINIAN, K., et al. 2015. Enhancing clinically-relevant shoulder function assessment using only essential movements. *Physiol Meas,* 36**,** 547-60.
(Note: article related to the MSc dissertation, closely related but not part of the PhD)

2. Article to be submitted in an international peer-reviewed journal

PICHONNAZ, C., BALTHAZARD, P. ANCEY, C., FARRON, A., JOLLES, B.M., COUTTS, F., GLEESON, N. Comparison of measurement properties of current questionnaires and movement analysis-based scores for shoulder function evaluation

3. Congress presentations

PICHONNAZ, C., ANCEY, C., DUC, C., LÉCUREUX, E., JACCARD, H., JOLLES, B., FARRON, A., AMINIAN, K., KANOUN, K. & GLEESON, N. Hands Up smartphone application for a quick and valid measurement of shoulder function. Mass Challenge, Personalised Healthcare, 3 October 2017 Renens.

PICHONNAZ C., DUC C., LÉCUREUX E., JACCARD H., ANCEY C., BALMELLI B., BOVEY A., AMINIAN K., FARRON A., JOLLES B.M. & N., G. Mesurer la fonction de l'épaule avec un Smartphone: une solution d'avenir? Colloque de recherche Hôpitaux Universitaires de Genève, 9 February 2016 Genève.

PICHONNAZ, C., DUC, C., ANCEY, C., JACCARD, H., LÉCUREUX, E., AMINIAN, K., FARRON A., JOLLES, B.M., GLEESON, N. Comparison of a dedicated body-worn inertial system and a smartphone for shoulder function and arm elevation evaluation. WCPT Congress, World Confederation for Physical Therapy, 2015, 1-4 May Singapore.

PICHONNAZ, C., DUC, C., LÉCUREUX, E., JACCARD , H., ANCEY, C., AMINIAN, K., et al. Simplification and validation of a kinematic shoulder function test using body-worn sensors. International Symposium: 3D Analysis of Human Movement 3D-AHM, July 14 - 17 2014 Lausanne.

JACCARD, H., PICHONNAZ, C., DUC, C., LÉCUREUX, E., AMINIAN, K., JOLLES-HAEBERLI, B. M., et al. Validation d'une application Smartphone pour l'évaluation de la fonction et de l'amplitude d'élévation de l'épaule. Physiocongress, 13 - 14 June, 2014 Bern.

BALMELLI, B., PICHONNAZ, C., LÉCUREUX, E., JACCARD , H., ANCEY, C., BASSIN, J.-P., et al. La « Subjective Shoulder Value » : un outil simple et valide pour évaluer la fonction de l'épaule. Physiocongress, 13 - 14 June, 2014 Bern.

DUC, C. & PICHONNAZ, C. Kinematic shoulder function evaluation: Toward a valid and simple kinematic shoulder function evaluation? 11th annual Research Day in Translational Orthopaedics, 14 June, 2013 Lausanne.

PICHONNAZ, C. & DUC, C. Validation de tests de la fonction de l'épaule avec un Smartphone. HESAV fait ses 400 coups: innovation et santé, 5 November 2012 Lausanne.

BASSIN, J.-P., PICHONNAZ, C., MARTIN, E., CHRISTE, G., DUC, C., DJAHANGIRI, A., et al. Fiabilité d'un score fonctionnel base sur l'analyse de deux mouvements fondamentaux de l'épaule. Physiocongress, 10 - 11 May 2012 Genève..

PICHONNAZ, C. Validation of a kinematic functional shoulder score including only essential movements. Journée scientifique du Réseau d'Etudes appliquées des Pratiques de Santé, de Réadaptation/Réinsertion (RéSaR), June 12 2012 Lausanne.

4. Published congress proceedings

PICHONNAZ, C., DUC, C., JACCARD, H., ANCEY, C., LÉCUREUX, E., AMINIAN, K., et al. 2015. Comparison of a dedicated body-worn inertial system and a smartphone for shoulder function and arm elevation evaluation. *Physiotherapy,* 101**,** e1205-e1206.

PICHONNAZ, C., DUC, C., JACCARD, H., ANCEY, C., LÉCUREUX, E., AMINIAN, K., et al. 2015. Validity of a straightforward shoulder function evaluation method using a smartphone. *Physiotherapy,* 101**,** e1206.

PICHONNAZ, C., HUU, F. N., PALLOT, A., BOLLA, B., MORICHON, A. & ANDRÉ-VERT, J. 2015. Pratique professionnelle et appareil locomoteur: le membre supérieur. *Kinésithérapie, la Revue,* 15**,** 28-34.

BALMELLI, B., PICHONNAZ, C., LÉCUREUX, E., JACCARD, H., ANCEY, C., BASSIN, J.-P., et al. 2014. La Subjective Shoulder Value: un outil simple et valide pour évaluer la fonction de l'épaule. *Kinesitherapie, la revue,* 150**,** 16.

JACCARD, H., PICHONNAZ, C., DUC, C., LÉCUREUX, E., ANCEY, C., BASSIN, J.-P., et al. 2014. Validation d'une application smartphone pour l'évaluation de la fonction et de l'amplitude d'élévation de l'épaule. *Kinesitherapie, la revue,* 150**,** 17-18.

BASSIN, J.-P., PICHONNAZ, C., MARTIN, E., CHRISTE, G., DUC, C., DJAHANGIRI, A., et al. 2012. Fiabilité d'un score fonctionnel base sur l'analyse de deux mouvements fondamentaux de l'épaule. *Kinésithérapie, la revue* 27.

5. Posters

PICHONNAZ, C., ANCEY, C., DUC, C., LÉCUREUX, E., JACCARD, H., JOLLES, B., FARRON, A., AMINIAN, K., KANOUN, K. & GLEESON, N. 2016, 6 December. Application HandsUp pour la mesure du B-B Score. Poster presented at:. *Rencontre Ingénierie-Santé : "Activité physique et sport. Médicaments de demain?".* Lausanne

PICHONNAZ, C., DUC, C., JACCARD, H., ANCEY, C., LÉCUREUX, E., AMINIAN, K., FARRON A., JOLLES, B.M., GLEESON, N. 2015, 1-4 May. Validity of a straightforward shoulder function evaluation method using a smartphone. Poster presented at: *WCPT Congress, World Confederation for Physical Therapy.* Singapore

JACCARD, H., PICHONNAZ, C., DUC, C., LÉCUREUX, E., ANCEY, C., BASSIN, J.-P., et al. 2013, 22-23 November. Validation d'une application smartphone pour l'évaluation de la fonction et de l'amplitude d'élévation de l'épaule. Poster presented at: *Symposium Romand de Physiothérapie.* Lausanne. [2nd Best poster award]

PICHONNAZ, C., BASSIN, J.-P., DUC, C., CHRISTE, G., JACCARD, H., HAEBERLI-JOLLES, B., et al. 2012, 5 November. Un test simple de la fonction de l'épaule. Poster presented at: *HESAV fait ses 400 coups: innovation et santé.* Lausanne.

PICHONNAZ, C., DUC, C., BASSIN, J. P., SAUVAGE PASCHE, N., DJAHANGIRI, A., JOLLES, B. M., et al. 2012, 27-29 June. Validation of a Smartphone application for shoulder elevation. Poster presented at: *Swiss Orthopaedics and Traumatology Society Congress.* Basel.

PICHONNAZ, C., BASSIN, J.-P., CHRISTE, G., DUC, C., DJAHANGIRI, A. & FARRON, A. 2011, 14-17 September. Reliability of a kinematic functional shoulder score including only two movements: a pilot study. Poster presented at: *23rd Congress of the European Society for Surgery of the shoulder and the Elbow SECEC-ESSSE.* Lyon.

6. Teaching

Use of PhD-related examples in the lecture on the properties of measurement tools, 1st year entry-level training, HESAV, Lausanne 2014 -2018.

Use of the HandsUp application in the workshop on the properties of measurement tools, 1st year entry-level training, HESAV, Lausanne 2014 – 2018.

Presentation of PhD project and results in the international exchange lectures to Bouvé College of Health Sciences, students, Northeastern University, Boston, HESAV, Lausanne 2015, 2017, 2018.

7. Diffusion on the clinical setting workplace

Presentation of the PhD project, continuing education symposium of musculoskeletal physiotherapy Department of Physical Therapy, 2012.

Presentation of the PhD results, continuing education symposium of musculoskeletal physiotherapy Department of Physical Therapy, University Hospital of Lausanne, Lausanne, 2016.

Presentation of the PhD results, medical continuing education symposium of Orthopaedics and Traumatology Department, University Hospital of Lausanne, Lausanne, 2016.

Abbreviations

ANOVA: Analysis of Variance

ASES: Shoulder Score American Shoulder And Elbow Surgeons Shoulder Score

AUC: Area Under the Curve

B&A graph: Bland and Altman graph

B-B Score: hand to the Back and hand to the ceiling as to change a Bulb Score

BMI: Body Mass Index

CHUV: Centre Hospitalier Universitaire de Lausanne

CI: Confidence Interval

COSMIN: COnsensus-based Standards for the selection of health Measurement Instruments

DAL: Département de l'Appareil Locomoteur

DASH: Disabilities of the Arm, Shoulder and Hand Questionnaire

Deg: Degree

EMG: ElectroMyoGraphy

EPFL: Ecole Polytechnique Fédérale de Lausanne

EQ-5D: European Quality of Life in 5 Dimensions scale

ES: effect size

GRS: Global Rating Scale

HES-SO: University of Applied Sciences of Western Switzerland

HESAV: Haute Ecole de Santé Vaud

ICC Intra-class Correlation Coefficient

ICF: International Classification of Functioning, Disability and Health

IQR: InterQuartile Range

IRT: Item Response Theory

IMU: inertial measurement unit

LMAM: Laboratory of Movement Analysis and Measurement

LoA: limits of agreement

MAB: Movement analysis-based

MCID: Minimally Clinically Important Difference

MCII: Minimally Clinically Important Improvement

MDC: Minimal Detectable Change

ME: Measurement Error

P Score: Power Score

PASS: Patient Acceptable Symptom State

PROM: Patient-reported Outcome Measure

QMU: Queen Margaret University

QuickDASH: Quick Disabilities of the Arm, Shoulder and Hand Questionnaire

ROC curve: Receiver Operating Characteristic curve

ROM: Range Of Motion

SD: Standard Deviation

SEM: Standard Error of Measurement

SNF: Swiss National Science Fundation

SPADI: Shoulder Pain and Disability Index

SST: Simple shoulder test

SRM: Standardised Response Mean

UCLA: Shoulder rating scale University of California, Los Angeles Shoulder rating scale

VAS: Visual Analogue Scale

WOSI: Western Ontario Shoulder Instability Index

WORC: Western Ontario Rotator Cuff Index

Yr.: Years

CHAPTER ONE

INTRODUCTION

1.1. Introduction

1.1.1. Epidemiology of shoulder conditions

Shoulder problems are a frequent cause of pain and disability. Prevalence of shoulder problems ranges in-between 7% and 35% in the general population (Yamamoto et al., 2010; Green et al., 2003), which represents the second most frequently affected musculoskeletal area in the body (Picavet and Schouten, 2003). This results in substantial disability at work or in daily living activities and impaired quality of life (Green et al., 2003). The quality of tools for the evaluation of shoulder function is of primary interest to adequately address the problems of this large population and therefore limit the impact of shoulder pathologies on patients and society.

1.1.1.1. Impact of main shoulder conditions on function

A large variety of conditions may lead to shoulder function alteration. However, each one of them has to be considered separately for evaluation, as each impairs the function of the shoulder specifically. Pain, stiffness or weakness might for example be present to a variable degree according to the pathology. Thus, the items of a patient-rated outcome measure (PROM) must be adequate to target the specific shoulder function alterations induced by a condition, and a kinematic outcome measure must account for the fact that each pathology affects the movement in a specific way. Therefore, measurement properties of a shoulder function measure are valid only in the population in which they were tested (Robertson et al., 2017; Collins and Roos, 2016; Riddle and Stratford, 2013).

In addition to issues related to shoulder conditions, the evaluation of surgically or conservatively treated populations should be differentiated for the same reason. The size of the conservatively treated population is much larger than that of the surgically treated one. Overall, only one in every 10 patients presenting with shoulder pain requires surgery (Colvin et al., 2012).

Patients with rotator cuff conditions, proximal humerus fractures, adhesive capsulitis, and shoulder instabilities are frequently encountered in shoulder consultations (van der Windt et al., 1996; Yamamoto et al., 2010; van der Windt et al., 1995; Court-Brown

and Caesar, 2006; Liavaag et al., 2011; Owens et al., 2007). It is thus essential to have efficient tools to evaluate shoulder function as a priority for these conditions, of which the main characteristics are developed hereafter.

1.1.1.1.1. Characteristics of rotator cuff conditions

Conditions associated with the shoulder's rotator cuff musculature are the most common source of shoulder pain (65%). The notion of a rotator cuff condition is non-specific, as the pain may come from several causes that are difficult to differentiate in practice, like rotator cuff tendinopathy, rotator cuff tears, subacromial impingement or subacromial bursitis (Mitchell et al., 2005). Rotator cuff tendinitis affects 29% of patients presenting with shoulder pain in general practice (van der Windt et al., 1995). Rotator cuff tear prevalence is also very high and is strongly related to age. Tears are present in 2.5% of the general population in their 30's, 25% in their 60's, and 50% in their 80's (Yamamoto et al., 2010). A painful arc during arm elevation is typical of rotator cuff conditions (O'Kane and Toresdahl, 2014). However, clinical presentation of rotator cuff conditions varies considerably. Range of motion (ROM) limitations may or may not be observed, and tears may remain asymptomatic despite the anatomical lesions (Yamaguchi et al., 2006; Yamamoto et al., 2010; Moosmayer et al., 2009).

1.1.1.1.2. Characteristics of adhesive capsulitis

Adhesive capsulitis, also named frozen shoulder, represents the second most prevalent cause of shoulder pain (22%) (Yamamoto et al., 2010). It is an idiopathic disease of the joint capsule causing mainly pain and stiffness (Mitchell et al., 2005). The adhesive capsulitis is usually considered a 12- to 18-month self-limiting process, but mild symptoms may persist longer (Kelley et al., 2013).

1.1.1.1.3. Characteristics of proximal humerus fractures

Proximal humeral fractures are also common, as they account for 6% of all adult fractures (Court-Brown and Caesar, 2006). The incidence of this type of fracture in Western countries is growing, due to the increasing age of the population. The movement is altered during the rehabilitation phase by pain, stiffness, and loss of

strength. The recovery at one year is generally good and equivalent for the conservative and the surgical approach (Handoll et al., 2012).

1.1.1.1.4. Characteristics of shoulder instability

Shoulder instability is also a frequent cause of medical consultation in younger populations. It is characterised by the inability to maintain the humeral head in the glenoid fossa of the scapula, so that the humerus slides partially or completely out of its socket. The shoulder instability's incidence rate reaches 56.3 per 100 000 person-years in the general population, but 2.8% in a physically active young population (Liavaag et al., 2011; Owens et al., 2007). Instability is problematic because it frequently leads to recurrent shoulder dislocation, apprehension, and loss of quality of life (Handoll et al., 2004; Rouleau et al., 2010). The movement is altered in the less stable positions of the glenohumeral joint. Typically, the patient experiences apprehension at the end of ROM, while undertaking combined movements, but can perform activities without problem in stable glenohumeral joint positions.

1.1.2. Evaluation of shoulder function

1.1.2.1. Patient-reported outcome measures

Shoulder function is most frequently evaluated using PROMs questionnaires. Up to thirty-nine shoulder function evaluation tools have been audited within reviews, but most have not undergone a full validation process that would be expected to underpin good quality research (Kirkley et al., 2003; Oh et al., 2009; Huang et al., 2015; Harvie et al., 2005). Thus, the measurement of the shoulder functional outcome using PROMs remains a contemporary and controversial issue. Consequently, no questionnaire has been widely recognised as a standard (Fayad et al., 2005; Oh et al., 2009; Placzek et al., 2004; Roy et al., 2009). The use of a large variety of outcome measurements tools and assessment tools in research limits the development of evidence about treatments of shoulder conditions, as the results are hardly comparable between studies that rely on different PROMs (Green et al., 2003; Harvie et al., 2005; Makhni et al., 2015; Page et al., 2015).

Clinical questionnaires have essentially the advantages of handiness and low cost. Conversely, they present intrinsic limitations related to language and cultural issues,

respondents' interpretations and content validity (Ragab, 2003; Olley and Carr, 2008). The validation of questionnaires' translations into various languages is a time-consuming and cumbersome process. Moreover, the delineation between objective and subjective aspects of evaluation is sometimes ambiguous in questionnaire-based assessment. This is all the more important as objective and subjective approaches generally produce different results (Krueger et al., 2011; Moustgaard et al., 2014).

Despite the questionnaires' limitations, PROMs represent the current standard in clinical shoulder function evaluation. Actually, as there has rarely been a direct critical comparison of PROMs and alternative measurement methods (e.g movement analysis, physical testing or observation), no concurrent measurement method has demonstrated its superiority or inferiority over PROMs to date. In the current context, the development of a new questionnaire based from its conception on recognised methods would probably have limited added value, as it would face the same difficulties as its predecessors in overcoming methodological pitfalls in order for it to be considered as a standard. There is therefore a need to investigate alternatives to provide clinicians and researchers with well-recognised and convenient measurement tools that would not present the same drawbacks as PROMs. This could ideally lead to the development of new clinimetrically-relevant measurement tools with the capability for delivery within a clinical environment. The role of these innovative approaches should also be explored to understand if they mainly concur with or complement the results of current approaches.

1.1.2.2. Movement analysis-based (MAB) assessment

Computerised movement analysis produces a purely objective outcome, and could potentially be recognised as a standard for shoulder function evaluation due to its accuracy and precision (Pandyan, 2002). It could also overcome limitations related to language and cultural issues, respondent interpretations and content validity associated with questionnaires (Ragab, 2003; Olley and Carr, 2008; Kirkley et al., 2003). It has thus been largely used in research studies aiming at the characterisation and evaluation of shoulder motion.

Although three-dimensional laboratory motion analysis systems have assumed a growing importance in research, their application in clinical settings has remained limited to date. Most motion laboratory analysis studies have mainly addressed the development of innovative measurement' models or have investigated differences between healthy and pathological participants' groups. This led to a better understanding of shoulder movement and its alterations, but has rarely resulted in the development of measurement tools that could be used in clinical research, let alone in clinical practice. No laboratory-based research had proposed a MAB outcome measure for shoulder function that could be possibly used to monitor patient's clinical change in routine practice, to the best of the author's knowledge.

Most laboratory-based approaches for shoulder movement analysis rely on infrared cameras, ultrasounds systems, electromagnetic systems or electromyography (Coley et al., 2009). Constraints of location, time, complexity and costs of laboratory measurement restrict its use in clinical practice, and research (Aminian and Najafi, 2004; Clark et al., 2017). Therefore, embedded systems, like inertial measurement units (IMU) using gyroscopes and accelerometers have also been developed for shoulder evaluation, as their portability and practicality facilitates the procedures for measurement. Ambulatory systems may represent a well-balanced compromise between practicality and reliability. While they are highly correlated to laboratory measurements and display adequate accuracy, measurement completion is easier and application is not restricted to laboratory-based environments (Coley et al., 2007a).

Embedded sensors have been applied with promising results to measure arm and shoulder movement in various conditions (Luinge et al., 2007; Wong et al., 2007; Coley et al., 2008b; Coley et al., 2008a; Teece et al., 2008; Ludewig and Cook, 2000; Borstad and Ludewig, 2002; Rundquist et al., 2003; Rundquist and Ludewig, 2004; Rundquist and Ludewig, 2005; Ludewig et al., 2009; Ludewig and Reynolds, 2009; Duc et al., 2013; Duc et al., 2014). These studies demonstrated the potential of movement analysis based on body-worn sensors to characterise healthy and pathological shoulder movement. Thus, several research teams have proposed scoring methods that could potentially be used to evaluate shoulder function in clinical settings (Korver et al., 2014a; Coley et al., 2007a; Duc et al., 2014; Yang et al., 2014; Jolles et al., 2011; Pichonnaz et al., 2015c) (please see literature review Chapter five).

Nevertheless, despite the simplification of the measurement procedures provided by body-worn sensors, their use for shoulder function evaluation has remained limited in clinical settings. Several barriers, including for example access to the device, time constraints, familiarity with the technology, still hinder the widespread use of such devices among health professionals. Though apparently self-evident, the requirements for the routine application in clinical practice are very demanding as, in addition to measurement properties, time, practicability, user-friendliness and cost are of higher concern than when used for research' purposes.

Several of the existing scoring methods are based on Coley's work, who proposed a relatively simple shoulder function score based on three dimensional measurements of a power-related metric by accelerometer and gyroscopes (P score) (Coley et al., 2007a). The procedure relied on a sequence of seven functional movements based on the Simple Shoulder Test functional score (Lippitt, 1993). This approach demonstrated clinical relevance, as the score was clearly capable of discriminating healthy from pathological subjects, was correlated to clinical questionnaires and displayed adequate responsiveness after shoulder surgery (Coley et al., 2007a). However, the full test procedure needed around 20 minutes to perform, which precluded routine application in clinical settings.

The latter limitation of the P Score to shoulder function assessment was addressed in a QMU MSc dissertation project in physiotherapy that investigated whether it was possible to simplify Coley's testing procedure (Pichonnaz, 2010; Pichonnaz et al., 2015c) (Appendix I). This preliminary work aimed at selecting only essential movements that should be performed during the measurement protocol, and this research ultimately acted as a forerunner for the research questions addressed within this PhD thesis. A simplified score was developed based on multivariate statistical approaches of principal component analysis and multiple regressions of P Score raw data at baseline and at 3, 6 and 12 months after surgery. Principal component analysis allowed identifying two main constituent dimensions: an "arm elevation" and an "arm rotation" dimensions. Therefore, simplified scoring systems were developed based on multiple regressions of two movements, representative of these dimensions, and focusing on their ability to predict the P Score. Several possibly relevant movement associations were investigated (hand to the back + reach back of head with hand; hand to the back + 90° abduction; hand to the back + touch opposite

shoulder with hand ; hand to the back + lift arm as if changing a bulb). The most efficient statistical model for a simplified score arising from the multiple regressions was found to be 16.71 + (0.32 x hand to the Back) + (0.45 x lift arm as if changing a Bulb). This two-movement combination was therefore selected as the best possible alternative to the P Score and named B-B Score (B-B Score meaning Back-Bulb Score). It was demonstrated that the testing procedure limited to only two essential movements instead of seven, did not induce any significant information loss ($R^2 > .97$).

The outcomes of this simplified scoring procedure were then compared to the P score outcomes in the same sample at baseline and at 3, 6 and 12 months following surgery. The mean results closely matched and the correlation between the simplified and the reference score was excellent regardless of stage of rehabilitation. The simplified score demonstrated measurement properties similar to those of the reference score for the study population and the responsiveness for both assessment approaches was comparable. Moreover, the discriminative power between patients and controls of the simplified approach was excellent with 97% sensitivity and 94% specificity, indicating that the score was able to detect the function loss in patients following rotator cuff surgery or shoulder arthroplasty.

All other things being equal, the main advantage of the simplified scoring system resides in its clinical practicality. Moreover, the simplified scoring procedure can easily be repeated, which can potentially contribute to increased reliability of measurement by taking the mean of several replications into consideration when calculating the score.

Concurrent studies investigated another two-movement combination including “arm to the back” and “arm behind the head” movements (Korver et al., 2014a; Korver et al., 2014b). These movements were selected because they represented motion tasks related to activities of daily living that are part of several standard clinical questionnaires. This score required less than 5 minutes to perform and demonstrated high intra- and inter-rater reliability, with intraclass coefficient of correlation (ICC) of 0.95 and 0.91, respectively. The diagnostic sensitivity was 98% and the specificity 81%. However, the relationship to shoulder function evaluation was limited, as correlations with the DASH (Disabilities of the Arm, Shoulder and Hand) questionnaire and SST (simple shoulder test) clinical score were weak (Pearson $r < 0.25$) (Lippitt, 1993; Hudak et al., 1996). As stated by the authors, this score’s outcome cannot thus

be considered to be representative of shoulder function as it has been conceptualised within these PROMs. So, these kinematic scores cannot be considered as a potential substitute to shoulder function PROMs.

Conversely to the scores developed by Korver et al., who used similar PROMs and inertial sensor system outcome measures for patients suffering shoulder disorders, the correlations of the B-B Score with current PROMs ranged from 0.51 to 0.77, indicating that the B-B Score had good criterion-based validity for shoulder function evaluation (Pichonnaz et al., 2015c). Despite these promising preliminary results, further research would nevertheless be required to establish extensively the measurement properties of the B-B Score. It would also be necessary to precisely standardise its measurement procedure, to determine healthy subjects' performance and to evaluate applicability to populations presenting with other shoulder conditions than rotator cuff surgery or arthroplasty surgery that were investigated in previous works (Coley et al., 2007a; Pichonnaz et al., 2015c).

Though the testing movements are kept to their simplest expression in the B-B Score, this simplicity might not prove sufficient for routine clinical application if a complex movement analysis device is needed for the score's completion. More research would therefore also be needed to investigate if the score can be usefully measured using an accessible and affordable device. Using a smartphone for evaluation purposes might contribute to meeting these requirements and facilitating the transfer of objective movement analysis-based functional outcome in current practice. This approach is conceivable nowadays because, like embedded measurement systems, most smartphones are fitted with built-in accelerometers and gyroscopes. If used in conjunction with a dedicated but as yet to be developed application, they could thus potentially be used for shoulder function analysis.

1.1.2.3. Smartphone applications for shoulder evaluation

The use of a smartphone for the B-B Score measurement might further improve the practicability of the evaluation procedure. In case the measurement properties are acceptable, smartphones may offer a cost-effective and straightforward clinical outcome measurement, provided that a simple measurement procedure is applied.

Cost and training might be considerably reduced and this could favour routine objective function measurement of the shoulder.

However, there are also limitations in the use of smartphones for scientific measurement. For instance, the precise features of the device are not fully disclosed due to commercial sensitivities. Furthermore, the smartphone results might possibly differ from inertial-based systems, as the sensors' features have not been specifically designed for scientific measurement. Users should also remain conscious that measurement properties might be device-dependent, because the characteristics will differ according to smartphone version and brand.

Smartphone-based evaluation in clinical conditions is thus valuable only provided that the measurement properties have been previously be verified to meet necessary clinical criteria - which was still to be completed for a possible smartphone version of the B-B Score. This is a prerequisite to any clinical implementation because important decisions are taken based on delivered clinical outcome, for example about treatment continuation, hospital stay or intervention needs (Roe et al., 2013; Michener, 2011). Considering these issues, the results on which the decision is based must previously have proven to be valid, responsive and reliable. Extensive verification studies of clinimetric utility would thus be needed before clinical implementation of a smartphone-based approach, whether it is in general or more specifically for shoulder function evaluation.

The exploration of the literature shows that smartphone applications are taking growing importance for patient evaluation, patient education or to assist health care professionals in their practice. Concerning the shoulder, most applications address the assessment of shoulder range of motion (ROM), generally finding reliable results (Werner et al., 2014; Shin et al., 2012; Mitchell et al., 2014; Cuesta-Vargas and Roldan-Jimenez, 2016; Johnson et al., 2015; Brophy et al., 2005). The results on shoulder function may possibly differ from these research outcomes, as ROM is only one component of shoulder function evaluation, which is a more complex concept than merely an end-range mobility evaluation. One study on healthy subjects showed that the analysis of accelerations was achieved generally with adequate precision when using a smartphone (Cuesta-Vargas and Roldan-Jimenez, 2016). However, to the best of the knowledge of this thesis' author, no smartphone-based software application to assess shoulder function is currently available, let alone validated. The

verification of a smartphone application for functional outcome measurement using the B-B Score would thus be novel and of paramount importance, especially when considering the high prevalence of shoulder conditions, the existing controversy about shoulder function questionnaires and the complexity of current computerized movement analysis methods.

1.1.2.4. Thesis aim

The combination of a score that includes only essential movements and a device whose use has entered into daily life reduces the testing procedure to its simplest expression. However, the transfer into practice is indicated only if this minimalist approach has previously proven its validity and has been compared with alternative approaches, that is PROMs questionnaires and measurement using movement analysis dedicated body-worn sensors.

The general aim of this thesis was thus to validate the simplest possible kinematic shoulder function scoring procedure applicable in clinical practice and research, and compare it with alternative approaches.

The research process included four phases: 1) Definition of the testing procedure, 2) Comparison of B-B Score measurements derived from a specifically developed smartphone application and an inertial measurement system 3) Validation of the smartphone B-B Score in current shoulder pathologies (rotator cuff conditions, humerus fracture, capsulitis, instability), 4) Benchmarking of the new approach with concurrent kinematic- and questionnaire-based methods

The Phase 1 study of the research programme centred on the precise definition of the research plan and the determination of the most efficient score calculation method of the B-B Score measured with an inertial measurement unit (IMU). At this stage, the variability in measurement was analysed and sources of variability in scores were tracked. Attention was focused notably on the influence of the measurement device, the subjects' characteristics, the feasibility issues and the inconsistencies in the measurement protocol. Recommendations were made accordingly about the research plan, measurement procedure and score calculation for the following phases.

The thesis' Phase 2 study aimed at the comparison of the outcomes and measurement properties of the B-B Score acquired using a dedicated IMU (inertial measurement unit) Physilog II system (Physilog®, Gait Up, Lausanne Switzerland) and an iPod (iPod®, Apple, Cupertino, USA) with a dedicated software measurement application.

The thesis' Phase 3 study aimed at establishing the measurement properties of the B-B Score derived from a smartphone, as delivered within the thesis' Phase 2 study, by critically evaluating the scope and effectiveness of its application within four prevalent shoulder pathologies encountered in physiotherapy: rotator cuff condition treated conservatively, shoulder instability treated conservatively, proximal humerus fracture treated surgically or conservatively, and capsulitis treated conservatively. Normal performance and score reliability over 6 months were investigated in a healthy population. At the end of the Phase 3 study, the convergent validity of the B-B Score in comparison with the current clinical function questionnaires was established for each shoulder pathology, as well as its discriminative power between healthy and pathological participants, intra- and inter-rater reliability, responsiveness, measurement error and interpretability aspects .

Data for Phase 2 and Phase 3 studies were collected simultaneously. The measurement method defined in the Phase 1 study was used in these phases to establish the measurement properties of B-B Score based on the calculation method that we had found to be the most efficient.

With the measurement properties of the B-B Score using a smartphone having been established, a benchmarking of the new approach with concurrent kinematic- and questionnaire-based methods was made in Phase 4, to contextualise the B-B Score measurement properties with regard to other methods used for shoulder function measurement.

The generic development process of a measurement instrument is presented in Figure 1.1.

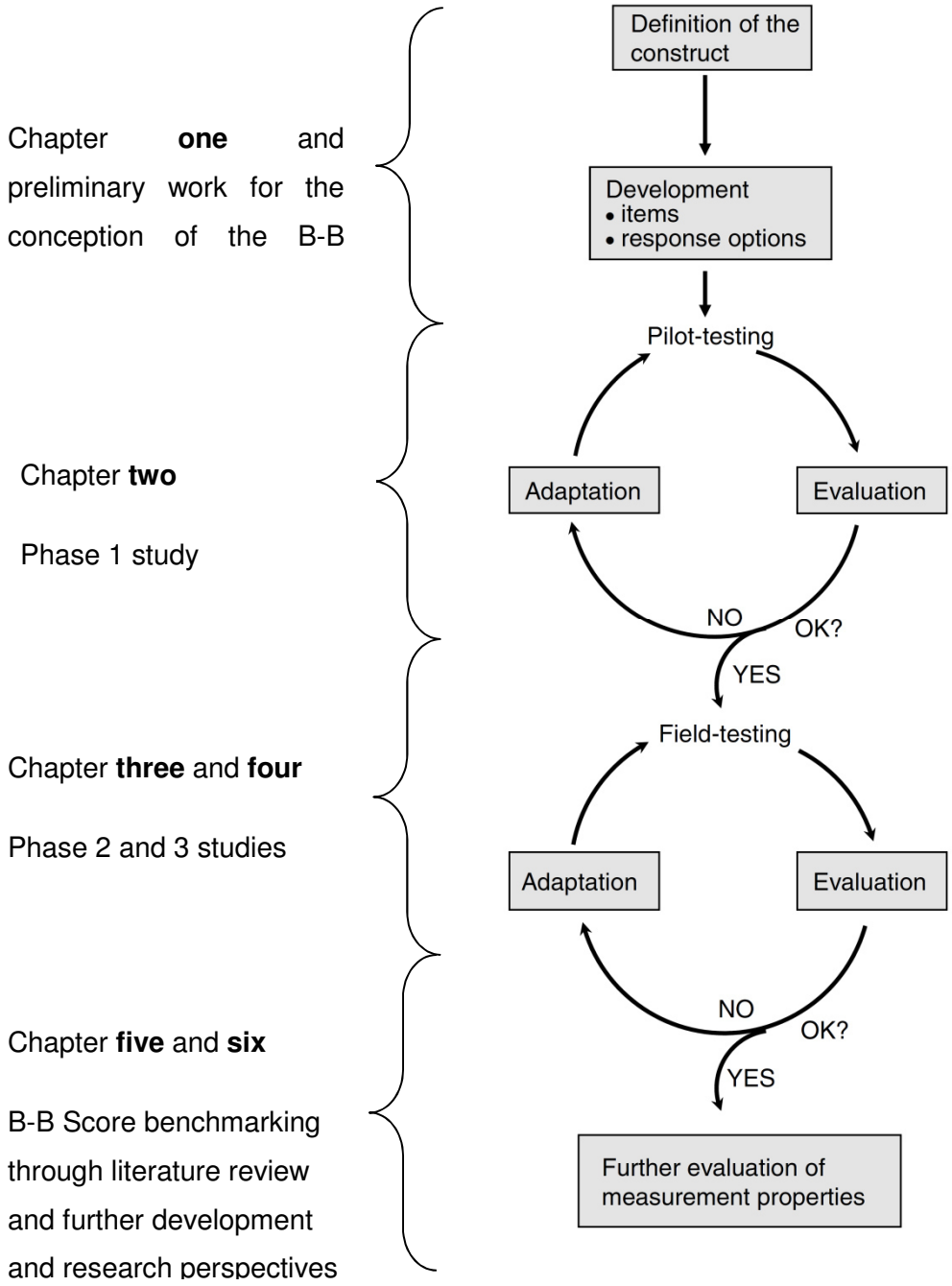


Figure 1.1: Steps in the development process of a measurement instrument. From: DE VET, H. C., TERWEE, C. B., MOKKINK, L. B. & KNOL, D. L. 2011. Development of a measurement instrument. *In:* TERWEE, C. B., KNOL, D. L., DE VET, H. C. W. & MOKKINK, L. B. (eds.) *Measurement in Medicine: A Practical Guide.* Cambridge: Cambridge University Press.

1.1.3. Definition of central concepts

1.1.3.1. Framework for the definition of shoulder function

Shoulder function is not a straightforward concept and no unequivocal definition is commonly accepted (Roe et al., 2013). In contrast to lower limb function that has locomotion as its main purpose, delineating the domains that describe upper limb function is much more complex because of the increased diversity of its possible actions.

The International Classification of Functioning, Disability and Health (ICF) could serve as a reference to delineate what encompasses the notion of shoulder function (World Health Organization, 2001; Michener, 2011; Roe et al., 2013). The ICF that was developed in 2001 by the World Health Organisation to serve as the international standard to describe and measure health and disability, classifies functioning within the components of body functions, body structures, activities & participation and environmental and personal factors. It promotes then an approach from a bio-psycho-social perspective and is largely recognised as a reference to conceptualise rehabilitation using a shared framework.

As such, using the ICF for shoulder function evaluation implies considering the problem from within a large perspective. The latter would include investigating impairments at the origin of dysfunction, while recognising the influence on the activities undertaken and the global consequences on a patient's life. A review about the measures of shoulder pain and function showed that the most currently addressed concepts were related to activities and participation, that is the execution of a task or action by an individual and his/her involvement in life situations. (Roe et al., 2013). Conversely, psychosocial functioning and environmental factors were more scarcely investigated. Overall, the most frequent items covered in questionnaires concerned pain, movement related body functions and structures, sleep, hand and arm use, self-care, household tasks, work and employment, and leisure activities.

These aspects are thus commonly considered as important components of shoulder function. However, they obviously do not encompass all aspects covered by the ICF. Though suitable from a conceptual point of view, covering exhaustively the ICF

framework would imply investigating a very large panel of items, because body structures and environmental factors should be included. While this extensive process may be required to investigate the determinants of shoulder function fully, it might not be necessary when only the functional outcome is the parameter of interest. The extent of the domains that should be covered by shoulder function evaluation tools remains a focus for debate (Beaton et al., 2001b; Michener, 2011). However, when designing a shoulder function evaluation based on the ICF framework, it seems reasonable to focus on the specific domains that characterise the person functioning (that is body function, activities and participation), like most instruments do. A broader rationale that accounts for all ICF domains would be necessary only when the determinant and consequences of dysfunction are of concern.

In the absence of a precise and universally accepted definition of shoulder function, it was nevertheless necessary to produce one that might be acceptable for the purpose of this thesis. Based on the aforementioned considerations, shoulder function should be understood in this thesis to be the ability of the shoulder to perform the movement and hold the positions required for the management of activities and life situations that are significant for the person.

This is an operational definition to facilitate the PhD's goals being pursued effectively, i.e. to develop a kinematic score for shoulder function evaluation. It stems from a logical reasoning based on the ICF framework and as such, remains focused on this reference. Other frameworks like the Disability Creation Process (Fougeyrollas et al., 1998) could have been considered as references. However, they are far less accepted worldwide, and it was more appropriate to propose a consensual than an innovative definition of shoulder function in the context of this thesis.

1.1.3.2. PROMS and clinical questionnaires

1.1.3.2.1. Definition of function in clinical questionnaires

It is challenging to envisage a measurement tool that remains easily usable while successfully encompassing all important and requisite clinimetric aspects. The absence of a universally accepted definition of function partly explains the controversy

surrounding the content of clinical questionnaires. In a review that investigated the aspects investigated in measures of shoulder pain and functioning, Roe et al. concluded that there are huge differences in the content of the condition-specific multi-item measures (Roe et al., 2013). The lack of an unequivocal definition has thus contributed to the manifold attempts made to create questionnaires that are more efficient in encompassing shoulder function (Fayad et al., 2005; Oh et al., 2009; Placzek et al., 2004; Roy et al., 2009). Consequently, the variety of rationales that sustained the conception of clinical questionnaires have led to the creation of a multitude of instruments that aim at the measurement of shoulder function (Huang et al., 2015; Harvie et al., 2005).

In the field of shoulder function evaluation, almost all clinical questionnaires are PROMs, i.e. patient-reported outcome measures, a type of clinical questionnaire in which the patient acts as his/her own rater. Note that for the sake of simplification, PROMs (for example SST, DASH, UCLA, WOSI) and composite scores that include patient-reported and clinical measurements items (for example Constant, ASES) will be mentioned as “PROMs” within the rest of this work. Indeed, both types of questionnaires adopt the same approach, which consists in collating the necessary amount of relevant information in a pool of items chosen to encapsulate shoulder function as it has been conceived.

In spite of this shared approach within PROMs, variations exist in the underlying conceptual-frameworks leading to the inclusion of items. Differences in PROMS lie in the fact that the evaluation may have condition-generic/specific, population generic/specific, shoulder-specific/upper limb, subjective/mix of subjective and objective or patient centred/standardised emphases. Each approach has his advantages and drawbacks, which are discussed hereafter.

This variety hinders the comparison between studies using different instruments and the syntheses within meta-analyses but nevertheless allows the choice of a targeted instrument in line with the measurement purpose (Green et al., 2003; Harvie et al., 2005; Makhni et al., 2015).

1.1.3.2.2. Shoulder function PROMs types

The degree of specificity of a questionnaire has an influence on its validity for shoulder function evaluation and its measurement properties. Typically, the SF-36 quality of life generic instrument has been used in several studies in conjunction with shoulder function PROMs (Hudak et al., 1996). It was consistently shown to be to be less responsive than shoulder PROMs and compared to the strength of correlation amongst the PROMs, showed lower magnitude relationships (Beaton and Richards, 1996; Angst et al., 2008; MacDermid et al., 2006). This illustrates that a generic instrument, which is only marginally affected by variations in shoulder function, does not effectively target this outcome. Conversely, it can prove useful to assess the broader impact of shoulder dysfunction on patients' lives.

A very specific instrument may have the advantage of circumscribing very precisely the condition and functional needs in a given population of patients. This results in greater validity for the instrument within the population of interest. Conversely, PROMs adapted to the general population may suffer from a marked ceiling effect when applied to an athlete population for example, which precludes the functional performance differentiation between them. This type of problem was illustrated by the SPORTS score for shoulder instability in athletes showing a lower ceiling effect than shoulder function PROMs designed for the general population (Blonna et al., 2014). However, very specific PROMs suffer from low adaptability, as they are adapted for the evaluation of precisely defined patients' populations. As these instruments are reserved for situations in which current PROMs have demonstrated their limitations, their use, and consequently the experience acquired about them, has remained marginal to date (Makhni et al., 2015; Gartsman et al., 2015). The development of a variety of specific PROMs for each condition, patient population and intervention (for example conservative and surgical) is conceptually sound, as instruments are valid in the population for which they were tested. However, the latter may result in the creation of an overwhelming number of new instruments for clinicians and researchers (Slobogean et al., 2011).

Actually, the most frequently used PROMs are situated in-between the two latterly described extremes, though to different degrees (Makhni et al., 2015; Gartsman et al., 2015). Some instruments for the general population were designed either for

shoulder function in specific pathologies (for example WORC, WOSI), generic shoulder function (for example Constant, ASES) or generic upper extremity function (for example DASH). However, despite their possible applicability in different clinical contexts, users should remain conscious that measurement properties found in one context are not transferable into a different one (Slobogean and Slobogean, 2011). Though the instrument is generic to some extent, the measurement validity associated with it remains specific to the context in which it was tested.

Another issue in shoulder function evaluation is the degree of subjectivity or objectivity of the instrument. The differentiation is apparently simple, subjective measurement being influenced by feelings and ideas, whereas objective measurement is not. However, the delineation is actually more complex to establish, as the definition of "subjective" is ambiguous in the scientific literature, the term being used to mean either rater-dependent, patient-reported or only assessable by the patient (Moustgaard et al., 2014).

Concerning shoulder PROMs, some are clearly subjective, for example investigating the perceived difficulty to successfully complete tasks, while other ones are composed of a mix of subjective and objective items, like the Constant that gathers perceived limitations and clinical measurements. Sometimes, the delineation between subjective and objective is ambiguous, as for example in the SST in which the patient is asked if he/she is capable of successfully undertaking an activity. The answer is subjective if the patient thinks that he is able to achieve it and objective if he can really perform it. The subjective or objective character of the results is important to define as both approaches investigate different aspects of function that are complementary (Matsen et al., 2017; Krueger et al., 2011; de los Reyes-Guzman et al., 2014). Thus, the delineation between subjective and objective outcomes is important for the understanding of the issue at stake and for correctly combining measurement tools in an evaluation.

Patient-centred evaluation tools, like for example the Patient Specific Functional Scale (PFSF) are clearly subjective, as they investigate the activities that are specifically of interest to the person. They have the advantage of focusing the questions on the ones that make sense in the context of the person being treated, but the generalisation of the results is therefore difficult, because the items investigated are different for each patient. Although there are some exceptions, patient-centred

evaluation tools have been rarely used in shoulder function research (Hurd et al., 2017; Hefford et al., 2012; Horn et al., 2012).

1.1.3.2.3. Implications for the thesis

In summary, though all PROMs are generally considered together and contrast with clinical or laboratory measurements, a large variety of approaches have been used to substantiate their conception. Therefore, it must be kept in mind that each PROM investigates a particular aspect of function that is slightly different from other PROMs and that to date, none of these approaches has demonstrated its superiority over any other.

This situation is suboptimal for clinicians and researchers, as it is complex and hardly applicable to choose several complementary PROMs to get a broad view of all aspects of shoulder function (Christie et al., 2009).

This discussion about shoulder function PROMs has implications for the PhD design. The criterion validity evaluation of the B-B Score, i.e. the demonstration that it actually measures shoulder function similarly to an established reference instrument, can only be relative to the conditions and environment in which it was assessed. As no gold standard PROM exists, no comparison can be made between the new kinematic score and a unique and strongly established reference. The best that can be done is to evaluate its convergent validity, that is the relationship with several other instruments that aim at the evaluation of the same outcome, though the comparators may also have their own limitations (McDowell, 2006). To be able to draw the most robust conclusions from the research on the validity of the B-B Score to truly measure shoulder function, it will be necessary to challenge it with several currently used shoulder function PROMs with different characteristics. Similarly, a comparison with the most frequently used PROMs will be needed in the literature review that will complete the PhD programme of research with a benchmarking of the calculated B-B Score's measurement properties against those of contemporary assessment tools with which it might represent an alternative.

It might be that the relationship varies according to the characteristics of each PROM, but these variations will also be informative about the aspects of shoulder functions that are measured by the B-B Score and the respective advantages of each approach.

1.1.3.3. Movement analysis-based methods

1.1.3.3.1. Validity issues

Computerised movement analysis offers useful practical advantages over PROMs. Notably, they overcome the intrinsic limitations of PROMs related to content validity, language and cultural issues and respondents' interpretations (Olley and Carr, 2008; Ragab, 2003; McDowell, 2006). Unlike questionnaires, no time-consuming and cumbersome process is needed for the translations into various languages. It has therefore a better potential for universal recognition than PROMs.

Moreover, the advantage of the computerised movement analysis approach - of which the different variations in methods will be presented hereafter - is to measure movement objectively. While PROMs capture by essence an interpretation of what happens, movement analysis captures the movement as it is, provided that the measurement error is contained. Fundamentally, computerised movement analysis translates the primary function of the shoulder, which is to orient the upper limb in the visual work space, in terms of biomechanical parameters (for example range of motion, accelerations, speed, power…)(Culham and Peat, 1993).

Nevertheless, and similarly to questionnaire-based approaches, it is a challenge to evaluate shoulder function using computerised movement analysis. As for PROMs, the imprecise definition of the term "function" means that various notions referring to different levels of the ICF classification have been used by researchers, leading to an absence of consensus about what should be measured to adequately reflect shoulder function (De Baets et al., 2017).

Actually, the fact that some movement parameters are captured does not imply that they are representative of function as defined in this thesis[1]. A biomechanical parameter cannot be considered to reflect shoulder function as defined above until its

[1] Shoulder function: the ability of the shoulder to perform the movement and hold the positons required for the management of activities and life situations that are significant for the person.

convergent validity has been demonstrated by an adequate correlation between it and recognised shoulder function measurement tools (de los Reyes-Guzman et al., 2014). This is usually done in contemporary research by investigating its correlation with PROMs that target the same outcome.

Due to the aforementioned controversies, convergent validity but no gold standard validity, can be established in the absence of a universally recognised PROM for shoulder function evaluation (McDowell, 2006). However, they currently represent the best available references, as similarly to PROMs, no movement analysis-based method has established itself as a reference for the assessment of shoulder function. Thus, when a new computerised movement analysis-based method shows an adequate correlation to PROMs, it is considered as valid for shoulder function evaluation. Conversely, when the correlation is weak, the unclear definition of function also makes for a situation in which it may be considered that complementary dimensions of function to those associated with PROMs are being investigated by movement analysis parameters (Korver et al., 2014a; Matsen et al., 2017). This illustrates that in the present situation, more research is needed to understand the degree to which shoulder function evaluation using PROMs or computerised movement analysis produce concurrent or complementary outcomes.

1.1.3.3.1.1. Data collection approaches

The capture of shoulder function using computerised movement analysis implies successfully identifying which representative movements and which relevant parameters should be measured to describe accurately the limits of the shoulder's capacity for functional motion. Two approaches have mainly been used for this purpose, an approach aimed at synthesis using short-time measurements and an extensive approach using long-time measurements (De Baets et al., 2017).

The rationale of the approach aimed at synthesis has been to identify a parameter indicative of shoulder function and analyse it over a limited number of movements (Coley et al., 2007a; Korver et al., 2014a; Jolles et al., 2010). The assumption of this approach is that the alterations of selected movements measured over a limited time can be representative of the variety of difficulties encountered in functional tasks. If so, the advantage of such an approach would be that rapidly acquired measurements

are able to provide information on a large range of functional difficulties faced by the patient in his/her activities of interest.

Conversely, the extensive approach has aimed at the acquisition of a revealing parameter over several hours to identify how the shoulder operates in unconstrained conditions of daily life, taking advantage of the portability of a dedicated measurement system (Coley et al., 2008b; Duc et al., 2014; Wylie et al., 2016). The assumption of this approach is that, provided that the parameters revealing the difficulty has previously been identified, the long measurement time makes it possible to capture the patient's functional difficulties when they happen. If so, such an approach would provide an objective picture of shoulder function in conditions that are very close to those within the patient's life. However, it is challenging to identify parameters that consistently reveal the functional alterations of patients independently of the great variety of situations potentially encountered within an unconstrained environment (Duc et al., 2013).

1.1.3.3.2. Definition of normal movement

A difficulty of upper limb movement analysis resides in the fact that each person has his own dynamic of movement (Khadilkar et al., 2014; Wickham et al., 2010; Linkel et al., 2017). Among others, the speed, range of motion and developed power are highly dependent of the person physical and psychosocial characteristics, as well as the conditions in which the task is executed. Movement, notably mobility and accelerations, is affected by age (Patel et al., 2007; Cutti et al., 2014; Roldan-Jimenez and Cuesta-Vargas, 2016). This dependency makes it difficult to parameter values from one person to the other and to determine with precision from which threshold an alteration is problematic.

The use of the healthy side as the reference may help overcoming this shortcoming. In this case, the healthy side is considered as the reference for normal movement and the magnitude of the difference between sides is considered as representative of the dysfunction, with the patient acting as his/her own control. An advantage of between-sides comparison is that age-related modification of movement is accounted for by the reference, as the physiological decline is reflected in the performance of the heathy side.

A condition for the comparison with the healthy side to be valid is that the range of the normal difference between sides must be previously known for the results to be interpreted. Importantly, the difference between the dominant and non-dominant side has to be established, or demonstrated to be negligible, for the comparison between sides to be valid. Another condition is that one of the sides has to be healthy. In case of bilateral problems, no shoulder can represent the normal performance that serves as a reference. This may be particularly limiting in older people, due to the increasing prevalence of shoulder rotator cuff tears with age (Yamaguchi et al., 2006; Yamamoto et al., 2010; Moosmayer et al., 2009).

1.1.3.3.3. Objective measurement vs. patient perception

The purely objective character of computerised movement analysis might be considered as an advantage, as the real performance and not an interpretation of the performance, is recorded. However, it is now commonly accepted that subjective and objective outcomes should be considered as complementary aspects of function evaluation, without hierarchy between them (Matsen et al., 2017; de los Reyes-Guzman et al., 2014). Subjective measures give insights into matters of human concern such as pain and suffering, while physical measures allow the quantification of function (McDowell, 2006). Moreover, the fact that the measurement is objective does not mean that the patient's subjectivity has no influence on the outcome. For example, kinesiophobia (which is a subjective feeling) may influence the course of the movement and consequently, the measured outcome. Thus, some subjective aspects that influence shoulder function are also accounted for when proceeding to an objective measurement. This is more an advantage than a disadvantage when measuring shoulder function, as the subjective aspects that influence the function will be also reflected in the outcome.

Conversely, it should be considered that biomechanical parameters do not reflect the perceived functional importance of the movement for the person until that has been formerly specified. A challenge in movement analysis is the translation from a technical to a clinically valuable tool that is relevant for both the therapist's and the patient's perspectives (De Baets et al., 2017). For example, it might be highly important to be able to lift the arm above the head in the professional occupation of

one person and of little importance for someone else. Hence, the same objective result may have quite a different meaning for each of them.

1.1.3.3.4. Kinetics and kinematics

Movement analysis is usually separated into two branches of mechanics, involving kinetic or kinematic analyses. The kinetics is the branch of mechanics that concerns the effect of forces and torques on the motion of bodies having mass (Encyclopædia Britannica Online), while the kinematics concerns the description of the motion of a body or system of bodies that is geometrically possible without consideration of the forces involved (that is, without focusing on causes and effects of the motions) (Encyclopædia Britannica Online). Kinetics addresses either forces when considering the linear movement or torque when considering the angular movement. Conversely, kinematics addresses positions, linear velocities and accelerations, or angular velocities and accelerations.

Kinematics has been used more often than kinetics for shoulder function analysis, as functional outcome is related to the ability to perform a movement rather than about causational explanations. Kinetics may offer a supplementary insight when the reasons for alterations in movement are of concern, but offer little added value when the aim is to evaluate function as an outcome. Moreover, they are more complex to acquire, as they require additional information compared to kinematics, such as mass or intensity of muscular activity. Thus, the possibility of applying a 'lighter' measurement procedure that is sufficient to analyse shoulder function explains why kinematics is more frequently used in the literature for the evaluation of shoulder function. In this context, kinematic analysis based on inertial sensor devices that record accelerations and angular velocities is increasingly used in the assessment of shoulder characteristics, due to their portability, and relative ease of use (Cutti et al., 2008).

1.1.3.3.5. Issues in the measurement of shoulder function

The shoulder's primary function is to orient the upper limb in visual work space, which will then allow, in contribution with the other upper limb joints, to place the hand in a

favourable condition for the execution of tasks (Culham and Peat, 1993). The respective role of the shoulder, elbow, wrist and fingers are the general upper limb orientation, distance adjustment, hand orientation and the handling of objects (Kapandji, 1971).

The shoulder structure is designed to meet two apparently conflicting conditions for the great variety of possible tasks in human activity to be achievable: to have sufficient upper limb mobility for the target point to be reached and sufficient stability for the shoulder to remain steady when it is put under physical constraints and stress during a task's execution (Veeger and van der Helm, 2007). This is made possible by the distribution of the movement over a complex structure composed of four different joints, the glenohumeral, acromioclavicular, sternocostoclavicular and scapulothoracic joints (Culham and Peat, 1993). The glenohumeral joints and surrounding muscles are in charge of the motion of the arm in relation to the shoulder girdle, while the four other joints regulate the motion of the shoulder girdle in relation to the trunk.

This construction has the advantage of adding the mobility to the movement of the scapula over the thorax, which contributes to approximately one third of the motion, to the movement of the humerus in relation to the scapula. The thoracohumeral mobility, which reflects the global mobility of the arm in relation to the thorax, results from the addition of these two mobilities (Veeger and van der Helm, 2007). A harmonious scapulohumeral rhythm, that is distribution of movement between the humerus and the scapula, is necessary to reach the full possible motion without overloading the joints and surrounding soft tissues (Ludewig and Reynolds, 2009). Alterations of the scapulohumeral rhythm are frequently described in the case of pathologies, because pain or decreased mobility of one of the shoulder joints induces compensatory movements in the other ones.

The shoulder complex enables large movements in the three space dimensions that is flexion-extension, abduction-adduction, medial-lateral rotations. Importantly, the spontaneous movements are rarely executed in one of the orthogonal planes, as the joints' physiological orientations and shapes predispose to the execution of three dimensional movements. When the shoulder motion is not sufficient, additional upper limb mobility can be found by adding trunk movements to shoulder movements. Large mobility being intrinsically linked to increased instability, the coordination of the

seventeen muscles surrounding the shoulder is very important to avoid dislocation while providing power for three dimensional motions (Nordin and Frankel, 2001). Notably, the rotator cuff muscles acts as a stabilising sleeve around the glenohumeral joint.

The complexity of the shoulder has meant that several approaches have been envisioned for its functional evaluation. Some authors underline the importance of separate analysis of each segment to be able to identify the source of a movement alteration (De Baets et al., 2017). This is sustained by the fact that coordination of the scapular and humeral movements is of importance for shoulder good functioning and is frequently altered in shoulder pathologies (Kibler et al., 2009; Ludewig et al., 2009; Ludewig and Reynolds, 2009; Lopes et al., 2015). However, this approach has technical implications, as it implies the need to use a complex model that accounts for the trunk, scapula and humerus movements, as well as their intersegmental coordination.

Running this kind of analysis is challenging because of the controversies about the reliability of scapula movements (van den Noort et al., 2014; De Baets et al., 2013) and of the relation between pattern variations and clinical symptoms (Littlewood and Cools, 2017; Kibler et al., 2013). From a practical point of view, the correct placement of the markers on the flat surface of the scapula, which is surrounded by muscles, remains a limitation to reliable movement analysis due to the difficulty of managing skin-movement artifacts (Lefèvre-Colau et al., 2017; Matsui et al., 2006). Moreover, asymptomatic dyskinesia or compensatory movements between segments may mean that the observed alterations in a single joint do not automatically induce an alteration in upper limb function (Littlewood and Cools, 2017).

Thus, a multisegmental model offers an insight into the intersegmental biomechanics that may be useful to understand some causes of altered function. Conversely, when only the functional outcome is of interest, without consideration for the underlying causes, a multisegmental model also induces an increased complexity that does not necessarily produce useful information for this purpose.

Therefore, some authors have adopted a minimalist approach aiming at the design of a simple model based on the minimal number of markers or sensors. In this way, they have relied on the measurement of the thoracohumeral motion, a virtual joint between

the shoulder and the trunk that does not consider the involvement of a complex multisegmental structure in the movement (Coley, 2007; Korver et al., 2014a; Duc et al., 2013). This model has obvious limitations for the precise investigations of the causes and locations of shoulder problems. Conversely, it may be efficient in capturing the arm motion in relation to the trunk, which is the resultant of the motion produced by each shoulder joint, when shoulder function is the outcome of interest. In this case, the investigations of the integrity of each single joint are omitted in favour of the investigation of the shoulder function, defined as the ability to perform the shoulder movements required for the management of activities. This model targets solely the functional consequences of the problems that affect the shoulder function, without consideration for the origins or for the intrinsic biomechanical alterations. Despite these limitations, this concept of shoulder evaluation has advantages for routine clinical assessment of outcome, as it relies on a limited number of markers/sensors and produces data that is less complex to analyse than a multisegmental model.

1.1.3.3.6. Implications for the thesis

The aforementioned discussions about computerised movement analysis have theoretical implications for the scope of the results and practical implications for the research methods.

In the absence of a universally recognised definition of shoulder function, it was necessary to define one that serves as a common thread for the thesis [2]. Thus, the statement that the B-B Score is or is not a measurement of shoulder function will be made with reference to this definition that is an operational one but one which could be challenged, as any definition might be.

As the kinematic approach analysis captures the movement by its essence, it will then be necessary to investigate if the B-B Score is actually representative of function. If

[2] Shoulder function: the ability of the shoulder to perform the movement and hold the positons required for the management of activities and life situations that are significant for the person.

the correlations with reference PROMs are low, it will imply that it does not measure function as conceptualised in these scores.

In the absence of a gold standard for shoulder function evaluation, either as a PROM or as a computerised movement analysis method, it will be possible to assess the convergent validity of the B-B Score but not its gold standard validity. The convergent validity will thus be estimated using the correlations of the B-B Score with currently used PROMs, which represent the most established references to date for assessing shoulder function.

Choices will have to be made in line with the thesis' aim to validate a score that is applicable within routine practice. However, the pursuit of this aim also implies that the score will be incorporating some of the intrinsic limitations of the method that will be used.

The B-B Score belongs within the category of the approach aimed at synthesis of shoulder function outcome measurements, which attempts to identify a limited number of movements and parameters that are representative of function. In pursuing the design of a simple approach to assessment, a thoracohumeral measurement is used in this score to minimise the sensors' configuration. As previous researchers have shown that one sensor was sufficient for function evaluation within measurement conditions in which the trunk movements can be controlled, only one IMU or smartphone will be fixed on the arm segment (Coley et al., 2007a; Jolles et al., 2011; Korver et al., 2014a). A limitation of this approach lies in the fact that no insight will be possible into the causes and precise location of the shoulder movement alterations.

The B-B Score is based on the comparison of a power-related metric (multiplication of angular velocities by accelerations) between sides (Pichonnaz et al., 2015c). This represents the patient's ability to control the humerus' velocity by its acceleration during the execution of the movement (Coley et al., 2007a).

The patient acts as his own control within this approach, with the healthy shoulder representing the normal performance of the person. This implies that the score will not be applicable in cases of bilateral shoulder pathologies, as no normal performance can be determined on either side when both sides are affected.

The B-B Score is based on accelerations and angular velocities, because it is preferable to refer to dynamic rather than static kinematics like end ROM, to capture shoulder function (Coley, 2007; Lopez-Pascual et al., 2017a). This may be explained by the fact that in some patients, the full ROM can be reached, but with difficulty. In these cases, the end ROM is normal, while the difficulty in executing the movement is captured by dynamic parameters, which are therefore indicative of the altered function of the shoulder. This approach proved to be discriminative for the P Score, which uses the same metric as the B-B Score (Coley et al., 2007a). Due to the three-dimensional aspect of functional shoulder movements, a 3D data capture will be needed.

A technical advantage of the use of IMU fitted with accelerometers and gyroscopes is that accelerations and angular velocities are measured directly, without the need for differentiation or integration calculations. Thus, the drift that affects inertial measurement remains negligible, as no calculation process amplifies it. This directly contrasts to the situation for angular measurements made using an IMU, which may be effected by drift in this case, but will not be used for the purpose of shoulder function assessment in this thesis (Amasay et al., 2009; Rowe, 1999). To ensure the soundness of this argument, this point was investigated and confirmed by preliminary measurements that preceded the start of the thesis.

1.1.3.4. Clinimetrics

To be recognised as validated, any new score needs to undergo an extensive validation process based on current requirements. The scientific discipline that deals with the establishment of the measurement quality has been referred to by various terms with corresponding definitions, including psychometrics, metrology or clinimetrics.

Subtle nuances that are subject to controversies exist between the delineations of these terms, about which further discussion is beyond the scope of this thesis (Streiner, 2003). For the purpose of this thesis, the term clinimetrics was adopted for its focus on clinical measurement. Clinimetrics is a methodological discipline that deals with the quality of clinical measurement (Feinstein, 1983). It thus encompasses both the quality of the instruments and the quality of the actual measurements, while

also accounting for errors induced by human factors and the environment (de Vet et al., 2003).

Multiple qualities are expected from a measurement instrument to ensure that the result gives a correct representation of the reality. These qualities are encapsulated by the concepts of validity, reliability and responsiveness (Mokkink et al., 2010d). In addition, the determination of normal performance and interpretability aspects is of importance for the interpretation of the results (Tubach et al., 2007). Some practical aspects like accessibility, interpretability and affordability are also of importance for clinical implementation. These notions will be defined immediately hereafter, as well as a discussion of their implications for validating a measurement tool.

The definitions and relations between concepts for measurement are difficult to obtain in essence and are thus subject to controversies. Although every definition might be considered disputable, the definitions provided by the COSMIN study represent a sustainable taxonomy of measurement properties of HR-PROs (health-related patient-reported outcomes), as they are based on an international consensus of experts using a systematic methodology (Mokkink et al., 2010d). This COSMIN terminology will thus be used to structure this chapter and to determine the measurement properties to be used in the measurement property study. An overview of the COSMIN classification that summarises the domains, measurement properties and aspects of measurement properties that define the quality of an instrument is presented in Figure 1.2.

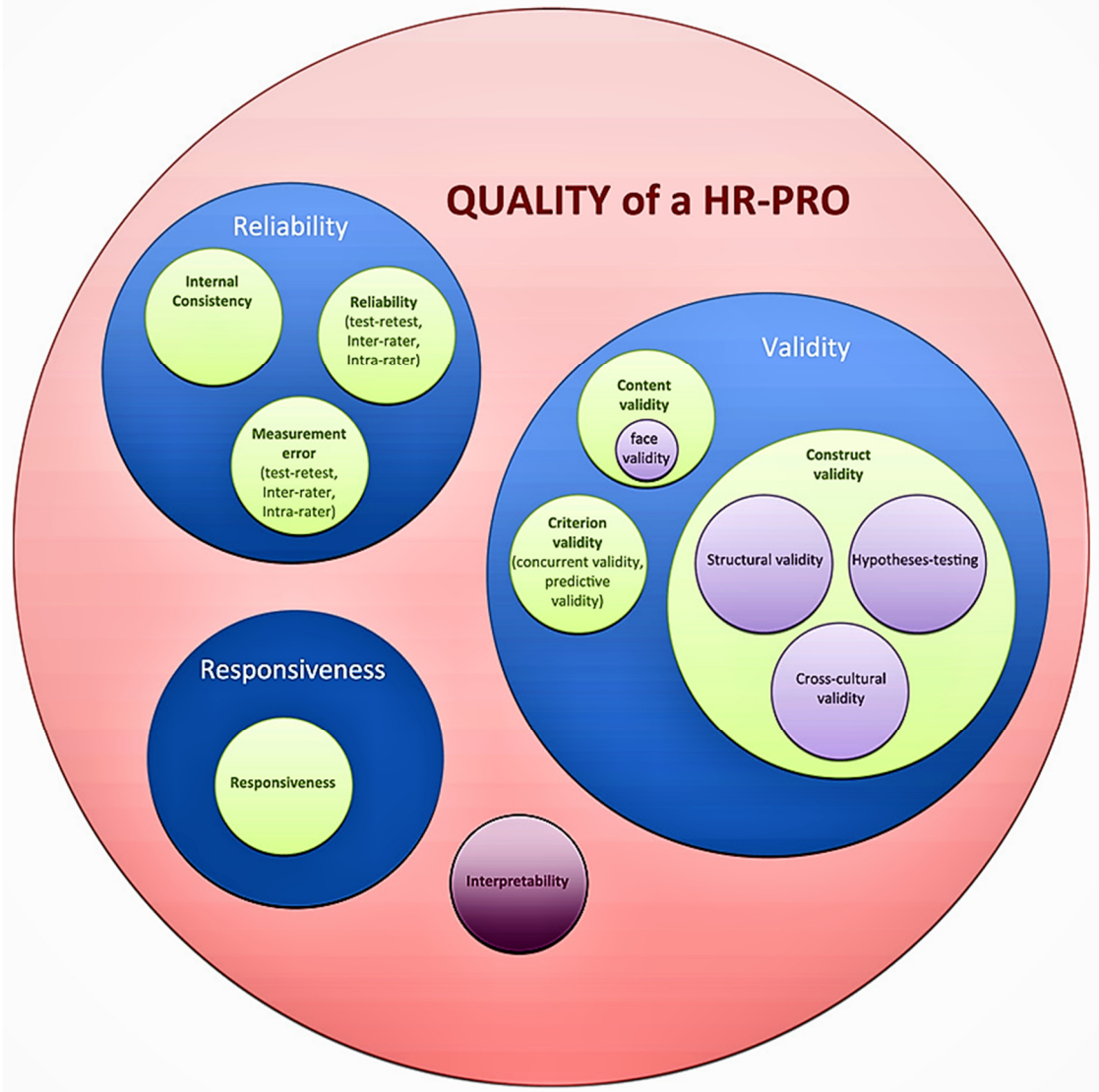


Figure 1.2: COSMIN classification that summarises the domains, measurement properties and aspects of measurement properties that define the quality of an instrument. Source: MOKKINK, L. B., TERWEE, C. B., PATRICK, D. L., ALONSO, J., STRATFORD, P. W., KNOL, D. L., BOUTER, L. M. & DE VET, H. C. 2010. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol,* 63, 737-45.

1.1.3.4.1. Validity

The validity is the degree to which an instrument measures the construct(s) it purports to measure (Mokkink et al., 2010d). It is thus a fundamental quality, because there is no point in running further studies to establish the measurement properties of an instrument that would not effectively measure the outcome it is intended to measure. Validity includes three measurement properties, which are content validity, construct validity and criterion validity.

1.1.3.4.1.1. Content validity

Content validity refers to the degree to which the content of an instrument is an adequate reflection of the construct to be measured (Mokkink et al., 2010d).

Content validity also encompasses the aspect of face validity, which is the degree to which (the items of) an instrument indeed look as though they are an adequate reflection of the construct to be measured (Mokkink et al., 2010d). The establishment of face validity is based on a subjective assessment of the content of the instrument (De Vet et al., 2011e). It relies on the impression of persons that are recognised as knowledgeable about the concept to be measured, considering their experience and the related literature in their field of competence. Though rather basic in its conception, face validity is a fundamental initial step to consider before the initiation of complex validation studies (De Vet et al., 2011e).

Content validity implies the need to investigate the content of the instrument in more detail than face validity to assess whether it adequately represents the construct being scrutinised (De Vet et al., 2011e). It requires the assessment of the relevance and the comprehensiveness of the items for the construct to be measured (McDowell, 2006). Thus, the clear definition of the concept of interest is a prerequisite for content validity evaluation. After the concept has been circumscribed, it can be assessed if the items of the instruments are in close relationship to it and cover all its dimensions. The content validity is assessed by the persons who are concerned with the instrument. For PROMs, these persons can be patients, who have become knowledgeable through their experience of the disease, or health professionals, who have become informed through their training and encounters with patients. So, content validity is based on expert opinion rather than on statistical testing. This measurement property

is relative to the context of the measurement and the population, and should therefore be assessed specifically for each population in which the instrument is used. A positive rating for content validity can be given if a clear description is provided about the instrument's wider development process. This includes the reporting of the measurement aim, the target population, the concepts that are being measured, and the items' selection. Furthermore, the target population and/or experts should have been involved in the instrument's construction process (Terwee et al., 2007).

The determination of content validity is central for clinical measurement as it relates to the fundamental issue of what is measured. In some cases, the content validity is rather straightforward to establish, as the relation between the concept and the instrument is obvious. For example, the relationship between the knee joint mobility and its range of motion is evident in an osteoarthritic population. Conversely, some concepts are more difficult to define are they give more scope for subjective interpretation. The latter concern has been addressed previously within this introduction to the thesis when issues and controversies related to the measurement of shoulder function had been discussed. Nevertheless, measures of subjective aspects are sometimes irreplaceable in clinical evaluations, as they provide an insight into matters of human concern that cannot be investigated from physical measurements (McDowell, 2006).

Specifically concerning the B-B Score, its face validity has been determined by its close relationship to the P Score, which itself measured objectively the movements described in the SST, a commonly used shoulder function PROM (Pichonnaz et al., 2015c). The two movements "hand to the back" and "hand to the ceiling as to change a bulb" are also reported as being problematic by patients suffering shoulder function loss (van der Windt et al., 1995; Magermans et al., 2005). This endorses the face validity of the B-B Score, as it shows an *a priori* link to shoulder function.

The content validity of the B-B Score has not been determined by a systematic investigation of expert opinion, though the two selected movements are currently used for clinical evaluation of the mobility of the shoulder. Conversely, the choice of the movements was justified by statistical methods that relate to construct validity (Pichonnaz et al., 2015c). Further investigations conducted within this thesis will examine to which degree the score content – that is the measurement of a power-related metric for the two movements – is congruent with the measurement aim that

is to grasp shoulder function in various pathologies. The conclusions will have to be drawn separately for each pathology, because the content validity is specific for each population (De Vet et al., 2011e). Therefore, specifically for this score, the content validity of the score is sustainable, but has been determined using different methods than the ones that are currently used for item selection in PROMs questionnaires (i.e. statistical approach instead of expert opinion).

1.1.3.4.1.2. Floor and ceiling effect

Floor and ceiling effects are classified within “Content validity” in the COSMIN taxomony, as these effects are related to the distribution of the items over the scale. They are also related to interpretability aspects, as information about the floor and ceiling effects are important for the interpretation of the performance or the change of a score.

The sensitivity of an outcome measure may not be homogeneous along all the possible scale values, especially at the scale’s end ranges. Such a phenomenon is indicative of a limitation in content validity of a scale (Terwee, 2007). For example, a scale containing overly challenging items will not be responsive to the deterioration of patients with low performance, if they have already exhibited the minimum score before the deterioration. Such a scale offers insufficient scoring sensitivity at its lower echelons in particular. This phenomenon is called the floor effect. Conversely, a ceiling effect is observed when the outcome measure’s items are not challenging enough for the evaluated population. In this case, the improvement of patients who perform high will remain undetected, as they have already achieved the maximum outcome measure value, before displaying further improvement. Such a scale offers insufficient scoring sensitivity at its higher echelons in particular. Therefore, when floor effect is present, patients with the lowest possible score cannot be distinguished from each other, even though their performances may differ. The same reasoning apply for patients with the highest possible score when ceiling effect is present. For example, an outcome measure designed for patients, will typically show a ceiling effect in athletes, and will thus be unable to discriminate the performance level among the latter group. Floor and ceiling effects are thus contextual to the population for which they were determined (Terwee et al., 2007).

It is generally considered that floor and ceiling effects exist when more than 15% of the patients get the minimum or maximum value on the score, respectively (Terwee et al., 2007; McHorney and Tarlov, 1995). However, this approach does not account for possible measurement errors. Therefore, it has also been proposed to use the minimum scale value + MDC (minimal detectable change), as a threshold for floor effect, and the maximum scale value - MDC, as a threshold for ceiling effect. This approach accounts for the fact that a change below the MDC value is unlikely to be detected if it is close to one of the extremes (van der Linde et al., 2015; van der Linde et al., 2014).

The commonly accepted approach, which consists in taking 15% of patients as a threshold for floor and ceiling effect determination, is somewhat arbitrary and implies a dichotomous conception of floor and ceiling effects that can only be considered “present” or “absent”. However, when the measurement properties of several outcome measures are compared within a study, the proportion of outcomes at the minimum and maximum scale values is also informative about the respective trend of each outcome measure towards floor and ceiling effects, regardless of the 15% threshold.

1.1.3.4.1.3. Construct validity

In contrast, to content validity, construct validity implies the need for statistical analyses to determine objectively to what extend the instrument is coherent with the construct of be measured. It encompasses several measurement properties, which are structural validity, hypotheses testing, and cross-cultural validity (Terwee et al., 2007).

Structural validity is concerned with the degree to which the scores of an instrument are an adequate reflection of the dimensionality of the construct to be measured (Mokkink et al., 2010d). For abstract concepts, gold standards do not exist and thus validity testing is more challenging. Some analyses need to be done to investigate if a single score can really summarise several variables into a coherent single result (McDowell, 2006).

When an instrument is unidimensional, it uses several items or measurements that should all be related to the targeted concept. In cases where a measurement contains

several scales to produce a resulting score that combines different subscores, each of those should be unidimensional. For example, a score that aims to measure a complex clinical phenomenon might have a physical, a mental and a psychosocial dimension, which should be clearly differentiated in a multidimensional instrument. The scale dimensionality can be assessed using factor analysis (FA), which is an advanced statistical approach used to reduce a large number of variables into fewer numbers of factors corresponding to dimensions. It calculates the maximum common variance from all variables and indicates if it is relevant to condense them into a single score. Exploratory factor analysis is used in case no *a priori* hypothesis is made about the concept. It shows how the measured variables cluster together to represent an underlying construct (McDowell, 2006). Exploratory FA aims thus to identify groups of variables that form a dimension that is related to the concept. Confirmatory FA is used to test hypotheses regarding the factor structure that has been previously formulated based on a theoretical approach (Terwee et al., 2007).

Hypothesis testing is another aspect of construct validity. Some theoretical relations are hypothesised, when a score is developed based on a construct. These theoretical hypotheses have to be confirmed to be real by statistical analyses for the score to be considered valid. Hypothesis testing relates thus to the degree to which the scores of an instrument are consistent with hypotheses based on the assumption that the instrument validly measures the construct to be measured (Mokkink et al., 2010e; McDowell, 2006).

Many hypotheses requiring various research designs can be envisaged when an outcome measure is to be validated. Most important types of hypotheses can be grouped into the notions of convergent, divergent or known-groups validity. Contextualising these notions in the context of shoulder function, a new shoulder function outcome measure would be expected to be correlated to other outcome measures pursuing the same aim (convergent validity), and negatively correlated to dysfunction outcome measures (divergent validity). It would also be expected that the results are related to the shoulder health status, that is, there would typically be significant differences between patients and healthy controls (known-groups validity) (Mokkink et al., 2010e; McDowell, 2006).

The translation of a questionnaire into various languages is not a straightforward issue, because some subtle nuances may be hard to translate and because the idea

underlying a question can be diversely interpreted according to the cultural background and lifestyle of different populations. A rigorous translation process with forward and backward translation by several translators is required to ensure the language and cultural equivalency of the translated and original version of an outcome measure (Wild et al., 2005; De Vet et al., 2011e). Nevertheless, a thorough translation is not sufficient to ensure that the equivalency is reached. Cross-cultural validity (that is the degree to which the performance of the items on a translated or culturally adapted instrument are an adequate reflection of the performance of the items of the original version) has to be established. This implies that by using various statistical approaches (confirmatory factor analysis, logistic regression and item response theory), it must be demonstrated that the questionnaire structure, item difficulty and measured performance are similar to those of the original version for the measurement of similar populations (De Vet et al., 2011e). It may also be sound to check that reliability and responsiveness are equivalent between versions.

The confirmation of the outcome measures' equivalency is of first importance for the realisation of meta-analyses that compile results from various countries. Yet, it is a very cumbersome process that limits the possibility of using a questionnaire universally. A long time is needed until valid translations are available in major languages and there is a high risk of excluding the numerous populations that speak local idioms. Quantified evaluation has a clear advantage over questionnaires on this issue, as numbers are more universally shared language.

1.1.3.4.1.4. Criterion validity

Criterion validity is the degree to which the outcomes acquired using an instrument are an adequate reflection of a "gold standard" (Mokkink et al., 2010e). It considers thus whether outcomes on the instrument agree with another measurement of the same outcome that is an undisputable reference. This property is typically investigated when a new instrument is developed as an alternative that could potentially be simpler, cheaper or more convenient to use than an established measurement. Criterion validity is sometimes called *concurrent validity* when the criterion refers to a current state, and *predictive* validity when it refers to the anticipation of a future state (McDowell, 2006).

In the situation of shoulder function, no criterion validity can be established due to the lack of a gold standard, because no indisputable measurement of shoulder function exists and the definition of one in the future is hardly conceivable. In this scenario, the establishment of convergent validity, that is the testing of hypotheses that state that the results are correlated positively with the results of other instruments that measure the same concept, is the best approach that can be envisaged (McDowell, 2006). Thus, for shoulder function, its evaluation requires the calculation of the strength of correlation of the tested instrument with other recognised measurements of shoulder function, in order to explore how far the tested instrument actually reflects shoulder function.

For diagnostic tests, criterion validity comes from discriminating correctly those who have from those who do not have a disease, as would be demonstrated by a gold standard that classifies relevant people without mistake. A test is considered as *sensitive* when it identifies all the people with the condition of interest, and it is *specific* when the people identified by the test as having the condition, really have it. Therefore, if the test lacks sensitivity some people with the condition will miss being identified. A negative sensitive test is particularly useful for ruling out with minimal doubts the people without the disease, and thus for avoiding unnecessary interventions. A positive specific test is particularly useful for ruling in with minimal doubts, the people with the disease, and thus for undertaking necessary measures for them (Nendaz and Perrier, 2004; Christe, 2017).

Considering computerised shoulder function analysis, movement alterations have not been shown to be pathognomonic of specific shoulder conditions to date. So, this approach cannot be used to diagnose shoulder conditions. Conversely, it can be used to discriminate patients with shoulder function alteration from patients with healthy shoulders (Korver et al., 2014a; Pichonnaz et al., 2015c). Though unable to identify precisely the medical diagnosis, a good outcome measure of the shoulder function should be able to separate people who probably have a functional decrease from those who have not.

In contrast to a disease diagnosis, which is dichotomous (the disease is either present or absent), a shoulder function outcome measure fits in a continuum that ranges from a completely absent to a normal function. Thus, there is a requirement to have identified a cut-off score that represents the best balance between sensitivity and

specificity, knowing that an increase in sensitivity is almost always associated with a decrease in specificity (McDowell, 2006). The receiver operating characteristic (ROC) curve that plots true-positive (sensitivity) against false-positive (1-specificity) results can be calculated to illustrate the trade-off between sensitivity and specificity. Therefore, the cut-off score that represents the optimal balance between sensitivity and specificity and the area under the curve (AUC), indicates the amount of information provided by the test, which can be calculated. An AUC value of 0.5 indicates that the outcome measure is no better than merely guessing to identify if someone has a shoulder function loss, while a value of 1 means that the outcome measure discriminates without mistake those who have from those who do not have a shoulder function loss (McDowell, 2006; Hanley and McNeil, 1982).

1.1.3.4.2. Reliability and agreement

The notion of reliability relates to the degree to which the measurement is free from measurement error (Mokkink et al., 2010e). It is defined by the proportion of the total variance in the measurements, which is due to “true” differences between the patients. It expresses thus how well patients can be distinguished from each other despite the presence of measurement error (McDowell, 2006). Ideally, an instrument is expected to produce the same results for repeated measurements of patients who are stable. This should be the case when using different sets of items from the same instrument (internal consistency), over time (test-retest), by different persons on the same occasion (inter-rater) or by the same person (that is, rater or responder) on different occasions (intra-rater) (Mokkink et al., 2010e)

1.1.3.4.2.1.Internal consistency

Internal consistency refers to the interrelatedness of the items (Cortina, 1993). This measurement property is evaluated using a statistical construct involving the computation of Cronbach’s alpha, which is indicative of the degree of inter-correlation between the items, and thus their consistency in measuring a latent trait. A 0.70 to 0.90 value is generally considered as a measure of good internal consistency. A lower Cronbach’s alpha indicates a lack of homogeneity between the items, which makes summarising them into a single score unjustified. Conversely, a higher Cronbach’s alpha is an indication of redundancy between the items (Terwee et al., 2007).

Despite an existing controversy about its ability to determine the internal structure of an outcome measure, Cronbach's alpha remains the most frequently used statistics for this purpose and is considered an adequate evaluation by the COSMIN (COnsensus-based Standards for the selection of health Measurement Instruments) initiative (Mokkink et al., 2010c). Cronbach's alpha refers to Classical Item Theory (CTT), but an alternative approach is to use Item Response Theory (IRT). IRT is a more complex approach that does not assume that each item is equally difficult and incorporates this information in the analysis using item characteristic curves that reveal the item's difficulty (van Alphen et al., 1994). A different reliability coefficient will thus be calculated for each item, implying that no single reliability result exist for a measurement (McDowell, 2006). This topic will not be considered further as it has limited implications within the context of the thesis.

1.1.3.4.2.2. Test-retest, intra- and inter-rater reliability

Conversely, test-retest, intra- and inter-rater reliability have strong implications for the work to be done. These three aspects of reliability have in common that they share a focus relating to the degree of error in repeated measurements. Classical test theory considers that the result is a combination of the underlying true score and error to some degree (McDowell, 2006). In test-retest reliability, the errors are only due to day-to-day variations or to the instruments. In intra-rater reliability, the error introduced by the variations that a rater makes between his/her measurements is added to the previously mentioned sources of error. In inter-rater reliability, the error introduced by the variations between raters is added to all previously mentioned sources of error (De Vet et al., 2011b).

Intraclass correlation should be ≥ 0.75 to be considered as good and should be ≥0.90 to ensure reliability in clinical measurements (Portney and Watkins, 2015).

Intraclass coefficient of correlations (ICCs) is the most frequently used statistic to evaluate the reliability of continuous variables, while the Kappa coefficient is used for dichotomous variables and the weighted Kappa for ordinal variables (De Vet et al., 2011b; Kottner et al., 2011).

ICCs will be used to evaluate continuous variables within this thesis. ICC is indicative of the ability of a test to differentiate between individuals (Weir, 2005). It has the

advantage of being sensitive to systematic differences, in contrast to the Spearman correlation, and thus, it's an estimator of agreement and not just consistency (i.e. the actual similarity of measurements rather than just an estimate of their association) (McDowell, 2006). This might be of importance in the detection of training- or fatigue-related effects when proceeding to repeated measurements, for example. Several forms of ICCs have been described to adapt to various testing conditions, for example the number of replications and raters (Shrout and Fleiss, 1979). All formulas for ICCs consist of a ratio of the variance due to systematic differences between the "true" scores of patients and the total variance (summing true and error variance) (De Vet et al., 2011b). Thus, ICCs have been criticised because of their tendency to be low when the sample variance is low and high when the sample variance is high, independently of the measurement error (Russek, 2004). Thus it implies that reliability is a characteristic of an instrument used in a population, and not just an intrinsic property of an instrument (McDowell, 2006). Another limitation of ICCs is that they provide a global indicator of reliability but do not give indications on the potential error magnitude between measurements (Bland and Altman, 1986b).

An alternative to the calculation of ICCs, is the calculation of the concordance correlation coefficient (CCC), which indicates the agreement between the observed data and a 45° slope (line of identity) (McDowell, 2006). Both methods are considered as equivalent and produced results are comparable (Feng et al., 2014; Carrasco and Jover, 2003)

1.1.3.4.2.3. SEM

The standard error of measurement (SEM) is an indication of the precision of an outcome measure, that allows the construction of confidence intervals around the measured values (Weir, 2005). Though SEM is abbreviated similarly to the standard error of the mean, it should not be confused with it. The standard error of the mean is not related to reliability, as it is defined as the standard deviation of the sampling distribution around the mean (McDowell, 2006).

The SEM is representative of the "typical error" of a measurement, as it quantifies the precision of individual outcome measures on a test. It defines the boundaries within which a subject's true outcome probably lies (Weir, 2005). It is indicative of the amount of error that may be expected, due to chance alone. Ideally, the SEM would be zero

when using a perfectly reliable instrument and all variation would reflect true differences (McDowell, 2006). The SEM_{95} is generally reported (indicating 95% confidence limits within which the true outcome is expected to lie), though some authors report it using the less stringent 90% interval. Using the SEM_{95}, a clinician can be 95% confident that the patient's true outcome lies within the '±' error boundaries specified for this parameter. As such, this is an important indication of the margin of error of a result in constructing the clinical interpretation of an outcome. For example, a rater can be reasonably confident that a patient improved only if the difference between the initial and follow-up measurement is larger than the SEM_{95} (Michener, 2011).

The SEM is calculated using the standard deviation of errors amongst repeated measurements (De Vet et al., 2011b). There are $SEM_{agreement}$ and $SEM_{consistency}$ versions of the SEM that accounts or does not account for systematic errors, respectively. However, as a limited amount of repetitions of measurements is generally available in practice, the SEM is rarely determined using the SD of repeated measurement. It more frequently estimated using SD of the difference between two raters ($SEM_{consistency} = {SD_{difference}}/{\sqrt{2}}$) or the formula based on the ICC ($SEM_{consistency} = pooled\ SD \times \sqrt{1 - ICC}$), which represent an estimate of SD of errors for the data available (De Vet et al., 2011b; Portney and Watkins, 2015). Applying the second formula, it should also be kept in mind that, for the SEM to be estimated and relevant, the ICC used in the calculation should originate from the same population as the one in which the SEM will be used (De Vet et al., 2011b).

1.1.3.4.2.4. Minimal Detectable Change

When a difference is observed between two measurements, the issue for the rater is to differentiate between the difference caused by error in the value measured by the instrument, or by a real difference between measurements, knowing that both are combined to a variable and unknown extent. The Minimal Detectable Change (MDC), also sometimes called SDC (smallest detectable change), MDD (minimal detectable difference) or SDD (smallest detectable difference), can be calculated to evaluate the value beyond which the difference can be considered as true (Beaton et al., 2001a). The MDC is linked to the SEM value as the mathematical expression to calculate MDC is: MDC (95% confidence level) = 1.96 * $\sqrt{2}$* SEM

The 95% confidence interval is generally used, though the less stringent MDC at 90% confidence interval (MDC_{90}) is sometimes reported (Beaton et al., 2001a; Membrilla-Mesa et al., 2015a; Michener, 2011). It is considered that values larger than the MDC at 95% confidence level (MDC_{95}) have 95% probability to be due to a real difference (van Kampen et al., 2013).

The MDC is an important property for the interpretation of differences for the clinician. However, it should be kept in mind that, though real, a change could be of little importance for the patient's subjective state (de Vet et al., 2006a; Michener, 2011). It should also be considered that the MDC is population dependent (Schuller et al., 2014).

1.1.3.4.2.5. Bland and Altman analysis

Bland and Altman (B&A) have proposed a procedure to plot the values of the differences between measurements against the measured value, as well as to calculate the limits of agreement (LoA) and the bias (Bland and Altman, 1986b). They proposed this approach to overcome some shortcomings of the correlation and regression analyses that are indicative of the strength of the relationship but do not provide values on the systematic and random error of measurements. Conversely, the 95% limits of agreement and bias inform the user on the range that contains 95% of random measurement differences and the systematic measurement difference, respectively. In addition, while calculated correlations tend to be higher when the study sample heterogeneity is high, agreement parameters are independent of the data dispersion (de Vet et al., 2006b; Russek, 2004). The B&A analysis can be performed for test-retest, intra-rater, and inter-rater measurements. The magnitude of the LoA is closely related to the magnitude of the SEM, as the LoA represent 1.96 * SD of the difference between measurements, and the SD of the difference can be estimated using the formula: $SD_{diff} = (\sqrt{2} * SEMconsistency)$.

When performing a B&A analysis, it is important to check the assumption that the differences between the measurements do not change as a function of measured values (De Vet et al., 2011b; Giavarina, 2015). For example, it can be useful to check if the errors increase with the measured value. In addition to using calculations, Bland and Altman proposed that graphical analyses of the differences be performed to allow for a visual inspection of their characteristics (Bland and Altman, 1986b). The B&A

graph consists in plotting the values of the differences between measurements against the measured value, and tracing the lines that indicate the bias (mean value of the differences between measurements) and the 95% LoA (range from the bias ± 1.96 standard deviation of the differences). When further analysis is required, the relationship between errors and measured values can also be characterised using regression analysis and the randomness of their distribution can be checked using graphs and inferential statistics.

LoA and bias can be expressed as absolute values when the error on the scale is of interest, or as percentages when the proportion of error is of interest. The degree of precision of the bias and LoA estimations can also be determined calculating the 95% confidence interval (Giavarina, 2015). Illustrations of Bland and Altman plots with the representation of the limits of agreement are available in Figure 1.3.

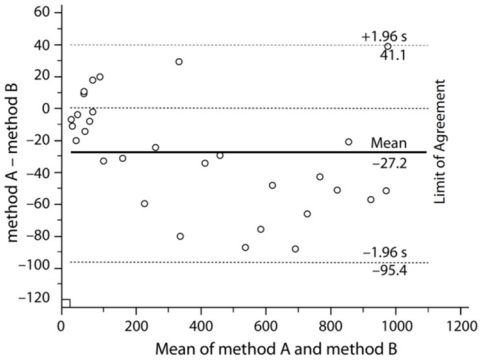


FIGURE 5. Bland and Altman plot for data from the table 1, with the representation of the limits of agreement (doted line), from -1.96*s* to +1.96*s*.

a)

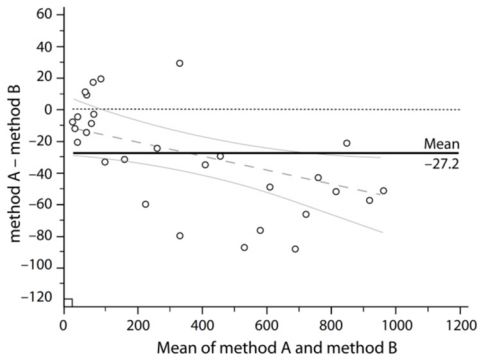


FIGURE 3. The same plot as Figure 1 including regression line and confidence interval limits.

b)

Figure 1.3: a) Bland and Altman plot with the representation of the limits of agreement (dotted line), from -1.96 standard deviation to +1.96 standard deviation and bias representing the mean of the differences between measurements. b) Bland and Altman plot including regression line and its confidence interval limits. From: Giavarina, D. (2015). Understanding Bland Altman analysis. Biochemia Medica, 25(2), 141-151.

1.1.3.4.3. Responsiveness

The responsiveness is defined as the ability of an instrument to detect change over time in the construct to be measured (Mokkink et al., 2010e). This definition implies that the instrument must not only be able to measure a change that happened, but also that the measured change has to be in close relationship with the outcome

targeted by the instrument. Considerable controversies surround the delineation of what responsiveness encompasses and which are the appropriate methods to measure it (Terwee et al., 2003; Mokkink et al., 2010e; Angst, 2011). The above definition was adopted in this thesis because it results from a consensus. Methods for the evaluation of responsiveness in measurement instruments that have been used in current practice will be described, without entering into the conceptual debate on the definition of responsiveness.

Responsiveness is an important measurement property, considering that the assessment of the change in patient's status is crucial for health interventions aiming at improving the patient's condition (De Vet et al., 2011c). Two facets are of interest when investigating the responsiveness of an instrument: its ability to detect a treatment-induced change over a given time period, and also the relationship between the change that the instrument measures and the change in an external standard (Husted et al., 2000; Terwee et al., 2003). Based on the latter stated expectations and the methods currently used in the literature (as stated in the literature review performed in this thesis), the following characteristics for responsiveness' evaluation will be presented: the statistical difference between groups/stages, effect size, standardised response mean, correlation between change scores and ROC curves analysis.

1.1.3.4.3.1. Statistical difference between groups or stages

The calculation of the statistical significance of the difference between groups, when differences are expected between the groups, or of the difference between measurement times for treatments of known efficacy, are currently used to evaluate the ability of an instrument to detect differences. This constitutes a fundamental step for responsiveness evaluation, as an instrument that would fail this test would have limited value in measuring the patient's state and change. However, the collected information is limited because the significance of the differences provides neither information on the magnitude of the change, nor on the quality of the tested instrument compared to an external standard (De Vet et al., 2011c; Mokkink et al., 2010b). The ability of an instrument to discriminate between groups and between measurement times is merely a prerequisite of a responsiveness evaluation, and is also sometimes

classified within hypothesis testing for the evaluation of construct validity (Mokkink et al., 2010e).

1.1.3.4.3.2. Effect size

The Cohen's effect size (ES) is a relative and without unit indicator of responsiveness that is calculated as the mean change score in a group, divided by the pooled standard deviation (SD) (McDowell, 2006). It is thus influenced by the magnitude of the change and the variance in scores. An effect size of ≤ 0.20 represents a small, 0.50 represents a moderate and ≥ 0.80 represents a large change (Portney and Watkins, 2015; Husted et al., 2000). However, these values should not be considered as standards applicable regardless of the context, as is discussed hereafter.

The criteria for the qualification of effects sizes are useful to provide an insight about the efficacy of a treatment. For example, a statistically significant difference compared to baseline status might have limited clinical interest if the ES is small or lower than that of a concurrent treatment. The comparison of the effect sizes of several measurement instruments that measure the same construct in the same conditions is also instructive about their respective sensitivity to capture the change that happened. The most responsive instrument will have a higher ES than the other ones for the measurement of the same phenomenon. Conversely, the ES of a single instrument or the longitudinal comparison of effect sizes across several studies has little interest for the evaluation of measurement properties, because it is as much influenced by the treatment effect as by the quality of the instrument (Angst, 2011; De Vet et al., 2011c).

The determination of the ES is also of interest to calculate the sample size that is required for a study. In general, the higher the effect size, then, the smaller the sample size need to be in order to reach the desired study power (generally 0.80) (Portney and Watkins, 2015; McDowell, 2006).

1.1.3.4.3.3. Standardised response mean

The standardised response mean (SRM) is based on a statistical approach that is close to that of the ES. Its calculation is based on the mean change score in a group, divided by the SD of this change. As a ratio of change relative to the standard deviation of the change in scores, it is thus influenced by the variability in the degree

of change, rather than by the sample's degree of homogeneity both prior to and after an intervention (as the ES is). Cohen's criteria for small, moderate and large effect sizes apply for this index as well (Portney and Watkins, 2015).

1.1.3.4.3.4. Correlation between change scores

When a reference instrument exists, the correlation between the change score measured on this reference and on the tested instrument can be used as an indicator of responsiveness. A significant correlation means in this case that the sensitivity to change of the tested outcome measure is related to that of an outcome measure that is known to be responsive for the same construct. As would be expected for a construct validity evaluation, *a priori* relevant hypotheses should be formulated on the level of correlation, although in this case, evaluations should be focussed on change scores (Mokkink et al., 2010b). The magnitude of the correlation amongst change scores is generally lower than the correlation between scores at a given timepoint, because each measurement has a certain degree of measurement error (De Vet et al., 2011c).

A limitation of correlation amongst change scores lies in the fact that frequently, no gold standard exists for a measurement. A solution is to measure the change correlation simultaneously alongside another previously validated instrument that aims to measure the same construct, though it might not perfectly measure it. A good correlation demonstrates that the measured change of the tested instrument is related to that of an instrument that has previously demonstrated to be responsive (De Vet et al., 2011c). However, the degree of correlation will be relative to the reference instrument only, and it might even tend to decrease when the tested instrument is more responsive than the reference instrument. An alternative would be to calculate the change correlation in comparison with a global rating scale (GRS). At follow-up, patients are then asked in a single question, to indicate how much they have changed on the construct of interest. However, the reliability and validity of such retrospective measures of change is debated (De Vet et al., 2011c).

1.1.3.4.3.5. Receiver operating characteristic curves

A receiver operating characteristic (ROC) curve can be used to evaluate the responsiveness when the gold standard is a dichotomous variable. In this context, the

curve illustrates the trade-off between sensitivity and specificity for the classification of patients as improved or non-improved. The specificity (that is, the probability of the measure correctly classifying patients who do not demonstrate change on the external criterion, in this context) and sensitivity (i.e. probability of the measure correctly classifying patients who do not demonstrate change on the external criterion, in this context) can also be assessed for each score value, and the optimal detection threshold (cut-off presenting the highest sensitivity-specificity ratio) can also be determined using a ROC curve (McDowell, 2006; Husted et al., 2000).

The area under the ROC curve (AUC) is used to measure the ability of an instrument to discriminate between participants who are considered to be improved and those who did not improve, according to the gold standard (De Vet et al., 2011c). An AUC value of 1 would be found for an instrument that perfectly discriminates improved from non-improved participants, and a 0.50 value for an instrument that would not help discriminate amongst them at all. AUC values of 0.6 to 0.7 represent thus poor accuracy, 0.7 to 0.8 fair, 0.80 to 0.90 good and >0.90 excellent accuracy (Pines et al., 2012; Terwee et al., 2007; De Vet et al., 2011c). A score of 0.70 is usually considered appropriate (De Vet et al., 2011c; McDowell, 2006; Jimerson, 2007).

A disadvantage of the ROC curve analysis is that the external clinical change score must be dichotomised between improved and unimproved. So, the information, which is provided about the magnitude of change by the external criterion, is lost in the process of dichotomisation (Husted et al., 2000).

1.1.3.4.3.6. MCID/MCII

Many measurement properties are determined based on statistical calculations that do not account for the patient's point of view. It might happen that a treatment makes a significant difference from a statistical point of view while the patients consider that the treatment effect is not large enough to induce a meaningful change for them. Conversely, the Minimal Clinically Important Difference (MCID) is a measurement property that indicates from beyond which pre-post treatment difference, the change of his/her state is meaningful for the patient (Michener, 2011; de Vet et al., 2006a).

MCID includes patients who improved and patients who worsened, though the extent of change that patients consider clinically important is not the same in these two

populations. Thus, the concept of MCII (Minimal Clinically Important Improvement) is more specific, as it provides information about the magnitude of the improvement on the scale expected by the patient, for the treatment to be considered as valuable by him (Tubach et al., 2012).

There is a controversy about the best method to use for the determination of MCII/MCID (Tubach et al., 2005c). This is problematic as the use of different methods leads to the determination of varying MCII/MCID values (Beaton et al., 2011). Several distribution-based methods, which are related to a distribution of scores and several anchor-based methods, which use an external criterion to define clinical importance, have been used to define important change (Portney and Watkins, 2015). Anchor-based methods are generally preferred, as they imply that what is considered as minimally important has previously been defined (de Vet et al., 2006a; Tubach et al., 2012).

An example of a calculation process on which a consensus has been reached for the determination of MCID/MCII (Tubach et al., 2007) is presented hereafter. The definition of MCID/MCII implies discriminating the patients who improved from those who remained unchanged and those who worsened, using a simple question. Focusing attention on only patients who report improvement, those patients are then asked to rate the importance of the improvement on a Likert scale that uses standardised wording. The MCID/MCII is then calculated based on the 75th percentile of those who consider themselves as at least slightly improved, as reported by themselves on the Likert Scale (Tubach et al., 2005c).

The MCID/MCII is relative to the population in which it was calculated (Schuller et al., 2014; King, 2011). It must be larger than the MDC to be considered as valid, as it would be contradictory to define a value that is supposedly important but is below the change detection threshold for an individual patient (van der Linde et al., 2017; De Vet et al., 2011a).

1.1.3.4.3.7.PASS

Another measurement property that accounts for the patient point of view is the Patient Acceptable Symptom State (PASS). Despite the effect of the treatment, it might happen that the change is not sufficient for the patient to think that the level of

symptoms is sufficiently satisfying for him/her to feel well. The PASS indicates from which value the patients estimate that the result is acceptable, according to their standard (Tubach et al., 2005a). This is important information to fix treatment objectives or for deciding about the continuation of therapy, for example.

As for the MCID/MCII, the best approach for this measurement property's use has been debated. Relying on an established consensus, the PASS is based on the calculation of the 75th percentile among patients who report an acceptable level of symptoms (Tubach et al., 2007; Tubach et al., 2005c; Tubach et al., 2005b).

1.1.3.4.4. Synthesis on clinimetrics

A considerable quantity of information needs to be generated before it can be asserted that a new instrument has undergone an exhaustive validation process. This has important implications for this thesis, as it should be anticipated that there will be an investigation of many of the expected properties within its component related research projects, including the clinical validation projects that will aim at the investigation of many relevant aspects of the validity, reliability and responsiveness of the B-B Score, using a smartphone and a dedicated IMU system (Chapter two: Optimisation of scoring procedure and measurement method development; Chapter three: devices' comparison; Chapter four: B-B Score measurement properties study). Consecutively, there will be a literature review project that will compare the newly investigated measurement properties of the B-B Score with those of well-established contemporary PROMs (Chapter five: literature systematic review challenging the measurement properties of patient-reported and movement analysis-based outcome measures for shoulder function evaluation).

Despite efforts made to standardise the approaches to determine the measurement properties, it has been frequently mentioned above that there are controversies about the appropriate methods to use. This is problematic for users of assessment tools and researchers because the comparisons of results between studies may be subject to amplified caution, due to the dependency of results on the applied and potentially idiosyncratic methods of calculation.

Moreover, it was also frequently mentioned within this review of introductory concepts that results are population-dependent. This implies that any new measurement

method proposed within this thesis will have to be validated separately for several populations of patients with shoulder complaints and that similarly, the measurement properties of shoulder function will have to be analysed separately for each population of interest within the systematic literature review. Though the population dependency of results is largely acknowledged, no precise definition of the degree of similarity amongst populations has been found. The populations considered in validation articles in the literature offer very diversified characteristics, for example a common pain location (e.g. shoulder pain), specific shoulder condition (e.g. shoulder instability), treatment approach (e.g. shoulder surgery/conservative treatment), stage of treatment, and so on. Therefore, it is difficult for users to estimate to what degree a particular set of results apply to their specific situation of interest. A corollary of this is the importance of studies that compare the results from several instruments within the same population, to allow for a benchmarking. The latter aspects will thus be taken into consideration in this thesis.

1.1.4. Practical issues

While the quality of measurement properties is fundamental in order to guarantee the soundness and trustworthiness of measurements, some validated outcome measures may be rarely used due to practical barriers. It is thus important to consider accessibility, cost, feasibility and interpretability at the outcome measure's stage of inception, to account for the fact that most measurements are realised within contexts in which time, cost and burden matter (Valderas et al., 2008).

1.1.4.1. Accessibility and cost

Accessibility and cost are frequently related. The accessibility of questionnaires can easily be handled today, by presenting them within the original publications and/or on dedicated websites. Numerous questionnaires are thus immediately accessible at no cost. However, their unrestricted use may be limited by existing copyrights in some cases. Their access is conditional on a simple request for authorisation or a payment, as the case may be.

The accessibility to measurement devices is more problematic because it implies a physical access to the device. The accessibility may be limited by the cost of the device or the absence of availability and its diffusion throughout the countries of

potential use. When the device is not transportable, the access to the device location may also be complicated for the participants, especially for people with reduced mobility.

1.1.4.2. Practicalities

If a measurement instrument is accessible, practicalities enter then into consideration. Time, number of items or steps, administrative burden, complexity of instructions, availability of language-translated versions may be barriers to routine questionnaire use (De Vet et al., 2011d). Specifically in relation to measurement devices, issues of maintenance, breakdowns, compatibility and obsolescence also enter into consideration. Some instruments are straightforward to use while others require training before they can be used by patient and/or professionals.

1.1.4.3. Interpretability

Every measurement instrument produces by essence a result, but the meaning of this result may not be straightforward to understand. The outcome measure's interpretation, that is "the degree to which one can assign qualitative meaning to an instrument's score or change score" is thus an important characteristic to consider (Mokkink et al., 2010e; De Vet et al., 2011a). Ideally, a result should be readily available and interpretable by the user for clinical use. More complex data management is possible in a research context.

The ability of an outcome measure to be readily interpreted and placed in context (interpretability) relies also on the prior determination of several measurement properties that have been previously presented (sub-sections 1.1.3.4.2.1 *ff* p. 42 - 50), for example, in the defining of MDC, MCID/MCII, PASS, LoA and bias, floor/ceiling effect and the determination of a normal performance (De Vet et al., 2011a). Importantly, these properties must have been established in the population of interest for the interpretation to be meaningful in a given context.

1.1.4.4. Implication of practical issues for the thesis

The objectives of this thesis centred not only on developing and testing an innovative measurement method, but also on ensuring that practical issues do not hinder the routine use of this instrument. The quality of the measurement properties are nevertheless of prime importance, as a practical but otherwise invalid measurement would be useless.

A preliminary step along the procedure simplification has been accomplished by the development of the B-B Score, which includes only two upper limb movements. This score has been developed using inertial sensors that are much less cumbersome to use than laboratory-based devices for movement analysis, but are still not easily accessible and affordable for clinicians. Therefore, there is an intention to test to what extent a smartphone might replace inertial sensors for the measurement of the B-B Score in patients. Accordingly, a concurrent evaluation of the measurement properties of assessment approaches using inertial sensors and smartphones was envisaged. Importantly, in the scenario where the measurement properties are deemed equivalent between the two approaches, the smartphone will inevitably be considered superior, due to its greater practicality.

The overview of the planned thesis process issued from the notions presented in sub-section 1.1.3 “Definition of central concepts” p. 14 - 51 and section 1.1.4 “Practical issues” p. 51 - 53, within this Chapter is available in Figure 1.4

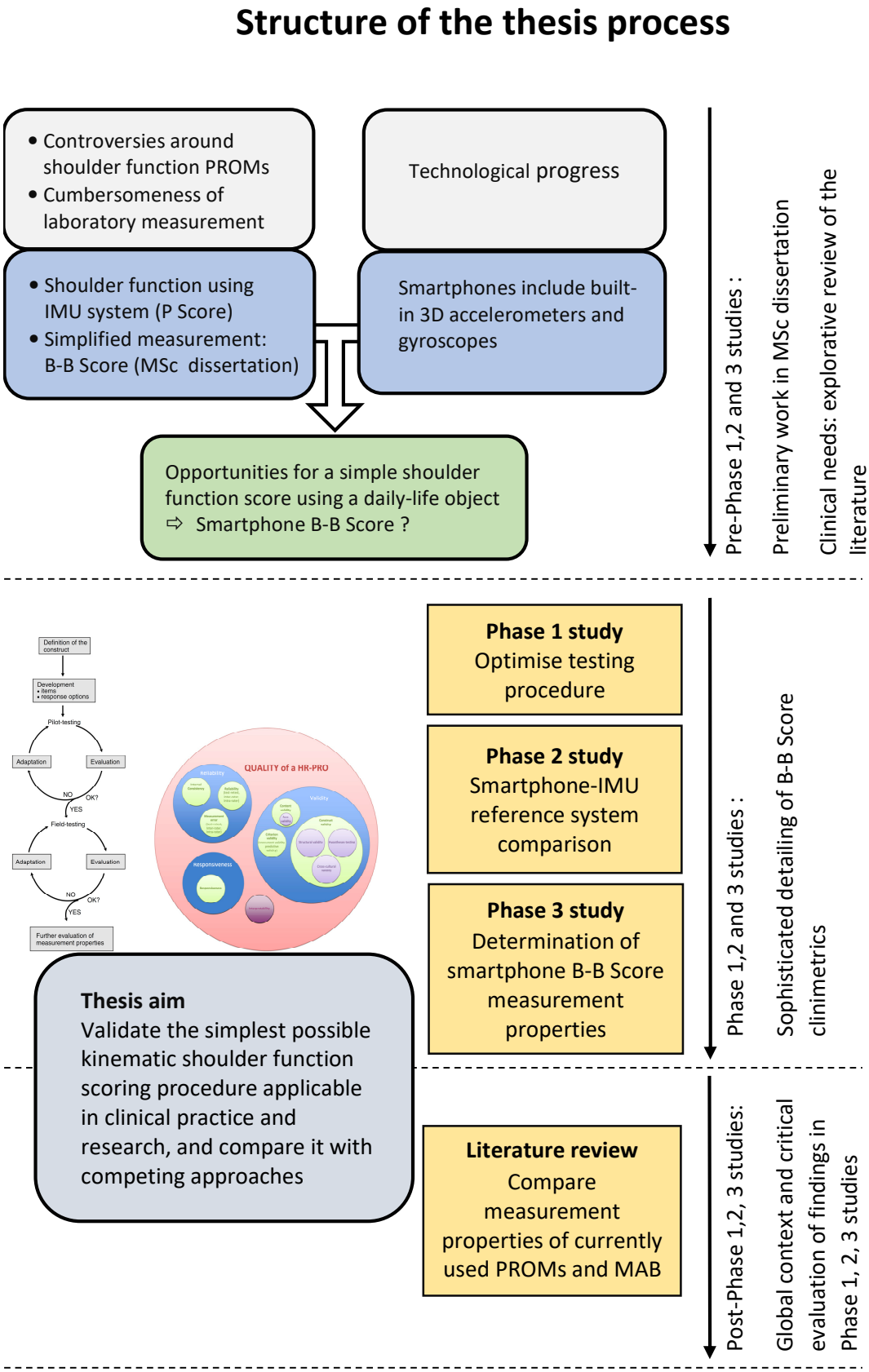


Figure 1.4: Overview of the planned thesis process

1.1.5. Potential impact of the results

1.1.5.1. Scientific significance of results

This study might contribute to the further use of movement analysis methods in clinical research and even in clinical practice. This could possibly facilitate routine application of more efficient measurement approaches for delivering objective outcomes of shoulder treatment in physiotherapy, surgery and rehabilitation.

The project is related to the latest technological development in embedded movement analysis systems. The conjunction of the simplification of testing procedure with the recent development of wireless ambulatory movement analysis systems, makes measurement much easier to perform, while keeping sound measurement properties.

Complexity of technology of movement analysis systems, time in setting them up, training in their correct use and their high cost have prevented their routine application within movement analysis to date. However, several of these barriers can probably well be overcome with the progress of wireless technology, lessening of the cost of electronic componentry and the development of user-friendly software. Typically, most middle segment smartphones are fitted with built-in accelerometers and gyroscopes, which makes technology that had previously been confined to use within leading scientific laboratories, accessible to almost everyone nowadays.

In terms of future developments, applications within telemedicine may also be envisaged for a patient's routine follow-up and surgery complications detection, as the testing procedure is quite simple to execute and has been well tolerated by patients (Jolles et al., 2011). Therefore, the study topic anticipates possible future developments in healthcare.

1.1.5.2. Significance for health professionals

The reliability of measurement methods is of importance for physiotherapists and medical doctors. Effective methods are needed to evaluate if therapeutic interventions are economical and efficient (LaMal, 1994). The development of evidence-based practice also relies on efficient measurement tools.

As stated within this introduction, shoulder function measurement remains a controversial issue. Researchers and clinicians have to face the dilemma of attempting the selection of a measurement tool in the absence of a gold standard. This situation has ongoing consequences on the health professionals' capacity to produce therapeutic evidence of treatment effectiveness in shoulder conditions (Green et al., 2003; Harvie et al., 2005). Undetermined validity and a proliferation of outcome measures contribute to the deficit in scientific evidence supporting some shoulder physiotherapy treatments (Green et al., 2003; Harvie et al., 2005; Page et al., 2015). Therefore, there is a need for research to provide clinicians and researchers with extensively validated and convenient measurement tools.

1.1.5.3. Significance for patients

Improvements in the quality of measurement tools is of interest for the patient, as important decisions concerning him or her are taken on the basis of outcome measures. Quality of outcome measurement influences fairness and equity of decisions toward patients. For example, the decisions to continue or stop the patient's treatment, or for him/her to return to work, are linked to measured functional outcome. Therefore, validity and reliability of measurement is a prerequisite for fair decision-making concerning the patient. Correct evaluation also contributes to the allocation of relevant resources according to patients' needs.

Consequently, trustworthy and straightforward measurement methods are needed to assist clinicians and clinical scientists in their decisions concerning patients.

1.1.5.4. Significance of results for clinical partner

Shoulder conditions are frequently encountered in orthopedic practice. Around 250 patients attend a medical consultation every month at the Département de l'Appareil Locomoteur (DAL), due to shoulder conditions. Therefore, the development of valid, reliable and convenient functional outcome measurement methods is of primary interest for the department. This is in direct relationship with the ambitions of the present study that aims to validate a straightforward measurement method. The score could potentially be integrated within the routine patient assessment procedure of the specialised medical consultation.

Around 15 - 20 patients suffering shoulder conditions are also treated every month in the physiotherapy department of the DAL. The developed measurement method is therefore also of interest in this field. It could contribute to patient follow-up and development of evidence-based practice in physiotherapy. As part of a university hospital, the physiotherapy department has the mission to participate actively in research. The project is a contribution to the fulfilment of this mission.

The DAL-CHUV has been active for more than ten years in the development of clinical evaluation using ambulatory measurement analysis, in partnership with the Laboratory of Movement Analysis and Measurement of the EPFL. The present project plays a strategic role in the pursuit of this long-term research orientation.

1.1.6. Study resources and implementation

Paradoxically, the validation of a simple kinematic approach to measure shoulder function involves a complex multistage process that rely on a great variety of resources and competencies. This includes methodological and statistical guidance, technological support, patient access and funding access.

Besides the resources available at Queen Margaret University (QMU), the required resources were accessible in the candidate's environment. In addition to his MSc study on the simplification of kinematic shoulder scores, which is related to this thesis, the author has had previous opportunities to collaborate in several projects of the Laboratory of Movements Analysis and Measurement of the Swiss Institute of Technology (LMAM-EPFL) that were related to shoulder function analysis (Coley et al., 2007a; Coley et al., 2008b; Duc et al., 2013; Duc et al., 2014; Jolles et al., 2010; Jolles et al., 2011; Pichonnaz et al., 2015b).

In conjunction to his main 70% employment as an assistant professor in the Physiotherapy Department of the Haute Ecole de Santé Vaud, a school of the University of Applied Sciences of Western Switzerland (HESAV//HES-SO), the author also worked as a clinical specialist physiotherapist at a 30% employment rate in the Physiotherapy Service of the Department of Musculoskeletal Medicine of the University Hospital of Lausanne (DAL-CHUV). Therefore, access to the required methodological, technological and clinical competencies was available in the author's

environment. The author is grateful to both of his employers for having agreed to provide partial time and financial support towards the completion of this thesis.

Thus, it was reasonable to have expected that the resources accessible through the author's work environment and the network of clinical colleagues of the applicant were compatible with the thesis' requirements.

The access to the required patient population was possible through the author's position at DAL-CHUV. An arrangement was concluded with the medical doctors in charge of the specialised shoulder consultation within the hospital and also, with the physiotherapy department of the hospital. The author's position as a staff member in the physiotherapy department was also useful to get the involvement of several colleagues who facilitated the delivery of the numerous clinical tests required in the validation process.

Through his work at HESAV, the author was entitled to apply for research funding to the Swiss National Science Foundation (SNF). At the time of the thesis' conception, an access to a launch fund of the SNF (DORE fund), dedicated to the development of research in universities of applied sciences, was possible.

First, a funding was obtained from the HES-SO University of applied sciences "RéSAR" fund to support the preparation work for the submission to the SNF DORE fund (Ré-Sa-R 17-10) (Appendix II). Then, a successful application for funding was made to the SNF DORE fund (SNF n° 135061). This ensured the financing of the clinical research (Phase 2 and 3) and allowed to consolidate the agreements between the research partners (HESAV, CHUV and EPFL) (Appendix III and URL http://p3.snf.ch/project-135061).

The Phase 1 study project was submitted and approved by the Ethical Commission of the Faculty of biology and medicine of the University of Lausanne (Protocol 205/10) (Appendix IV). An amendment to the original protocol was accepted to adapt the details of the Phase 2 and 3 protocols to the conclusion of the Phase 1 study (Appendix V).

The study was declared on the ClinicalTrials site (N° NCT01281085) (Appendix VI and URL https://clinicaltrials.gov/ct2/show/NCT01281085). It is required that clinical trials are registered to prevent selective reporting, identify publication bias caused by

unpublished negative results and avoid unnecessary duplication of trials (Costa et al., 2012). This step is now a prerequisite to publication in most physiotherapy and medicine journals.

CHAPTER TWO

OPTIMISATION OF SCORING PROCEDURE AND CALCULATION

2.1. Introduction

2.1.1. Phase 1 study general context

Some preparatory work had previously been done by Coley et al. and by the author in his MSc dissertation to design and simplify a relevant procedure for the kinematic evaluation of shoulder function (Coley, 2007; Coley et al., 2007a; Pichonnaz et al., 2015c). These works led to the proposal for a two-movement score based on the side-to-side comparison of a power-related metric, as an adequate approach for the measurement of shoulder function.

Nevertheless, several issues needed to be evaluated in the initial phase of the project to define the optimal testing procedure for the B-B Score that was to be used in the main measurement properties study. The Phase 1 study was also necessary to test the applicability and acceptability of the research protocol for patients and colleagues, as well as testing for any implementation issues (patient recruitment process, partners contribution, administrative process, database implementation, burden and practical issues) (Thabane et al., 2010). It also aimed at training all collaborators in the correct use of the measurement instruments (inertial sensors, smartphones and clinical questionnaires) and towards mastering the study protocol. Finally, the database was implemented and tested at this stage of the project.

2.1.2. Technical issues to explore in the Phase 1 study

Though previous studies had given a promising insight into the measurement properties of the P Score, and by extension to those of the B-B Score that predicts 97% of the P Score from which it is derived, several issues were still needing to be considered in order to optimise the measurement procedure. Notably, it was shown that the B-B Score had an excessive variability for single measurements, with LoA with the P Score, taken as a reference, reaching up to ± 21.6% at 6 months post-surgery (Pichonnaz et al., 2015c). This implies that the measured result on the B-B Score of a patient might occasionally differ by more than ± 20% from the performance measured by the P Score. It was thus necessary to explore approaches that could potentially reduce single measurement variability.

As the variability and error in the mean score of several measurements decreases with the square root of the number of repetitions (assuming a normal distribution of errors), it was thought that test replication and averaging over repeated intra-individual trials may decrease the possible variability in individual measurements (Winer, 1991; Gleeson and Mercer, 1996; Pichonnaz et al., 2015c). One pathway to explore was thus to take advantage of the simplicity of the B-B Score procedure to acquire the mean score from a series of several intra-individual scores for the two movements, which should decrease the B-B Score's statistical variability. Further exploration to see whether or not the B-B Score would be more stable by taking the median or the mean of the score replications was then required. Concomitantly, it would be necessary to verify if the above approach would be fully applicable, by investigating systematic intrusions from carry-over effects such as fatigue, warm-up or learning effects, during repeated intra-individual trials. The Phase 1 data would thus be used to evaluate the influence of the repetition number on the B-B Score's reliability and to get a first insight into its measurement properties.

It was also possible that a reason for the variability observed in previous studies was linked to the calculation method used for the determination of P and B-B Scores. These scores were based on the computation of the product of accelerations by angular velocities, to obtain a power-related metric [(deg/s)*(m/s^2)] (Coley et al., 2007a; Coley, 2007; Pichonnaz et al., 2015c). The values used for the calculation were determined by the whole range of accelerations and angular velocities during the measured movement, calculated for each axis and added to obtain a power-related parameter called Pr. This approach corresponds to the calculation for each dimension of the surface of rectangles that would circumscribe the curve representing a whole range of measured values (Figure 2.1). This parameter was then compared between the healthy and the painful side.

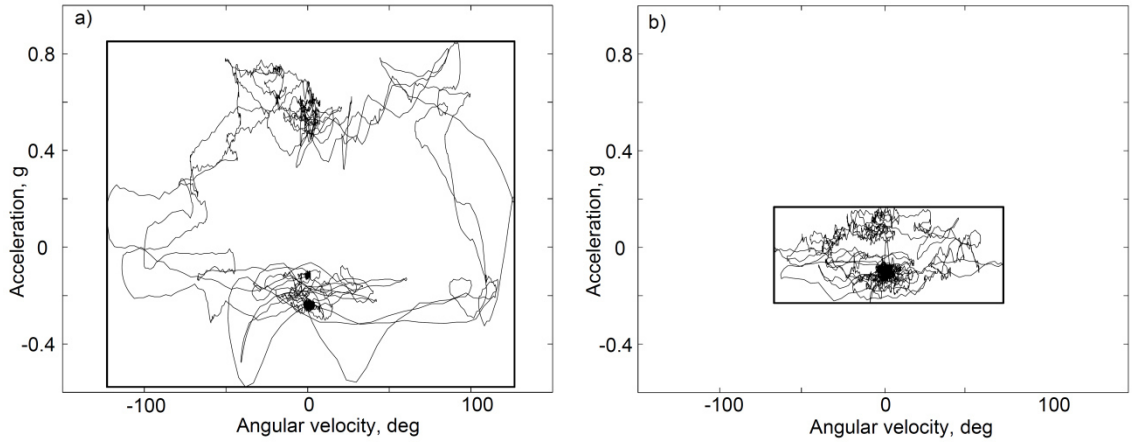


Figure 2.1: Humerus acceleration as a function of its angular velocity for the patient. a) The trace represents the humerus acceleration vs. angular velocity for the healthy side. b) The trace represents the humerus acceleration vs. angular velocity for the affected (painful) side. The rectangle that circumscribes the curve corresponds to the product. From: COLEY, B., JOLLES, B. M., FARRON, A., BOURGEOIS, A., NUSSBAUMER, F., PICHONNAZ, C. & AMINIAN, K. 2007. Outcome evaluation in shoulder surgery using 3D kinematics sensors. *Gait Posture*, 25, 523-32.

It was suspected that calculating a rectangle, of which surface is markedly influenced by the maximal and minimal peak values, instead of calculating the effective area inside the curve of measured values that is less influenced by peak values, might increase the variability. Thus, the variability taking the range or the effective area of measured values was compared in the Phase 1 study using both methods, with the support of the engineers of the LMAM-EPFL. The surface that is taken into consideration using the area calculation method is shown in Figure 2.2.

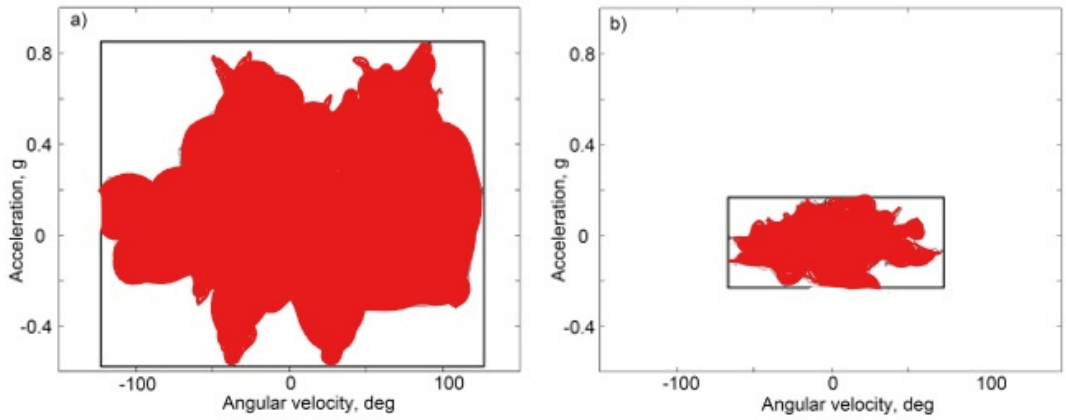


Figure 2.2: Surface (red area) taken into consideration when using the area calculation method for the B-B Score.

Conversely, no new try-outs were conducted to explore alternative testing procedures and technical features for the score measurement. Exploratory measurements that had been previously conducted at the conception stage of the P Score had shown no advantage in modifying the speed of movements or adding weights compared to a spontaneous movement at a self-selected speed within the pain free range of motion. Similarly, the technical features remained unchanged, as they had proven adequate in the previous studies. Thus, for the Phase 1 study, the sensors were placed as they had been in previous studies and the accelerations and angular velocities were amplified and low-pass filtered at a cut-off frequency of 17 Hz to remove any electronic noise, before being recorded by a Physilog data-logger (Physilog®, Gait Up, CH), at a 200 Hz sampling frequency (Coley, 2007).

Preliminary study try-outs had shown that the influence of errors in sensor' measurements was negligible in the study context, with an offset < 0.005g and static drift of 0.0038g over 5 minutes of measurement, with a maximum error < 0.028g (Pichonnaz, 2009). Therefore, the sensors accuracy was not re-tested in the Phase 1 study. Importantly on this aspect, the accelerometers and gyroscopes provided a direct measurement of accelerations and angular velocities, and thus the possible errors would not have been amplified by mathematical transformation (Luinge et al., 2007; Aminian and Najafi, 2004).

No smartphone measurements were collected at this initial stage. The tests were performed using the IMU approach only, which had previously been recognised as

relevant. Conversely, it would not have been possible to differentiate amongst the sources of variability using the smartphone approach, as the application was still under development at this stage of the project.

Another issue to explore before the start of the Phase 2 and 3 studies was the suitability of the B-B Score for the measurement of conservatively treated conditions, as the score had been developed based on surgically treated patients only. The promising results of the development study might not have been transferable to conservative treatment as measurement properties are context and population-dependent (Robertson et al., 2017; Collins and Roos, 2016; Riddle and Stratford, 2013). For example, the discrimination between patients and heathy controls might have been more difficult, or a ceiling effect might have been observed, as non-surgical patients are supposedly less severely affected than surgically treated patients are. Due to the small sample size, only the ability to discriminate patients from healthy controls was tested at this stage and no subgroup analysis based on pathologies, had been planned.

2.1.3. Aims

The aims of the Phase 1 study were:

- to define the optimal testing procedure for the subsequent Phase 2 and 3 studies, including the number of B-B Score replications
- to compare the respective advantages of an alternative score computation method (area calculation) to the original method (range calculation)
- to test the feasibility and applicability of the study protocol
- to test the organisational issues of the study's implementation

2.2. Methods

2.2.1. Study sample

A prospective cohort study was conducted at the Department of Traumatology and Orthopaedic Surgery of the University Hospital of Lausanne. Ethical approval was granted by the Human Research Ethics Committee of the Canton of Vaud (CER-VD), protocol number 205/10. Patients gave their signed informed consent for participation

in the study. The patient information sheet and consent form are available in Appendix VII that includes the complete baseline patient file. The study was registered under Clinicaltrials.gov Identifier NCT01281085 (Appendix VI).

The inclusion criteria were to be a > 18 year old adult and to present with one of the following shoulder conditions, as recorded during their first medical consultation at the specialised shoulder consultation unit of the hospital: rotator cuff condition, instability, adhesive capsulitis, and proximal humerus fracture.

Patients with various shoulder pathologies were included to test the applicability of protocol in the same populations of interest than in the subsequent Phase 2 and 3 studies. For the rotator cuff condition, instability or capsulitis, patients were selected who required only conservative treatment. As the B-B Score had previously been validated after rotator cuff and arthroplasty surgery (Pichonnaz et al., 2015c), it was of interest to explore its validity in different populations. Surgical and conservative fracture treatment were included in the same group as the expected progress and functional prognosis is similar in both populations (Handoll et al., 2012).

Exclusion criteria were bilateral shoulder conditions, any concomitant pain or condition involving the upper limb or cervical spine, medical contraindication to execute movements required for score completion, tumour, neurological condition interfering with the test and an insufficient local language level to give truly informed consent or to understand questionnaires. It was also required to proceed to a Mini Mental State score if a decrease in cognitive function was suspected, with exclusion criteria at 24 points/30 (ANAES, 2000).

The patients corresponding to the study criteria were recruited based on the notification to the thesis' author by the medical doctors in charge of the specialised consultation of the hospital-based patients. Following this first contact through the doctor, a telephone call was made by a PT collaborator of the research team to those who had previously agreed to be contacted about this study, as indicated by the doctor. Those who accepted to participate received then detailed information and an appointment time was arranged with them. The opportunity to ask for further clarifications was afforded to them before signing the consent form at the beginning of the measurement session.

A group of participants without history of shoulder condition/pain was also recruited within the professional environment of the applicant to evaluate the performance in a healthy population and the stability of the score. Inclusion criteria were to have no present pain or shoulder condition/pain on any side and to be < 50 years old. Exclusion criteria were any past or present shoulder condition/pain and to be > 50 years old. Those who accepted to participate followed then the same procedure than the patients.

At this stage, the sample size was determined based on a pragmatic approach determined by the needs of the study. It was estimated that around 15 patients and 5 healthy controls would be sufficient to reach the study objectives, which were to refine the testing procedure and test the study protocol without making further inferences about the score precise measurements properties. Nevertheless an *a priori* sample size calculation was conducted based on the results of the score development study to ensure that the statistical power was sufficient to test the discriminative power of the B-B Score, which is the most basic measurement property that is expected from an outcome measure. A score that would not be able to discriminate a pathological from a healthy group would very likely be useless.

This calculation showed that, considering that the patients of the Phase 1 study would reasonably have a performance approaching either the 3 or 6 months post-surgical state, in-between 4 and 5 patients per group were needed to reach a 0.80 power for the difference between the patient and the control groups, for a p value at $p < 0.05$ [Patients at 3 months mean (SD) 61.8 (16.8); Patients at 6 months mean (SD) 69.0 (15.9); Controls mean (SD) 102.9 (14.5)] (Pichonnaz et al., 2015c; Soper, 2004). A larger sample was included to acquire the required experience for the purpose of the subsequent Phase 2 and 3 studies and get estimations of the score properties with reasonable confidence intervals.

Patients residing in the canton were contacted by phone in the order in which they attended the medical consultation in the department, upon notification of eligibility by the consulting doctor. The patients underwent the baseline measurement session within two weeks following medical consultation, with the exception of patients with humerus fracture. Measurements were performed 6 weeks post-stabilisation for patients with humerus fracture, provided that the radiological control showed normal consolidation.

2.2.2. Measurement device

The system for body-worn movement analysis was a Physilog system composed of two inertial sensors modules and a datalogger system (Physilog®, Gait Up, Lausanne Switzerland). Each inertial sensor module included three dimensional accelerometers and gyroscopes (Accelerometers: Analog device, ADXL 210, ± 5 g, precision: ± 0.2% of Full Scale; Gyroscopes: Analog device, ADXRS 250, ± 400 deg/s, precision: ± 0.1% of Full Scale). The device resolution was 16 bits and the sampling frequency was 200 Hz. An inertial measurement system was used, preferably over concurrent measurements analysis systems, because inertial sensors provide direct measurements of angular velocities and accelerations used in the score calculation. As previously mentioned, preliminary try-outs had shown that the influence of measurement errors (offset, sensitivity or drift) was negligible in the study context. This try-outs had shown that the errors in sensor' measurements were < 0.005g for offset and < 0.0038g for static drift over 5 minutes of measurement, with a maximum error < 0.028g (Pichonnaz, 2009). The magnitude of the error was thus minor compared to the magnitude of the within group and between group differences observed in the previous study relying on the B-B Score (Pichonnaz, 2009).

2.2.3. Measurement procedure

The inertial sensors modules of the reference system were placed on each humerus, 3 cm above the midpoint of the line connecting the lateral epicondyle (EL) and medial epicondyle (EM). The sensor's axes were aligned to the anatomical frame of the humerus following the ISB recommendations (Wu et al., 2005; Coley et al., 2009): Yh on the line connecting the glenohumeral (GH) joint and the midpoint of EL and EM, pointing to GH; Xh on the line perpendicular to the plane formed by EL, EM and GH, pointing forward; Zh on the line perpendicular to Xh and Yh, pointing to the right (Figure 2.3). Similarly to previous work, angular velocities and accelerations in the sensor frame have been used to calculate the B-B Score (Pichonnaz et al., 2015c; Coley et al., 2007a).

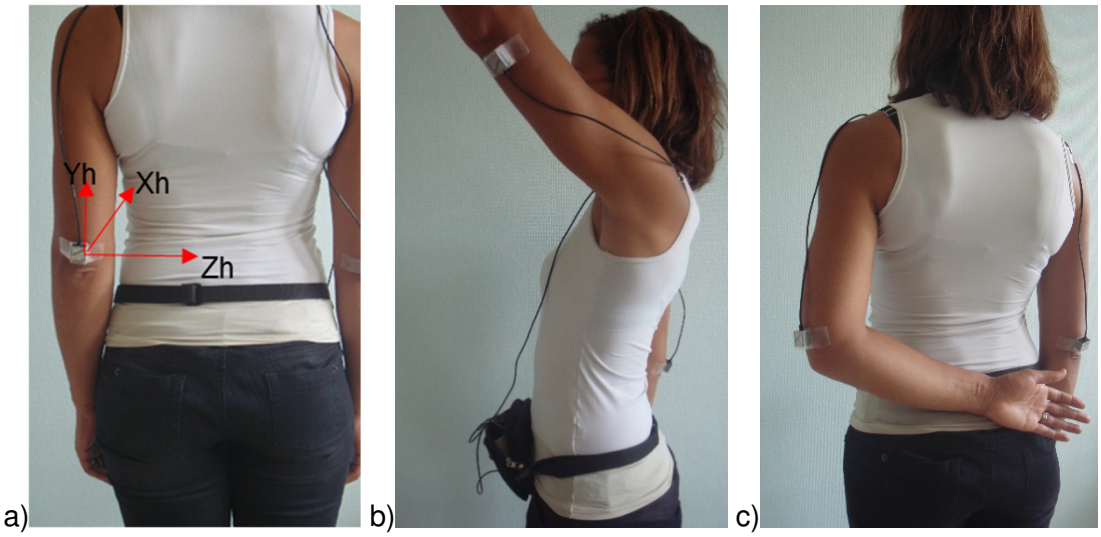


Figure 2.3: Inertial sensors placement and axes (a) The inertial sensor module (Physilog® reference system) attached to the arm with medical tape and connected by cable to the datalogger carried attached around the participant's waist. (b) Test completion of "hand to the ceiling" (c) Test completion of "hand to the back".

After setting-up of the system, the participants watched a video-recorded demonstration of the execution of the B-B Score. They were instructed to do the movements in the pain free ROM, at their self-selected speed and in their natural way. The starting position was the arm alongside the body, in a relaxed position. Movements were executed in a standing position following the rater's instructions. The patients undertook five repetitions of the two B-B Score movements on the healthy side (put hand to the back + hand to the ceiling as to change a bulb) and then repeated the task on the pathological side. The controls executed the same procedure beginning with the dominant side.

The measurement procedure was repeated twice alternating between two raters (the author and a physiotherapist colleague, previously trained in B-B Score methods). The first rater was randomly assigned. The measurement system was detached after each score measurement for inter-rater administration of assessments to account for the variability induced by possible inconsistent sensors' placement in clinics.

2.2.4. Clinical questionnaires

This phase essentially aimed at training all collaborators in the correct use of the clinical questionnaires and at testing the feasibility of the study protocol. The questionnaires were not interpreted at this stage of the PhD as limited useful information could have been drawn from them considering the limited sample size and the diversity of pathologies. Thus, the detailed questionnaires' description has been placed in the chapter in which the understanding of them is of most importance (Chapter five Literature review, p. 187 - 189).

Patient-reported outcome measures on shoulder function, pain and quality of life were also completed. Three PROMs were selected for shoulder function evaluation: the Quick Disabilities of the Arm and Shoulder score (QuickDASH), the Simple shoulder test (SST), the Constant Score and its variation, the Constant relative score (based on a percentage comparison of the measured value to an age- and sex-matched normal populations) (Lippitt, 1993; Constant and Murley, 1987; Fialka et al., 2005; American Academy of Orthopaedic Surgeons, 2009). The Constant Score was undertaken according to the modified guidelines of Constant (Constant et al., 2008). The EuroQol [EQ-5D] quality of life questionnaire and the pain visual analog scale (VAS) were also completed to capture a broader picture of patient clinical state (EuroQol, 2018).

The QuickDASH is an shortened version of the DASH, a self-assessment PROM of the entire upper extremity symptoms and function that provides a whole upper-extremity evaluation, including the shoulder (Hudak et al., 1996).

The SST is as shoulder function PROM that comprises binary 12 items (yes/no), among which two are about function related to pain, seven about function related to strength and three about range of motion (Lippitt, 1993; Beaton and Richards, 1998).

The Constant Score is a composite outcome measure that includes questions on pain and activity, and objective measures of range of motion and abduction strength (Constant and Murley, 1987). The relative Constant expresses the performance as a percentage of the expected value, based on the comparison of the patient's performance to a sex and age matched group (Constant, 1986; Yian et al., 2005; Katolik et al., 2005; Fialka et al., 2005; Constant et al., 2008).

The EQ-5D is a validated generic quality of life PROM that includes 15 items investigating 5 dimensions of the quality of life (mobility, self-care, usual activities, pain/discomfort and anxiety/depression) and a visual analogue scale to record the patient's self-rated health.

The VAS pain scale is a very widely used instrument on which the patient has to rate his/her pain intensity on a 10 cm scale representing the range between "no pain" and "the worst imaginable pain".

The PROMs and socio-demographic questionnaire are available in Appendix VII and a more detailed description of the selected PROMs is available in sub-section 5.2.7.3 "Characteristics of selected shoulder function PROMs", within Chapter five, p. 187 - 189**.**

2.2.5. B-B Score calculation

The B-B Score was calculated according to the method described in Pichonnaz et al. (2015c) and Coley at al. (2007a). A power-related parameter was extracted from the recorded signals: the range of acceleration was multiplied by the range of angular velocity, with a measurement unit of [(deg/s) × (m/s^2)], for each movement. This parameter was calculated for each axis and for each movement of the B-B Score ("hand to the Back" movement and "lift hand as to change a Bulb" movement) and added, separately for each side and for each movement. The ratio of the performance of the affected side relative to the healthy side (or the dominant side relative to the non-dominant side for healthy controls), expressed in percentage, was then calculated for each of the two movements. The values of the movements were then weighted using the equation: B-B Score = 16.71 + 0.32 x hand to the Back + 0.45 x lift hand as if changing a Bulb, based on the observed relationship between the B-B Score and the P Score which was considered as the reference score (Pichonnaz et al., 2015c).

One hundred percent represents a perfect balance in capability between sides and the score decreases in accordance with the severity of functional loss. For example, while a typical healthy person performs near to 100%, the average patient might reach 46% before surgery, 67% at 3 months and 71% at 6 months after surgery.

2.2.6. Feasibility analysis

The recruitment rate was recorded for this Phase 1 study and this would facilitate a projection of the numbers of patients that might realistically be recruited subsequently for the Phase 2 and 3 studies to ensure that it achieves appropriate experimental design sensitivity and statistical power. The latter would be constrained to some extent by the amount of funding that was available to underpin the delivery of the following studies within the thesis (Phases 2 and 3), which was based on a two year recruitment phase plus six months for follow-up.

The ethical, technical, clinical measurement, data management and communicational issues were recorded systematically and this would facilitate evidence-based and on-going adjustments to the protocols that would be used subsequently within the main research studies of the thesis (Phases 2 and 3), as required. The patients were asked to give their impressions about their pain and their difficulties during the testing procedure at the end of the measurement session.

2.2.7. Statistical analysis plan

Descriptive statistics including group mean with standard deviation (SD) and median with interquartile range were calculated for patients' characteristics, for each B-B Score replication and for the mean of each number of replications. Box plots were produced for the visual inspection of data's dispersion.

The significance of the differences between the scores obtained with different methods (range and area approach; mean and median of replications), different numbers of replications (one to five replications), different measurements by the same rater and measurements by different raters were calculated using non parametric tests (Wilcoxon signed rank test for two related samples, Wilcoxon rank sum test for independent samples and Friedman test for more than two related samples). The significance level was set at $p < 0.05$.

The Cohen's *d* effect size was calculated to estimate the magnitude of the differences, where significant differences were found. Effect sizes < 0.20 were considered as small, < 0.50 as medium and < 0.80 as large (Cohen, 1988).

Non-parametric tests were used because the assumption of a normal distribution of data was not expected to be met in a sample of patients with various pathologies that may imply variable levels of alteration of shoulder function.

The intraclass correlation coefficient (ICC) (2,1) was calculated to estimate single measurement reliability and the strength of the relationship amongst measurement replications of the B-B Score (test-retest reliability), between measurements made by the same rater (intra-rater reliability) and between measurements delivered by different raters (inter-rater reliability), respectively. The $ICC_{agreement}$ for a single measurement was recorded, because the absolute agreement for measurements by a single rater is of interest in this study's context (Koo and Li, 2016). The limits of agreement (LoA) and bias using the Bland and Altman (B&A) approach were calculated for intra- and inter-rater reliability, for the mean and the median of the number of replications.

Several criteria had to be met for the score to be considered as sufficiently efficient to employ within the following studies of the thesis (Phases 2 and 3), without any modifications to the manner in which the B-B Score is computed or the protocol for measurement is delivered. The experimental hypothesis for a difference between the B-B Scores for the control and the pathological group should be accepted as indicated by a statistically significant finding. However, group mean differences for measurements made by the same rater, or by different raters should be statistically similar, with retention of the corresponding null hypotheses involving no differences. The ICCs should be ≥ 0.90 for intra-rater and inter-rater reliability (Portney and Watkins, 2015). No *a priori* hypotheses were formulated about the other areas of data analysis due to their exploratory nature. No subgroup analysis was conducted due to the limited sample size.

For the sake of brevity and relevance, the results are hereafter reported following a progressive selection process, based on decisions taken at each step concerning the score optimisation. First, the results related to the choice of the range or area computation methods for the B-B Score will be reported. After a choice has been made based on these results, only the results related to the most suitable computation method will be reported in the next steps. Then, the results related to the choice of the mean or median of number of replications will be reported for the most suitable computation method only, for each number of replications. Finally, the results related

to the choice of the most efficient number of replications will be reported for the chosen computation method only, using only the most suitable method in-between the mean or median of replications.

2.3. Results

2.3.1. Feasibility

The recruitment of the planned number of participants took 6 months to complete successfully.

No test had to be cancelled and no data was lost due to technical or practical issues. The duration of measurement sessions ranged from 45 to 60 minutes, including the completion of questionnaires. The data were used to implement the study database and test the data extraction process that would be used in the Phase 2 and 3 studies.

The collaborations involving HESAV, the Lausanne University Hospital and the Swiss Institute of Technology were implemented as planned. Liaison meetings amongst staff involved in the research were held approximately once per month.

2.3.2. Study sample

Sixteen patients, i.e. one more patient than anticipated in study plan, and seven healthy controls i.e. two more than anticipated in study plan were enrolled in the Phase 1 study. Patients presenting with the four targeted conditions could be enrolled, but patients with rotator cuff conditions represented half of the patients' sample. The participants' characteristics are presented in Table 2.1.

Table 2.1 Participants characteristics in the patient and control groups.

	Patient group	Control group
Sample size, number	16	7
Gender, number male/female	10/6	4/3
Age mean (SD), years	56.2 (8.9)	37.1 (7.5)
Weight mean (SD), kg.	76.4 (17.8)	65.1 (11.5)
Height mean (SD), m.	171.9 (12.2)	171.6 (8.3)
Dominance (right-/left-handed)	15/1	6/1
Affected side (right/left)	15/1	-
Dominance of affected side (dominant/non-dominant)	13/2	-
Shoulder condition (n)	Rotator cuff:8 Instability: 3 Humerus fracture: 3 Capsulitis: 2	

Legend: SD: Standard Deviation; n: number

2.3.3. B-B Score outcomes

2.3.3.1. B-B Score by replication

The mean score with SD and the median score with interquartile range for each replicate of the B-B Score are presented in Table 2.2, and the box plots showing the outcomes for each repetition using the range computation method for the B-B Score are presented in Figure 2.4.

Table 2.2: Mean B-B Scores with standard deviations and median B-B Scores with interquartile range for the patient and the control group using the range and the area computation method, for each score replication (1 to 5).

Mean (SD) **Median (IQR), %**	**Patient (n=16)**		**Control (n=7)**	
	Range	**Area**	**Range**	**Area**
1st replication	65.2 (24.4) 70.2 (40.4-84.4)	68.0 (24.5) 70.5 (45.8-84.6)	102.7 (22.3) 98.7 (87.0-110.4)	107.4 (25.2) 100.7 (87.6-124.4)
2nd replication	66.2 (22.3) 71.2 (44.6-81.7)	69.1 (22.4) 71.9 (51.6-84.0)	100.1 (18.8) 97.0 (83.4-118.1)	103.6 (19.5)* 102.3 (86.1-116.2)
3rd replication	67.3 (24.2) 72.5 (41.8-84.8)	70.0 (22.6) 73.7 (47.3-86.0)	99.6 (17.6) 96.1 (88.4-112.5)	106.9 (22.0)* 102.7 (91.9-120.2)
4th replication	68.4 (23.9) 72.5 (45.2-86.0)	70.9 (23.2) 72.8 (51.3-87.3)	103.8 (18.6) 102.1 (91.4-123.8)	107.1 (23.6) 97.8 (91.0-133.3)
5th replication	68.7 (24.6) 73.6 (42.8-98.2)	71.1 (23.4) 76.6 (47.8-88.3)	97.6 (21.1) 91.1 (77.8-120.3)	106.8 (25.7)* 96.3 (83.0-133.6)

Legend: SD: standard deviation; IQR: interquartile range; * significant difference with B-B Score range computation method

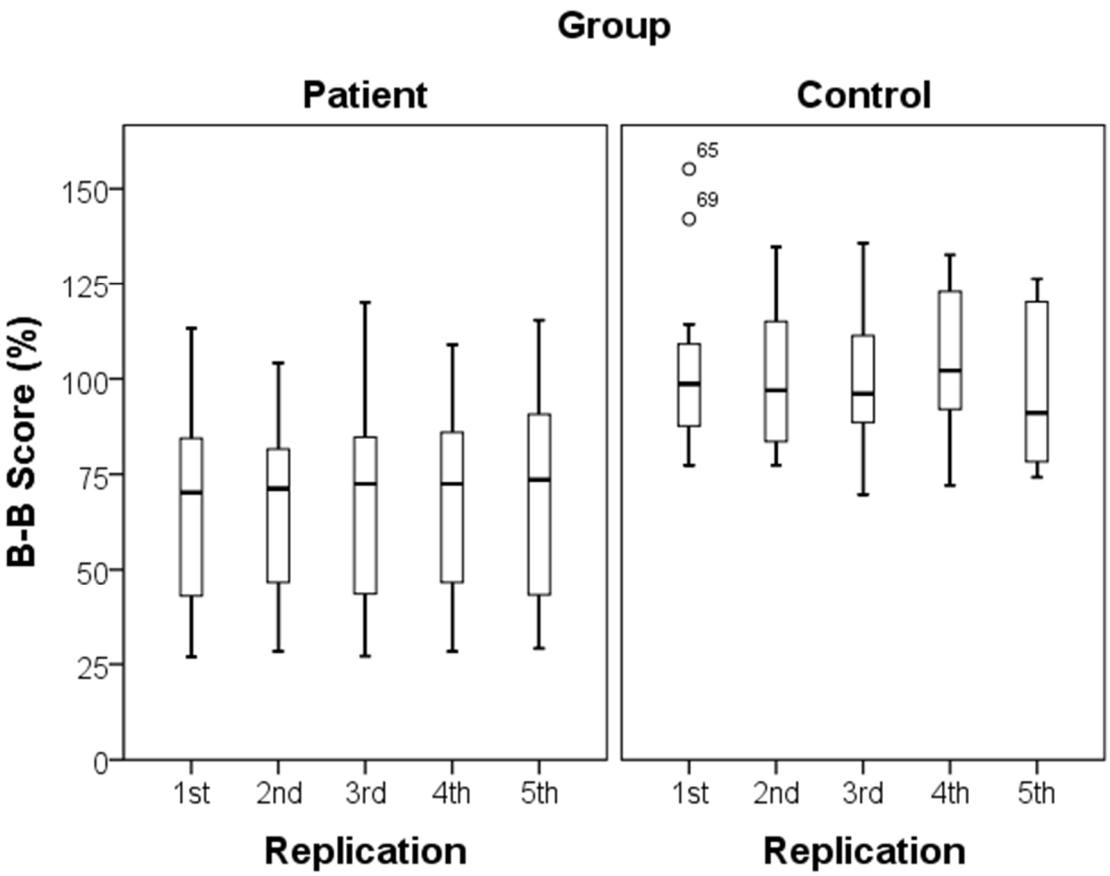


Figure 2.4: Traditional box plots showing median, lower and upper quartile, range and outliers (open circles, 1.5 interquartile range, with case numbers) B-B Scores, comparing the control (n= 7) and the patient (n=16) groups according to the number of intra-assessment replications (1 to 5*), with B-B Scores computed using the range method (* no significant differences across replicates; $p > 0.05$).

The Friedman test showed no differences between replications of the B-B Score, regardless of the group and the computation method: range computation method in the patient group ($\chi^2(4) = 9.15$, $p = 0.06$), area computation method in the patient group (also $\chi^2(4) = 9.15$, $p = 0.06$), range computation method in the control group ($\chi^2(4) = 6.62$, $p = 0.16$) and area computation method in the patient group ($\chi^2(4) = 0.55$, $p = 0.97$). This indicated that no systematic carry-over effects, such as warm up learning or fatigue had intruded during the execution of the manoeuvres associated with the B-B Score.

2.3.3.2. B-B Score determined by range or area of computation method

The Wilcoxon signed-rank test showed no significant difference between the range and the area computation methods in the patient group for 1 (Z = -1.71, p = 0.09), 2 (Z = -1.60, p = 0.11), 3 (Z = -1.55, p = 0.12), 4 (Z = -1.45, p = 0.15) and 5 (Z = -1.60, p = 0.11) replications. In the control group, the area B-B Score was significantly different from the range B-B Score for 2, 3 and 5 replications (Z = -2.19, p = 0.03 in all cases), but not for 1 (Z = -1.69, p = 0.09) and 4 (Z = -1.35, p = 0.18) replications. The one-sample Wilcoxon signed-rank test highlighted no significant difference with a median B-B Score of 100 in the control group, using either the range or area computation method (range computation method: 1st replication (Z = -0.28, p = 0.78), 2nd replication (Z = -0.28, p = 0.78), 3rd replication (Z = -0.34, p = 0.73), 4th replication (Z = -0.66, p = 0.51), 5th replication (Z = -0.28, p = 0.78); area computation method: 1st replication (Z = 0.72, p = 0.47), 2nd replication (Z = 0.47, p = 0.64), 3rd replication (Z = 1.04, p = 0.30), 4th replication (Z = 0.34, p = 0.73), 5th replication (Z = 0.72, p = 0.47). Thus, this indicated that none of the two methods detected a side-to-side asymmetry in healthy controls.

The effects sizes were of comparable magnitude regardless of the replication considered and the range or area computation method for the B-B Score calculation, with Cohen's d ranging from 1.26 to 1.65 (1st replication d = *1.60* for range, d = 1.59 for area; 2nd replication d = *1.64* for range and for area; 3rd replication d = 1.53 for range, d = 1.65 for area; 4th replication d = *1.65* for range, d = 1.55 for area; 5th replication d = *1.26* for range, d = 1.45 for area). These results highlight the ability of the B-B Score to discriminate correctly two groups that are anticipated to be different, regardless of the use of the range of area computation method.

The ICCs for the evaluation of the reliability between replications using the range and the area approaches for computation of B-B Scores showed comparable reliability between these approaches, for the patient and for the control group. The ICCs results are presented in Table 2.3.

Table 2.3: ICC values with interval at 95 level of confidence for the patient and control group for test-retest reliability between replications, using the range and area computation methods for the B-B Score calculation.

ICC [95% CI]	Patient (n=16)	Control (n=7)
Range method for the computation of B-B Score	0.90 [0.86 – 0.93]	0.71 [0.50 – 0.87]
Area approach for the computation of B-B Score	0.90 [0.86 – 0.93]	0.70 [0.50 – 0.87]

Legend: ICC intraclass coefficient of correlation; 95%CI: limits of interval at 95% confidence level

Based on the rationale for a selective reporting of results announced at the end of the statistical analysis plan, only the results obtained using the range B-B Score computation method are reported from this point. The exploration of the new area computation method showed no advantage on the range method in terms of reliability, responsiveness and discriminative power. As the range computation method was the original approach for B-B Score computation and as it had been tested in previous studies (Coley et al., 2007a; Jolles et al., 2011; Pichonnaz et al., 2015c), it was decided to continue to use it in the Phase 2 and 3 studies. These points are further detailed in the discussion section.

2.3.3.3. B-B Score determined by mean or median of replications

The B-B Score outcomes determined using the mean or the median of replications are presented for each number of replications in Table 2.4.

Table 2.4: Patient and control groups B-B Scores (mean with standard deviation and median with interquartile range) for each number of replications (1 to 5), for the mean and median of score replications computed using the range method.

Mean (SD) Median (IQR), %	**Patient (n=16)**		**Control (n=7)**	
	Using mean of replications	Using median of replications	Using mean of replications	Using median of replications
1 replication	65.2 (24.4)* 70.2 (40.3-84.4)		102.7 (22.3)† 98.7 (87.0-110.4)	
2 replications	65.7 (22.8) 71.5 (40.4-82.7)	-	101.4 (19.0)† 98.1 (86.4-114.7)	-
3 replications	66.2 (23.0) 72.0 (40.6-83.0)	65.8 (22.8) 72.0 (40.6-83.0)	100.8 (17.2)† 96.6 (88.0-113.5)	98.8 (15.4)† 97.6 (87.9-108.0)
4 replications	66.8 (22.9) 72.8 (41.1-85.1)	66.4 (22.8) 72.7 (40.7-84.0)	101.6 (17.0)† 98.5 (87.2-115.3)	101.0 (16.0)† 97.8 (86.7-114.7)
5 replications	67.0 (22.9) 73.2 (40.1-85.2)	66.7 (22.8) 73.4 (41.9-84.2)	100.8 (17.2)† 94.4 (87.9-116.6)	101.4 (17.8)† 97.6 (86.7-120.3)

Legend: SD: standard deviation; * significant difference with mean of 5 replications; † significant difference with the patient group

For the patient group, the Friedman test showed a significant difference between the B-B Scores obtained using 1 replication or the mean of 2 to 5 replications ($\chi^2(4) = 10.75$, $p = 0.03$). Post hoc analysis with Wilcoxon signed-rank tests showed a significant difference between the B-B Score obtained using 1 replication and the B-B Score obtained using the mean of 5 replications ($Z = -2.33$, $p = 0.02$). However, the Cohen's d was small ($d = 0.08$). Conversely, for the control group, the Friedman tests showed no significant difference between the B-B Scores obtained using 1 replication or the mean of 2, 3, 4 and 5 replications ($\chi^2(4) = 1.82$, $p = 0.77$).

These results indicated that using only one replication, the patient group outcome measured using the B-B Score might possibly differ from the performance measured using the mean of several replications, which was less likely to be influenced by occasionally divergent values. However, the magnitude of the difference remained small.

The Friedman test showed no significant difference between the B-B Scores obtained using 1 replication or the median of 3, 4 and 5 replications ($\chi^2(3) = 6.13$, $p = 0.10$) for the patient group. It also highlighted no significant difference between the B-B Scores obtained using 1 replication or the median of 3, 4 and 5 replications ($\chi^2(3) = 0.83$, $p = 0.84$) in the control group. This indicated that, for the patient and for the control group, using the median of replications, the measured group performance measured on the B-B Score was comparable, regardless of number of replications taken into consideration.

The Wilcoxon signed-rank test showed no significant differences between the B-B Score determined using the mean and the median of several replications, whether it is for 3 replications ($Z = -1.48$, $p = 0.14$), 4 replications ($Z = -1.20$, $p = 0.23$) or 5 replications ($Z = -1.00$, $p = 0.32$), for the patient group. This highlighted that the measured patient group outcome was not influenced by the choice of one or the other method for score averaging over replications.

Similarly, the Wilcoxon signed-rank test showed no significant differences for the control group between the B-B Score determined using the mean and the median of several replications, whether it is for 3 ($Z = -1.35$, $p = 0.18$), 4 ($Z = -1.15$, $p = 0.25$) or 5 replications ($Z = -0.84$, $p = 0.40$). This is indicative that the group performance of healthy participants is comparable using the mean or the median of scores replications.

Regardless of the number of replications, the Wilcoxon rank sum test showed significant differences between the patients and control groups using the mean or the median (for 1, mean of 2,3,4 and 5 replications, median of 3 and 4 replications: $Z = -3.1$, $p = 0.001$, for the median of 5 replications $Z = -2.9$, $p = 0.02$). The effects sizes for 1 replication and for the mean and median for the B-B Score replications were of comparable magnitudes, regardless of the replications considered, with Cohen's d ranging from 1.60 to 1.70. These results highlight the ability of the B-B Score to discriminate with large effect sizes two groups that are anticipated to be different, regardless of the use of the mean of the median as an averaging method.

The ICCs for single-measurement reliability of the B-B Score associated with the B-B Score measurements acquired by the same rater, for the mean or median scores across varying numbers of replications, are reported in Table 2.5. The intraclass

correlation coefficients for single-measurement reliability of the B-B Score associated with two separate raters, for the mean or median scores across the varying number of replications, are reported in Table 2.6.

Table 2.5. Comparison of ICC values with intervals at 95 level of confidence for intra-rater reliability of the measurements acquired by each rater, for each number of replications using mean or median of replications.

ICC (95%CI)	Rater	Mean	Median
5 replications	1st	0.96 (0.88 – 0.98)	0.96 (0.88 – 0.98)
	2nd	0.98 (0.95 – 0.99)	0.98 (0.93 – 0.99)
4 replications	1st	0.93 (0.83 – 0.98)	0.94 (0.84 – 0.98)
	2nd	0.98 (0.95 – 0.99)	0.97 (0.92 – 0.99)
3 replications	1st	0.93 (0.81 – 0.97)	0.93 (0.82 – 0.97)
	2nd	0.97 (0.91 – 0.99)	0.96 (0.88 – 0.98)
2 replications	1st	0.91 (0.76 – 0.97)	-
	2nd	0.96 (0.89 – 0.99)	-
1 replications	1st	0.83 (0.57 – 0.94)	
	2nd	0.91 (0.77 – 0.97)	

Legend: ICC intraclass coefficient of correlation; 95%CI: limits of ICCs at 95% confidence level; 1st: 1st measurement of rater 1 vs. 2nd measurement of rater 1 reliability; 2nd: 1st measurement of rater 2 vs. 2nd measurement of rater 2 reliability

There were no significant differences amongst measurements of the patient group B-B Scores recorded by the same rater using the Wilcoxon signed rank test, for the mean of replications: for 1 replications (Z = -0.84, p = 0.40), the mean of 2 (Z = -1.25, p = 0.21), 3 (Z = -0.35, p = 0.70), 4 (Z = -0.18, p = 0.85) and 5 (Z = 0.90, p = 0.37) replications. Similarly, no significant difference was found for the median of 3 (Z = 0.06, p = 0.96), 4 (Z = -0.18, p = 0.85) and 5 (Z = -0.15, p = 0.88) replications. The effects sizes were small in all cases, with Cohen'*d* ranging from 0.07 to 0.11.

In the control group, no significant difference was found for the differences between measurements of the same rater using the Wilcoxon signed rank test, for 1 replications (Z = -0.68, p = 0.50), the mean of 2 (Z = -0.85, p = 0.40), 3 (Z = -0.68, p = 0.50), 4 (Z = 0.00, p = 1.00) and 5 (Z = 0.00, p = 1.00) replications, and the median of 3 (Z = -0.17, p = 0.87), 4 (Z = -1.35, p = 0.17) and 5 (Z = -0.51, p = 0.61) replications.

The effects sizes for the differences between measurements of the same rater were small in all cases, with Cohen's *d* ranging from 0.01 to 0.15.

There were no significant differences amongst the patient group mean B-B Scores recorded by two raters for increasing numbers of score replications using the Wilcoxon signed rank test, for 1 replications (Z = 1.49, p = 0.14), the mean of 2 (Z = -0.48, p = 0.63), 3 (Z = -1.41, p = 0.25), 4 (Z = -1.50, p = 0.13) and 5 (Z = -1.62, p = 0.10) replications, and the median of 3 (Z = -1.38, p = 0.17), 4 (Z = -1.23, p = 0.22) and 5 (Z = -1.68, p = 0.09) replications. The differences between raters could not be calculated in the control group as only one rater proceeded to the measurements. The effects sizes of the differences between raters for the patient group were small in all cases, with Cohen*'s d* ranging from 0.07 to 0.11.

Thus, all the differences between measurements, acquired by the same rater or two different raters, were non-significant and of minor magnitude.

Table 2.6: Comparison of ICC values with intervals at 95 level of confidence for inter-rater reproducibility for the 1st and the 2nd measurement acquired by the two raters, for each number of replications using the mean or the median of replications.

ICC (95%CI)	Rater's measurement	Mean	Median
5 replications	1st	0.96 (0.89 – 0.99)	0.96 (0.90 – 0.99)
	2nd	0.96 (0.90 – 0.99)	0.95 (0.87 – 0.98)
4 replications	1st	0.95 (0.86 – 0.98)	0.95 (0.86 – 0.98)
	2nd	0.96 (0.89 – 0.99)	0.95 (0.87 – 0.98)
3 replications	1st	0.94 (0.83 – 0.98)	0.93 (0.83 – 0.98)
	2nd	0.96 (0.88 – 0.98)	0.96 (0.89 – 0.99)
2 replications	1st	0.96 (0.89 – 0.99)	-
	2nd	0.95 (0.86 – 0.98)	-
1 replications	1st	0.94 (0.85 – 0.98)	
	2nd	0.87 (0.68 – 0.95)	

Legend: ICC intraclass coefficient of correlation; 95%CI: limits of ICCs at 95% confidence level; 1st: 1st measurement of rater 1 vs. 1st measurement of rater 2 reliability; 2nd: 2nd measurement of rater 1 vs. 2nd measurement of rater 2 reliability

The bias and limits of agreement for intra- and inter-rater reliability are reported for the mean and the median of replications of the B-B Score in Table 2.7 and 2.8. The Bland and Altman graphs for the intra- and inter-rater reliability are reported in Figure 2.5 and 2.6.

Table 2.7: Bias and 95% limits of agreement (%, original units) as estimates of intra-rater reproducibility of B-B Scores (range method only), recorded for two measurements by the same rater (1st; 2nd) using mean or median scores from 1 to 5 replications

Bias ± 95% limits of agreement	Rater	Mean	Median
5 replications	1st	1.8 ± 13.8	0.2 ± 13.5
	2nd	1.2 ± 8.7	1.3 ± 10.4
4 replications	1st	2.1 ± 16.2	1.7 ± 15.6
	2nd	1.1 ± 9.5	1.3 ± 11.6
3 replications	1st	2.3 ± 16.7	1.5 ± 16.5
	2nd	1.2 ± 12.7	0.5 ± 14.6
2 replications	1st	3.3 ± 18.9	-
	2nd	- 1.0 ± 13.6	-
1 replications	1st	5.6 ± 25.8	
	2nd	- 3.5 ± 21.2	

Legend: 1st: first measurement of rater 1 vs. second measurement of rater 1 reliability; 2nd: first measurement of rater 2 vs. second measurement of rater 2 reliability

Table 2.8: Bias and 95% limits of agreement (%, original units) as estimates of inter-rater reproducibility of B-B Scores (range method only), recorded across two serial assessments (1st; 2nd) acquired by the two raters using mean or median scores from 1 to 5 replications.

Bias ± 95% limits of agreement	Rater's measurement	Mean	Median
5 replications	1st	- 3.0 ± 13.5	- 1.8 ± 12.2
	2nd	- 0.1 ± 12.3	0.3 ± 13.9
4 replications	1st	- 2.5 ± 14.8	- 2.5 ± 14.9
	2nd	- 0.7 ± 13.1	- 0.6 ± 14.2
3 replications	1st	- 2.6 ± 16.6	- 2.2 ± 17.2
	2nd	- 0.9 ± 13.3	0.2 ± 12.8
2 replications	1st	- 1.7 ± 14.1	-
	2nd	- 0.7 ± 14.2	-
1 replications	1st	1.2 ± 17.7	
	2nd	- 3.3 ± 23	

Legend: 1st: first measurement of rater 1 vs. first measurement of rater 2 reliability; 2nd: second measurement of rater 1 vs. second measurement of rater 2 reliability

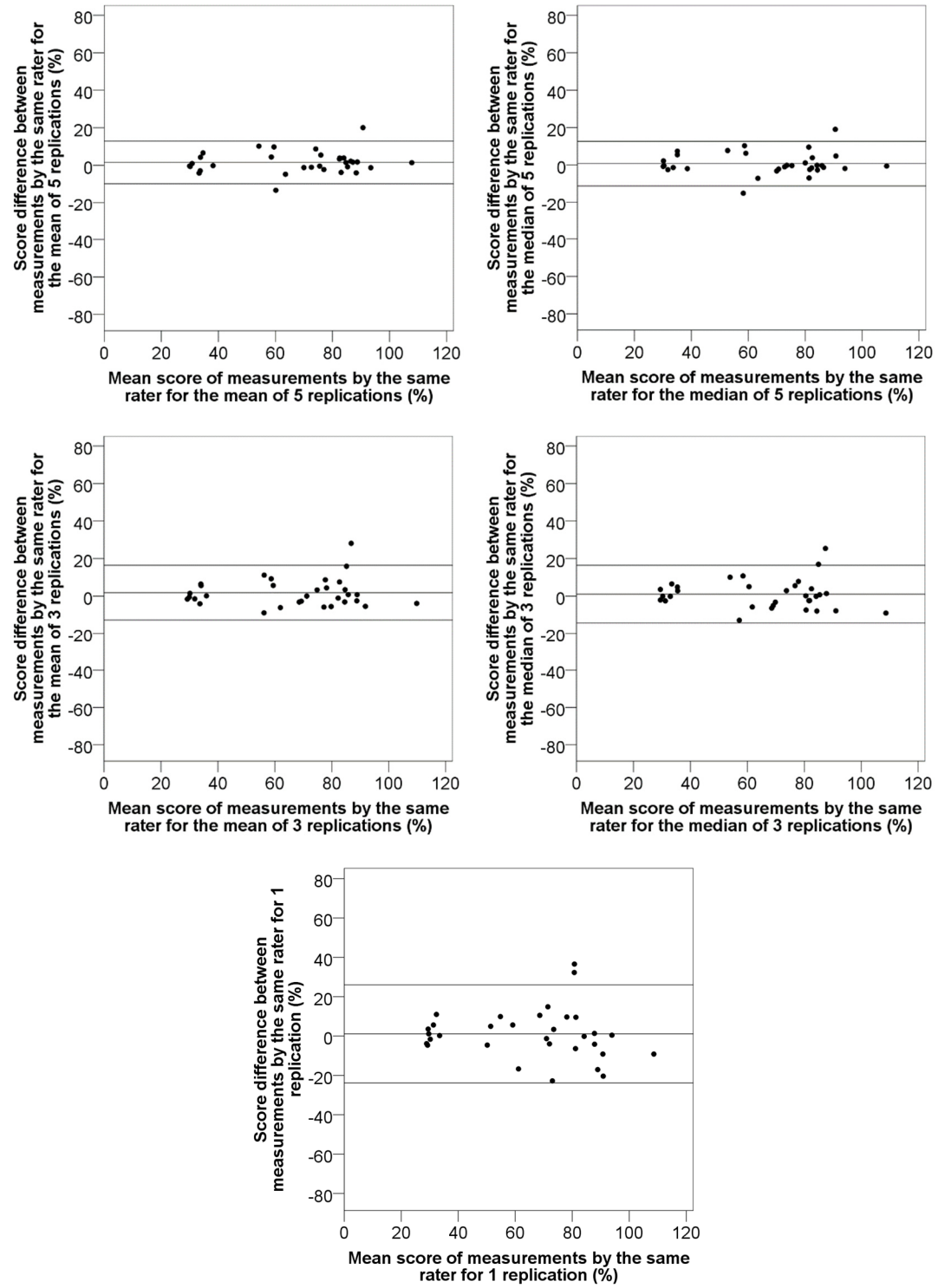


Figure 2.5: Bland and Altman plots of bias and 95% limits of agreement (%, original units) as estimates of intra-rater reliability of B-B Scores (range method only), recorded across two measurements acquired by the same rater, using mean and median scores (mean [left panel]; median [right panel]) from 1, 3 and 5 replications of score's movements across two serial assessments.

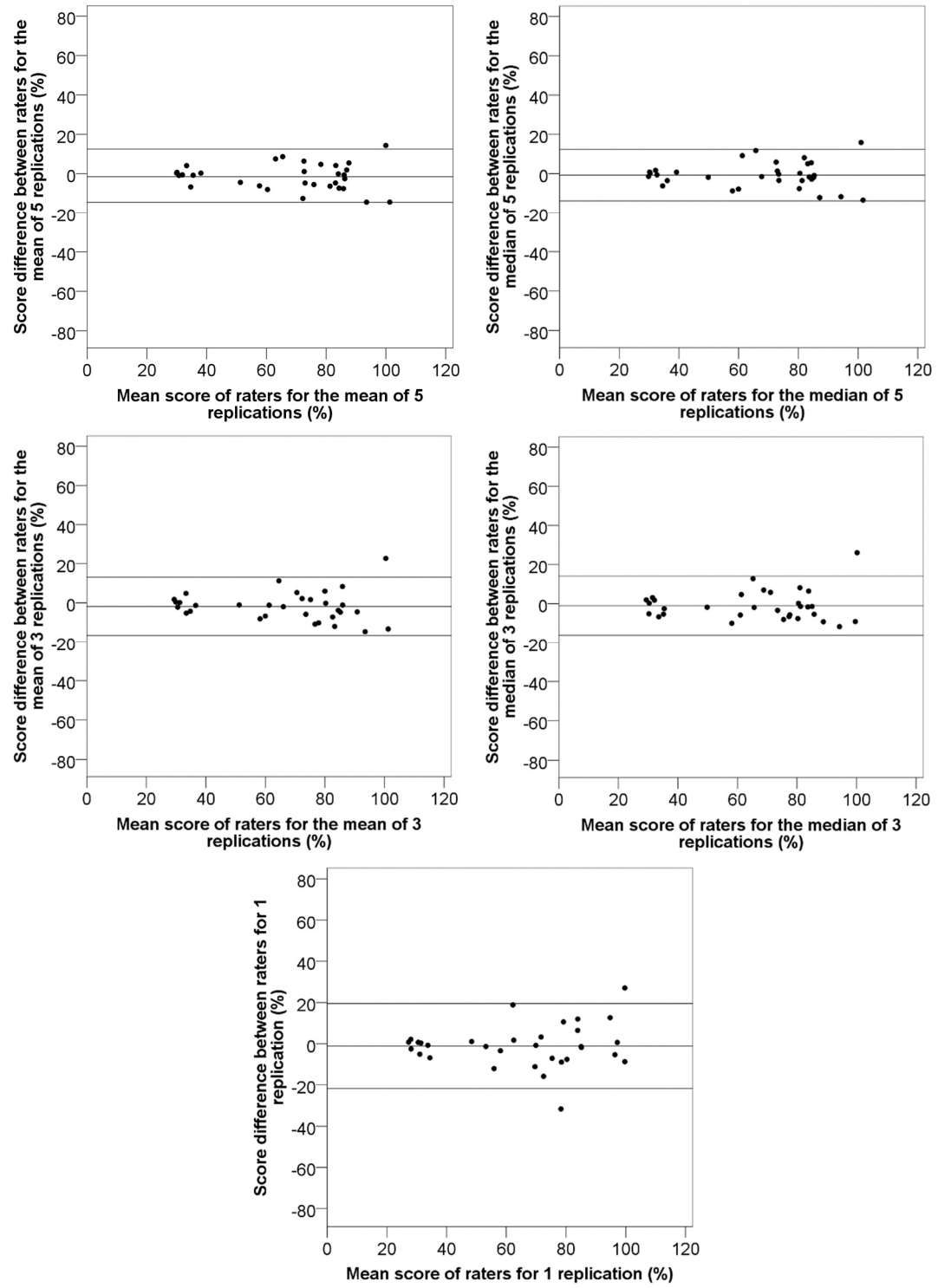


Figure 2.6: Bland and Altman plots of bias and 95% limits of agreement (%, original units) as estimates of inter-rater reliability of B-B Scores (range method only), recorded across two serial measurements by two separate raters, using mean and median scores (mean [left panel]; median [right panel]) from 1, 3 and 5 replications of score's movements.

Based on the rationale for the selective reporting of results announced at the end of the statistical analysis plan (sub-section 2.2.7 “Statistical analysis plan”, p. 72 - 74), only the results obtained using the mean of several replications using the range B-B Score computation method are reported from this point. Though the mean and median of replications produce equivalent results, the use of the mean of the replications was estimated to be somewhat more intuitive for potential users.

The evolution of the LoAs as a function of number of replications for the mean of replications using the range B-B Score computation is presented in Figure 2.7.

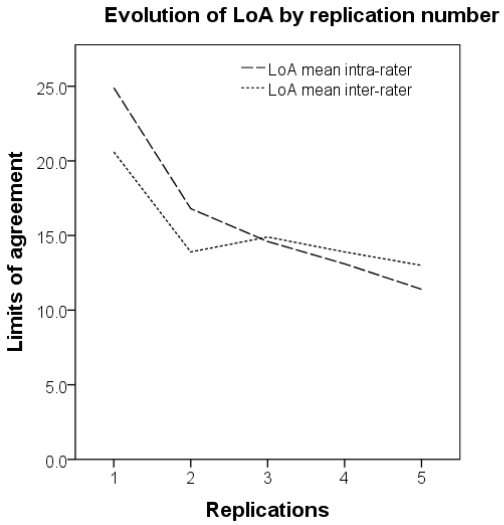


Figure 2.7: The graphical evolution of Bland and Altman 95% limits of agreement (%, original units) as estimates of intra- and inter-rater reliability of B-B Scores (range method only), as a function of mean score from 1 to 5 intra-measurement replications. Graphical plots show responses averaged mathematically over two separate raters and across two serial measurements by the same rater, respectively.

2.4. Discussion

2.4.1. Feasibility

The Phase 1 study showed that the access to patients presenting with the four targeted pathologies was possible. Nevertheless, the recruitment of patients with rotator cuff conditions was easier than for the other conditions and the latter sub-group predominated within the experimental sample.

The recruitment of patients had improved as the study progressed, and this was achieved by means of two strategies. Firstly, systematic recalls to the doctors in charge of the specialised shoulder consultations relating to potential patients of interest to the study were implemented following further liaison with and, secondly, following an addendum procedure that had been approved by the hospital's ethical commission, the recruitment focus was extended to include patients treated in the musculoskeletal physiotherapy department of the hospital (for details on the addendum pleases see sub-section 3.2.2 "Study sample", within Chapter three, p. 101 - 103).

The absence of any technical and practical issues during the measurement sessions confirmed the robustness of the measurement device and the applicability of the measurement procedure. The acquired data were all exploitable and contributed fully to subsequent statistical analyses.

Patients who had been assessed were able to undergo the scoring procedure within a reasonable amount of time, reaching 45 to 60 minutes. They tolerated the protocol well and reported no disagreement at the end of each of the measurement sessions.

Thus, the results from the Phase 1 study confirmed the feasibility of the subsequent Phase 2 and 3 studies of the thesis, with the exception of concerns about of the rate of patients' recruitment which would have been too slow overall to allow the Phase 2 and 3 studies completion within the anticipated and prescribed amount of time. Nevertheless, because the delays in patient recruitment had centred around issues of communication at the beginning of the study rather than around an insufficient population of targeted patients, it was expected that the adjustment to the process of recruitment that had been alluded to earlier (i.e. improved liaison with consultant

clinicians and a wider scope for recruitment within the hospital) would allow the recruitment for the Phase 2 and 3 studies to be successfully completed within two years, plus 6 months for follow-up.

2.4.2. Study sample

The control sample was younger than the patient sample but this had probably a minor influence on the B-B Scores. Its calculation is based on a side-to-side comparison, of which balance associated with ipsilateral-contralateral performance is not likely to be altered by age, in the absence of pathology. The enrolment of younger patients limited the risk to include patients with undetected rotator cuff conditions, of which frequency increases with age (Yamaguchi et al., 2006; Yamamoto et al., 2010; Moosmayer et al., 2009).

2.4.3. Score optimisation

2.4.3.1. Influence of number of replications on B-B Scoring

The analyses highlighted no carry-over effect, as was stated by the non-significant differences amongst replications for the patient and the control group, both for the range and area computation methods. However, the mean outcome of the patient group tended slightly to increase in the measured patient sample with each additional replication for both computation methods (e.g. from Table 2.2 +1%, +2.1%, +3.2% and +3.5% vs. mean of 1st replication for 2nd, 3rd, 4th and 5th replication for range approach, respectively). Though these results were non-significant *stricto sensu*, considering the small sample size (16 patients) and the *p* values in the patient group ($p = 0.06$), a Type II error cannot be fully excluded.

The box plots of the Figure 2.4 showed that the replication (1st to 5th replication) had little influence on the group score, but that extreme results were observable when only one replication was accounted for. Thus, performing only one repetition would be possible when investigating large samples, as it would have little influence on the groups score, but the results of single measurements might be misleading in some cases.

No robust conclusion could be drawn from the Phase 1 study on the potential influence of the number of replications on the score increase over replications. No carry-over effect was demonstrated based on the measured patient sample, but this effect could nevertheless not be completely excluded, because a trend toward increase was observed in the sample and a Type II error was possible. Should a Type II error have occurred and the carry-over effect really exist, it would then be a warm-up effect leading to a progressive increase of the performance, which would induce an overestimation of the real patient's performance. Should a Type II error have occurred and the carry-over effect really exist, it would then be a warm-up effect leading to a progressive increase of the performance, which would induce an overestimation of the real patient's performance. Nevertheless, since the magnitude of the increase remained small compared to the 33 to 39% difference between groups, the implications for the accuracy of the evaluation would remain limited.

It seemed thus reasonable to keep the number of replications to a minimum in the next phases of the thesis, provided that the reliability is sufficient using this selected number of replications, to avoid a possible artificial increase of the score results by the execution of overly high number of replications.

2.4.3.2. Comparison of the range and area methods

Overall group mean B-B Scores for the patient group computed using the area and range methods did not differ in the patient group ($p > 0.05$) (Table 2.2). The magnitude of the difference between range and area mean scores (2.4% – 2.9% difference from Table 2.2. results) was limited in comparison to the magnitude of the difference between the patient and control groups, which reached 28.9% to 39.4% amongst all replications. The responsiveness evaluated using the Cohen's *d* effect size was comparable using either method (range 1.26-1.65), with no systematic advantage for one method amongst replications.

The differences between the range and area B-B Score computation methods were higher in the control group (3.3% – 9.2% from Table 2.2), with significant differences ($p < 0.05$) for the 2nd, 3rd and 5th replication. Though these results highlight that both methods may produce different results in this group, none of the results found in this study indicated that this difference induced an advantage for either method. Indeed,

as presented above, this did not lead to larger effect sizes for the area method though the outcomes were higher in the control group for this computation method.

The difference with a median B-B Score of 100, indicating perfect side-to side symmetry, was non-significant ($p > 0.05$) using either the range or the area B-B Score computation method. This implied that there was no need to consider an adjustment of the B-B Score outcome as a function of the dominant/non dominant side being involved with the pathology, in general and subsequently in Phase 2 and 3 studies.

ICCs indicating single-measurement reliability for B-B Score measurements showed essentially equivalent reliability amongst scores computed using area and range methods (Table 2.3). The ICCs were at the threshold between good and excellent for scores recorded for the patient group, indicating an adequate reliability between replications based on accepted standards for clinical measurement (Portney and Watkins, 2015). The ICCs levels were moderate for the control group (Portney and Watkins, 2015). The ICCs were expected to be lower in this group, as correlations levels are known to be influenced by the data variance (Bland and Altman, 1986a). In this study, as in most clinical studies, the group heterogeneity was lower in the control than in the patient group, which leads to find higher levels of correlations in the latter group.

Overall, the range and the area approach demonstrated equivalent properties in terms of responsiveness, reliability (excellent for the patient group and moderate for the control group), side-to-side symmetry in the control group and discriminative power. These findings contradict the theoretical advantages of the area computation method, which was supposedly less sensitive to peak measurements. As a decision had to be taken in the absence of clear statistical advantages in favour of one or the other approach, the range approach was retained for the Phase 2 and 3 studies. This choice was made because no reason was found to abandon the range method that was the original approach used for the B-B Score computation, and because it had previously been utilised in several studies in which it had demonstrated acceptable measurement properties (Coley et al., 2007a; Jolles et al., 2011; Pichonnaz et al., 2015c).

2.4.3.3. Comparison of the mean and median of replications

The results obtained for each replication (1 to 5) (Table 2.2) and the visual inspection of box plots (Figure 2.4) showed that the score was stable over replications. However, the presence of outliers using only one replication (box plot of the 1st measurement of control group) and the significant difference for the patient group score between the 1st and 5th replication (Table 2.4) confirmed the importance of averaging the score over several replications to contain the influence of diverging measurements.

The use of the mean or the median of several replications of the B-B Score had no significant influence on the group score in the patient and control groups (Table 2.4). The differences between the patient and the control group were large (Cohen's *d* 1.60 – 1.70) and significant ($p < 0.05$), regardless of the number of replications and of the calculation of the mean or median of replications. Though no subgroup analysis by shoulder condition could be performed at this stage of the project, this result globally confirmed the discriminative power of the B-B Score in shoulder conditions, which was a prerequisite to any further measurement properties analysis.

The ICCs calculated for the assessment of the intra-rater reliability were excellent for the two raters (ICC ≥ 0.90), regardless of the number of replications, except for the reliability of the 1st rater for 1 replication, for which the ICC level was good (ICC = 0.83) (Table 2.5). Similarly, the ICCs for the assessment of inter-rater reliability were all ≥ 0.90, except for the reliability of the 2nd measurement of the raters for 1 replication, for which the ICC level was good (ICC = 0.87) (Table 2.6). Close data inspection showed that this lower ICC was essentially due to a single diverging measurement on the 1st replication. This statement, added to the fact that the ICCs increased with the number of replications, reinforced the hypotheses that several replications were needed to contain the potential influence of measurement variability when measuring a patient.

The calculation of LoAs and the generation of B&A graphs showed a limited bias between measurements of the same rater (Table 2.7 and Figure 2.5). It was < 3.3% when more than 1 replication was executed, which was minor compared to the 33.0% to 37.5% difference between the patient and control group (Table 2.4). The same

statements could be made for the inter-rater reliability (Table 2.8 and Figure 2.6), with bias always < 3.3%.

As expected if the measurement errors are randomly distributed, the measurement variability decreased with the increase of replications, though in higher proportion for the first replications (Winer, 1981, Gleeson and Mercer, 1996).

Detailed data inspection showed that a few extreme measures had a considerable influence on the LoA, especially when one measurement was used instead of the mean/median of several. The graph inspection confirmed the similarity of the bias and LoA using the mean for the median of scores. No obvious trend in data distribution was visually detected from the graphs, indicating that the variability was independent of the score value.

Though the median is theoretically less influenced by extreme values, the use of the median in practice did not positively influence the bias, the responsiveness or the variability of the measurement in any case. As using one or the other can be considered as strictly equivalent from a statistical point of view, it has been decided to use the mean of several replications, of which use is somewhat more intuitive for clinicians who are the target users of the B-B Score.

2.4.3.4. Influence of the number of replications on score variability

The Figure 2.7 showed that the magnitude of the LoAs decreased with the number of replications, but that most of the improvement was obtained during the first replications. Observing the curve trend, it can be stated that most of the decrease in LoA magnitude was obtained using two replications for the inter-rater LoA and three for the intra-rater LoA.

Considering this, it was decided to use three replications in the Phase 2 and 3 studies. This decision was driven by the intention to design an efficient score, presenting a good balance between limiting measurement constraints and containing measurement variability. However, it was stated that, even using 5 replications, the LoAs ranged from 8.7% to 13.9% (Table 2.7 and 2.8), which did not fully guarantee that the level of precision of single measurements would be acceptable in clinical

situations where the difference between measurements is small (e.g. when a patient's change at follow-up is limited), though the B-B Score reliability was excellent based on ICCs values.

2.5. Conclusion

2.5.1. Phase 1 study's impact on Phase 2 and 3 studies

The Phase 1 study showed that the main project was reasonably realistic and feasible. Adjustment of the recruitment procedure had to be made to prevent a slow recruitment rate, which was identified as the main potential risk of failure.

The Phase 1 study served as a basis to define the optimal testing procedure. The range and area score computation methods were found to be equivalent, despite the theoretical advantages of the latter one. In addition, none of the method implied to use a compensation factor to correct for side-to-side asymmetry between the dominant and non-dominant arm to be able to compare the results of patients affected on the dominant or on the non-dominant side. If it had been present, this asymmetry would have considerably complicated the application and interpretation of the B-B Score, (e.g. to compare the performance of a right-handed patient affected on the right side to that of a right-handed patient affected on the left side). Based on the absence of difference between the range and area computation methods, it was decided to use the original range computation method that had already been used in previous studies (Coley et al., 2007a; Jolles et al., 2011; Pichonnaz et al., 2015c).

Five replications of the B-B Score movements were executed in this study. It was stated that the score was stable over replications, but that the averaging of the replications, using the mean or the median of several replications decreased the influence of the measurement variability over replications, and had thus a positive influence on the reliability, as demonstrated by increasing ICCs values and decreasing magnitude of intra- and inter-rater LoAs over replications.

The use of the mean or the median had no influence on the measurement properties investigated in this study (i.e. LoA, ICCs, effect size between the patient and the control group), which were comparable. As most of the decrease of the LoA magnitude was obtained in the three first replications, it was decided to use the mean

of three score replications. This number appeared to represent good balance between measurement constraints and measurement precision.

A few measurements diverging from the others taken in similar conditions were observed in the Phase 1 study, as reported in Table 2.7 and 2.8 and as can be stated on the Bland and Altman graphs. This highlighted the need of precision in sensors' placement and patient's instruction to prevent the occurrence of some extremely diverging results in the Phase 2 and 3 studies.

2.5.2. General implication of the Phase 1 study

On a more global level, this study's results were in line with the measurement properties that had previously been reported about the B-B Score and its parent P Score from which it is derived (Coley et al., 2007a; Jolles et al., 2011; Pichonnaz et al., 2015c; Coley, 2007), though it was the first study that investigated the measurement property of the B-B Score in a non-surgical sample. This was of importance for the following studies of the thesis, which aimed at the investigation of the measurement properties of the B-B Score in conservatively treated patients' populations.

The score easily discriminated the patients from the control group, and the inter- and intra-rater reliability were excellent, provided that the average of more than one replication was taken into account for the score calculation. More investigations were needed to determine to which extend the score precision was sufficient for single measurements,

These results were nevertheless only indicative, due to the small sizes of the patient and control group, and the heterogeneity of the patient sample. Though there is no standard recommendation concerning the required sample size for the estimation of psychometric properties, the size of the sample has an important influence on the precision of the results, larger sample producing estimations that are more precise.

The sample sizes were sufficient for the exploratory analysis conducted in Phase 1, but did not allow drawing precise conclusions on the measurement properties of the B-B Score. The precision of the estimation was limited and no information could yet be drawn on the validity of the score in specific shoulder populations at this stage of the thesis.

CHAPTER THREE

MEASUREMENT METHOD DEVELOPMENT AND COMPARISON

3.1. Introduction

The aim of this thesis – validate the simplest possible kinematic shoulder function scoring procedure for clinical practice and research – focused on investigating the possibility of being able to rely on a simple and accessible device for shoulder function scoring, in addition to the investigations that were made to establish the scoring properties. Therefore, the relevance of the use of smartphones for the evaluation of the B-B Score was investigated and presented within this chapter of the thesis, because this device could greatly facilitate the practicality of measurements.

As a reminder of Chapter one 1.1.2.4 “Thesis aim” p. 11 - 13, the data of Phase 2 and 3 were collected simultaneously. The data collection protocol was designed to allow a two-step analysis, the first step aiming at the assessment of the smartphone measurement capacities compared to an inertial sensor system used as a reference device, regardless of pathologies (Phase 2, presented in this Chapter), the second step aiming at the extensive investigation of the B-B Score measurement properties twice at 6 months interval for several frequent shoulder conditions, using the most efficient device. This data collection approach only marginally increased the complexity of the measurement protocol and contributed to a rational use of resources, with respect to patients’, raters’ and ethics committee members’ respective solicitations, as well as with respect to the use of premises and measurement instruments.

Aspects of the findings of this Phase 2 study have been published in the peer-reviewed open-access journal Plos One (Thomson Reuters 2017 impact factor 2.77) (Pichonnaz et al., 2017) (Appendix VIII). Note that, though the results of Phase 2 logically should have been published before those of Phase 3, this was not the case due to respective contingencies related to the review process of the submitted articles.

The rationale for the investigation of a smartphone application were previously developed in the points 1.1.2.1 to 1.1.2.3 of the thesis’ Introduction p. 4 - 11. Briefly summarised, the evaluation of shoulder function using questionnaires has remained a controversial issue (Kirkley et al., 2003; Oh et al., 2009; Huang et al., 2015; Harvie et al., 2005). Movement analysis may be a possible alternative to questionnaires for shoulder function evaluation. Yet, its use in this respect within clinical practice has been limited to date by issues of cost, accessibility, practicality and training (Aminian

and Najafi, 2004; Clark et al., 2017). Relying on smartphones to overcome these limitations might be an option, as most of them are fitted with built-in movement sensors that can potentially be used for shoulder movement analysis (Mark, 2011). The studies exploring the smartphone approach have found a good reliability for range of motion measurements, but to the best of the author's knowledge, no study has evaluated specifically the measurement properties of smartphones for function evaluation, let alone specifically for the B-B Score (Werner et al., 2014; Shin et al., 2012; Mitchell et al., 2014; Cuesta-Vargas and Roldan-Jimenez, 2016; Johnson et al., 2015; Brophy et al., 2005).

A new dedicated application for shoulder function evaluation using the B-B Score was developed for the purpose of this study by the engineers of the Laboratory of Movement Analysis and Measurement (LMAM) of the Swiss Institute of Technology of Lausanne (EPFL), as a parallel venture to the thesis' Phase 1 study. This facilitated assessing the measurement properties and practicalities of a smartphone for the evaluation of shoulder function in Phase 2.

The application was developed using the same algorithm as the one used for the Physilog® IMU system (Physilog®, Gait Up, Lausanne Switzerland) deployed within the Phase 1 study. Laboratory simulations were conducted before clinical validation using the data previously collected from the Physilog to test the correct functioning of the application's algorithm. This initial testing confirmed that the smartphone produced similar scores to those from the Physilog reference system when using these data.

A preliminary clinical measurement properties study was then undertaken by the engineers on seven healthy controls from within the staff working within the LMAM, using the Physilog system and the application simultaneously (Oïhénart et al., 2012; Duc, 2013). The differences between the values of the shoulder function scores measured using the smartphone application and the Physilog reference system reached 0.2 ± 0.8% (max: 1.4%). It was concluded that the smartphone was able to measure, process, display and store the kinematic scores effectively and that the shoulder function score' values given by the smartphone were precise and accurate compared to the reference system.

The latter preliminary study was appropriate to explore if a smartphone could potentially be used for the B-B Score measurement, but would have always been insufficient to investigate precisely the usefulness of the smartphone for clinical measurements in patients with a shoulder condition. Thus, a larger scale study was conducted to determine to which degree the smartphone B-B Scores measured on patients and healthy controls were comparable to those of the Physilog inertial sensor system, used as a reference because it had demonstrated its suitability for this purpose in previous studies (Coley et al., 2007a; Jolles et al., 2011; Pichonnaz et al., 2015c; Duc et al., 2013; Duc et al., 2014; Pichonnaz et al., 2015b).

3.1.1. Study aim and hypotheses

The aims of this study were to investigate the validity and reliability of a smartphone-assessed kinematic shoulder function B-B Score, and to compare the performance of the smartphone to a reference inertial sensor system.

The results of this study were of clinical importance, as they contributed to explore to what extent a tool used in everyday life could be a reasonable substitute to a dedicated movement analysis inertial sensor system, potentially making objective evaluation of shoulder function more accessible to health professionals. Specifically to this thesis, the results were needed to determine which of these devices was the most efficient one to measure the B-B Score. It had been anticipated that the Phase 3 study, which aimed at the extensive validation of the B-B Score measurement properties in frequent shoulder pathologies, would then be conducted relying on the data acquired using the most efficient of the devices, as stated in Phase 2. It had been decided that if the study showed the equivalency of the two devices, the comparison would be considered to the advantage of the smartphone, which is more accessible, cheaper and more user-friendly.

The study's hypothesis was that the B-B Score measured with a smartphone would meet the requirements of a valid shoulder function score. This implies that the differences between the control and the patient group but not the difference between devices should be significant, the statistical reliability involving ICCs would be ≥ 0.90 for inter-device, intra-rater and inter-rater comparisons, and that the limits of agreement (LoA) between devices, raters and measurement would be ≤ ± 10% with the bias ≤ ± 5% (Walter et al., 1998; Portney and Watkins, 2015). The B-B Score

results should also be coherent with those of shoulder function PROMs, i.e. show similarity in the levels of functional deficiency for patients and show performances that would be close to normal as indicated by healthy controls.

3.2. Methods

For the sake of concision, only aspects of the methods that differ from Phase 1 are detailed hereafter. The full description of the aspects that have been omitted is available in sub-section 2.2, “Methods”, within Chapter two, p. 65-74

3.2.1. Ethical issues

Amendments had to be made to the Phase 1 protocol 205/10 that had been previously approved by the Human Research Ethics Committee of the Canton of Vaud (CER-VD), to account for the implications of the results of the Phase 1 study. These amendments had been approved by the Ethics Committee, as they were minor and did not raise new ethical issues (Appendix V). They could thus be incorporated within the current Phase 2 study’s methods.

The Phase 2 study was registered under ClinicalTrials.gov Identifier: NCT01431417, simultaneously to Phase 3 (Appendix IX).

3.2.2. Study sample

One patient and one control groups were enrolled for this study. The patients were recruited at the Department of Traumatology and Orthopaedic Surgery of the University Hospital of Lausanne and the healthy controls within the working environment of this researcher and those that had collaborated for this prospective cohort study.

Based on the difficulties of patient recruitment reported previously within the Phase 1 study, the recruitment area was widened to include both the patients addressed at the specialised shoulder consultations and the patients treated in the musculoskeletal physiotherapy department of the hospital. Additionally, a request was made at the local Ethics Committee to have the permission to screen the records of patients attending the specialised shoulder consultations, which improved the recruitment rate

compared to that in Phase 1's study (Protocol 270/12) (Appendix X Accord éthique accès Soarian) (please see sub-section 2.2.1, "Study sample", within Chapter two p. 65 - 67, for full details). Therefore, eligible patients residing in the canton, as indicated by the inspection of their medical records, were contacted by phone in the order in which they attended the medical consultation in the department.

The inclusion criteria of patients were the same as those for the study in Phase 1 of the research (patients with conservatively treated rotator cuff condition, adhesive capsulitis or shoulder instability, and conservatively of surgically treated proximal humerus fracture). Exclusion criteria were also unchanged. As presented in the thesis aims, the data for Phase 2 and Phase 3 were collected simultaneously, Phase 2 aiming at the investigation of the measurement performance of a smartphone compared to an inertial sensor system used as a reference system and Phase 3 aiming at the extensive determination of the B-B Score measurement properties for various shoulder pathologies using the most efficient of the two systems (please see sub-section 1.1.2.4, "Thesis aim", within Chapter one p. 10 - 12).

The results for patients with shoulder instability were purposely not included in the analyses reported hereafter because of retrospective considerations for the utility of the data and a desire to maintain coherence with the overall aims of the thesis focusing on the real-world clinical applicability of the B-B Score. It had been demonstrated retrospectively and described at a later stage of the thesis that the B-B Score has insufficient validity specifically in this population (please see Chapter four, p.125-165) (Pichonnaz et al., 2015a). It was thus estimated that it was more relevant for potential users of the B-B Score, to report a focused analysis of it on the populations for which the score can be used in the future.

It was specified as a delimitation for this study that the group of participants without history of shoulder condition/pain, which was included to evaluate the performance and stability of the B-B Score in a healthy population, had to be younger than 35 years-old. These participants, acting as healthy controls within this study, were selected purposefully to be younger than the patients in order to avoid bias related to the high prevalence of asymptomatic rotator cuff tear above 40 years old (Yamaguchi et al., 2006; Yamamoto et al., 2010; Moosmayer et al., 2009).

The sample size calculation was based on the data of the Phase 1 study that had included responses from 7 controls and 16 patients. The calculation was made so that, with a significance level at $p < 0.05$, the power of 0.80 was reached when the minimal standards for acceptable properties of the score were met. Forty-six patients were required considering a lowest acceptable estimate statistically of reliability for two measurements (ICC) of 0.80, and an expected estimate statistically of measurement ICC of 0.90 (Landis and Koch, 1977; Walter et al., 1998). Nine patients per group were required to get the expected power of 0.80 with a significance level at $p < 0.05$ for the difference in B-B Score between the patients and the control group (Soper, 2004; Lenth, 2010).

The number of patients to be enrolled according to these calculations represents the minimum sample size required to meet the standards for research design (Portney and Watkins, 2015; Soper, 2004). A considerably larger sample was enrolled, because the data were to be collected for Phase 2 and 3 together. It was therefore needed to anticipate the subsequent subgroup analyses by pathologies at baseline and 6 months that had been planned within the Phase 3 of the thesis. The phase 3 sample size calculations showed that at least 20 participants were required in each subgroup (please see sub-section 4.2.1 “Study sample”, within Chapter four p.132). Since patients with three pathologies were included in Phase 2, 60 patients and 20 healthy controls were enrolled, plus a few additional patients needed to account for drop-outs at the 6 months follow-up.

The use of a larger sample size than theoretically required, which had been previously approved by the ethics committee, represents therefore an appropriate use of the study’s resources and contributed to enhance the precision of the calculated estimates in Phase 2.

3.2.3. B-B Score calculation

The B-B Score was calculated using the “range calculation” method, as the Phase 1 study had demonstrated that the alternative “area calculation” method had no advantages over this methods, which had previously been used in previous studies (please see sub-section 2.3.3.2, “B-B Score determined by range or area of computation method”, p. 78 - 79; sub-section 2.4.3.2, “Comparison of the range and

area methods", p. 91 - 92; sub-section 2.5.1, "Phase 1 study's impact on Phase 2 and 3 studies", p. 95 - 96, within Chapter two).

The mean of three B-B Score replications was used as it was shown to represent a good balance between containing measurement variability and limiting measurements constraints (please see Chapter two: sub-section 2.3.3.1, "B-B Score by replication", p. 75 - 77; sub-section 2.3.3.3 "B-B Score determined by mean or median of replications", Tables 2.5, 2.6, 2.7 and 2.8 and Figures 2.4, 2.5 and 2.6 p. 79 - 88; sub-section 2.4.3.4, "Influence of the number of replications on score variability", p. 94- - 95; sub-section 2.5.1, "Phase 1 study's impact on Phase 2 and 3 studies", p. 95 - 96).

3.2.4. Experimental systems: smartphone and reference system

A smartphone (iPod®, Apple, Cupertino, USA) was chosen as the support device for the development of the application. This device was adapted to the measurement purposes, as it is fitted with 3D built-in sensors (Accelerometers : ± 2 g precision: ± 0.02 g; Gyroscopes: ± 500 deg./s precision: ± 0.2 deg./s; Sampling frequency: 100 Hz) (Mark, 2011). An application, called iShould (instrumented shoulder test) was programmed in Objective-C (Oïhénart et al., 2012; Laboratory of Movement Analysis and Measurement—Swiss Institute of Technology of Lausanne, 2016). This application enabled the acquisition of the acceleration and angular velocity signals during the movements of the B-B Score and the computation of the B-B Score value, as described in the Figure 3.1. The laboratory preliminary testing of the smartphone application had shown, previously to the research within the Phase 2 study, that this approach was viable, as the measurements were close to those obtained using the IMU Physilog system as reference device (Physilog®, Gait Up, Lausanne Switzerland) (Oïhénart et al., 2012; Duc, 2013).

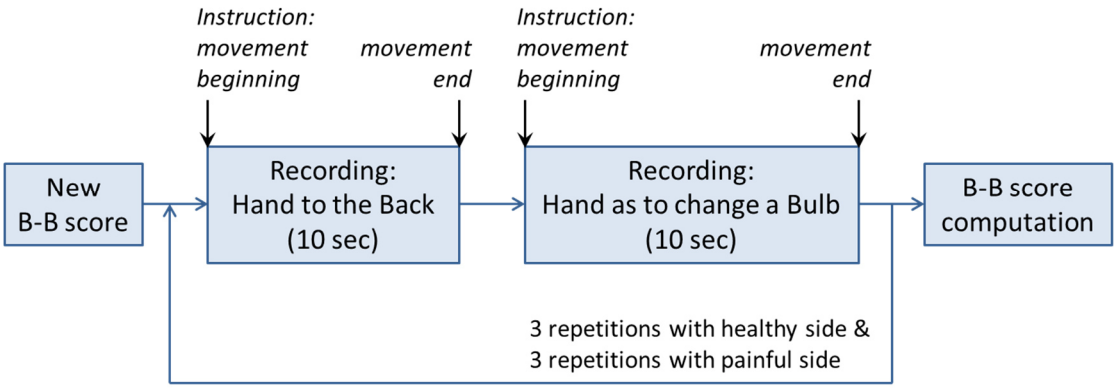


Figure 3.1: Schema of the application steps for the recording of a B-B Score (From: Pichonnaz C, Duc C, Gleeson N, Ancey C, Jaccard H, Lecureux E, et al. Measurement Properties of the Smartphone-Based B-B Score in Current Shoulder Pathologies. Sensors (Basel). 2015;15(10):26801-17).

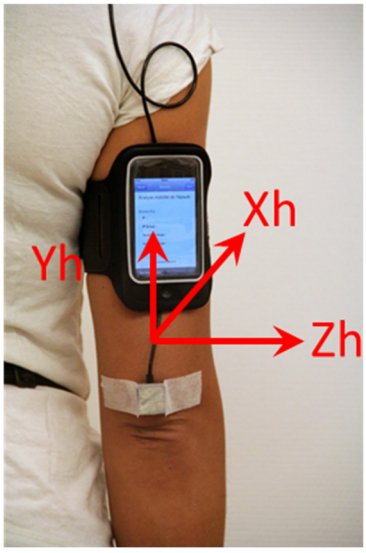
On 'launching' and initiating the software application, the smartphone provided instructions to the user, through the smartphone's loudspeaker, when to perform a B-B Score-related movement. For each movement, the application recorded the acceleration and angular velocity signals for a predefined period of 10 sec. The movements were first performed using the healthy side of the body and then repeated with the affected side. At the end of the test, the B-B Score was directly calculated, displayed on the smartphone screen and then stored within the smartphone's internal memory. The application enabled exporting of all saved data to an external computer for its direct comparison with the data from the inertial sensors of the reference system.

The Physilog measurement system (Physilog®, Gait Up, Lausanne Switzerland) was used as the reference system to which B-B Score smartphone measurements were compared. The reference system's set-up was the same as that used in the Phase 1 study (please see sub-section 2.2.3, "Measurement procedure", within Chapter two, p. 68 - 69).

3.2.4.1. Measurement procedure

The measurement procedure was identical to used within the Phase 1 study with the exception that the smartphone was attached to the back of the arm by means of an armband, while simultaneous recordings were made by the reference system. The

lower edge of the smartphone was set 3 cm above the upper edge of the inertial sensors' module, which were themselves placed on each humerus, 3 cm above the midpoint of the line connecting the lateral epicondyle (EL) and medial epicondyle (EM). Thus, the smartphone was on the back of the patient's arm, the screen facing backward, when he/she was standing in the initial testing position with his/her arms along the body. Special attention has been paid to ensure that both smartphones were strictly at the same height (Pichonnaz et al., 2015a) (Figure 3.2). Similarly to previous work, angular velocities and accelerations in the sensor frame were used to calculate the B-B Score [11, 28].

a) 

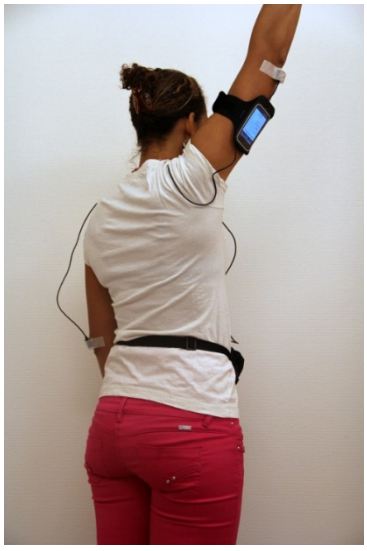
b) 

Figure 3.2: Inertial sensors and smartphone placement and axes (a) The inertial sensor module (Physilog® reference system) attached to the arm with medical tape and connected by cable to the datalogger carried on waist. The smartphone is attached to the arm by means of the armband. (b) Test completion of the "hand to the ceiling as to change a bulb" movement.

Movements were executed in a standing position following the smartphone-recorded instructions. The participants undertook first 3 repetitions of the two B-B Score movements on the healthy side (put hand to the back + hand to the ceiling as to change a bulb) and the task was then repeated on the pathological side. Preliminary trials showed that the smartphone B-B Score procedure completion took around 2-3 minutes, including smartphone set up, measurement, smartphone removal and results reading.

A team of six raters was constituted, from which pairs of evaluators performed the assessment sessions, while alternating the order of who undertook measurements. The first rater was randomly assigned. All measurement systems were detached for inter-rater administration of assessments to account for and incorporate the variability amongst measurements induced by possible inconsistent sensors' placement in clinics. All raters were experienced physiotherapists engaged in the project, who had been trained in the study protocol completion prior to their involvement in this Stage 2 study. The constitution of a relatively large team made it possible to adapt to controls' and patients' availability, and thus facilitated the recruitment.

Patient-reported outcome measures on shoulder function, pain and quality of life were also completed. Three PROMs were selected for shoulder function evaluation: the Quick Disabilities of the Arm and Shoulder score (QuickDASH), the Simple shoulder test (SST), the Constant Score and its variation, the Constant relative score (based on a percentage comparison of the measured value to an age- and sex-matched normal populations) (Lippitt, 1993; Constant and Murley, 1987; Fialka et al., 2005; American Academy of Orthopaedic Surgeons, 2009). The Constant Score was undertaken according to the modified guidelines of Constant (Constant et al., 2008). The PROMs and socio-demographic questionnaire are available in Appendix VII and a more detailed description of the selected PROMs is available in sub-section 5.2.7.3 "Characteristics of selected shoulder function PROMs", within Chapter five, p. 187 - 189**.**

This selection of shoulder function PROMs was made based on published literature reviews that investigated their frequency of use and the existence of a formal investigation process underlying the validity of the shoulder function outcome measures (Gartsman et al., 2015; Makhni et al., 2015; Fayad et al., 2004; Longo et al., 2011; Kirkley et al., 2003; Huang et al., 2015; Rouleau et al., 2010). The use of these PROMs allowed the evaluation of the convergent validity for the B-B Score but not of its validity against a gold standard, due to the controversy surrounding shoulder function evaluation.

The EuroQol [EQ-5D] and the pain visual analog scale (VAS) were also completed to capture a broader picture of patient clinical state (EuroQol, 2018). The EQ-5D is a validated generic quality of life PROM that includes 15 items investigating 5 dimensions of the quality of life (mobility, self-care, usual activities, pain/discomfort

and anxiety/depression) and a visual analogue scale to record the patient's self-rated health. The EQ-5D test-retest reliability is good to excellent. Some ceiling effects have been reported, particularly when used in general population surveys. A validated translation is available in the language of the study participants (French) and the form is easy to complete (Béthoux, 2003; EuroQol, 2018).

The VAS pain scale is an instrument on which the patient has to rate his/her pain intensity on a 10 cm scale representing the range between "no pain" and "the worst imaginable pain". It is a very widely used tool for subjective pain intensity evaluation, that demonstrated adequate reproducibly and responsiveness. The limitations of the pain VAS scale is that the pain ratings cannot be compared between patients and that the scale is not adapted to young children and to patients with cognitive impairments (Béthoux, 2003), which were considered to be minor drawbacks in the context of this study.

3.2.4.2. Analysis

Descriptive statistics including mean and standard deviation (SD) were performed for participants' characteristics and outcomes for the patient and the control group, using the reference device and the smartphone. These statistics were also calculated for the selected PROMs. In this phase of the thesis, the reporting of PROMs only intended to illustrate the performance level of included participants, the determination of the specific relationships amongst the PROMs and the B-B Score in each pathology being planned in the Phase 3 study. Box plots were also generated to illustrate the B-B Score outcomes for the patient and the control group, using the reference device and the smartphone.

Parametric tests were used to test the statistical significance of differences when the assumption of normality was met, as stated by the Kolmogorov-Smirnov test when $n > 50$ and the Shapiro-Wilk test when $n < 50$ (Yap and Sim, 2011), and the assumption of homoscedasticity was met, as stated by the Levene's test for equality of variance ($p > 0.05$). Non-parametric test were used when these assumptions were not met. The difference between the B-B Scores measured by each device was evaluated using the paired Student *t*-test. The difference between the patient and the control groups were evaluated using the Student *t*-test when assumption for the use of

parametric tests were meet, the Wilcoxon rank-sum test when they were not met and the Chi-square (χ^2) for nominal data.

The relationship between the B-B Scores of each device, and the intra- and inter-rater reliability were evaluated using the ICC, measurement error (ME: standard error of the mean difference), standard error of measurement [SEM: $(\text{pooled SD} \times \sqrt{1 - \text{ICC agreement}})$] and Bland and Altman analysis, including graphs generation and bias and LoA calculations. The $ICC_{agreement}$ for a single measurement was recorded, because the absolute agreement for measurements by a single rater is of interest in this study's context (Koo and Li, 2016). Intra-rater reliability was calculated comparing the 1st with the 2nd B-B Score obtained by the same rater, for the two raters. Inter-rater reliability was calculated comparing the score obtained by one rater with the score by the other rater, for the 1st and 2nd measurements made by each evaluator.

The discriminative power of the B-B Score for the detection of shoulder function loss was evaluated by the significance level for the differences between groups (Student *t*-test, $p < 0.05$). This calculation aimed at providing a first insight into the capacity of the B-B Score to detect differences that are likely to exist, which should be considered as a preliminary investigation. Further calculations related to discriminative power analysis were planned in the Phase 3 of the thesis to determine the magnitude of the change and the quality of the B-B Score compared to an external standard, specifically for each shoulder condition included in the study (De Vet et al., 2011c; Mokkink et al., 2010b).

3.3. Results

3.3.1. Study sample

Twenty healthy controls and 65 patients (20 with rotator cuff condition, 23 with fracture, 22 with capsulitis) were included.

The population characteristics and the significance of the differences between groups are described in Table 3.1.

Table 3.1: Participants' characteristics for the patient and the control group, with indication of the significant differences between groups.

	Patient (n = 65)	Control (n = 20)
Age mean (SD), years	58.5 (14.2)**	28.2 (6.2)
Sex (% women)	63	50
Weight mean (SD), kg	75.2 (15.8)	74.7 (17.4)
Body mass index mean (SD), kg/m^2	26.6 (5.8)	24.2 (3.9)
Size mean (SD), m.	1.68 (0.10)	1.75 (0.10)
Hand dominance (% right-handed)	92	90
Pathology (n)	Rotator cuff 20 Fracture 23 Capsulitis 22	-
Affected side (% dominant side)	43	-

Legend: ** Significant difference between groups with p-value < 0.01

The differences between groups were non-significant for the weight ($Z = 0.23$; $p = 0.81$), height ($Z = 1.94$; $p = 0.05$), BMI ($Z = 1.25$; $p = 0.81$) but significant for age $Z = 4.74$; $p < 0.01$), based on the Wilcoxon rank-sum test results. The difference between groups were non-significant for sex ($\chi^2(1) = 0.42$, $p = 0.52$) and hand dominance ($\chi^2(1) = 0.02$, $p = 0.90$), using the Chi-square test.

3.3.2. Score outcome

The B-B Scores outcomes of the control group and the patient group, for the smartphone and the reference system (Physilog), respectively, are presented in Table 3.2 and in Figure 3.3.

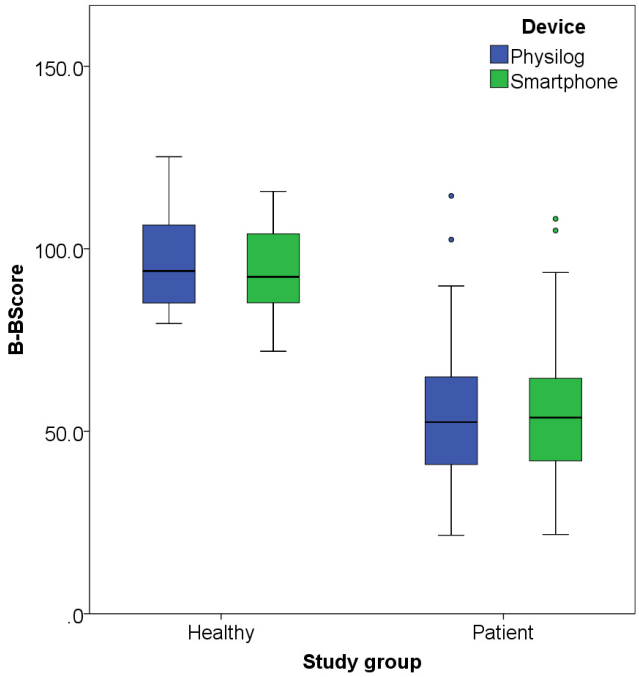


Figure 3.3: Traditional box plots showing median, lower and upper quartile, range and outliers (open circles, 1.5 interquartile range, B-B Scores, comparing the healthy control (n= 20) and the patient (n=65) groups using the reference system (Physilog, blue colour) and the smartphone (green colour)

Table 3.2: Mean and standard deviation of B-B Score using the smartphone and the reference system. Unit of scores are % representing the performance of the pathological side compared to the healthy side

Mean (SD), % **Min;max**	**Reference system**	**Smartphone**
Control	97.0 (13.8) 79.5 ; 125.2	94.1 (11.1) 71.9 ; 115.7
Patient	54.0 (19.0) 21.5 ; 114.5	54.1 (18.3) 21.7 ; 108.2

Legend: SD: standard deviation; Min: minimum measured value; Max: maximum measured value

The visual inspection of the box plots (Figure 3.3) highlighted the similarity between the outcomes obtained using the smartphone or the reference device for the patient and the control group, and the difference between the outcomes of the patient and the control groups, which were then further analysed using inferential statistics.

The Kolmogorov-Smirnov test confirmed the normal distribution of data, for the patient group using the reference device (df (65) = 0.06, p = 0.08) and the smartphone (df (65) = 0.07, p = 0.07), and the Shapiro-Wilk in the control group using the reference device (df (20) = 0.94, p = 0.21) and the smartphone (df (20) = 0.97, p = 0.78). All assumptions for parametric tests being met, these type of tests were used in further calculations.

The difference between the control and the patient group was statistically significant for the reference system (mean (SD) control group 97.0% (13.8) vs. patient group 54.0% (19.0) $t(83) = 9.41$, $p < 0.01$) and the smartphone (mean (SD) control group 94.1% (11.1) vs. patient group 54.1% (18.3), $t(83) = 9.23$, $p < 0.01$) (please see Table 3.2). These results confirmed the hypothesis that there would be a significant difference between the B-B Score outcomes of the patient and the control groups, (please see sub-section 3.1.1 “Study aim and hypothesis” of this Chapter, p. 100 - 101).

The difference between the reference system and the smartphone was non-significant for the control (mean (SD) 97.0% (13.8) for the reference system and 94.1% (11.1) for the smartphone, $t(19) = 1.39$, $p = 0.18$) and for the patient group (mean (SD) 54.0% (19.0) for the reference system and 54.1% (18.3) for the smartphone $t(64) = -0.18$, $p = 0.86$), as demonstrated by the result of the Student t-test (please see Table 3.2). This result confirmed the hypothesis that there would be no difference between the smartphone and the reference device when measuring groups (please see sub-section 3.1.1 “Study aim and hypothesis” of this Chapter, p. 100 - 101).

3.3.3. Measurement reliability

The numerical and graphical presentations of reliability of B-B Score measurements for inter-devices and intra- and inter-rater comparison are presented in Table 3.3 and Figure 3.4.

Table 3.3: Inter-devices and intra- and inter-rater reliability assessment using ICC, LoA, bias, ME and SEM for the B-B Score outcomes (%, original units) acquired using the smartphone or the reference system (n = 85).

	ICC (95% CI)	LoA (%)	Bias (95% CI)	ME (%)	SEM (%)
Inter-devices	0.97 (0.94 - 0.98)	-13.2 – 12.0	- 0.6 (-0.9 – 1.1)	0.7	4.0
Intra-rater					
Smartphone	0.92 (0.89 - 0.94)	-17.4 – 20.3	1.5 (0.0 – 2.9)	0.7	6.6
Reference System	0.92 (0.89 - 0.94)	-19.3 – 19.6	0.1 (-1.4 – 1.6)	0.8	6.6
Inter-rater					
Smartphone	0.92 (0.90 - 0.94)	- 16.9 – 20.0	1.5 (0.1 – 3.0)	0.7	6.6
Reference System	0.93 (0.91 - 0.95)	- 18.1 – 20.0	1.0 (-0.5 – 2.4)	0.7	6.4

Legend: ICC: intraclass coefficient of correlation; 95%CI: limits of interval at 95% confidence level; LoA: limits of agreement; ME: measurement error; SEM: standard error of measurement

Intraclass correlation coefficients for inter-device, intra-rater and inter-rater assessment presented in Table 3.3 were all above the threshold defined for clinical utility (ICC ≥ 0.90; please see sub-section 1.1.3.4.2.2 “Test-retest, intra- and inter-rater reliability”, within Chapter one, p. 40 - 41). Therefore, the hypotheses that the B-B Score would have adequate reliability for clinical measurements, regardless of the rater, the measurement and the use of a smartphone or an inertial sensor system were met (please see sub-section 3.1.1 “Study aim and hypothesis” of this Chapter, p. 100).

The measurement error, that represents the standard error of the mean difference, showed that the differences between devices, raters or measurements had minor influence on the variability of group scores, as they ranged from 0.7% to 0.8% (Table 3.3).

The biases i.e. mean values of the differences between the measurements, indicated that the magnitudes of the systematic errors between devices, raters or measurements were also minor, as they ranged from -0.6% to 1.5% (Table 3.3 and

Figure 3.4. Thus, the hypothesis that the bias would be ≤ ± 5% was met (please see sub-section 3.1.1 “Study aim and hypothesis” of this Chapter, p. 100 - 101).

The LoAs, which represent the ranges that contains 95% of random measurement differences, were lower for inter-device (-13.2 – 12.0%) than for intra- or inter-rater comparisons (upper limit up to 20.3%) (Table 3.3 and Figure 3.4). Nevertheless, all the LoAs were higher than the ≤ ± 10% threshold that had been defined for adequate agreement (please see sub-section 3.1.1 “Study aim and hypothesis” of this Chapter, p. 100 - 101). The range of error associated with a single measurement should thus be taken into consideration when repeated assessments of the B-B Score are performed, as the magnitude of error may have a clinically significant influence on the measured outcome in some cases.

The visual inspection of the Bland and Altman graphs showed an increase of the error at higher scores.

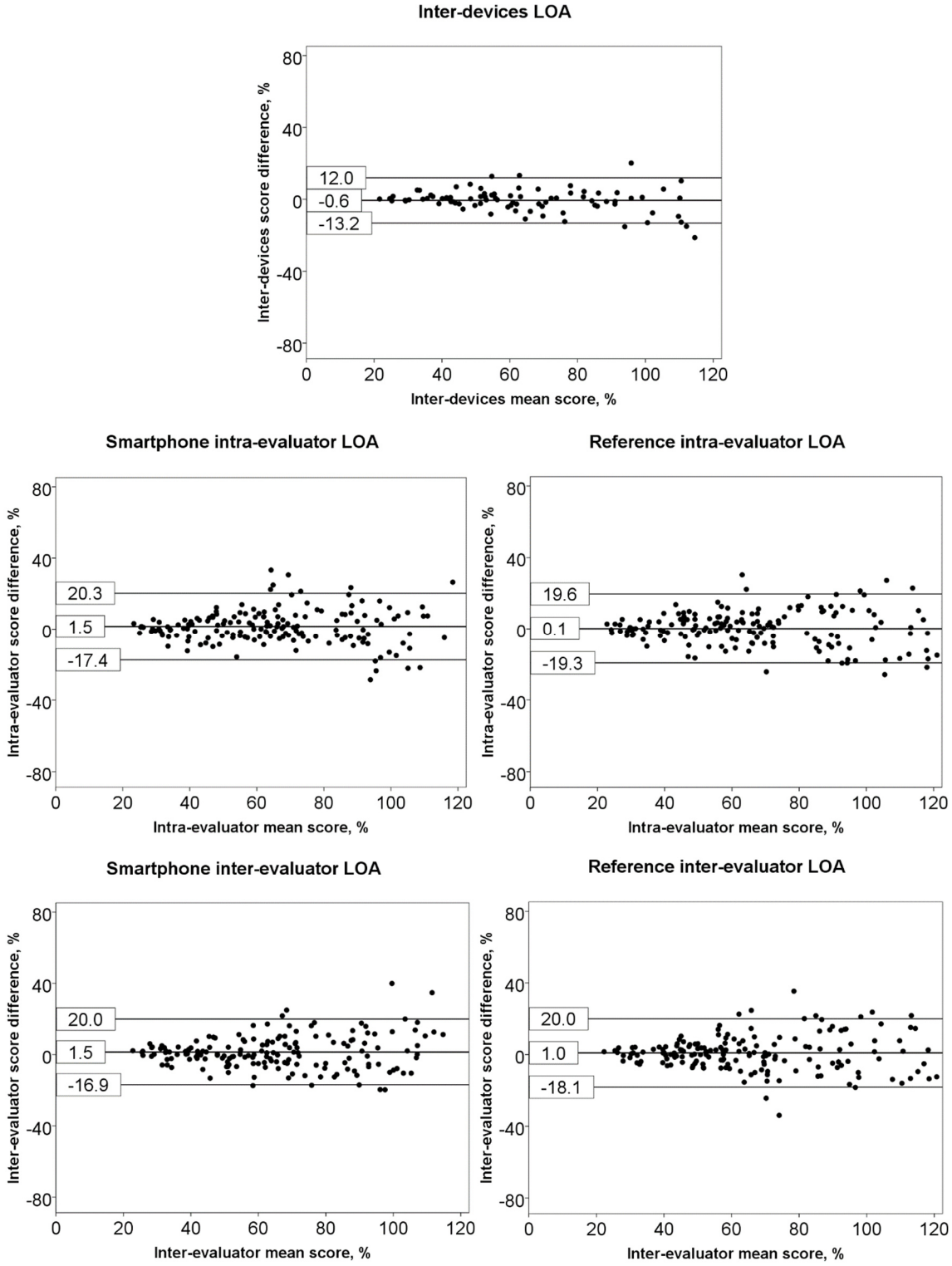


Figure 3.4: Bland and Altman plots of bias and 95% limits of agreement (%, original units) as estimates of inter-devices, intra- rater and inter-rater limits of agreement of B-B Scores, recorded across two serial measurements by two separate raters, using the reference device and the smartphone.
Legend: LoA: limits of agreement.

3.3.4. Patient-rated outcome measures

The results of shoulder function, pain and quality of life PROMs are presented in Table 3.4.

Table 3.4: Mean group outcomes of patient-reported outcome measures for the patient and the control group, with standard deviations, minimum and maximum measured values.

PROMs mean (SD) * Min; max	Patient (n = 65)	Control (n = 20)
Constant Score (SD), points	42.8 (17.9)	93.7 (6.6)
	10; 85	80; 100
Relative Constant Score (SD), %	55.5 (23.9)	97.6 (7.5)
	12; 110	82; 108
SST (SD), points	4.6 (3.1)	11.9 (0.2)
	0; 12	11; 12
QuickDASH (SD), %	42.8	1.1 (2.5)
	0.0; 86.4	0.0; 6.8
VAS pain (SD), mm	40.5 (24.2)	0.9 (2.7)
	0; 81	0.0; 10
EQ-5D (SD), index	0.70 (0.19)	1.00 (0.00)
	- 0.18; 1.00	1.00; 1.00
EQ-5D VAS (SD), points	74.3 (18.0)	98.4 (44.9)
	10.0; 100.0	85.0; 100.0

Legend: * Best possible scores: Constant Score100 points, relative Constant Score theoretically no limit (scores in % based on an age-and sex-matched normal population for Constant Score), SST 12 points; QuickDASH 0, VAS pain 0, EQ-5D 1.00 (index score of a value set derived from the general population sample), EQ-5D VAS 100.

The inspection of Table 3.4 outcomes showed that the patient and control groups were representative of the populations that they are supposed to represent. The outcome of the patients on the PROMs demonstrated that the patient sample level of shoulder dysfunction was realistic, with regard to that encountered in clinical practice and in the literature. The outcome of the control group was near from the maximum score, indicating that the controls had, as expected, no shoulder dysfunction.

3.4. Discussion

This study investigated the validity and reliability of a smartphone-assessed kinematic shoulder function B-B Score, and compared the performance of the smartphone to a reference inertial sensor system. Using shoulder function scores derived from a dedicated smartphone application, the study aimed at the technical and, to some extent at the clinical validation of them within a sample including various shoulder pathologies. Provided that the score was valid, it could offer a valuable alternative to concurrent MAB outcomes measures of shoulder function, as it was accessible and quickly performed.

3.4.1. Study sample

No significant difference was observed between the characteristics of the patient group and control group participants, except for age. The control group was purposefully younger than the patient group as it was of primary importance that the reference population had healthy shoulders. The characteristics of the patient group were representative of the population commonly treated for shoulder pain (Picavet and Schouten, 2003; van der Windt et al., 1995).

There were no deviations away from the planned sampling for this study. The study sample was sufficient to obtain precise results, as indicated by the narrow 95% confidence interval (maximum ± 0.025 for ICC; maximum ± 1.5 for bias) (Table 3.3).

3.4.2. Devices' comparison

The reference system (Physilog®) and the smartphone produced comparable B-B Score outcomes regarding group measurements. Although the specificities of the measurement systems were different, for example sensors' noise, sensors' ranges and sampling frequency, the performance of the smartphone appeared to be sufficient for the B-B Scores' proper measurement (for technical features please see sub-section 2.2.2 "Measurement device", within Chapter two, p. 68 for the Physilog and sub-section 3.2.4 "Experimental systems: smartphone and reference system" within this Chapter, p.103 for the smartphone).

The mean differences between the devices were non-significant in the patient (p = 0.86) and the control group (p = 0.18), and of limited magnitude (0.1% for the patient group and 2.9% for the control group). These differences are unlikely to have a clinically significant influence on the measured outcome, as they are minor in proportion to the 42.9% and 40% difference in B-B Score between the patient and the control group, for the reference system and the smartphone, respectively (Table 3.2).

Based on these results the hypothesis that there would be no difference between the smartphone and the reference device when measuring groups was confirmed, making the smartphone a possible substitute to inertial sensor systems for the evaluation of groups' outcomes using the B-B Score.

An excellent relationship was found between measurements from the devices (ICC 0.97) (Table 3.3). Moreover, the Bland and Altman analysis demonstrated that the systematic error of the smartphone was minor. The ME (0.7%) and SEM (4%) were proportionally small compared to the difference observed between the patient and the control group, using the reference system (43%) or the smartphone (40%) (Table 3.2). Thus, the measurement error related to the device was not expected to interfere importantly with the capacity of the B-B Score to classify correctly the participants as patients or healthy controls. Conversely, the LoA exceeded the 10% criterion that had been selected to define an acceptable threshold (please see sub-section 3.1.1 "Study aim and hypotheses", p. 100-101) (Walter et al., 1998; Portney and Watkins, 2015). Thus, the Physilog and the iPod are interchangeable for group measurement, but the magnitude of the LoA might preclude the devices' routine exchange when measurements concern individual participants.

3.4.3. Groups' comparison

The B-B Score difference between the control and the patient groups was highly significant ($p < 0.01$) and large, regardless of the devices (mean (SD) control group 97.0% (13.8) vs. patient group 54.0% (19.0) for the reference device and mean (SD) control group 94.1% (11.1) vs. patient group 54.1% (18.3) for the smartphone) (Table 3.2). Hence, the B-B Score clearly discriminated the performance of the patient group from that of the healthy group.

These results are in line with the hypothesis that there would be a significant difference between the B-B Score outcomes of the patient and the control groups, which confirmed the capability of the Score to discriminate groups for which a difference is expected (please see sub-section 3.1.1 “Study aim and hypotheses”, p. 100-101).

3.4.4. B-B Score intra- and inter-rater reliability

The intra- and inter-rater reliability was excellent (0.92 – 0.93) and comparable between devices (Table 3.3). The hypothesis that the B-B Score ICCs would be higher than the level recommended for clinical measurement (≥ 0.90) for inter-device, intra-rater and inter-rater comparisons was met, regardless of the device used for the measurement (Portney and Watkins, 2015).

As shown by the small bias derived from the Bland and Altman analyses (≤ 1.5%, while the threshold defined for clinical utility was ≤ ± 5% in this study) (Table 3.3 and Figure 3.4), the B-B Score’s biases related to the device, measurement and the rater were relatively minor, indicating that the systematic errors were not likely to interfere significantly with clinical measurements.

Conversely, for both devices, the LoA for the intra- and inter-agreement of the B-B Score had exceeded an arbitrary ≤ ± 10% threshold defining the upper 95% confidence limit for measurement error associated with its clinical utility (please see sub-section 3.1.1 “Study aim and hypotheses”, p. 100-101). The limits of agreement ranged from ± 18.5% to ± 19.4% on both sides of the bias Table 3.3 and Figure 3.4). The visual inspection of the Bland and Altman graphs showed an increase of the error at higher scoring, due to the technical aspects of scoring. This is related to the formula used for the determination of the B-B Score, which expresses the shoulder function as the ratio of the performance of the affected side relative to the healthy side (or the dominant side relative to the non-dominant side for healthy controls), reported as a percentage. Thus, variations in the affected shoulder (denominator in the formula) have proportionally more influence on the score when its performance is near from that of the healthy shoulder (numerator). Thus, the B-B Score tends to be more stable for patients who perform at a low functional level.

Thus, the results are comparable between replications and between raters for measurements focusing on the performance of a group, but excessive variations and divergence amongst repeated assessments of the B-B Score are possible even when the outcome is derived from the mean of three repetitions, as has been used in this study's protocol. Measurements relating to the assessment of a single patient is still feasible but would be expected to require acquiring the mean of more than three replications in order to counteract inflated error and establish the requisite precision of measurement (Mercer and Gleeson, 2002), as the variability and error in a measurement mean score decreases with the square root of the repetitions' number (assuming a normal distribution of error). The simplicity of the procedure for assessing the B-B Score facilitates measurement repetition and largely overcomes this limitation.

3.4.5. Comparison with PROMs for criterion validity determination

The kinematic measurements were also compared to currently-used PROMs for benchmarking. The PROMs included estimates of shoulder function (Constant, Relative Constant, SST and QuickDASH), pain (VAS) and quality of life (EQ-5D).

In healthy participants, both the PROMs and the kinematic B-B Score had indicated near to the maximum performance, showing that the reference population had almost perfect shoulder function. For patients, the observed importance of shoulder function loss was globally comparable between the PROMs and the B-B Score, with all scores indicating a substantial function' loss in the measured sample (from Table 3.4, 42.8/100 points for the Constant Score, 55.5%/100% for the relative Constant Score, 4.6/12 points for the SST and 42.8/100 points for the QuickDASH). Thus, it appeared that in this study the B-B Score (54.0% using the reference system and 54.1 using the smartphone) produces coherent results to those from the shoulder function PROMs in terms of measured loss of function, regardless of the device used.

These results were in line with previously published results on the relationship between the B-B Score and PROMs, which showed moderate to high correlations of the B-B Score with scores from the Constant and SST and moderate correlations with the QuickDASH for various shoulder pathologies (Pichonnaz et al., 2015a). The

relationship between the B-B Score and the PROMs will be further explored and detailed in Phase 3 specifically for each included shoulder pathology.

3.4.6. Shoulder function evaluation by body-worn sensors in the literature

Most previous studies that had investigated the measurement properties of body-worn sensors for shoulder function scores used dedicated inertial-based systems (Coley et al., 2007a; Korver et al., 2014a; Korver et al., 2014b; Pichonnaz et al., 2015c; Jolles et al., 2011; Duc et al., 2013; Duc et al., 2014; Luinge and Veltink, 2005; Cutti et al., 2008; de Vries et al., 2016). All of these studies concluded that the inertial-based systems produced a valid evaluation of shoulder function. However, no comparison with a concurrent wearable system has been reported. To the best of our knowledge, the present study was the first to investigate the concordance and the relationship of a smartphone-based and a reference inertial-based system for shoulder function evaluation. The results are valuable for research and clinics, as they demonstrate that the validity of the B-B Score measurement is not altered when using a simple and accessible device.

3.4.7. Study limitations and further developments

This study provided a novel comparison of a smartphone with a reference device for the measurement of the B-B Score but did not yet provide a complete insight into the measurement properties of the B-B Score, with the exception of its focus on the reliability of measurements associated with intra- and inter-evaluator assessments. Although both devices (reference and smartphone) might have been deemed capable of offering equivalent measurement reliability characteristics for the assessment of the B-B Score, based on the results from the current Phase 2 study, it was still plausible that the B-B Score might not appropriately reflect shoulder' function status and its change over time. Furthermore, the definition of a norm and interpretability aspects was still lacking to support the correct interpretation of the B-B Score and of its change over time. Thus, further research was conducted within the next phase of the thesis to investigate the latter issues and to compare the B-B Score with alternative measurement methods.

The results thus far apply to a situation in which the measurements had been performed under supervision and at the patient's self-selected speed of movement. Further investigations are needed to determine the validity of the B-B Score in other conditions. For example, the relationship between assessment devices might be different if the patients perform movements associated with the B-B Score at their maximum speed, due to the difference in sensors' characteristics. Measurement' reliability might also be different if the patient performs the test without supervision, as would be the case in telemedicine applications.

The results were not detailed separately for each pathological subgroup within this Phase 2 study. This might be considered a minor limitation with regard to the study's objectives, as the relationship between devices is likely to be far more influenced by the testing conditions rather than by the pathology. Conversely, the use of a larger patient group had the advantage of providing more precise estimations of the reliability of the B-B Score's measurements.

Despite the widespread use and the convenience of smartphones, there are also limitations in their use for scientific measurement. The precise features of the device are not fully disclosed by manufacturers due to commercial sensitivities. The users should remain conscious that the characteristics may differ according to the smartphone version and brand. In view of these potential issues, it seemed reasonable to have chosen an accessible middle-segment smartphone model in order to offer insight into its performance' characteristics for the type of measurements that the B-B Score requires. The B-B Score would probably remain robust when faced with minor variations in smartphone' technology, as it would have compared the performance of the affected shoulder with that of the healthy one, with the score unaffected by systematic errors in measurement affecting both sides (Pichonnaz et al., 2015c). However, a study comparing the performance amongst smartphones should be conducted to investigate this assumption.

Based on the findings from this Phase 2 study and the body of literature on the subject, it appears that smartphones most likely offer measurement properties that are compatible with research requirements for measurements comparing both sides and for range of motion measurements (Shin et al., 2012; Werner et al., 2014; Mitchell et al., 2014). Nevertheless, the validity of using smartphones for more complex measurements, for example those associated with 3D kinematic analysis of sport

activities, remains unknown to date. In addition, the aforementioned variations in smartphones' features imply that further research is needed to investigate and quantify the influence of variations in the measurement context on the B-B Score's outcome before its clinical implementation.

The duration required to conduct the whole procedure to assess a B-B Score using the smartphone was around two minutes. All things being equal, the advantage of the measurement approach used in this study mainly resides in its clinical practicality and low cost. Further research may extensively investigate the smartphone B-B Score's specific measurement properties including convergent validity, responsiveness and interpretability aspects specifically in different shoulder pathologies. Thus, it was planned to address these issues in the Phase 3 study of the thesis, in order to provide a broad overview of the measurement properties of the smartphone B-B Score for potential users.

As part of a general approach by the B-B Score's research team within the DAL-CHUV and Laboratory of Movements Analysis and Measurement of the Swiss institute of Technology (LMAM-EPFL) to improve access to this approach to the assessment of shoulder function by clinicians and patients, an android version of the application has been developed and made available to the public (Gait Up, 2018). The latter offered an important adjunct to this thesis, facilitating research into the further development of the smartphone approach to assessing shoulder' function in order to accrue maximum benefits from it in situations where that might be warranted. A presentation of the B-B Score application features is available in Appendix XI.

This type of smartphone application might also underpin future developments facilitating the communication of clinically-relevant results between stakeholders, producing progression curves of functional improvements and comparing the patient's change of performance during care-pathways to benchmark results on a routine basis.

For recall, Phase 2 study aimed to investigate the validity and reliability of a smartphone-assessed kinematic shoulder function B-B Score, and to compare the performance of the smartphone to that of a reference inertial sensor system. Further developments that will be conducted in Phase 3 study will aim at the determination of the measurement properties of the smartphone B-B Score for the assessment of

current shoulder pathologies (rotator cuff condition, capsulitis, proximal humerus fracture and shoulder instability).

3.5. Conclusion

This study aimed at the technical and clinical validation of a B-B Score smartphone application for the evaluation of the functional capabilities of the shoulder. Either the assessments acquired using a smartphone or a reference inertial sensor system displayed comparable measurement properties across a wide-range of clinimetrics. This comparison is to the advantage of the smartphone, which is more accessible, cheaper and more user-friendly than dedicated movement analysis inertial sensor systems.

The results showed that the B-B Score acquired by means of a smartphone, was valid, reliable and reproducible for the measurement of shoulder function of groups of patients including those presenting with rotator cuff conditions, proximal humerus fractures or adhesive capsulitis. It displayed excellent intra- and inter-rater reliability and discriminative power. Conversely, single assessments of the B-B Score, even when involving the mean of three measurements, may offer reduced precision in some circumstances.

Thus, the B-B Score measured with a smartphone allows valid, user-friendly and low-cost evaluation of shoulder function for research and clinical work. This could facilitate the use of objective measurement methods for shoulder function evaluation in routine practice and thus improve the quality of patient follow-up. Further research is needed to investigate extensively the smartphone B-B Score's specific measurement properties in various patient populations, which will be addressed in the next phase of this thesis. It may also investigate the influence of the specific characteristics of various smartphone' models on results. Further technological developments are also required to achieve maximum benefit from the smartphone approach.

CHAPTER FOUR

SCORE MEASUREMENT PROPERTIES STUDY

4.1. Introduction

4.1.1. Study context

Research results are strongly influenced by the quality of measurements. In addition, important decisions concerning patients are taken based on measured outcomes. Thus, the establishment of the measurement properties of an evaluation tool is paramount before it is used in clinical conditions.

The Phase 2 study demonstrated that the transfer of the B-B Score to a smartphone did not alter the measured score or its reliability compared to the score measured using a dedicated inertial measurement system. In isolation, these results support the use of smartphones over dedicated movement analysis IMU systems, due to their accessibility, user-friendliness and low-cost.

Nevertheless, the fact that the measurements were comparable between devices does not necessarily imply that the B-B Score had acceptable measurement properties in the target populations of patients exhibiting different types of frequent shoulder conditions such as rotator cuff conditions, capsulitis, shoulder instability or proximal humerus fracture. It was therefore necessary to undertake further analyses to investigate if the smartphone B-B Score measurement properties were acceptable under a wider range of assessment challenges.

The Phase 3 study was undertaken to investigate the measurement properties of the B-B Score acquired using a smartphone and focused attention on the important issue of the latter's capability for delivering high-quality assessments of shoulder function amongst varied types of clinical conditions. Considering that measurement properties are population-dependent, they were to be established specifically for each of the main shoulder pathologies encountered in physiotherapy (Robertson et al., 2017; Riddle and Stratford, 2013; Collins and Roos, 2016).

As a reminder of information offered in Chapter one (subsection 1.1.2.4 "Thesis aim" p. 11 – 12), the data of Phase 2 and 3 were collected simultaneously. The first step, corresponding to Phase 2, presented in Chapter three, aimed at the assessment of the smartphone measurement capacities compared to an inertial sensor system used as a reference device, regardless of pathology. The second step, encompassing the

work of this Chapter, aimed at the extensive investigation of the B-B Score measurement properties for several frequent shoulder conditions, using the most efficient device, i.e. the smartphone as concluded from Phase 2 results. This phase implied a more detailed and targeted data analysis and the collection of follow-up data in order to investigate the B-B Score change over time.

Aspects of the findings of this Phase 3 study have been published in the peer-reviewed open-access journal Sensors (Thomson Reuters 2017 impact factor 2.48) (Pichonnaz et al., 2015a) (Appendix XII).

4.1.2. Definition of the target populations

The targeted populations included patients with rotator cuff conditions treated conservatively, shoulder instability treated conservatively, proximal humerus fracture treated surgically or conservatively, and capsulitis treated conservatively.

Conservatively treated populations were investigated because they represent much larger populations than the surgically treated ones. Overall, only one in every 10 patients presenting with shoulder pain requires surgery (Colvin et al., 2012). Moreover, some results were already available for the postsurgical context, as the B-B Score was developed in a population who had undergone rotator cuff and arthroplasty surgery (Pichonnaz et al., 2015c). In addition, it had been previously established that the B-B Score produced comparable results to the kinematic P Score, which has itself demonstrated to be valid and responsive following shoulder surgery (Coley et al., 2007a; Coley, 2007; Jolles et al., 2011).

Patients with rotator cuff conditions, proximal humerus fractures, adhesive capsulitis, and shoulder instabilities are frequently encountered in shoulder consultations (van der Windt et al., 1996; Yamamoto et al., 2010; van der Windt et al., 1995; Court-Brown and Caesar, 2006; Liavaag et al., 2011; Owens et al., 2007). It was thus essential to investigate the measurement properties of the B-B Score for these conditions. The characteristics of these conditions have been previously developed in sub-section 1.1.1.1. “Impact of main shoulder conditions on function”, within Chapter one, p. 2-4.

4.1.3. Measurement properties to be investigated

Multiple qualities are expected from a measurement instrument to ensure that the result gives a correct representation of the reality. These qualities are encapsulated by the concepts of validity, reliability and responsiveness, which all encompass several aspects that contain specific measurements properties (Mokkink et al., 2010d). In addition, the determination of normal performance and interpretability aspects is of importance for the interpretation of the results (Tubach et al., 2007).

These notions are not hereby detailed, as they have been extensively developed in sub-section 1.1.3.4 “Clinimetrics”, within Chapter one, p. 29 - 50.

Content and construct validities were not addressed in this thesis, because the rationale underlying the design of the B-B Score had been investigated and justified in a previous research, that aimed at the selection of essential movements for the evaluation of movement analysis-based shoulder function (Pichonnaz, 2010; Pichonnaz et al., 2015c). Conversely, criterion validity was investigated in this Phase 3 study, as it was still to be established in the targeted populations. In the absence of a universally recognised PROM for shoulder function evaluation, criterion validity but no gold standard validity could be established (McDowell, 2006).

Notwithstanding the aforementioned exceptions, the Phase 3 study protocol was designed to investigate as extensively as possible the B-B Score’s measurement properties and to provide users with the information they need to apply the Score with critical hindsight. The investigated properties are summarised in Table 4.1.

Table 4.1: Investigated measurements properties and their aspects (where applicable) with applied method.

Measurement property	Aspects of measurement property	Method
Validity	Concurrent	Correlation with PROMs
	Discriminative power	Difference between groups
		Difference between stages
		Area under the ROC curve, sensitivity-sensibility for the discrimination between patients and controls
Responsiveness	Responsiveness	Area under the ROC curve, sensitivity-sensibility for shoulder function change detection
		Change score correlation
		Effect size (comparison between outcome measures)
		Standardised response mean (comparison between outcome measures)
Measurement error		SEM
		MDC
Interpretability		Normal performance range
		MCII/MCID
		PASS
		LoA

Abbreviations: SEM: standard error of measurement; MDC: Minimal detectable change; MCII: Minimal clinically important improvement, MCII: Minimal clinically important difference PASS: Patient acceptable symptom state; LoA limits of agreement

4.1.4. Study aim and hypotheses

This study was aimed at the determination of the measurement properties of the smartphone B-B Score for the assessment or the progression of current shoulder pathologies (rotator cuff condition, capsulitis, proximal humerus fracture and shoulder instability).

Based on two assessments acquired over a six-month period, it was hypothesised that:

- the B-B Score would remain stable in the control group ($p > 0.05$), while it would progress significantly ($p < 0.05$) over time in each pathological subgroup,
- the responsiveness assessed using effect sizes (ES) and standardised response means (SRM) would be comparable to that of validated PROMs,
- the area under the receiver operating characteristic (ROC) curve indicative of discriminative power between patients and controls, and the ROC curve indicative of discriminative power between improved and unimproved patients at the 6 months follow-up, would be at least adequate (AUC ≥ 0.70) and comparable to that of validated PROMs (De Vet et al., 2011c; McDowell, 2006; Jimerson, 2007),
- the correlations with PROMs and the correlation between change scores would be at least moderate ($r \geq 0.50$) (Munro, 2005; Portney and Watkins, 2015),
- no floor and ceiling effect would be detected

No hypothesis was made about the MDC, MCII, and PASS values as these investigations primarily aimed at the determination of these values for the needs of clinical evaluation. For the definition of the used methods and the rationale that underpin their use in this study, please see sub-section 1.3.4.1.3 Construct validity p.35 - 37 and 1.1.3.4.3 “Responsiveness” p. 44 - 49.

4.2. Methods

As data for Phase 2 and Phase 3 were collected in the same time and had an intrinsic commonality, the measurement protocol used in both Phases were identical to that of Phase 2, which was reported within Chapter three section 3.2 “Methods” p. 101 – 109. Only the data collected using the smartphone have been analysed and are reported

hereafter, as this device has previously demonstrated its efficiency for the B-B Score calculation in the Phase 2 study. An additional measurement session was conducted six months after the baseline measurement, using the same measurement protocol as the Phase 2 baseline session, with the exception that only one rater collected the data and that the patient had to assess his or her progress on an anchoring questionnaire designed for the determination of the MCII and the PASS.

The 6 months' time interval had been chosen, as it constituted elapsed time that could realistically be considered as sufficient for most of the patients to have an evolution of their shoulder condition, whether spontaneous or induced by treatment. This was required to enable the assessment of the responsiveness to change of health state over time. No standardisation (e.g. of treatment or patient's activity) was implemented between the initial and follow-up evaluation session. This was not considered as being necessary, as these elements are not expected to have an important influence on the measurement properties of a score, which were the focus of the thesis.

On this anchor questionnaire, the patient had to rate the state of his/her shoulder in the last 48 hours compared to 6 months earlier, as worse, unchanged or better. If the answer was "better", he/she had to rate the change as unimportant, light, moderate or very important. He/she had then to rate whether he/she considered his/her present state acceptable or unacceptable.

Eligible patients residing in the canton, as indicated by the inspection of their medical records, were contacted by phone in the order in which they attended the medical consultation in the department (for authorisation to screen patients' medical records, please see Appendix X Accord éthique accès Soarian). With the exception of patients with humerus fractures, patients who gave their consent underwent a baseline measurement session within two weeks following the medical consultation, and a second session six months later. For patients with humerus fractures, measurements were performed six weeks post-stabilisation and six months later, provided that the radiological control showed normal healing.

The Phase 3 study was registered under ClinicalTrials.gov Identifier: NCT01431417 simultaneously to Phase 2 study (Appendix IX).

Conversely to the aims of the Phase 2 study, involving *inter alia*, the capability of the smartphone B-B Score to discriminate globally between patients for which the Score can be used in the future and healthy controls, the results of patients with shoulder instability have also been specifically reported hereafter within the Phase 3 study, in addition to the specific results for patients with rotator cuff conditions, proximal humerus fracture and capsulitis. The measurement procedure was strictly the same for patients with shoulder instability than for other patients, with the exception that they completed in addition the WOSI, a specific shoulder function PROM for shoulder instability (Kirkley et al., 1998). The selection of this PROM was made based on the same criteria than other PROMs, i.e. the fact that published literature reviews investigated the frequency of its use and the existence of a formal investigation process underlying the PROM validity. The WOSI was preferred to the Rowe score, which is frequently used but did not meet the second criterion (Gartsman et al., 2015; Makhni et al., 2015; Fayad et al., 2004; Longo et al., 2011; Kirkley et al., 2003; Huang et al., 2015; Rouleau et al., 2010). Moreover, several versions of the Rowe Score have been produced, without it being clearly established which version should be presently used (Jensen et al., 2009).

4.2.1. Study sample

A specific sample size calculation was made for the Phase 3, to sustain the soundness of the calculation of subgroups by conditions. Calculations were based on the data of the Phase 1 study that had included seven controls and 16 patients.

The rationale underpinning the power calculation was to include a sufficiently high number of patients to ensure a 0.80 power for each one of the statistical tests of hypotheses at the study's primary end-point, when a sample size calculation method existed and data were available to estimate the sample size. Thus, the sample size calculations were made for correlations, for difference between groups and for ROC curves.

The calculation was made so that, with a significance level at $p < 0.05$, the power of 0.80 was reached when the minimal standards for acceptable properties of the B-B Score were met. For convergent validity, 18 patients per group were needed for a significant correlation when the correlation was moderate ($r \geq 0.50$), as expected in the study hypotheses (sub-section 4.1.4 "Study aim and hypothesis, p.130). For

discriminative power for improvement and for diagnostic purposes, 11 patients were required for an area under a ROC curve of 0.80 with a standard error of 0.1 ensuring that the power was at least adequate (AUC ≥ 0.70), as expected in the study hypotheses (Chang, 2014). For discriminative power between groups, nine patients were required for a significant difference between the patients and the control group, based on the same 'pilot' effect sizes shown in the Phase 1 study (Soper, 2004; Lenth, 2010). According to these estimations, 20 patients were enrolled in each subgroup of pathology and 20 healthy controls in the control group. As these estimations applied to baseline and to 6 months measurements, patients lost at follow-up were compensated by including an equivalent number of additional patients to reach the required sample size at 6 months.

4.2.2. Analysis

Descriptive statistics including mean, standard deviation (SD) were performed for participants' characteristics and outcomes (B-B Score and PROMs) for the control group and each subgroup of patients at baseline and at six months. Box plots were also generated to illustrate the B-B Score outcomes for the control group and each subgroup of patients at baseline and at six months.

The assumptions for the use of parametric tests were checked, using the Shapiro-Wilk test for the assumption of normal distribution and the Levene's test for equality of variance for the assumption of homoscedasticity (Yap and Sim, 2011). Based on these verifications, non-parametric tests were used because the assumption of normal distribution was not met in several cases ($p < 0.05$). The differences between pathological subgroups and the control group were analysed using the Mann-Whitney (Wilcoxon rank-sum test) or the chi-square tests as applicable, and the differences between stages were tested for each pathological subgroup and the control group using the Wilcoxon signed-rank test.

The responsiveness for the baseline- six months change was calculated using the Cohen's *d* effect sizes (ES) with a 95% confidence interval, the standardised response mean (SRM) with a 95% confidence interval, the Spearman correlations between change scores and the area under the receiver operating characteristic curve (AUC). The sensitivity, specificity and optimal detection threshold (highest sensitivity-specificity ratio) were also derived from the ROC curve analysis. The discriminative

power between patients and controls was calculated using the same ROC curve analyses.

It was considered that a floor effect existed if > 15.0% of patients scored lower than a threshold set at 0 + MDC at baseline. This threshold account for the fact that patients scoring slightly above zero but within the error limits around zero might possibly have got the lowest possible score, and should therefore be taken into account in the calculation of the floor effect (Terwee et al., 2007; McHorney and Tarlov, 1995). Considering the ceiling effect the B-B Score has theoretically no upper limit. However, we investigated the number of patients reaching more than 100%, because, though some patients might exceed this result following treatment, this is not likely to be frequent.

For convergent validity assessment, the Spearman correlations were used to assess the strength of relationship between the B-B Score and the PROMs for each of the pathologies. Concerning the interpretability aspects, the MCII and PASS were determined for the patient group using the anchor-based method as described in Tubach *et al.* 2007) and presented in sub-section 4.2 “Methods” within this Chapter p. 130, and sub-section 1.1.3.4.3.6 “MCID/MCII” and 1.3.4.3.7 PASS, within Chapter one p. 48 - 50.

Concerning measurement errors, the MDC_{95} (was calculated using the formula $MDC_{(95\% \text{ confidence level})} = 1.96 * \sqrt{2} * SEM$, (Beaton et al., 2001a), where the SEM was determined using the formula SEM = pooled SD * $\sqrt{}$1-ICC. The pooled SD was calculated from the data of the four measurements that were done for each patient at baseline, which represented the most precise evaluation of the real performance of each pathological subgroup. The ICC value used in the calculation was 0.92, this value being valid both for intra- and inter-evaluator reliability, as determined in Phase 2 study.

The results were reported separately for each pathology and for the control group. When relevant for the comparison with the existing literature, the results were also reported for the whole patients sample, called “All patients” group (n = 88) and for the sample of patients with pathologies for which the use of the B-B Score is indicated, called “Indicated pathologies” subgroup (n = 65) (i.e. rotator cuff conditions, capsulitis

and proximal humerus fractures, but not instability, as will be demonstrated by the discriminative power analysis made in this phase). The ROC curve for improved/unimproved patients, the MCII value and the PASS value were calculated for the whole patient group only, because their calculation methods imply to dichotomise the group into two smaller groups (improved/unimproved for the MCII and ROC curve, acceptable/not acceptable state for the PASS). Therefore, the calculation of these values for each pathological subgroup would have been based on too few patients, especially in the unimproved group, to be precise.

4.3. Results

4.3.1. Study sample

One hundred and eight participants were tested at baseline (20 healthy controls, 20 patients with rotator cuff condition, 23 with fracture, 22 with capsulitis and 23 with shoulder instability). The participants were measured again six months after the baseline measurement. Four patients could not be contacted at six months and four refused to participate for reasons without relationship with the study (1 patient with rotator cuff condition, 3 with fracture, 1 with capsulitis and 3 with instability). Dropout rate was low (7%). Recruitment continued until the planned sample was enrolled at for the 6 months measurement, so that the expected statistical power could be reached.

The population characteristics and the significance of the differences between groups are described in Table 4.2.

Table 4.2: Participants' characteristics for each pathological subgroup and the control group, with indications of significant difference with the control group.

	Rotator Cuff (n = 20)	Fracture (n = 23)	Capsulitis (n = 22)	Instability (n = 23)	Control (n = 20)
Age mean (SD), years	63.5 * (10.6)	60.1 * (15.6)	52.5 * (13.8)	32.1 (14.1)	28.2 (6.2)
Sex, % Women	50	78	60	43	50
Weight mean (SD), kg	78.3 (18.2)	69.6 (15.1)	78.3 (15.1)	70.8 (12.9)	74.7 (17.4)
Body mass index mean (SD), kg/m^2	29.0* (6.4)	24.6 (4.2)	26.7 (6.4)	23.7(3.2)	24.2 (3.9)
Height mean (SD), m.	164.0* (7.4)	167.7* (9.7)	172.4 (10.9)	172.6 (9.4)	175.0 (10.3)
Hand dominance, % Right-handed	90	87	100	87	90
Affected side, % Dominant side	70	25	45	52	-

Legend: * significant difference with control group.

Significant differences were found for age between the control group and the rotator cuff ($Z = 5.30$, $p < 0.01$), fracture ($Z = 5.37$, $p < 0.01$) and capsulitis ($Z = 4.85$, $p < 0.01$) subgroups of patients, using the Wilcoxon rank-sum test. Significant differences were also found for height between the control group and the rotator cuff ($Z = - 3.02$, $p < 0.01$) and the fracture ($Z = - 2.14$, $p < 0.05$) subgroups, and for BMI between the control group and the rotator cuff ($Z = 2.69$, $p < 0.01$) subgroup.

Non-significant differences were found for weight and sex between the control group and the pathological subgroups, though the p value for sex was at the threshold for the fracture group ($\chi^2(1) = 3.76$, $p = 0.05$).

4.3.2. Discriminative power

The outcomes of the B-B Score for the control group and for the patient subgroups by pathologies are presented in Table 4.3 and Figure 4.1. Significant differences were found at baseline between the B-B Score performances of the control group [mean (SD) 94.1 (11.1)] and of the rotator cuff condition [mean (SD) 63.1 (19.7), $Z = - 4.24$, $p < 0.01$], fracture [mean (SD) 46.3 (17.5), $Z = - 5.36$, $p < 0.01$] and capsulitis [mean (SD) 54.4 (14.6), $Z = - 5.49$, $p < 0.01$] patient subgroups. The difference between the shoulder instability subgroup and the control group was non-significant [mean (SD)

84.5 (22.6), Z = - 1.88, p = 0.06]. Similar results, not detailed here for the sake of concision, were found at 6 months, despite the positive change in the pathological subgroups.

Table 4.3: Mean and standard deviation of the B-B Score, with the number of participants measured in the control group and each pathological subgroup, at baseline and 6 months. Unit of scores are % representing the performance of the pathological side compared to the healthy side.

Pathology		**Control**	**Rotator Cuff**	**Fracture**	**Capsulitis**	**Instability**
Baseline	Mean (SD)	94.1 (11.1)	63.1 (19.7)*	46.3 (17.5)*	54.4 (14.6)*	84.5 (22.6)
	Sample size (n)	20	20	23	22	23
6 months	Mean (SD)	96.0 (8.3)	77.6 (21.1)*,†	78.9 (15.1)*,†	75.3 (20.5)*,†	91.2 (15.6)
	Sample size (n)	20	19	20	21	20

Legend: SD: Standard Deviation; n: number; * Significant difference with the control group ($p < 0.01$); † Significant difference with baseline ($p < 0.01$).

The difference between the baseline [mean (SD) 94.1 (11.1)] and 6 months [mean (SD) 96.0 (8.3)] control group B-B Score was non-significant ($Z = 0.80$, $p = 0.42$) using the Wilcoxon signed-rank test. This confirmed the hypothesis that the B-B Score outcome would be stable over time in the control group that is not expected to have changed between the baseline and 6 months measurement sessions.

Conversely, significant differences were found between the baseline and the 6 months outcomes in the rotator cuff condition [mean (SD) baseline 63.1 (19.7), 6 months 77.6 (21.1), $Z = 2.63$, $p < 0.01$], fracture [mean (SD) baseline 46.3 (17.5), 6 months 78.9 (15.1), $Z = 3.82$, $p < 0.01$] and capsulitis [mean (SD) baseline 54.4 (14.6), 6 months 75.3 (20.5), $Z = 3.98$, $p < 0.01$] subgroups, but not in the shoulder instability subgroup [mean (SD) baseline 84.5 (22.6), 6 months 91.2 (15.6), $Z = 0.64$, $p = 0.53$], using the Wilcoxon signed-rank test. This confirmed the hypothesis that the B-B Score outcome would change over time in populations that are expected to have changed between the baseline and 6 months measurement sessions, for the rotator cuff condition, humerus fracture and capsulitis subgroups, but not for the shoulder instability subgroup.

The first statements concerning the shoulder instability subgroup highlighted that the B-B Score was not efficient to assess the function loss for this pathology. Therefore,

some calculations were made for a subgroup of patients that included only the pathologies for which the B-B Score could potentially be efficient, i.e. rotator cuff conditions, humerus fractures and capsulitis. This subgroup was called “Indicated pathologies” in the continuation of this work.

Also, as a reminder, the Phase 2 study reported in Chapter three of this thesis included only patients for whom the score may assess efficiently their shoulder function. This option had been chosen because it would have been inconsistent to report results on the B-B Score that could have been influenced by the results of patients for whom the Score should not be used.

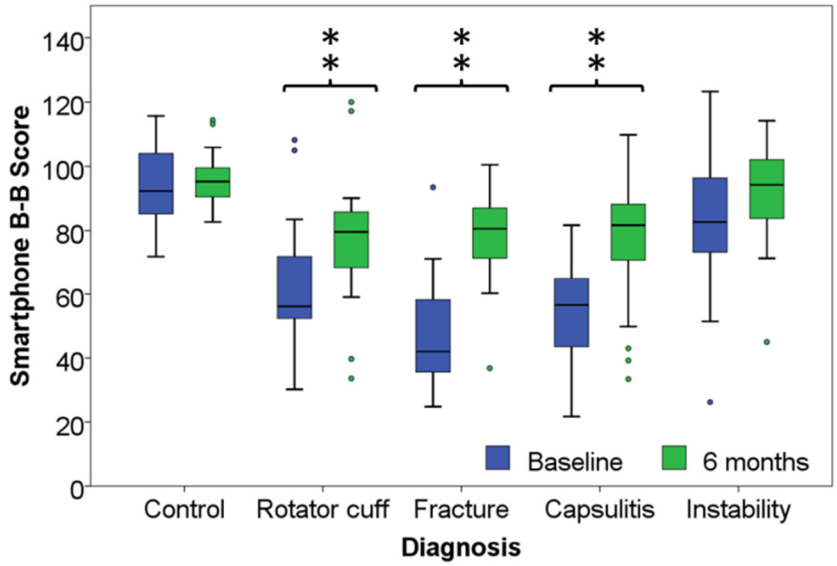


Figure 4.1 Traditional box plots showing median, lower and upper quartile, range and outliers (open circles, 1.5 interquartile range) B-B Scores, comparing the baseline and the six months outcomes for the control (n= 20), the rotator cuff (n=19), fracture (n = 20), capsulitis (n = 21) and instability (n= 20) subgroups. **: significant difference with the control group ($p < 0.01$).

The visual inspection of box plots (Figure 4.1) confirmed that the B-B Score was stable between baseline and 6 months in the control group, while it changed positively in the rotator cuff, fracture and capsulitis, and to a lower extent in the instability subgroup. The smaller difference between the control group and the shoulder instability subgroup was also visible. The presence of outliers showed that the outcome could vary considerably between patients with the same pathology.

The discriminative power analyses using the area under the curve (AUC) with 95% CI and the cut-off for optimal sensitivity-specificity ratio are presented in Figure 4.2 and Table 4.4.

Table 4.4: ROC curve analysis results for the discriminative power between patients and controls, with AUC, optimal B-B Score threshold for patients vs. controls discrimination, and sensitivity and specificity at the optimal threshold value in each study groups.

	AUC (95% CI)	B-B Score Threshold (%)	Sensitivity (%)	Specificity (%)
All patients (n = 88)	0.88 (0.82–0.95)	82.1	95	82
Indicated pathologies (n = 60)	0.96 (0.92–1.00)	82.1	95	94
Rotator Cuff (n = 20)	0.90 (0.78–1.00)	83.6	90	90
Humerus Fracture (n = 23)	0.98 (0.94–1.00)	71.6	100	96
Capsulitis (n = 22)	0.99 (0.98–1.00)	82.1	95	100
Shoulder Instability (n = 23)	0.67 (0.50–0.84)	81.6	95	48

Legend: AUC Area Under the receiver operating characteristic Curve.

The AUC, indicative of discriminative power between patients and controls, was excellent (AUC ≥ 0.90) for the “Indicated pathologies”, rotator cuff, humerus facture and capsulitis subgroups and good for the “all patients” group. Conversely, it was below the required standard (AUC ≥ 0.70) for the instability subgroup. This weakness was mainly related to a lack of specificity. This implied that the B-B Score was not efficient in correctly identifying the patient with shoulder instability, because of an excessive proportion of false positive results. The hypothesis that the B-B Score discriminated adequately the patients from the controls was refuted for this pathology only, and accepted in all other analysed cases. The B-B Score thresholds, indicative of the outcome level from which the functional outcome can be considered normal, were close to each other (81.6% - 83.6%), with the exception of humerus fracture for which the threshold was lower (71.6%).

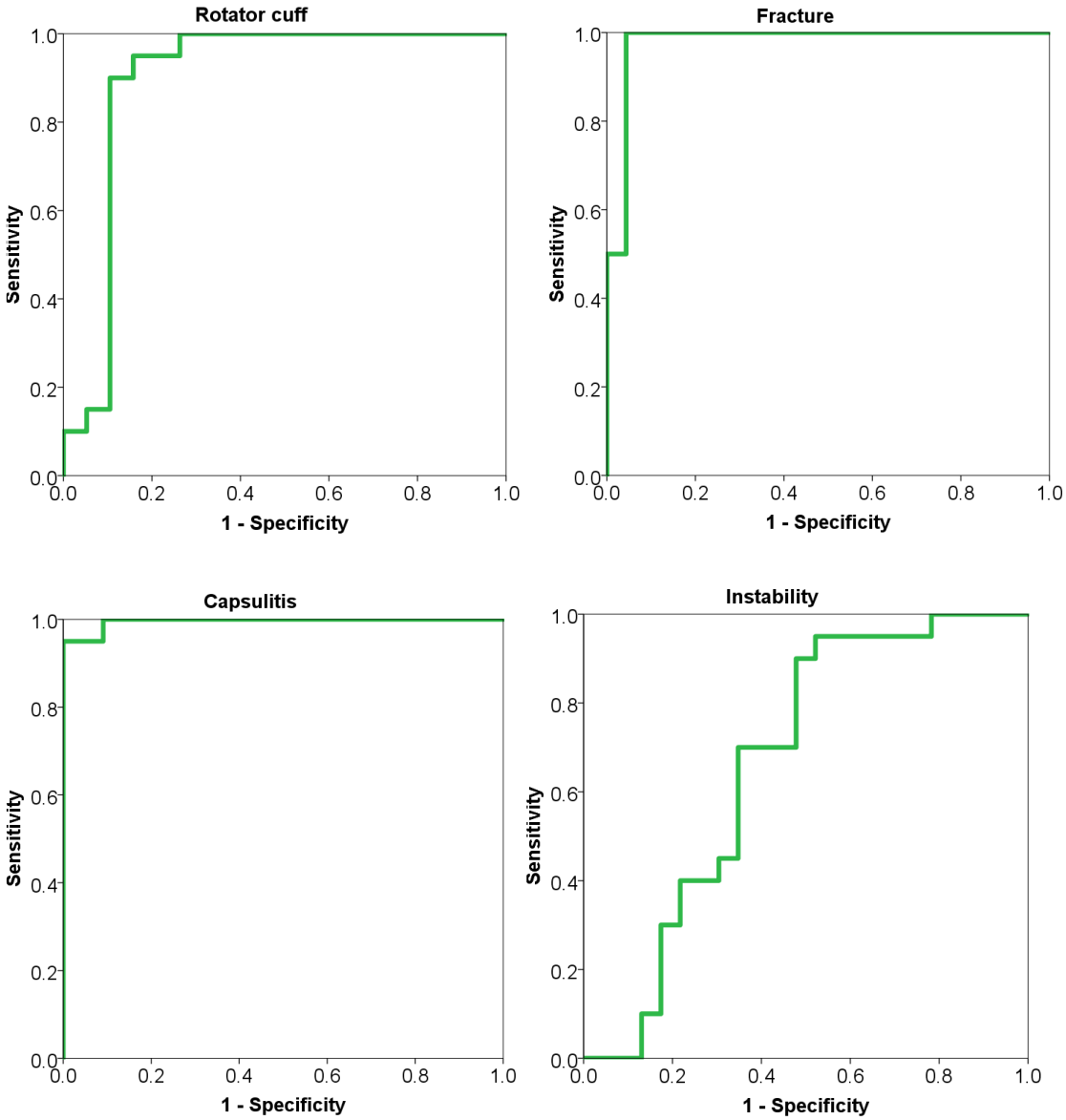


Figure 4.2: ROC curves representing the discriminative power between patients and controls of the smartphone B-B Score (green line), specifically for the rotator cuff conditions (n = 20), proximal humeral fracture (n = 23), capsulitis (n = 22) and shoulder instability (n = 23) subgroups of patients.

4.3.3. Convergent validity

The correlations amongst the shoulder function PROMs are presented for each of the pathologies in Table 4.5.

Table 4.5: Spearman correlation coefficients amongst the B-B Score and the PROMs, for each pathology.

	Rotator Cuff (n = 20)	Humerus Fracture (n = 23)	Capsulitis (n = 22)	Shoulder instability (n = 23)
Constant	0.82 **	0.70 **	0.68 **	0.46 *
Constant relative	0.84 **	0.69 **	0.69 **	0.43 *
SST	0.63 **	0.66 **	0.76 **	0.52 *
QuickDASH	-0.55 *	-0.40	-0.64 **	-0.57 **
WOSI	-	-	-	0.58**
VAS pain	-0.50 *	-0.07	-0.39	-0.19
EQ-5D	0.33	0.18	0.63 **	0.46 *
EQ-5D VAS	0.16	-0.30	0.44 *	0.47 *

Legend: SST: Simple Shoulder Test; QuickDASH: Quick Disabilities of the Arm, Shoulder and Hand score; WOSI: Western Ontario Shoulder Instability Index; SSV: Subjective Shoulder Value; VAS: Visual Analog Scale; EQ-5D: EuroQOL quality of life scale in five dimensions; * significant correlation ($p < 0.05$); ** significant correlation ($p < 0.01$).

The correlations between the B-B Score and the PROMs were higher than the hypothesised level ($r \geq 0.50$), except for the QuickDASH for humerus fractures, and the Constant and relative Constant. They were generally lower between the B-B Score and the pain VAS, EQ-5D and EQ-5D VAS.

4.3.4. Responsiveness

The effect size and SRM with 95% confidence intervals for the B-B Score, Constant and Constant relative score, SST, QuickDASH, and WOSI are presented in Table 4.6 and 4.7, respectively.

The magnitude of the effect sizes varied from one subgroup to the other, as a function of the importance of change over time, which is pathology-dependent. Thus, the comparison of the ES within the same pathology was more informative of the responsiveness of the outcome measures. Overall, the ESs of outcome measures

specific to shoulder function was higher than those of generic PROMs were (VAS pain and EQ-5D), with the exception of pain for the rotator cuff subgroup.

The statements made for the ES also apply for the SRM, showing that these two calculations are founded on close bases.

Table 4.6: Comparison of the effect sizes of scores' changes between the baseline and the 6 months measurements (95% confidence intervals) for the B-B Score and each PROM in each pathological subgroup.

Outcome Measure (95% CI)	Rotator Cuff (*n* = 19)	Fracture (n = 20)	Capsulitis (n = 21)	Instability (n = 20)	All patients (n = 80)	Indicated pathologies (n = 60)
B-B Score	0.69 (0.02–1.33)	1.94 (1.14–2.67)	1.16 (0.49–1.79)	0.10 (-0.52–0.72)	0.81 (0.48–1.13)	1.21 (0.81–1.59)
Constant	0.54 (-0.12–1.18)	2.09 (1.26–2.83)	1.05 (0.38–1.67)	0.21 (-0.42–0.82)	0.79 (0.46–1.11)	1.17 (0.78–1.56)
Constant relative	0.50 (-0.15–1.14)	2.10 (1.27–2.84)	1.04 (0.38–1.67)	0.27 (-0.36–0.89)	0.93 (0.60–1.26)	1.18 (0.80–1.57)
SST	0.52 (-0.13–1.16)	1.65 (0.89–2.35)	0.86 (0.22–1.48)	0.10 (-0.53–0.71)	0.75 (0.43–1.07)	1.02 (0.63–1.39)
QuickDASH	0.35 (-0.30–0.98)	1.25 (0.53–1.91)	0.55 (-0.08–1.16)	0.01 (-0.61–0.63)	0.55 (0.23–0.86)	0.70 (0.33–1.07)
WOSI	-	-	-	0.47 (0.17–1.09)	-	-
VAS pain	0.71 (0.05–1.35)	0.87 (0.23–1.48)	0.69 (0.06–1.29)	0.37 (-0.25–0.97)	0.58 (0.27–0.88)	0.72 (0.39–1.10)
EQ-5D	0.23 (-0.42–0.86)	0.76 (0.09–1.40)	0.34 (-0.27–0.94)	0.49 (-0.14–1.09)	0.41 (0.09–0.72)	0.41 (0.09–0.72)
EQ-5D VAS	0.07 (-0.57–0.70)	0.37 (-0.26–0.99)	0.06 (-0.55–0.66)	0.11 (-0.51–0.73)	0.14 (-0.17–0.45)	0.14 (-0.17–0.45)

Legend: SST: simple shoulder test; QuickDASH: Quick Disabilities of the Arm, Shoulder and Hand score; WOSI: Western Ontario Shoulder Instability Index; SSV: Subjective Shoulder Value; VAS: Visual Analog Scale; EQ-5D: EuroQOL quality of life scale in five dimensions

Table 4.7: Comparison of the standardised response means of scores' changes between the baseline and the 6 months measurements (95% confidence intervals) for the B-B Score and each PROM in each pathological subgroup.

Outcome Measure (95% CI)	Rotator Cuff (n = 19)	Fracture (n = 20)	Capsulitis (n = 21)	Instability (n = 20)	All patients (n = 80)	Indicated pathologies (n = 60)
B-B Score	0.69 (0.03–1.33)	1.98 (1.19–2.69)	1.68 (0.95–2.35)	0.13 (-0.49–0.75)	0.90 (0.57–1.22)	1.26 (0.86–1.64)
Constant	0.58 (-0.08–1.21)	2.02 (1.22–2.73)	1.98 (1.21–2.68)	0.19 (-0.43–0.81)	0.90 (0.57–1.23)	1.23 (0.83–1.61)
Constant relative	0.57 (-0.09–1.21)	2.09 (1.28–2.81)	2.02 (1.24–2.72)	0.22 (-0.40–0.84)	0.91 (0.58–1.23)	1.22 (0.82–1.60)
SST	0.48 (-0.18–1.11)	1.70 (0.95–2.39)	1.24 (0.56–1.87)	0.08 (-0.54–0.70)	0.75 (0.43–1.07)	1.00 (0.61–1.37)
QuickDASH	0.47 (-0.18–1.11)	1.45 (0.73–2.11)	1.07 (-0.40–1.69)	0.01 (-0.61–0.63)	0.67 (0.35–0.99)	0.89 (0.51–1.26)
WOSI	-	-	-	0.41 (0.23–1.03)	-	-
VAS pain	0.78 (0.10–1.42)	0.81 (0.15–1.44)	0.60 (0.02–1.21)	0.25 (-0.38–0.86)	0.62 (0.30–0.94)	0.74 (0.36–1.10)
EQ-5D	0.39 (-0.26–1.03)	0.52 (-0.12–1.14)	0.31 (-0.31–0.91)	0.33 (-0.30–0.95)	0.38 (0.07–0.69)	0.40 (0.05–0.76)
EQ-5D VAS	0.11 (-0.53–0.75)	0.33 (-0.30–0.94)	0.05 (-0.56–0.65)	0.14 (-0.49–0.76)	0.15 (-0.17–0.46)	0.15 (-0.20–0.51)

Legend: SST: simple shoulder test; QuickDASH: Quick Disabilities of the Arm, Shoulder and Hand score; WOSI: Western Ontario Shoulder Instability Index; SSV: Subjective Shoulder Value; VAS: Visual Analog Scale; EQ-5D: EuroQOL quality of life scale in five dimensions.

The Spearman change correlation is presented in Table 4.8 for each PROM and each pathology.

Table 4.8: Spearman correlation coefficients for baseline to 6 months change between the B-B Score and the shoulder function PROMs.

	Rotator Cuff (n=19)	Humerus Fracture (n=20)	Capsulitis (n=21)	Shoulder Instability (n=20)	All patients (n=80)	Indicated pathologies (n=60)
Constant	0.50 *	0.59 **	0.41	0.47 *	0.70 **	0.67 **
Relative Constant	0.55 *	0.66 **	0.47 *	0.50 *	0.71 **	0.69 **
SST	0.37	0.75 **	0.21	0.48 *	0.67 **	0.65 **
QuickDASH	-0.19	-0.56 **	-0.30	-0.28	-0.55 **	-0.47**
WOSI	-	-	-	0.32		

Legend: SST: simple shoulder test; QuickDASH: Quick Disabilities of the Arm, Shoulder and Hand score; WOSI: Western Ontario Shoulder Instability Index; SSV: Subjective Shoulder Value; VAS: Visual Analog Scale.
* significant correlation ($p < 0.05$); ** significant correlation ($p < 0.01$).
Indicated pathologies: pathologies for which the B-B Score showed sufficient validity and discriminative power to be reasonably used.

The correlation coefficients for change were above the hypothesised level for the humerus fractures subgroup, the “Indicated pathologies” subgroup and the “All patients” group. They were lower for capsulitis and the results were mixed for rotator cuff and shoulder instability.

The ROC curves analysis including the area under the ROC curve, sensitivity-sensibility and threshold for the discrimination between improved and unimproved patients are reported in Table 4.9 and Figure 4.3 for the “All patients” group and for the “Indicated pathologies” subgroup.

Table 4.9: ROC curve analysis results for the discriminative power between patients who consider themselves as improved or unimproved at the 6 months follow-up, with AUC, optimal threshold for improved vs. unimproved discrimination, and sensitivity and specificity at the optimal threshold value for the B-B Score and PROMs.

		AUC (95% CI)	Outcome measure threshold (%)	Sensitivity (%)	Specificity (%)
All patients (n = 80)	B-B Score (%)	0.73 (0.61–0.86)	9.5	0.76	0.65
	Constant (points)	0.82 (0.71–0.92)	10.0	0.80	0.78
	Constant relative (%)	0.83 (0.73–0.93)	13.5	0.80	0.78
	SST (points)	0.80 (0.67–0.89)	1.5	0.71	0.78
	QuickDASH (%)	0.78 (0.67–0.89)	4.5	0.76	0.69
Indicated pathologies (n = 60)	B-B Score (%)	0.70 (0.50–0.90)	15.9	0.66	0.73
	Constant (points)	0.81 (0.64–0.98)	10.0	0.87	0.73
	Relat. Constant (%)	0.83 (0.67–0.98)	17.5	0.81	0.81
	SST (points)	0.77 (0.61–0.94)	1.5	0.74	0.64
	QuickDASH (%)	0.73 (0.58–0.88)	6.8	0.77	0.63

Legend: AUC: area under the curve; 95%CI: limits of interval at 95% confidence level; Relat. Constant: relative Constant Score; SST: simple shoulder test; QuickDASH: Quick Disabilities of the Arm, Shoulder and Hand score; WOSI: Western Ontario Shoulder Instability Index; SSV: Subjective Shoulder Value; VAS: Visual Analog Scale. * significant correlation ($p < 0.05$); ** significant correlation ($p < 0.01$).

The AUC, indicative of discriminative power between the patients who consider themselves as improved and those who consider themselves as unimproved was adequate (AUC ≥ 0.70) for the “All patients” group and exactly at the threshold for the “Indicated pathologies” subgroup. The hypothesis that the B-B Score would met this standard could thus be accepted. The B-B Score AUC values were somewhat lower than those of the shoulder function PROMs were, but were situated within the PROMs confidence intervals of the AUCs.

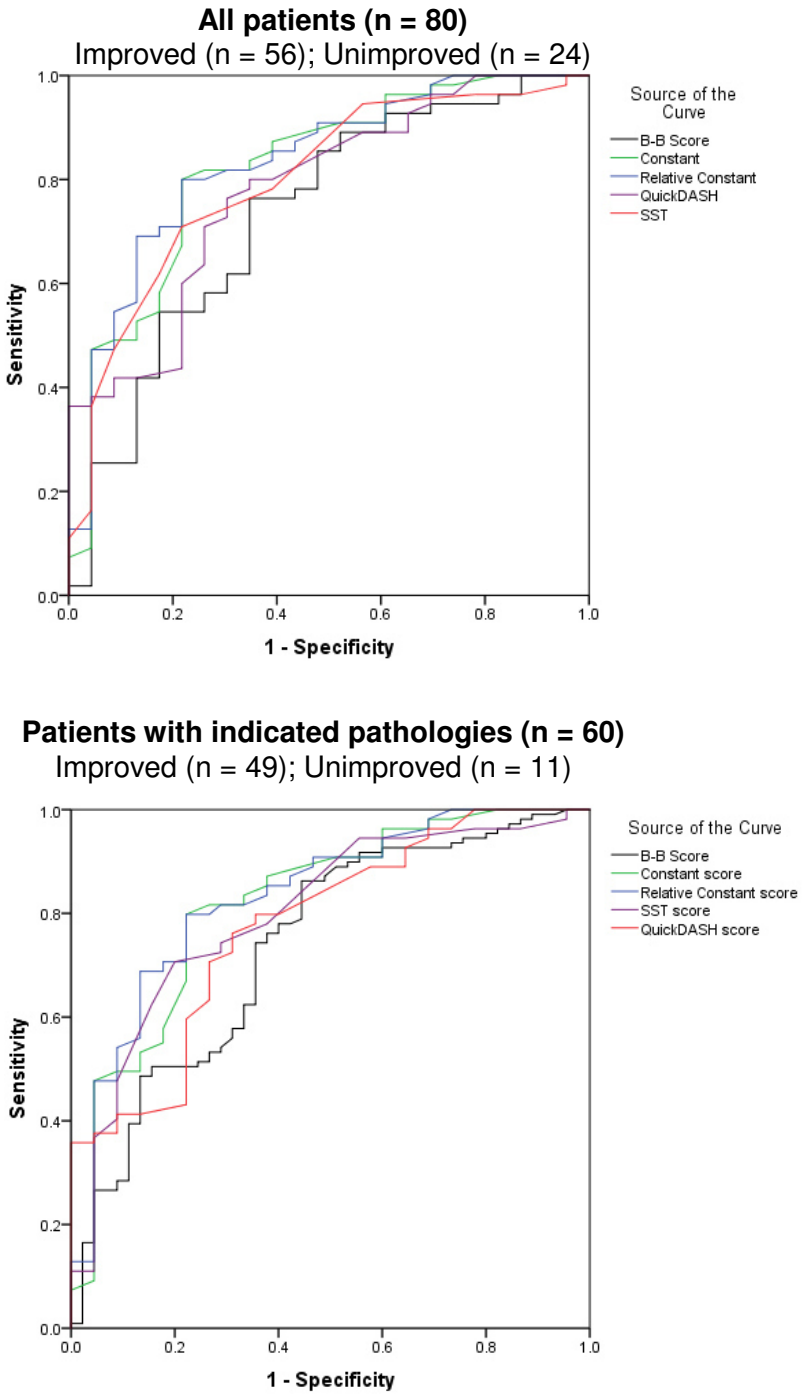


Figure 4.3: ROC curves representing the discriminative power between the patients who consider themselves as improved or unimproved, for the smartphone B-B Score (black line), Constant Score (green line), relative Constant Score (blue line), SST score (purple line) and QuickDASH score (red line). Legend: SST Simple Shoulder Test; QuickDASH: Quick Disabilities of the Arm, Shoulder and Hand Score.

4.3.5. Floor and ceiling effect

No floor effect was observed, as no patients performed lower than the threshold defined for this measurement property, i.e. 0 + MDC (please see sub-section of this Chapter, 4.2.2 “Analysis” p. 135 and sub-section 1.1.3.4.3.6 “Minimal Detectable Change” within Chapter one, p. 48). The hypotheses that less than 15% of the patients would reach a score lower than 0 + MDC (floor effect) was met. Seven patients obtained a score > 100% at baseline, of which 5 had shoulder instability (22% of the subgroup), ten patients obtained a score > 100% at 6 months, of which five had a shoulder instability (25% of the subgroup). The hypotheses that less than 15.0% of the patients would reach a score > 100% (ceiling effect) was met, as 7.9% reached this performance level considering the “All patients” group and 2.2% considering the “Indicated pathologies” subgroup at baseline. This percentage was 12.5% considering the “All patients” group and 6.2% considering the “Indicated pathologies” subgroup at 6 months, but it cannot be excluded that some patients had actually fully recovered at this stage.

4.3.6. Interpretability aspects

Based on the observed values in the control group at baseline (94.1 ± 11.1) and 6 months (96 ± 8.3), the typical performance of healthy controls can be situated at 95%. This value could be of use to determine if the performance of a group is consistent with what can be expected.

The MDC was 15.7% for the rotator cuff subgroup, 17.5% for the fracture subgroup, 14.6% for the capsulitis subgroup and 22.6 for the instability subgroup. These values indicate the level above which a measured difference can reasonably be considered as real, specifically for each pathology.

The MCII of the B-B Score, determined for the “All patients” group using the anchor-based method, was 25.2%. This indicate that patients whose change is higher than this value will consider this change as meaningful.

The PASS of the B-B Score, determined for the “All patients” group using the anchor-based method, was 77.6%. This indicate that patients whose score is higher than this value will consider their shoulder function level as acceptable.

4.4. Discussion

This study aimed at the determination of the measurement properties of the smartphone B-B Score in current shoulder pathologies (rotator cuff conditions, capsulitis, proximal humerus fracture and shoulder instability.

4.4.1. Interpretation of the results

4.4.1.1. Study sample

Participants younger than 40 years old were purposefully enrolled in the control group to prevent the inclusion of people with undetected rotator cuff conditions (Yamaguchi et al., 2006; Yamamoto et al., 2010; Moosmayer et al., 2009). As a consequence of this difference in age, the significant differences in patient size and BMI (size: between rotator cuff subgroup and control group, and between humerus fracture subgroup and control group; BMI: between rotator cuff subgroup and control group) reflected the known age-related tendencies to decrease in size and to increase in weight (Cline et al., 1989; Center for Disease Control, 2012).

The influence of the observed significant differences in age, BMI or size is hardly evaluable. However, based on logical reasoning, they are not likely to have an important impact on this study's results, as they are not likely to influence the side-to side symmetry of the power developed during arm movements, which is the parameter measured by the B-B Score. Conversely, the enrolment of healthy participants the same age than the patients' subgroups, with possible age-related rotator cuff tears, could have had a significant impact on the determination of the normal B-B Score performance from the control group and on the relevance of the comparisons between the control group and the pathological subgroups.

The high, though non-significant, proportion of women in the fracture subgroup is representative of gender prevalence in the wider population affected by this shoulder disorders (Court-Brown and Caesar, 2006). The low proportion of patients affected on the dominant side in the same subgroup can be considered of minor importance, as the shoulder fracture functional outcome is not influenced by the fracture side (Torrens et al., 2015). Further, the influence of dominance on the B-B Score is minimal, as

observed in the control group, in the Phase 1 study and in a previous study (Pichonnaz et al., 2015c).

Due to the lack of discriminative power of the B-B Score specifically for shoulder instability (p. 136 - 140 within this Chapter) that will be discussed hereafter, an analysis was conducted in a sample including only patients with rotator cuff condition, humerus fracture and capsulitis, which was called “Indicated pathologies” subgroup, in addition to the “All patient” group and the subgroups by pathologies.

4.4.1.2. Discriminative power

4.4.1.2.1. Discrimination between groups

The B-B Score differences between the control and the patient groups were highly significant ($p < 0.01$) with the exception of the shoulder instability subgroup (from Table 4.3: 9.6% difference with the control group, $p = 0.06$). The functional loss was, in order of importance, more marked for patient with a fracture (47.8%), a capsulitis (39.7%), and a rotator cuff condition (31.0%) than for instability (9.6%). Hence, the B-B Score clearly discriminated the three first subgroups from the healthy group but displayed a lower discriminative power for shoulder instability. Thus, the most basic and essential measurement property, i.e. the capacity to make a difference between affected and healthy populations, was not adequate for the B-B Score in this pathology, while it was for other included pathologies.

Shoulder instability is characterised by apprehension in the arm positions that exposes the patient to a glenohumeral dislocation risk (Rouleau et al., 2010). It might be that the B-B Score is not challenging enough for these patients, as it is executed in the pain-free ROM and relied upon a self-chosen speed. Thus, the movement of the involved shoulder is not affected by the instability in the normal testing conditions of the B-B Score. Consequently, the functional loss may remain undetected. Nevertheless, a more challenging version of the B-B Score inducing apprehension is hardly conceivable for reasons of ethics, as it might put the patient in a situation of actual dislocation likelihood. These results highlight that shoulder instability affects movement in a different way than other shoulder pathologies and should, thus, be evaluated using a specific tool, like the WOSI, for example.

4.4.1.2.2. Discrimination between stages

The non-significant baseline to 6 months progression in the control group indicated that the B-B Score was stable over time during which the healthy participant's performance can reasonably be expected to have remained unchanged (from Table 4.3 : 1.9% change, $p = 0.42$). Based on this result, the stability of the score for the measurement of a healthy population was demonstrated. The norm for a healthy population (~ 95.0%, based on the baseline (94.1%) and 6 months (96%) values) was also determined, although its value still needs to be refined using a larger sample.

The significant changes in the mean B-B Score over time observed in the rotator cuff condition (14.5%, $p < 0.01$), humerus fracture (32.6%, $p < 0.01$), and capsulitis (20.9%, $p < 0.01$) subgroups indicate that it discriminated amongst clinical stages for these pathologies. Conversely, no significant change over time was found in the shoulder instability subgroup (6.7%, $p = 0.53$). Therefore, the capacity of the B-B Score to capture group change was demonstrated in all pathologies except for shoulder instability.

The 6 months' time interval had been chosen, as it constituted elapsed time that could realistically be considered as sufficient for most of the patients to have an evolution of their shoulder condition, whether spontaneous or induced by treatment. It should be noted that the treatments were not standardised in this study, as the aim was to evaluate the B-B Score's properties but not the treatment's efficacy. Standardisation of events between measurements was not considered as being necessary, as these elements are not expected to have an important influence on the measurement properties of a score, which were the focus of the thesis. Thus, the observed results reflect the combination of the natural progress and of the individualised treatment received by the patients. The results of this thesis' investigations should therefore not be used to characterise the typical evolution of shoulder conditions, as could be done under controlled testing conditions.

4.4.1.2.3. Discrimination between patients and controls

The AUC of the ROC curves for detection of shoulder conditions were adequate (≥0.70) for all pathologies, except for shoulder instability. It was even excellent for the "Indicated pathologies", rotator cuff, humerus fracture, capsulitis and shoulder instability subgroups (≥ 0.90) (Jimerson, 2007; De Vet et al., 2011c; McDowell, 2006). The discriminative power between patients and controls of the B-B Score was higher for fractures and capsulitis (0.98 – 0.99) than for rotator cuff conditions (0.90). The sensitivity and specificity at the optimal threshold were excellent for these three pathologies (≥ 0.90) (Table 4.4). Conversely, the discriminative power between patients and controls was insufficient in the instability subgroup, as the AUC was lower than the 0.70 threshold, mainly due to a lack of specificity. This implies that the B-B Score was not efficient in correctly identifying the patient with shoulder instability, because of an excessive proportion of false positive results. (Portney and Watkins, 2015).

Consequently, the hypothesis that the B-B Score would have adequate discriminative power between patients and controls was met for all pathologies, with the exception of shoulder instability. It was highly efficient for detecting loss of shoulder function in rotator cuff, fracture, and capsulitis disorders. However, although the B-B Score is capable of discriminating whether or not a pathology alters the function of the shoulder, it is not possible to infer a diagnosis of the pathology based on the outcome measured by the Score. Further research should investigate to what extent alterations in specific movement patterns might allow discrimination amongst pathologies.

4.4.1.2.4. Synthesis on discriminative power

The discriminative power of the B-B Score was adequate in all respects for rotator cuff condition, proximal humerus fracture and capsulitis. Conversely, a lack of discriminative power of the B-B Score for shoulder instability was demonstrated, as it was neither able to discriminate the patient group from the control group performance, nor the baseline from 6 months follow-up shoulder instability subgroup performance, nor the patients from the controls. This implies that the B-B Score did not meet the

most basic measurement property for this pathology, contrary to the other pathologies included in this study.

The B-B Score can thus not be recommended to evaluate function in shoulder instability. The limitation of the score for this pathology was further confirmed by the analyses on convergent validity and responsiveness performed in this Phase 3 study. For this reason, the results with mixed pathologies were reported both for all four included pathologies (to account for the whole sample performance; “All patients group”) and excluding patients with shoulder instabilities (i.e. only for all patients for which the B-B Score was likely to be used in practice; “Indicated pathologies” subgroup).

4.4.1.3. Convergent validity

The correlations of the B-B Score with the Constant, Constant relative and SST were moderate to high (r = 0.63 – 0.82) for rotator cuff conditions, factures, and capsulitis (Table 4.5) (Munro, 2005). In contrast, the relationship with the QuickDASH was generally lower (r = -0.55 – -0.64 and non-significant for humerus fracture). The merely objective nature of the B-B Score and the merely subjective nature of the QuickDASH may explain the lower relation with this PROM. The lower correlations with the VAS pain scale (significant correlation only for the rotator cuff subgroup, r = 0.50) and EQ-5D quality of life PROM indicated that the B-B Score is essentially a measure of shoulder function.

Moderate to low correlations were found between the B-B Score and shoulder function PROMs when considering instability. These results indicated that the relation to function was limited for this pathology. Conversely, the B-B Score actually assessed the shoulder function of patients with rotator cuff, fracture, and capsulitis disorders, as demonstrated by the moderate to high correlations between the B-B Score and the Constant, Constant relative and SST scores.

The level of correlation found for these pathologies demonstrated that the B-B Score can be used to investigate shoulder function according to the same concept as that investigated by these PROMs, which supports the convergent validity of the B-B Score with regard to them (McDowell, 2006) (for convergent validity please see sub-section 1.1.3.4.1.4 “Criterion validity”, within Chapter one p. 37 - 39). The hypothesis

that the correlation would be ≥ 0.50 was therefore met for these pathologies. It suggests that, though it is an objective measurement, the B-B Score is influenced by subjective aspects like e.g. kinesiophobia (fear of movement) or patient level of self-confidence when moving, which are also investigated by PROMs.

Based on the literature, this level was not expected because objective and subjective evaluations are generally claimed to produce different results, and because low correlations were found between PROMs and the AR-score, which has similarities with the B-B Score (Krueger et al., 2011; Moustgaard et al., 2014; De Baets et al., 2017; Portney and Watkins, 2015; Korver 2014a). Nevertheless, these results are coherent with previous results found for the B-B Score and the P Score during their development in surgically treated populations (Pichonnaz et al., 2015c; Jolles et al., 2011; Coley, 2007). Thus, this study's results confirmed the stronger link of the B-B Score with function than with pain or quality of life, which was expected from an assessment tool designed for shoulder function evaluation.

4.4.1.4. Responsiveness

Several methods (ES, SRM, correlation coefficients for change, AUC) were used to assess the responsiveness of the smartphone B-B Score for the "All patients" group, the "Indicated pathologies" subgroup and for each specific pathological subgroup. This approach provided a large overview of this measurement property but also reflected the controversies surrounding the best methods to evaluate responsiveness and the fact that the result and the conclusion of a study on measurement properties are dependent on the method used to assess responsiveness. (Terwee et al., 2003; Mokkink et al., 2010e; Angst, 2011; Stratford and Riddle, 2005).

4.4.1.4.1. Effect size and standardised response mean

The effect sizes (ESs) (Table 4.6) and standardised response means (SRM) (Table 4.7) measured in this study should be considered as approximate indications, as their confidence intervals were large. As both methods produced results that lead to the same conclusions, their interpretation is presented jointly hereafter. The ES and SRM were larger, in decreasing order of magnitude, for the humerus fracture (d = 1.25 –

2.10; SRM = 1.45 – 2.09), capsulitis (*d* = 0.55 – 1.16; SRM = 1.07– 2.02) and rotator cuff conditions (*d* = 0.35 – 0.69; SRM = 0.47 – 0.69), than for the shoulder instability condition (*d* = 0.01 – 0.47; SRM = 0.01– 0.41). These differences of magnitudes amongst groups were essentially related to the respective baseline to 6 months progression in each one of these pathologies. The absolute size of the ES and SRM should not be considered as an appropriate indicator of responsiveness, because it is relative to the context of measurement (e.g. importance of the change and follow-up time) (Baguley, 2009; Husted et al., 2000). Therefore, comparison between ES/SRM of outcome measures were made within each group, but not across groups.

The comparison of the ESs and SRMs to concurrent measurement methods for a given condition is informative towards the respective responsiveness of several outcome measures. Based on comparisons amongst measurements of shoulder function, the B-B Score and Constant Score were the most responsive outcome measures within the “All patients” group and the ”Indicated pathologies” subgroup (Table 4.6 and 4.7). Considering specific pathologies, the ES of the B-B Score was highest for the rotator cuff (*d* = 0.69 vs. 0.35 – 0.54 for PROMs) and capsulitis (*d* = 1. 16 vs. 0.55 – 1.05 for PROMs) subgroups and the SRM for the rotator cuff subgroup only (SRM = 0.69 vs. 0.47 – 0.58 for PROMs). The Constant and Constant relative score displayed the highest ES and SRM for humerus fracture, followed by the B-B Score (*d* = 2.09 and 2.10, respectively vs. 1.94; SRM = 2.02 and 2.09, respectively vs. 1.98). The B-B Score nevertheless constitutes a reasonable alternative to the Constant Score for fracture evaluation, when the patient is unable to perform the strength measurement (as is the case before full fracture consolidation, and more generally in 51.9% of patients referred for shoulder surgery), and when the administrative burden is of concern (Christie et al., 2009).

The QuickDASH and, to a lesser extent, the SST globally performed lower than other shoulder function outcome measures in all subgroups. All shoulder function evaluation methods showed better responsiveness than the EQ-5D generic quality of life PROM. This was expected, as this generic quality of this life-focused PROM is only marginally influenced by the change in shoulder conditions. The suitable effect sizes found for the B-B Score in this study were expected, as the B-B Score or the P Score from which it is derived had previously shown comparable or better effect sizes

than PROMs in surgically treated shoulder populations (Pichonnaz, 2010; Jolles et al., 2011).

Similarly to the Constant (d = 0.21; SRM 0.19), DASH (d = 0.01; SRM 0.01) and SST (d = 0.10; SRM 0.08), the B-B Score demonstrated a poor responsiveness for shoulder instability based on ES and SRM analyses. The WOSI displayed the best responsiveness for the evaluation of the shoulder instability condition (d = 0.47; SRM 0.41). The limited responsiveness of the Constant, DASH, and SST for this patient population had previously been reported in the literature (Godfrey et al., 2007; Kirkley et al., 1998; Dawson et al., 1999). Further comparisons between the ES/SRM of the outcome scores used in this study and the ones reported in the existing literature on shoulder disorders cannot reasonably be made, due to the high diversity of treatments, timeframes, patients' characteristics and patients' change that led to the reporting of heterogeneous ES and SRM across studies (please see Chapter five literature review on this subject).

4.4.1.4.2. Correlations between change scores

Considering the responsiveness assessment of the B-B Score based on its correlations with the PROMs change in performance scores from baseline to 6 months, the hypothesis that the correlation value would be $r \geq 0.50$ and statistically significant was met in most but not all cases (Table 4.8). This level of correlation was met for all PROMs when the strength of correlation had been assessed within the "All patients" group (absolute r = 0.55 – 0.70) and in the humerus fracture subgroup (absolute r = 0.56 – 0.75). The results were mixed for the "Indicated pathologies" subgroup (absolute r = 0.47 – 0.69) and the rotator cuff subgroup (absolute r = 0.19 – 0.55), with some correlations higher and some correlations lower than the hypothesised threshold. The correlations between change scores were below the threshold for capsulitis (absolute r = 0.21 – 0.47) and instability (absolute r = 0.28 – 0.50) (Table 4.8).

The correlation coefficients between change scores found in this study can hardly be compared to those of the literature, because of the heterogeneity of the reported results (please see Chapter five literature review for detailed comparisons). The correlation coefficients between change scores observed in this Phase 3 study in samples including mixed pathologies ("All patients" and "Indicated pathologies"

tended to be higher than those of shoulder function PROMs in relatively similar samples (Lundquist et al., 2014; Negahban et al., 2015; Fayad et al., 2008b; Mintken et al., 2009; Schmitt and Di Fabio, 2004). The correlation coefficients between change scores tended to be comparable or lower than those found in the literature for conservatively treated rotator cuff (de Witte et al., 2012; Rysstad et al., 2017), and lower for capsulitis (Staples et al., 2010). However, it should be noted that studies frequently rely on Pearson correlations that would have produced higher correlations in this study, as stated in exploratory analyses run for this thesis purpose (van de Water et al., 2014; van de Water et al., 2016b; Staples et al., 2010; Holtby and Razmjou, 2005; Rysstad et al., 2017; de Witte et al., 2012). Pearson correlations were not used in this study, because it was estimated that their use was not adequate for ordinal data, such as those produced by the selected PROMs.

These results concerning correlations between change scores have limitations because the subgroup sample sizes were too small to get precise values. They were nevertheless sufficient to provide realistic estimations that allow a global insight into the relation between the B-B Score and the selected PROMs.

The use of the correlation coefficient change itself has limitations for the assessment of responsiveness, especially when none of the instruments for which the change score correlation is calculated is a gold standard, as is the case in this study (Angst, 2011). High change score correlations essentially show that two instruments, of which none is perfect but one is considered as a reference criterion, measured change in a related way. Low change score correlations may therefore be found both in case the investigated instrument is *more* sensitive or *less* sensitive to status change than the reference instrument. This implies that low change score correlations will be found when an instrument under investigation had better responsiveness than the reference instrument.

The correlations associated with the “All patients” groups and the “Indicated pathologies” subgroup were higher than those associated with subgroups reflecting specific pathologies were. However, it is important to consider that the magnitude of correlations tends to increase with data dispersion, which had become larger when the pathological subgroups had been amalgamated and the pathologies were considered as a single population with “shoulder disorders”.

The correlation coefficients between change scores found in this study can hardly be compared to those of the literature, because of the heterogeneity of the reported results (please see Chapter 5 literature review for detailed comparisons). The correlation coefficients between change scores observed in this Phase 3 study in samples including mixed pathologies ("All patients" and "Indicated pathologies" tended to be higher than that of shoulder function PROMs in relatively similar samples (Lundquist et al., 2014; Negahban et al., 2015; Fayad et al., 2008b; Mintken et al., 2009; Schmitt and Di Fabio, 2004). The correlation coefficients between change scores tended to be comparable or lower than those found in the literature for conservatively treated rotator cuff (de Witte et al., 2012; Rysstad et al., 2017), and lower for capsulitis (Staples et al., 2010) However, it should be noted that studies frequently rely on Pearson correlations that would have produced higher correlations in this study, as stated in exploratory analyses run for this thesis purpose (van de Water et al., 2014; van de Water et al., 2016b; Staples et al., 2010; Holtby and Razmjou, 2005; Rysstad et al., 2017; de Witte et al., 2012). Pearson correlations were not used in this study, because it was estimated that their use was not adequate for ordinal data, such as those produced by the selected PROMs.

It can mainly be concluded from these analyses that the patients' change measured using the B-B Score is globally related to that of currently used and supposedly responsive shoulder function PROMs, but that this relationship is variable across shoulder conditions. The hypothesis that the correlation value would be $r \geq 0.50$ was met for the "All patients" group and the humerus fracture subgroup. It was partially met for the "Indicated pathologies" and the rotator cuff subgroups, and rejected for the for capsulitis and instability subgroups.

4.4.1.4.3. ROC curves analysis

Considering the ROC curve analysis for the discrimination between patients considering themselves as improved or unimproved, the hypothesis that the AUC would be adequate (AUC ≥ 0.70) was met in the "All patients" group (AUC = 0.73) and just met in the "Indicated pathologies" subgroup (AUC = 0.70) (Table 4.9 and Figure 4.3). However, the AUC was lower than that of other shoulder function PROMs both in the "All patients" group (AUC = 0.78 – 0.82 for PROMs) and the "Indicated pathologies" subgroup (AUC = 0.73 – 0.83 for PROMs). Therefore, though adequate

according to established standard, the responsiveness of the B-B Score assessed using the AUC had a competitive disadvantage with regard to the PROMs selected in this study. The values found in this study should be considered as realistic but not precise estimations of the true AUC values, since they rely on small numbers of patients who considered themselves as unimproved, especially in the “Indicated pathologies” subgroup (24 unimproved in the “All patients” group; 11 unimproved in the “Indicated pathologies” subgroup) (Figure 4.3).

The slightly better responsiveness when all patients are included in the analysis (AUC = 0.73) than when only patients with indicated pathologies are included (AUC = 0.70) was not expected. The visual inspection of the “Indicated pathologies” ROC curves shows that the B-B Score curve is indented in its middle portion, what indicates that at this point the lack of specificity (specificity = capacity to detect correctly the improvement when it happens) importantly increases, while the sensitivity (sensitivity = capacity to rule out correctly the improvement when it did not happen) is not improved, when raising the improved/unimproved discrimination threshold. The indentation is considerably less marked when all patients are taken into consideration.

Thus, the proportion of patients that were correctly classified as improved was slightly better when all patients were included. This indicated that, despite its weaknesses for the evaluation of function in shoulder instabilities, the B-B Score correctly classified as improved, patients with shoulder instability who had actually improved. When only patients with indicated pathologies were considered, a somewhat larger proportion of patients who did not consider themselves as improved were misclassified as improved using the B-B Score, which decreased the specificity of the score. The number considering themselves as unimproved in the dedicated pathologies group being small (n = 11), any misclassification strongly affects the results. This might explain why the AUC was slightly better when all patients were included in the analysis.

The threshold of Table 4.9 represents the values for which the balance between sensitivity and specificity is optimal for the discrimination between the patients who estimated to have improved and those who did not. These threshold values were quite different between the “All patients” (9.5% improvement on the B-B Score required to consider an improvement) and the “Indicated pathologies” sample (15.9% improvement on the B-B Score required) (Table 4.9). Using the second threshold

would be more recommended, as it has been established in a sample that is more related to the population for which the B-B Score is likely to be used.

Note that these threshold values are sometimes considered as representative of the minimal clinically significant improvement (MCII) according to the patient's perception. However, this method is not the most widely accepted because it relies on statistics rather than directly on the perception of the patient. This is why the 75th percentile MCII, which is the subject of a broader consensus, will be presented below (Tubach et al., 2005a; Kvien et al., 2007) (please see sub-section 1.1.3.4.3.2. "MCID/MCII" within Chapter one, p. 48 - 49). As reported in the literature, the MCII values obtained using the ROC curve method or the 75th percentile method differed significantly in this study (please see sub-section 4.4.2 Interpretability aspects", within this Chapter, p. 161 - 162).

4.4.1.4.4. Synthesis on responsiveness

In summary, the B-B Score met most but not all of the standards for adequate responsiveness. The AUC values criteria (AUC ≥ 0.70) was met with a small margin. The criteria for change score ($r \geq 0.50$) was met in the "All patients" group and in the humerus fracture subgroup, partially met in the "Indicated pathologies" and the rotator cuff condition subgroups and unmet in the capsulitis and the shoulder instability subgroups. Considering the ES/SRM, the Constant Score and the B-B score were the two most reactive outcome measures, with an advantage for the former or the latter depending on the sample analysed. No floor or ceiling effects issues were detected.

Despite these globally adequate measurement performances of the B-B score with regard to established standards, the results did not demonstrate clearly whether its responsiveness was superior or not from that of the PROMs. The B-B Score compared either favourably or unfavourably with the PROMs selected in this study, depending on the method used for the responsiveness assessment. As an illustration, the ES/SRM methods were rather favourable to the B-B score, while the AUC values favoured the PROMs.

These controversial results between methods are not surprising, because the methods that were used in this study address different aspects of responsiveness: the proportion of real change vs. noise for ES and SRM, the capacity to perform a

dichotomous classification for AUC, or the relationship with another allegedly responsive measurement method for change score correlation. It has already been demonstrated in the literature that different methods of assessing responsiveness tend to produce different results, which is problematic for researchers and clinicians because they cannot rely on consistent scientific information (Stratford and Riddle, 2005; Husted et al., 2000; Beaton et al., 1997). The analysis conducted in this Phase 3 study demonstrated that the responsiveness of B-B Score measurements was adequate but did not allow drawing conclusions on the superiority or inferiority of the B-B Score responsiveness over currently-used shoulder function PROMs.

A 6 months' time interval without events standardisation had been chosen between measurements. This period was required to enable the assessment of the responsiveness to change of health state over time. Events in between measurements may have a major influence on the patient's evolution. Thus, the detailed results of this thesis' investigations about responsiveness should not be generalised to other testing conditions. Nevertheless, the comparison of the responsiveness between outcome measures that were conducted within this thesis were valid for the determination of their relative responsiveness, as all outcome measures were evaluated under the same testing conditions.

4.4.1.1. Floor and ceiling effects

No floor effect was observed for the B-B Score, as no patient performed lower than the threshold defined for this measurement property, i.e. 0 + MDC (please see sub-section of this Chapter, 4.2.2 "Analysis" p. 133 - 134 for floor effect threshold definition), indicating that the measurement' responsiveness was not reduced for patients performing at a low functional level.

Similarly, no problematic issue was observed with ceiling effects. The proportion of patients scoring more than 100% on the B-B Score was below the 15% threshold at baseline (2.2%) and at six months (6.2%) for the "Indicated pathologies" subgroup. The proportion of patients who scored more than 100% was logically higher at 6 months, because it was possible that the previously affected shoulder had recovered beyond the healthy shoulder performance after treatment in some patients.

The proportion of patients scoring above 100% was higher in the “All patients” group (7.9% at baseline and 12.5% at 6 months). However, most patients who reached a B-B Score of more than 100% had shoulder instability (seven at baseline and ten at six months), which again highlights the limitations of the B-B Score to capture alterations in the patient’s capacity for movement in this pathology, conversely to other investigated conditions. Based on these results, it can be considered that the hypotheses that no floor or ceiling effect would be detected were met.

4.4.1.2. Interpretability aspects

Some values useful for the interpretation of clinical results (normal performance, MDC, MCII and PASS) were also calculated in this study, and were grouped under the term "interpretability aspects". Due to the limited subgroups sample sizes, no differentiation between pathologies was made for the establishment of these values.

The results of the control group showed that the mean norm for performance (~ 95%) was close to 100%, indicating that healthy controls have a good balance between the dominant and non-dominant side considering the power-related parameter used in the B-B Score’s calculation. Comparing the magnitude of the difference (5.9% at baseline and 4.0% at 6 months) with perfect balance, with regard to the balance deficit in patients (36.9% for rotator cuff, 45.6% for capsulitis, 53.7% for fracture), it was considered that no adjustment was additionally necessary for the B-B Score to operate effectively in side-to-side comparisons of functional capability.

The MDC reflects the magnitude of change that is needed to consider that the change is greater than the measurement error for an instrument (Beaton et al., 2001a). The MDC of the B-B Score using a smartphone indicated that the score difference needs to be greater than 18.1% to ensure that it is a real variation of a patient’s state.

The MCII characterises which level of improvement in an outcome measure reflects a meaningful progress for the patient (Tubach et al., 2005a). Based on the MCII value determined using the 75% percentile method, the B-B Score improvement between two stages (in this Phase 3 study, it reflects the period between baseline and 6 months of treatment) needs to be greater than 25.2% for the patient’s improvement to be considered as meaningful by him/her.

The MCII values based on the thresholds for perceived improvement obtained using the ROC curve analysis were smaller than those obtained using the 75% percentile method and were quite different from each other depending on whether all patients were included or only those with indicated pathologies (Table 4.9: 9.5% for the “All patients” group, 15.9% for the “Indicated pathologies” subgroup). This discrepancy between MCII values determined using one method or the other had previously been reported (Beaton et al., 2011). The thresholds identified using the ROC curve analysis method cannot be considered as valid indicators of MCII, because both were smaller than the 18.1% MDC value. Indeed, the MCII must be larger than the MDC to be considered valid, as it would be contradictory to define a value that supposedly is important but is actually below the threshold for detecting changes in performance capabilities (van der Linde et al., 2017; De Vet et al., 2011a).

The PASS is the value beyond which patients consider themselves well (Tubach et al., 2005b). Patients performing above a level of 77.6% on the B-B Score will usually consider that the function loss is acceptable.

4.4.2. Limitations and further developments

Limitations are related to the limited sample size of each patient group. Though the group size was sufficient to compare the measurement properties of the B-B Score with those of concurrent outcome measures, larger sample sizes would be needed to get more precise estimations of measurement properties by pathologies and to be able to perform subgroup analyses for all methods used in this study. Notably, the AUC for improvement discrimination, MDC, MCII, and PASS could not be calculated realistically and separately for each pathology subgroup in this study.

Though the B-B Score was compared to frequently-used shoulder function PROMs, none of them is considered as a gold standard for shoulder function evaluation. Thus, the results of this study could only investigate the convergent validity but not the validity of the B-B Score by comparison to a gold standard. The use of other outcome measures than the selected PROMs would have provided a different benchmark for the comparisons. It can nevertheless be considered that the PROMs used in this study are fair comparators as no other concurrent PROM has demonstrated its superiority over them (Huang et al., 2015).

The results found in this study demonstrated that the B-B Score has limitations for the evaluation of patients with shoulder instability. The Score discriminated neither the instability subgroup from the control group, nor the baseline to 6 months change of the disorder within the instability subgroup. Additionally, the responsiveness of the B-B Score was lower than that of the WOSI and the discriminative power between patients and controls was poor (McDowell, 2006). Based on these results, the B-B Score should not be used for the evaluation of shoulder function in a shoulder instability population. Conversely, all minimum requirements were met for rotator cuff conditions, proximal humerus fractures, and adhesive capsulitis.

Based on the results of this Phase 3 study, it could be considered that the most clinically important measurement properties of the smartphone-based B-B Score had been defined, but that some still needed to be specified with more precision in homogenous pathological populations. The determination of the interpretability aspects for the shoulder pathologies considered in this study provided a background for adequate interpretation of the results in research and clinics. Future studies are needed in patient populations that were not investigated in this study. For example, robust validation of the B-B Score is needed within populations experiencing glenohumeral osteoarthritis, shoulder arthroplasty, and rotator cuff surgery that have been the focus of initial validation studies in the past (Pichonnaz et al., 2015c).

A middle segment smartphone model was chosen to have an insight into the performance of an accessible model. As a wide range of smartphones has similar or even better quality sensors, the results from these models should, theoretically, be at least comparable to those found in this study. The B-B Score is probably robust to variations in devices, as it compares the performance of the affected shoulder with that of the healthy one. Thus, systematic errors in measurement affecting both sides should not importantly affect the B-B Score. However, the influence of the characteristics of each smartphone on the outcome has to be investigated and quantified before clinical implementation.

The scientific value of a novel and objective test of shoulder function, the smartphone B-B Score technique, has been endorsed by the findings of this study, but no cost analysis was conducted at this stage of development. Further studies reproducing routine working conditions should evaluate this aspect. Given the reasonable material

costs and the simplicity of the procedure, there would be a reasonable expectation for a favourable outcome following scrutiny by a formal cost-analysis.

Information and communication technologies developments were not considered in this study but may be possible at a later stage. The use of a smartphone makes the measurement much more accessible for clinicians or even for patients. Thus, larger scale data collection could be performed by more raters at a lower cost. The smartphone B-B Score measurement might, for example, be used in telemedicine due to its simplicity and accessibility. It could also facilitate the centralisation of data collected in a large number of settings at an acceptable cost, thus facilitating data collection for multicentre studies and registries.

4.5. Conclusions

The smartphone B-B Score demonstrated adequate measurement properties in populations with a rotator cuff condition, proximal humerus fracture, and capsulitis. The diagnostic and discriminative powers were excellent for these populations. The correlations with the PROMs indicated that the B-B Score is valid for shoulder function evaluation. The responsiveness was globally comparable to that of PROMs although the results varied according to the method used to assess this clinimetric characteristic. No issues relating to floor or ceiling effects were detected. The determination of the MDC, MCII, and PASS for the B-B Score provided a robust basis for the clinical interpretation of the outcome. Though adequate, the measurement properties were not demonstrated to be superior to those of the selected PROMs. The advantage of the smartphone B-B Score resides mainly in the fact that is provides an objective measurement of shoulder function that is not affected by the translation, culture and items' interpretation issues, in contrast to clinical questionnaires.

All of these conclusions about the smartphone B-B Score open interesting perspectives for the routine objective shoulder function measurement in clinics, as this validated score can quickly be performed using an inexpensive device. The affordable measurement of large cohorts of participants may also be facilitated. Further investigation is needed to devise a movement analysis-based score for the evaluation of shoulder instability in situations where the B-B Score did not meet the minimal clinimetric requirements for clinical deployment. Moreover, the measurement

properties of the B-B Score should be further investigated in patient populations presenting other shoulder conditions such as osteoarthritis, rotator cuff repair, arthroplasty or clavicle fracture and in larger homogenous samples for the pathologies investigated in this Phase 3 study. Studies could also explore the possibility of using the smartphone B-B Score for remote follow-ups and for early detection of suboptimal recovery.

4.5.1. Further developments within the thesis

Phase 2 study demonstrated the equivalency of a smartphone and an inertial sensor system dedicated to the analysis of human movement, while Phase 3 investigated a broad range of measurement properties of the B-B Score in frequent shoulder pathologies. These were important steps to increase the body of knowledge on the measurement properties of the B-B Score, but they did not allow determining whether or not the B-B Score should be preferred to alternative outcome measures for the measurement of the shoulder function in various clinical situations. Although the B-B Score measurement properties were found to be adequate, it might be that other outcome measures have better measurement properties than the B-B Score.

This issue was addressed in the Chapter five of this thesis, in which a literature review was conducted with the aim to challenge the B-B Score clinimetric performances with those of alternative PROMs or movement analysis-based outcome measures.

CHAPTER FIVE

CHALLENGING THE MEASUREMENT PROPERTIES OF PATIENT-REPORTED AND MOVEMENT ANALYSIS-BASED OUTCOME MEASURES FOR SHOULDER FUNCTION EVALUATION: A SYSTEMATIC REVIEW

5.1. Introduction

5.1.1. Rationale for conducting a literature review

5.1.1.1. Contribution of the literature review to the achievement of the thesis objectives

For an outcome measure to be recommended it must have shown adequate measurement properties and must it must stand up to comparison with alternative tools. The adequacy of the measurement properties of the B-B Score measured using a smartphone was demonstrated in the previous chapter of this thesis (Phase 1, 2 and 3 studies), but no comparison had been made with the properties of other outcome measures at this stage.

Therefore, a benchmarking for the measurement properties of the B-B Score and its alternative outcomes measures is provided in this Chapter five, through the means of a systematic literature review. The measurement properties being context-dependent, they were compared separately for various shoulder disorders, either surgically or conservatively treated (Robertson et al., 2017; Riddle and Stratford, 2013; Collins and Roos, 2016; El Gaafary, 2016). As the alternative instruments can be PROMs or, similarly to the B-B Score, movement' analysis-based (MAB) outcome measures, the Score was compared to outcomes measures from these two approaches. This review was also undertaken because it was estimated that the comparison between PROMs and MAB outcome measures for the evaluation of shoulder function would add to the innovative aspects of this thesis, as no review had previously been carried out on this issue, to the best of this thesis' author knowledge.

Importantly, given that aspects of the findings from the preceding studies in this thesis have been published within the peer-reviewed literature (Phase 2 study in Pichonnaz et al., 2017, Phase 3 study in Pichonnaz et al., 2015a), it was anticipated that these articles would be included within the retrieved literature on the subject of interest. It was therefore thought that this novel systematic review would help to further highlight the characteristics of the B-B Score, as a MAB approach to assessing shoulder function and act as a culmination for the aims of the thesis in this respect.

As the Phase 2 and 3 results were extensively but not exhaustively published, the comparisons of the B-B Score with alternative outcomes measures based on the literature data were complemented by the comparisons including unpublished data from the thesis. Conducting this literature review also contributed to an appreciation of the methodological quality within the thesis' studies for evaluating clinimetric properties of health outcome measures, which will be contextualised by reference to the quality of the studies from the literature.

5.1.1.2. Present situation in shoulder function evaluation

To meet the patients' and societal expectations, clinicians are expected to treat patients with optimal efficiency i.e. with maximum efficacy that is matched to both affordable financial and temporal investments. They have thus to rely on efficient measurement tools to evaluate their patients' status and to draw appropriate conclusions about the relevance of their intended approaches to treatment. PROMs (patient-reported outcome measures) and MAB methods are the most frequently-used approaches to evaluate shoulder function performance. Both approaches have proponents that robustly put forward the advantages of each method. However, to the authors' knowledge, the measurement properties of these approaches have never been directly compared within a literature review. Such a comparison would help clinicians and researchers to opt for the most suitable tool matching the needs of their situation, and to highlight the most promising pathways for future developments in the evaluation of the functional and performance capabilities of the shoulder.

5.1.1.3. Challenges to PROMs and movement analysis based methods

The shoulder is the second most frequently-treated body region in rehabilitation (Picavet and Schouten, 2003). Clinicians are thus very regularly called upon to evaluate shoulder's function in their practices. This situation is challenging as there is a plethora of PROMs for assessing the shoulder, but none has been recognised as a "gold standard" (Fayad et al., 2004; Harvie et al., 2005; Huang et al., 2015; Wright and Baumgarten, 2010). As such, it might be difficult for them to choose the PROM offering appropriate clinimetric qualities within a given situation. Moreover, the

evidence about the important measurement properties for each measurement tool must be synthesised and easily accessible in order for it to be exploitable in current clinical practice.

With simplifications to the measurement process afforded by technological progress, and with increasing people' literacy in computer' manipulation, there's an imperative to investigate whether computerised movement analysis-based (MAB) methods could represent a viable alternative to traditional questionnaire-based approaches, which to date, have been used routinely in clinical settings. Considering the ongoing debates on the validity and other measurement properties of PROMs (Roe et al., 2013; Makhni et al., 2015; De Baets et al., 2017; Oh et al., 2009; Bot et al., 2004; Fayad et al., 2004), and recent developments in shoulder movement analysis, a review comparing their respective merits would provide useful knowledge to clarify to which degree both approaches' properties are comparable.

Thus, the measurement properties of PROM and MAB scoring systems were investigated using the contemporary scientific and clinical literature, to evaluate the state of the evidence for both approaches and compare the adequacy of their measurement properties. This will help understand to which degree, in their present stage of development, the MAB evaluation methods are able to complement or replace PROMs, and provide orientations for future research that aims at their improvement (including the B-B Score, depending on its associated research studies meeting inclusion criteria for this review).

5.1.2. Literature review scope

5.1.2.1. Limitations of contemporary field-based reviews of literature

A considerable selection of reviews has already accumulated in the literature focusing on shoulder function evaluation using PROMS. Thirty of them were retrieved during the preliminary bibliographic researches of this review. Most of them addressed validity issues but did not differentiate the measurement properties for different patients' populations.

The body of knowledge of previous reviews remained very heterogeneous and difficult to synthesise, due to the variety of approaches and of quality levels of reviews, which lead to inconsistent conclusions amongst articles. Globally, some reviews had concluded that no shoulder function PROM was superior to the other ones, while other had recommended the use of one or several PROMs, without a common trend across reviews emerging in favour of one of them.

Most reviews did not display clearly the rationale for choosing the included PROMs. Moreover, those that included patients with different pathologies did no differentiate them in the analysis, though measurement properties are known to be context-dependent.

Thus, it was considered useful for raters to have a focused review on measurement issues in the various populations currently treated for shoulder disorders, as measurement properties are known to be context-dependent (Robertson et al., 2017; Riddle and Stratford, 2013; Collins and Roos, 2016; El Gaafary, 2016). As sufficient information is available on the validity of shoulder-focused PROMs, it was estimated that investigating this topic again would have had little added value, and would probably not solve the controversies surrounding the validity of shoulder function PROMs (Bot et al., 2004; Fayad et al., 2004). Thus, a pragmatic approach was adopted focused on measurement properties only, with the thinking that the clinicians cannot wait for a “perfectly valid” outcome measure, and have to rely on existing measurement methods to face today’s challenges.

In contrast to the large number of reviews focused on shoulder PROMs, only one recent review was found on the validity and reliability of shoulder function evaluation using computerised movement analysis, and more specifically inertial measurement units (De Baets et al., 2017). The properties of movement analysis-based measurements were investigated in this latter review, but they were not compared with those of PROMs. As movement analysis is a growing and promising field in the literature, it was estimated that a literature review challenging the traditional approach based on PROMs and the innovative approach based on movement analysis would be of great use to clarify the respective merits of each approach for the various patient populations encountered in clinical practice.

Reviews on the measurement properties that are more recent report the methods with greater precision. Nevertheless, a large majority of them did not evaluate the quality of the literature. Among those who did, only three used the COSMIN checklist (Kennedy et al., 2013; Thoomes-De Graaf et al., 2016; Sahinoglu et al. 2019).

5.1.2.2. Scope of included shoulder conditions

The chosen pathologies were rotator cuff conditions, humerus fracture, adhesive capsulitis and shoulder instability, due to the frequency of these conditions in rehabilitation.

Several conditions, which are hardly clinically distinguishable from each other, are associated with the shoulder's rotator cuff, including rotator cuff tendonitis, rotator cuff tears, subacromial impingement or bursitis (Mitchell et al., 2005). Altogether, they represent the most common source of shoulder pain (65%).

Shoulder osteoarthritis (OA) is a common cause of shoulder pain and disability, particularly in the aging population, which is characterised by radiological narrowing of the glenohumeral joint. It affects 5% – 21% of the adult population in the United States and Europe (Singh et al., 2010). It may be conservatively treated using active and passive joint mobilisations, strengthening and proprioceptive rehabilitation methods. The main surgical options are total shoulder arthroplasty (TSA) and hemiarthroplasty, with TSA performed in 80% of interventions. Due to aging of the population and improvement in surgical outcomes, there was a 3.7-fold increase in TSA intervention rate in the last decade (Trofa et al., 2014).

With a recorded incidence of 22% in the literature, adhesive capsulitis (also frequently called frozen shoulder) represents the second most prevalent cause of shoulder pain (Yamamoto et al., 2010). This idiopathic pathology of the joint capsule causes mainly pain and stiffness that progressively resolves within 12- to 18-months (Kelley et al., 2013; Mitchell et al., 2005).

Proximal humeral fractures is another shoulder disorder that is frequently treated in rehabilitation. Proximal humeral fractures account for 6% of adult fractures (Court-Brown and Caesar, 2006). Their incidence is growing due to the increasing age of the population in Western countries.

Finally, the shoulder instability concerns mainly young adults and is a common cause of medical consultation in this population. It is characterised by the tendency of the humeral head to slide partially or completely out of its socket in the glenoid fossa. Its incidence reaches 2.8% in a physically active young population (Liavaag et al., 2011; Owens et al., 2007).

5.1.2.3. Scope of included measurement properties

This chapter's clinically-orientated literature review also focused on the properties that are of direct interest for measurement interpretation in contemporary real-world treatment situations. Thus, the analysed properties were selected as a function of the clinical demand that the clinicians have to face, and the issues that are of concern in current practice i.e.:

- Evaluate the present status of the patient:
    - Reference norm for healthy subjects: how far is the patient from normal status?
    - PASS (patient acceptable symptoms state): which is the value from which the patient considers his/her state as acceptable?
- Evaluate the patient's change at follow-up:
    - Effect size and standardised response mean: to what extent does the tool capture the status change over time?
    - Specificity, sensitivity, area under the ROC curve for perceived change of status: is the tool able to discriminate those who felt that they evolved from those who do not?
    - Change correlation: is the change score of the outcome measure under investigation related to that of a reference that is assumed to be responsive?
    - Minimal clinically important difference (MCID)/minimal clinically important improvement (MCII): which is the value of the status change/improvement beyond which the improvement becomes meaningful for the patient?
    - Floor and ceiling effects: does the tool capture the differences in performance over time at end range values?
- Estimate the influence of the measurement variability on an outcome measure's reliability, accounting for the repetition of the measurement, the rater or the instrument:

  - Intra-rater, inter-rater and test-retest reliability: to what extent would measurements taken at several occasion produce closely related results?
  - Limits of agreement (LoA) using Bland and Altman graphs: what range of error is associated with a single measurement? Are repeated measurements affected by systematic errors such as carry-over effects?
- Estimate the influence of measurement error on measured change:
  - Standard error of measurement (SEM): what is the typical margin of error of the outcome measure?
  - Minimal detectable change (MDC): beyond which threshold can the measured change be considered as real, and not caused by random measurement variability?

5.1.3. Study aim and hypotheses

This review aimed to collate and compare the measurement properties of currently used patient-reported and MAB outcome measures of function in frequent shoulder pathologies. This will contribute determining if an approach has advantages over the other one, considering their respective measurement properties. It will also help identify paths for future research, based on any detected shortcomings and promising orientations for the systems of measurement.

More specifically to this thesis, the literature review aimed at challenging the measurement properties of the B-B Score with the measurement properties of alternative outcome measures, considering both PROMs and MAB outcome measures. This comparison of the B-B Score clinimetric performances with those of other outcome measures pursuing the same purpose may contribute to circumstantiated recommendations on its use in various clinical contexts.

It was hypothesised that the measurement properties of the PROMs and the MAB outcome measure, including the B-B Score, would comply with recognized standards for the adequacy of measurement properties (see sub-section 5.2.6 “Interpretation delimitation” and Table 5.2, within this Chapter p. 182 - 185 for detailed definition of standards). Based on previous reviews comparing PROMs and in the absence of any previous formal comparison between MAB outcome measures, it was also hypothesised that the measurement properties of the alternative scores, including

those of the B-B Score, would be comparable within in each one of the investigated pathologies (Fayad et al., 2004; Harvie et al., 2005; Huang et al., 2015; Wright and Baumgarten, 2010).

Based on these literature review's aims, the PICOS (Participant, Intervention, Comparison, Outcome, Study design) can be formulated as follows:

- Participants: patients with rotator cuff conditions, osteoarthritis, proximal humerus fracture, capsulitis and shoulder instability,
- Intervention: any kind of surgical or conservative treatment for the aforementioned shoulder disorders
- Comparison: measurement properties of PROMs outcome measures with measurement properties of MAB outcome measures, including the B-B Score
- Outcomes: statistical results for each one of the investigated measurement properties (sub-section 5.1.2.3. "Scope of included measurement properties", within this Chapter 172 - 173).
- Study design: any kind of validation studies.

Thus, this literature review aimed to answer the following question:

- What are the specific measurement properties of the shoulder function outcome measures for patients with rotator cuff conditions, humerus fracture, capsulitis and shoulder instability for PROMS and MAB outcome measures, respectively
- Are the measurement properties of PROMs and MAB outcome measures comparable, for each one of the included pathologies?

5.2. Methods

5.2.1. Formal issues

The Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guidelines were used as a reference for the methodological conception and the reporting of this review, as far as items of the list apply for a literature review without meta-analysis. Reasons for not undertaking a meta-analysis are developed in sub-section 5.2.6 "Interpretation delimitations", p. 182 - 185).

The literature review was registered in PROSPERO under registration number CRD42018104508) (Appendix XIII). Prospero is an international database of prospectively registered systematic reviews in health care and related fields, where there is a health related outcome. It aims to provide a comprehensive listing of systematic reviews registered at inception to help avoid duplication and reduce opportunity for reporting bias by enabling comparison of the completed review with what was planned in the protocol (http://www.crd.york.ac.uk/ PROSPERO/).

5.2.2. Search strategy

The review was constructed in four steps. The first step comprised the identification of the existing outcome measures of shoulder function. All MAB outcome measures were retained, while selection procedures were used to focus only on currently used and valid PROMs, which was necessary for reasons of feasibility given a multitude of approaches (Huang et al., 2015) (please see sub-section 1.1.2.1 “Patient-reported outcome measures”, within Chapter one, p. 4 - 5 for the presentation of the contemporary situation concerning shoulder function PROMs and sub-section 5.2.7 “Preliminary bibliographic search of the selection of PROMs”, within this Chapter, p. 186 - 187 for the detailed selection process and results).

Then, in step two, bibliographic search strategies for relevant databases (Medline, Embase, CINAHL, Web of Science, Pedro) were constructed to retrieve the selected tools measurement properties, for PROMS and MAB outcome measures. The bibliographic search was then completed by a manual search inspecting the references list of included articles. Data concerning measurement properties were then extracted on an excel spreadsheet and compiled on as a third step. A fourth and final step focused on an interpretation of the results based on recognised threshold values for sound measurement properties and on benchmarking for the PROMs and MAB outcome measures of shoulder function.

A double-check was operated by a senior physiotherapist and lecturer colleague of the thesis’ author at the Haute Ecole de Santé Vaud (HESAV), Pierre Balthazard. The checking focused on the terms of the bibliographic strategies on all investigated databases, the retrieved references, the retained articles, the extracted data and the definition of the levels of evidence. At each stage, the differences were discussed and resolved by consensus, taking the objectives and methods of the study described in

this protocol as references. It had been planned that this thesis' supervisor (Prof. Nigel Gleeson) would act as an arbitrator in case of disagreements between the two involved authors, which was finally not necessary. Finally, the data interpretation was discussed and approved by the colleague auditor.

For feasibility reasons, no article exclusion was made based on literature ratings and no meta-analysis was conducted. The specific reasons for these decisions are developed in sub-sections 5.2.5 "Rating quality within the literature" p. 180 - 182 and sub-section 5.2.6, " Interpretation delimitations", p. 182 - 184).

The literature review selection process, for the PROMs and MAB outcome measures is summarized in Figure 5.1

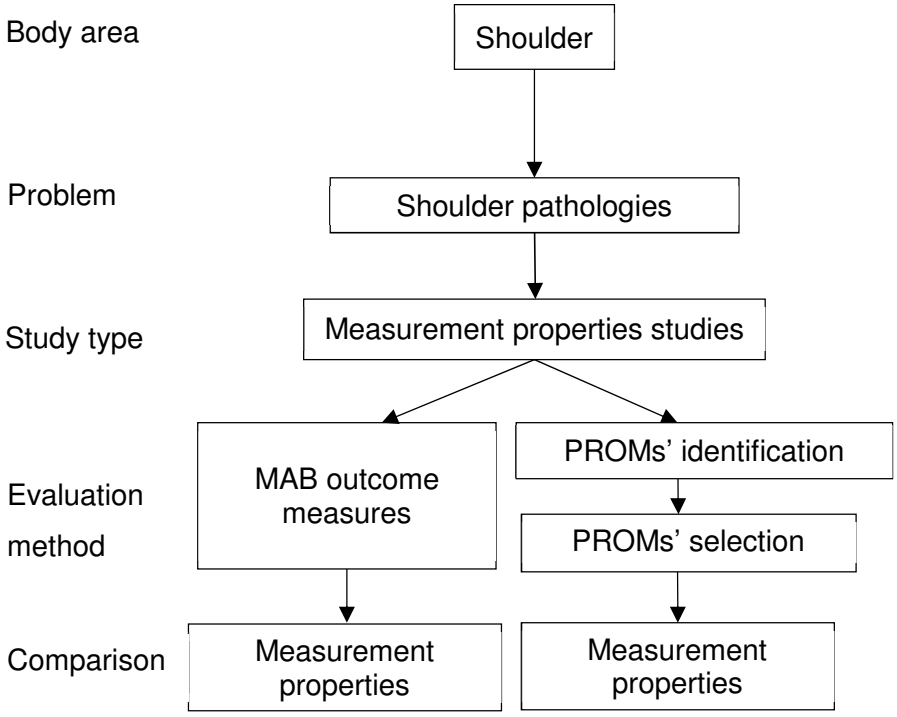


Figure 5.1: Literature review selection process, for the PROMs and MAB outcome measures.

5.2.3. Selection of shoulder function outcome measures

For shoulder function PROMs, the selection was based on frequency of use and the fact that a validation process had previously been completed. As the validity will not be re-evaluated in the present review, it was a necessary to put this second condition, because it would not make sense to evaluate measurement properties of a tool that may potentially have presented with conceptual shortcomings. A preselection was performed based on previous articles that had investigated the frequency of use of shoulder function PROMs (Gartsman et al., 2015; Makhni et al., 2015). Then a preliminary bibliographic search was conducted to assess the frequency of use of the pre-selected articles and therefore be able to proceed to a final selection (please see sub-section 5.2.7 “Preliminary bibliographic search of the selection of PROMs”, within this Chapter, p. 186 - 187 for the detailed selection process and results).

For MAB outcome measures assessing shoulder function, the aim had been to retrieve all of those that allowed the functional performance of the shoulder to be assessed using a scale system. Thus, the articles mentioning solely a difference between a healthy control group and a pathological group in one or several parameters, were not retained, due to their limited utility for monitoring patients’ change, as the clinicians need a scale that allows rating of their patients’ performance from totally non-functional to fully functional. The MAB outcome measures were considered only if the purpose of the tool was to measure shoulder function as a main outcome, e.g. tools that were intended to measure shoulder ROM only were not included, because ROM is not sufficient to reflect shoulder function extensively.

The same measurement properties were extracted for shoulder PROMs and MAB outcome measures, to allow for comparison. Additionally, the correlations amongst shoulder function PROMs and MAB outcome measures were also extracted, as they reflect the degree to which a MAB outcome measure is related to outcome measures currently used for shoulder function evaluation, and thus captures actually shoulder function.

Inclusion criteria:

- Any measurement properties study indexed in relevant databases until 05.05.2017 that investigated the measurement properties of the selected PROMs or any MAB outcome measure designed to assess shoulder function. PROMs selection was based on a preliminary bibliographic search that aimed to determine which were the most commonly used PROMS within the last five years, amongst those pre-selected at the beginning of this sub-section 5.2.3.
- The translated versions of PROMs were included provided that the translation process complied with recommendations for the translation of PROMs (Eremenco et al., 2017), as reported in the article or stated by an ascertainable reference (please see sub-section 5.1.2.3 “Scope of included measurement properties” p. 172 - 173 for detailed description of investigated measurement properties, and sub-section 5.2.7 “Preliminary bibliographic search of the selection of PROMs” for the PROMs selection process, p. 186 - 187).

Exclusion criteria:

- Studies that only addressed the discriminative power between a healthy and a pathological group, without investigating any other measurement property.
- MAB outcome measures measurements whose objective was not to measure the function of the shoulder.
- Studies that included patients with shoulder disorders within a broader upper limb sample of patients, without providing a separate analysis for shoulder disorders.
- Studies including paediatric patients.

5.2.4. Bibliographic search process

Bibliographic research strategies were built to retrieve the measurement properties of shoulder function outcome measures in the four selected current shoulder pathologies presented in sub-section 5.1.2.2 “Scope of included shoulder conditions”, i.e. rotator cuff condition, humerus fracture, capsulitis, shoulder instability and glenohumeral osteoarthritis. Articles concerning conservative and surgical treatments were included to account extensively for the types of patients’ scenarios commonly encountered in physiotherapy practice. However, they were analysed separately to account for the fact that the populations and context of patient follow-up cannot be aggregated because of their inherent differences.

The strategy was also built to retrieve all important measurement properties that are mentioned in point 1.2.3.” Scope of included measurement properties”

These measurement properties were retrieved only for the PROMs selected at the first step of the review, while they were retrieved for all MAB outcome measures.

The search was conducted in the main biomedical (Medline, Embase), allied health (CINAHL) and interdisciplinary databases (Web of Science). The final search included all articles indexed before 05.05.2017 without an inferior time limit, so that all articles about an outcome measure could be taken into account. Strategies for all databases are available in Appendix XV. In summary, strategies for PROMs properties evaluation were constructed to retrieve articles on: shoulder AND selected conditions AND measurement properties AND each selected PROMs. The equation for MAB outcome measure was similar except for the last operator, which targeted movement analysis-based methods applicable for shoulder function evaluation.

5.2.5. Rating quality within the literature

5.2.5.1. Possible checklists considered

5.2.5.1.1. COSMIN checklist and its shortcomings for this review

The use of a rating scale was initially considered to evaluate the quality of the literature, but this approach had to be abandoned for reasons of applicability and equity in the specific context of this review. The use COSMIN checklist had been initially considered for this purpose, as it had specifically been developed to evaluate the methodological quality of studies on the measurement properties of health measurement instruments (COSMIN, 2010). It was finally not used based on the tests conducted in the preliminary try-outs.

The COSMIN checklist was not used mainly because it was not adapted to rate the quality of the methods used to determine several measurement properties considered in this review (MIC, MID, floor/ceiling effects and normative values). It would thus have been inappropriate to interpret the results of some of the studies based on their methodological rating according to the COSMIN checklist, when this could not have

been done for the studies that addressed the properties that could not be assessed using the checklist.

Another problem is related to the fact that a considerable number of articles presented their research with several aspects of varying methodological quality. For example, the same study could have been highly rated for one aspect of the research and poorly for another, which could have compromised the interpretability of the results. Also, the COSMIN approach for responsiveness evaluation is controversial (Angst, 2011). Most studies published to date would have been poorly rated, because ES and SRM calculations, which are widely used, are considered in the checklist as inappropriate methods for the assessment of responsiveness.

Another point to consider is the low inter-rater agreement of the checklist, with Kappa coefficient below 0.40 for 61% of the checklist items (Mokkink et al., 2010a). Therefore, it was decided to proceed to a qualitative analysis of potential biases when results were controversial between studies, but not to present quantitative quality ratings in this review.

Some shortcomings of the original COSMIN checklist have been reported by users and recognised by its developers (Mokkink et al., 2018). Therefore, a new version of the "Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures" have recently been released, although at a date too late for it to be taken into account in this thesis (July 2018) (COSMIN, 2018a). The content of the revised checklist has been targeted more specifically on the issues related to the risk of bias and the rating procedure have has been clarified to some extent. However, the issue mentioned above still remains.

5.2.5.1.2. Contributions of the COSMIN checklist for this review

Although the COSMIN checklist was not used to assess quantitatively the quality of the included studies, its items' questions were used where relevant for the qualitative quality assessment. The considered items addressed the appropriateness of sample size and sample characteristics, stability of patients, time interval between measurements, statistical approaches, similarity of testing conditions and identification of important flaws (COSMIN, 2018a).

Based on these criteria, the quality of the evidence for each measurement property was estimated using a version of the GRADE approach that had been adapted by the COSMIN group to be specific to the evaluation of measurement properties (GRADE Handbook, 2013; Prinsen et al., 2018). The criteria used for the assessment of the degree of evidence were the risk of bias (i.e., the methodological quality of the studies), the inconsistency (i.e., unexplained inconsistency of results across studies), the imprecision (i.e., sample size of the available studies) and the indirectness (i.e. evidence from a different population than the one of interest). The evidence was graded as low when it relied on one study only. When grading the quality of the evidence, the overall rating was initially assumed to be of high quality and was subsequently downgraded to moderate, low or very low by one or two levels per criteria when shortcomings are stated (COSMIN, 2018b) (Table 5.1).

Table 5.1: Modified GRADE approach for grading the quality of evidence with reasons for downgrading the level of evidence. Adapted from: PRINSEN, C. A. C., MOKKINK, L. B., BOUTER, L. M., ALONSO, J., PATRICK, D. L., DE VET, H. C. W. & TERWEE, C. B. 2018. COSMIN guideline for systematic reviews of patient-reported outcome measures. *Qual Life Res,* 27, 1147-1157[3].

Study design	**Quality of evidence**	**Lower if**
At least one measurement properties study	High	Risk of bias -1 Serious -2 Very serious -3 Extremely serious Inconsistency -1 Serious -2 Very serious Indirectedness -1 Serious -2 Very serious
No measurement properties study	Moderate	
	Low	
	Very low	

[3] COSMIN materials on this site may be reproduced in whole or in part in any form for educational or non-profit purposes without special permission.

So, when applying these recommendations for grading the quality of evidence, the evidence is initially considered as high. After analysing if there are available articles, the evidence remains high if there is at least one high quality article. It is downgraded by one (high ⇨ moderate), two (high ⇨ low) or thee levels (high ⇨very low) each time a weakness is stated due to bias, insconsistency or indirectedness.

As proposed in the guidelines for grading the level of evidence, the measurement properties clinimetric performances were rated as "+" sufficient, "-" insufficient, +/- "undetermined" in the tables of results. A question mark (?) was used when a measurement property had never been investigated (please see Table 5.2 Rating criteria of measurement properties p. 185 for the rating criteria) (COSMIN, 2018b).

5.2.5.2. Considerations about the Evaluating Measures of Patient-Reported Outcomes (EMPRO) tool

The Evaluating Measures of Patient-Reported Outcomes (EMPRO) tool was also considered for the screening of the measurement properties of the outcome measures (Valderas et al., 2008). This tool has not been frequently used to date, as only ten publications mentioning its name were retrieved on Medline when searching information on this literature rating approach.

Several limitations were identified regarding the use of the EMPRO for this study's purpose. Although it provides a broad overview of methodological issues, it would have not allowed the assessment of all the measurement properties of interest in this thesis, similarly to the COMIN checklist. In addition, the EMPRO combines the assessment of the methodological quality of the studies with the clinimetric performances of outcomes measures, without making a clear distinction between these issues. The scoring would therefore have been problematic for this thesis' literature review, in which both aspects needed to be clearly differentiated. Finally, the transition from the items' rating to the final overall recommendation for the use of an outcome measure is essentially based on a qualitative appreciation by the rater. The EMPRO tools would therefore not have contributed to improve the objectivity of the recommendations, compared to a qualitative interpretation of the results without relying on this tool.

5.2.6. Interpretation delimitations

All articles were retrieved in English or French. Translations of PROMs in other languages were included provided that the related article had been published in English or French and was based on a validated translation of the questionnaire.

When various upper extremity conditions were analysed and differentiated, the articles were retained only if the data for the conditions relating to the shoulder were separately reported.

The measurement properties were extracted separately for each shoulder condition of interest (rotator cuff condition, humerus fracture, adhesive capsulitis, shoulder instability and glenohumeral osteoarthritis), and differentiated for surgical and non-surgical interventions. These differentiations were made to account for the fact that measurement properties are context and population-dependent (Robertson et al., 2017; Riddle and Stratford, 2013; Collins and Roos, 2016). For example, the responsiveness may be different for patients with shoulder instability or capsulitis, which affects function quite differently, as each condition has a very different progress pattern over time. Similarly, the properties might be different with or without surgical shoulder stabilisation aiming to restore glenohumeral stability. Though studies that include a sample with various shoulder pathologies do not account for the context-dependency of measurement properties, they were nevertheless included in this review. This decision was taken because of the frequency of such studies in the literature and because studies that include various shoulder pathologies represent nevertheless a feasible and useful research option to provide an initial insight into measurement properties, until more precise investigations are conducted.

Similarly to previous authors who had addressed the topic, no meta-analysis was conducted within the data of the current selection of studies included in the systematic review, because of the heterogeneity of the methods, timeframes and sample composition (Kirkley et al., 2003; Oh et al., 2009; Huang et al., 2015; Harvie et al., 2005; Fayad et al., 2005; Placzek et al., 2004; Roy et al., 2009). An example of several problems relating to heterogeneity was that no relevant mean MCID was capable of being calculated from any two studies reporting results that differ because of variations in methods (e.g. distribution-based or anchor-based approach), follow-up

time, pathologies and populations ages. These issues apply equally to the meta-analysis of the other measurement properties considered in this review.

An interpretation of the differences between measurement properties was made only when a direct comparison between tools was conducted within the same research. The differences in measurement properties across studies were not accounted for, as the variations in populations, treatment and follow-up period limited the possibility to proceed to valuable comparisons.

Cut-off values were used when they were available to ensure fair interpretation of results based on common standards and to allow benchmarking for the measurement properties of outcome measures. This was the case for the area under the curve (AUC) of the receiver operating characteristic (ROC) curve, correlations, ICCs and floor/ceiling effects. As no recognised threshold was found in the literature for the limits of agreement (LoA) and the bias, the thresholds used in the Phase 1 and 2 studies of the thesis were applied (Table 5.2). Therefore, an outcome measure's measurement properties were interpreted according to their adequacy in comparison to established standards and to their comparison with concurrent tools. Comparisons were made between outcome measures within a study or between several studies only when the testing conditions were equivalent.

Table 5.2: Rating criteria of measurement properties.

Outcome	Cut-off values	Comments and references
Area under the curve	0.90-1.00: excellent 0.80-0.90: good 0.70-0.80: fair 0.60-0.70: poor 0.50: no discriminating ability	Responsiveness was considered as adequate when AUC was ≥ 0.70 (Pines et al., 2012; Terwee et al., 2007; De Vet et al., 2011c)
Limits of agreement (LoA)	≤ ± 10% and ≤ ± 5% bias	LoAs were considered as adequate when ≤ ± 10% and bias ≤ ± 5% Based on clinical utility, no available reference was found
Correlation	0.00 to 0.30 negligible 0.30 to 0.50 low 0.50 to 0.70 moderate 0.70 to 0.90 high 0.90 to 1.00 very high	Correlations between outcome measures and between change scores were considered as adequate when *r* was ≥ 0.50. (Hinkle et al., 2003)
ICC	≥ 0.70 minimum acceptable threshold ≥ 0.90 expected threshold for clinical use	Reliability was considered as adequate when ICC was ≥ 0.90 (Terwee et al., 2007; Portney and Watkins, 2015; Prinsen et al., 2018)
Floor/ceiling effect	The effect is present when ≥ 15% of the respondents achieved the highest or lowest possible outcome measures	Percentage of patients reaching the maximum or minimum scores was discussed in the review when several outcomes had been investigated in the same study (Terwee et al., 2007)

5.2.7. Preliminary bibliographic search of the selection of PROMs

5.2.7.1. Selection process

A pre-selection of current tools was made based on those that had been identified in previous literature reviews investigating their frequency of use (Gartsman et al., 2015; Makhni et al., 2015). As a result, the DASH (Disabilities of the Arm, Shoulder and Hand) and QuickDASH, Constant Score [4], ASES (American Shoulder and Elbow Score), SST (Simple Shoulder Test), SPADI (Shoulder Pain and Disability Index) , UCLA (University of California Los Angeles), Shoulder rating scale, Rowe score, WOSI (Western Ontario Shoulder Instability Index) and WORC (Western Ontario Rotator Cuff Index) were preselected and their frequency of use estimated based on the number of articles indexed in Medline for the last 5 years in which the tool had been used for shoulder function evaluation (please see Appendix XIV for bibliographic strategies)

The literature search was limited to the last five years to reflect the recent practice of shoulder function evaluation. Abstract and Medline data of retrieved references were inspected to ensure that the preselected scores were actually used in the articles. Despite their frequency of use, the UCLA shoulder score (Fayad et al., 2004; Longo et al., 2011; Kirkley et al., 2003; Huang et al., 2015; Gartsman et al., 2015) and the ROWE score (Rouleau et al., 2010; Fayad et al., 2004; Kirkley et al., 2003) were not retained as it had been consistently stated in hereby mentioned publications that they had not undergone a formal validation process.

[4] Constant Score: unless otherwise specified, “Constant Score” refers to the absolute Constant Score. It will be specified “relative Constant Score” when the Constant result is expressed as a percentage of the expected performance from a gender and age-matched group.

5.2.7.2. Results for the selection of PROMs

The following number of occurrence for the preselected PROMs was found to be:

1) Constant: 1070
2) ASES: 605
3) DASH or QuickDASH: 452 (among them 98 QuickDASH)
4) SST: 348
5) SPADI: 199
6) WOSI: 95
7) WORC: 79

The four most frequent PROMs were selected based on these results. Although the Constant Score includes a clinical examination, it was incorporated within the category of PROMs (Patient-Reported Outcome Measures), as it serves the same purpose and is a very current outcome measure of shoulder function. The WOSI was also added to the PROMs that would be used within the thesis' systematic review. This was because, similarly to a previous review on shoulder outcomes, it was estimated that investigating at least one validated instrument for shoulder instability was necessary, due to the specificity of this condition and the poor performance of generic shoulder function PROMs for this pathology (Angst et al., 2011).

5.2.7.3. Characteristics of selected shoulder function PROMs

5.2.7.3.1. DASH and QuickDASH scores

The DASH is a self-assessment PROM of the entire upper extremity symptoms and function (Hudak et al., 1996). It provides a whole upper-extremity evaluation including the shoulder. Only studies on the measurement properties for shoulder evaluation were considered in this review. The original version comprises 30 items among which 6 are about symptoms (3 pain, 1 tingling, 1 weakness, 1 stiffness) and 24 about function (21 physical function, 3 social function) (Angst et al., 2011). Two optional additional modules for work and sports/performing arts, exist for specific evaluation of manual workers and athletes. Items are scored on 5-point Likert scales ranging from

no difficulty to extreme difficulty or symptoms, with a highest score of 100 indicating the worst disability.

A shortened version, the QuickDASH that comprises 11 items only, has been developed to limit the evaluation' burden. The QuickDASH has been designed to measure the same concept as the DASH, but its developers estimate that the full DASH should be preferred when more precision is needed (American Academy of Orthopaedic Surgeons, 2009).

5.2.7.3.2. Constant Score

The Constant Score is a composite outcome measure that includes questions on pain and activity, and objective measures of range of motion and abduction strength (Constant and Murley, 1987). It can be used in various shoulder pathologies. The score rates the shoulder function on a 100-point scale, with a higher score indicating a better outcome. The relative Constant has been proposed to overcome the gender dependency and the decline with increasing age that were observed using the original approach of the Constant Score. The relative Constant expresses the performance as a percentage of the expected value, based on the comparison of the patient's performance to a sex and age matched group, which facilitates validity when comparisons of this type are undertaken (Constant, 1986; Yian et al., 2005; Katolik et al., 2005; Fialka et al., 2005; Constant et al., 2008).

5.2.7.3.3. ASES score

The ASES is a composite shoulder evaluation tool that can be used in various shoulder pathologies. The original version was published in 1994 (Richards 1994), and a modified version mASES in 1998 (Beaton and Richards, 1998), to provide a more comprehensive evaluation of upper extremity function (Angst et al., 2011; Fayad et al., 2004). The ASES includes a physician-assessed part and a patient self-assessment part. However, the patient-reported section only is generally taken into consideration in the scoring (Hettrich CM, 2007). The latter section comprises questions on pain, activities of daily living and instability. The patient-reported ASES rates the shoulder function on a 100-point scale, with a higher score indicating a better outcome. Function is evaluated based on a series of ten 4-point scales for each arm, and pain using a 10-point VAS (Slobogean and Slobogean, 2011).

5.2.7.3.4. SST score

The SST is as shoulder function PROM that comprises binary 12 items (yes/no), among which two are about function related to pain, seven about function related to strength and three about range of motion (Lippitt, 1993; Beaton and Richards, 1998). The SST rates the shoulder function on 12 points, which can be converted into percentage of “yes” responses, with a higher score indicating a better outcome.

5.2.7.3.5. WOSI score

The WOSI is a specific shoulder outcome measure designed for disease-specific quality-of-life evaluation in patients with shoulder instability (Kirkley et al., 1998). It comprises 21 items in four domains that are scored on a 100-mm visual analog scale: ten items on physical symptoms and pain, four items on sports/recreation/work, four items on lifestyle and 3 on emotions. The lower score represents the better outcome. The score can be reported either as the sum of 21 unweighted items (0 – 2100) or as a percentage (0 – 100%).

5.3. Results

The results of the main bibliographic search, which aimed to retrieve the articles in which the measurement properties of PROMS and MAB outcome measures are investigated is reported in this section.

Concerning PROMs, 4537 references, among which 13 were found by manual search, were identified. One thousand eight hundred references were screened after removal of duplicates. Following title and abstract reading, 1668 articles were excluded. Of the 132 remaining articles, 86 were finally retained after full-text reading. The thesis’ authors had initially selected 82 articles and the colleague auditor 58 articles, of which six had not been retained by the thesis’ author. Most of the articles that were not selected by the colleague auditor addressed mainly the measurement properties of a PROM that had not been selected for this review but nevertheless contained information on one of the selected PROMs. Following discussions, all of the thesis’ authors selected articles were retained and four more were added out of the six articles that had been selected by the colleague auditor only, which resulted in the total of 86 selected articles.

Flowchart with detailed reasons for exclusion is available in Figure 5.2. Please see Appendix XVI for the references of selected articles.

Concerning MAB outcome measures, 4996 references were identified of which 1642 were screened after duplicates removal. Following reading of titles and abstracts, 1626 articles were excluded. Of the 17 remaining articles, nine were finally retained after full-text reading. The thesis' author had initially selected seven articles. Two more articles that addressed the convergent validity between MAB outcome measures and one of the selected PROMs were added following the colleague auditor check.

Flowchart with detailed reasons for exclusion is available in Figure 5.3**.** Please see Appendix XVI for the references of selected articles.

Figure 5.2: PRISMA 2009 Flow Diagram PROMs.

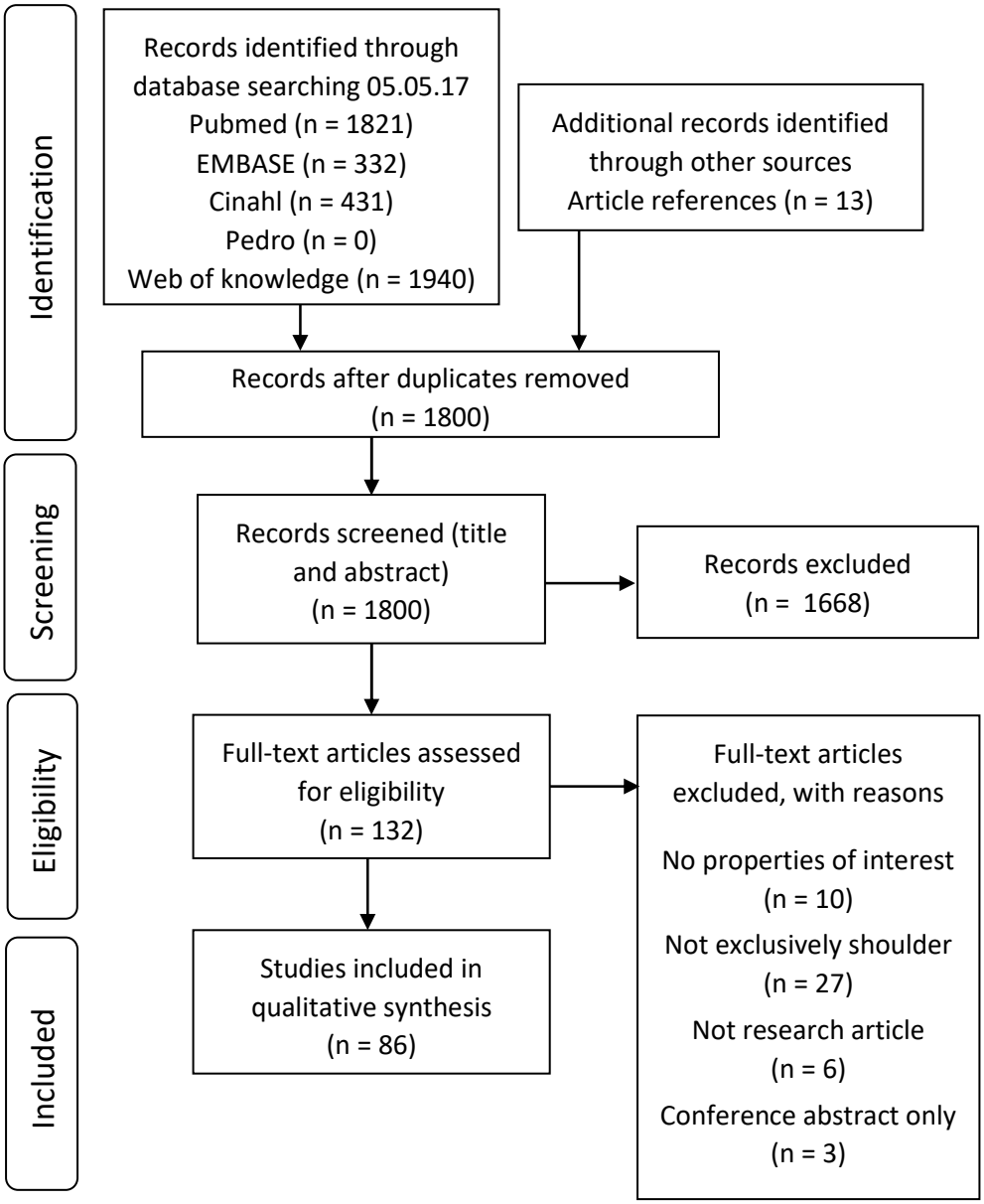
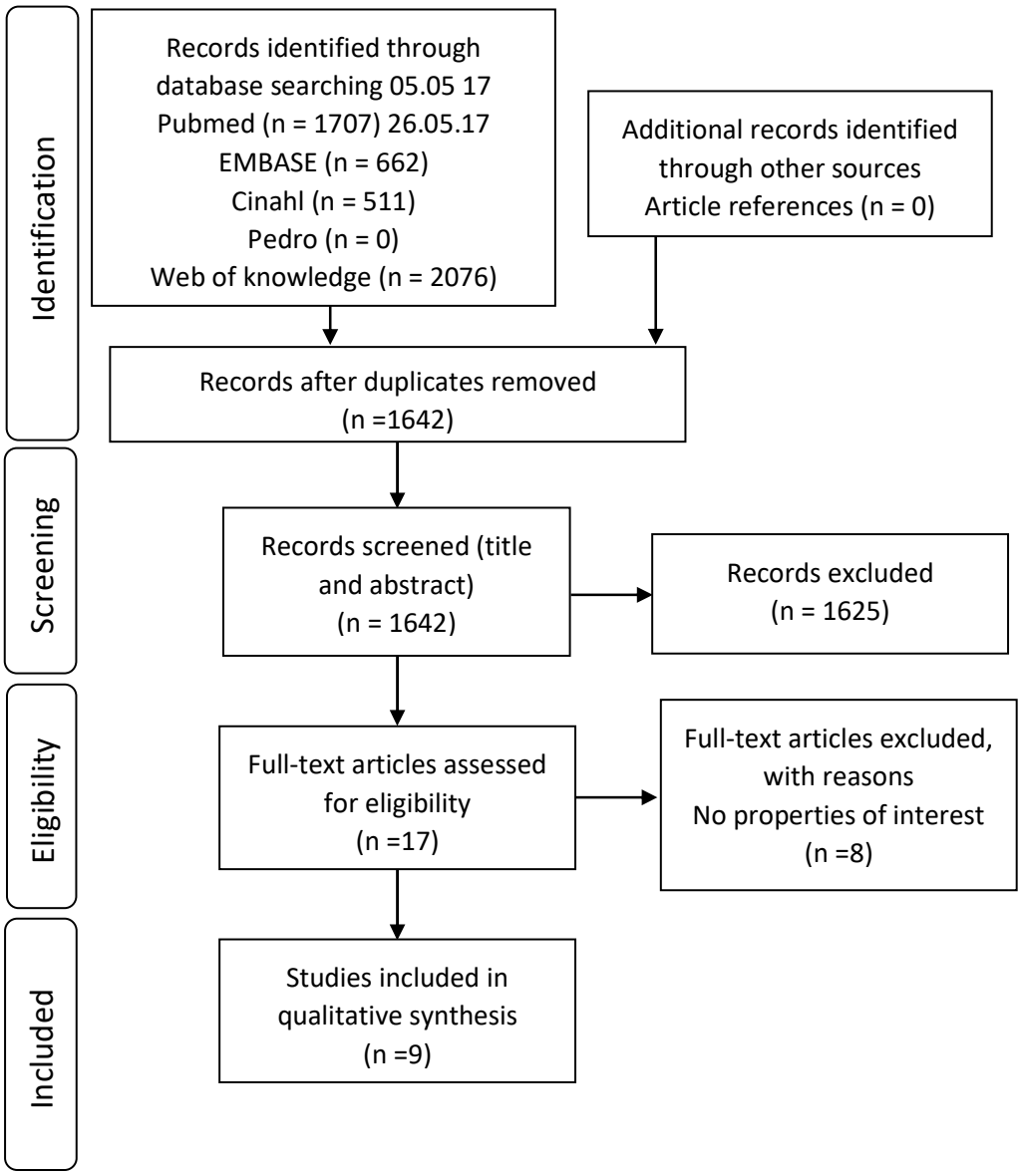


Figure 5.3: PRISMA 2009 Flow Diagram MAB outcome measures.



5.3.1. PROMs measurement properties

5.3.1.1. PROMs normal performance in a healthy population

ASES	
Sallay 2003	95.8 (SD 9.0 points)
Constant	
Yian 2005	Varying according to age and sex from 94 (male 21 - 40 yr.) to 84 points (female 71 - 80 yr.) For all details see table in original publication
Constant 1986	Varying according to age and sex from 99 (male 21-31 yr. left side) to 50 points (female 91 - 100 yr. left side) For all details see table in original publication
Katolik 2005	Varying according to age and sex from 96 (male 40 - 49 yr.) to 81 points (female ≥71 - 80 yr.) For all details see table in original publication
DASH	
Angst 2011	Cut-off scores for “no problem” : < 15
Aasheim 2014	Whole population mean 13; women (SD) 15 (3), increasing from 5 (9) in their 20s. to 36 (26) in their 80s; men 11 (2), increasing from 5 (9) in their 20s. to 22 (23) in their 80s
Hunsaker 2002	Whole population mean (SD) 10.1 (14.7); women (SD) 12.0 (12.0) increasing from 8.4 (13.5) in 19-34 yr. old population to 22.3 (20.3) in > 75 yr. old population; men 7.4 (12.1), increasing from 1.9 (3.9) in 19-34 yr. old population to 16.1 (16.5) in > 75 yr. old population
QuickDASH	
Aasheim 2014	Whole population mean 13; women (SD) 15 (3), increasing from 6 (9) in their 20s. to 36 (27) in their 80s; men 11 (2), increasing from 5 (10) in their 20s. to 23 (23) in their 80s
Hunsaker 2002	Whole population mean (SD) 10.9 (15.3); women (SD) 12.0 (12.0) increasing from 8.9 (14.8) in 19-34 yr. old population to 22.0 (19.1) in > 75 yr. old population; men 7.4 (12.1), increasing from 2.2 (4.7) in 19-34 yr. old population to 14.6 (17.7) in > 75 yr. old population
WOSI	
Salomonsson 2009	mean 96%

Notes for all tables:

- Range possible values and units of the PROMs: ASES: 0 – 100 points; 1 – 100 Constant: points (relative Constant: 0 – 100%); DASH: 0 – 100 points; SST 0 – 12 points or 0 – 100% (as both rating methods are used, the unit is specified in the table for this PROM); WOSI: 0 – 2100 points or 0 – 100% (as both rating methods are used, the unit is specified in the table for this PROM).

- \+ above the threshold required to be considered adequate
  +/- not clear whether above or below the threshold required to be considered adequate
  \- below threshold the threshold required to be considered adequate
- "Change correlations" are coefficient of correlation with change scores of a Global rating of change scale, unless otherwise specified

- The range of reported values found in the literature is reported when several studies investigated a measurement property, in order to avoid an overwhelming level of details in the tables

- When studies compared the properties of several scales, the results are reported in the "Direct comparison" column. The interpretability aspects were not compared between tools, because their comparisons between scales that do not rely on the same rating system would not have been relevant

ES and SRM were reported only when comparisons between outcome measures were made within a study. As ES and SRM values are relative to the magnitude of the change and follow-up time, there were of importance for responsiveness assessment only when a comparison between outcome measures was made (Baguley, 2009; Husted et al., 2000).

5.3.1.1.1. Normal performance for PROMs

General normative values have been reported for the Constant (3 studies), DASH (3 studies), QuickDASH (3 studies), ASES (1 study) and WOSI (1 study), but without consideration for the potential influence of age and gender for the latter two PROMs (Sallay and Reed, 2003; Salomonsson et al., 2009). No normative values were found for the SST (results sub-section 5.3.1.1 “PROMs normal performance in a healthy population”, p. 195).

Normative values in the US and in the Norwegian general population stratified by age and sex have been defined for the DASH and QuickDASH (Hunsaker et al., 2002; Aasheim and Finsen, 2014). Both studies found convergent results and close values for these two outcome measures. They also highlighted the dependency of their norms to age and sex, with higher scores in females and in older patients (higher score meaning more disability). Additionally, Angst has defined a cut-off value between healthy and pathological subjects (Angst et al., 2011).

The norms of the Constant Score have also been shown to be age and gender dependent and have been debated (Yian et al., 2005). This dependency had already been mentioned in the original Constant’s work (Constant and Murley, 1987) and has also been reported in later publications (Katolik et al., 2005; Yian et al., 2005).

Consequently, a relative Constant Score, that classifies the patients based on an age-and gender-matched normal population, has been developed to limit the impact of these factors on the Constant outcome. When using the relative Constant, the origin of the reference values that are used should be reported in manuscripts, c are available (Yian et al., 2005). It should also be noted that a Constant Score revised testing procedure has been recommended to improve the precision of the evaluation, but that the norms available in the literature might not be fully applicable, because they had been established based on the original procedure (Constant et al., 2008).

5.3.1.2. PROMs measurement properties in a diversified shoulder conditions sample

Measurement properties established on a diversified shoulder conditions sample				
	Reliability and measurement error	**Interpretability**	**Responsiveness**	**Direct comparison**
Non-surgical				
Constant				DASH ICC 0.86 vs. ASES ICC 0.75 (Moser 2012) 7.5% maximum and 7.5% minimum score for SST vs. 2.5% and 0% for ASES (Robins 2017)
Blonna 2012 Celik 2016 Conboy 1996	Intra-rater: Expert ICC 0.94 - 0.97 Non expert ICC 0.80 - 0.95 Inter-rater : ICC 0.84 - 0.95* SEM 6 MDC 16.4 Intra-rater LoA* (bias:) 8.6 - 18.6 (0.4 - 4.3) Inter-rater LoA* (bias): 11.0 - 28.0 (1.0 - 5.0)	No floor/No ceiling effect		
DASH				
Lundquist 2014; Moser 2012; Negahban 2015	- Test-retest ICC 0.86	MCID 11.7 No floor/No ceiling effect	+ AUC (0.76 - 0.77) Change correlation: 0.52 - 0.59	
QuickDASH				
Fayad 2009; Mintken 2009	+ Test-retest ICC 0.90 - 0.94 LoA (bias) 11.8 (3.4) SEM : 4.8; MDC 11.2;	MCII: 8	+ AUC 0.82 Change correlation: 0.45 - 0.57	

* depends on expertise and use of Constant Score classical, standardised or relative version

Measurement properties established on a diversified shoulder conditions sample (continued)				
	Reliability and measurement error	**Interpretability**	**Responsiveness**	**Direct comparison**
SST				
Ebrahimzadeh 2016; Neto 2013; Membrilla-Mesa 2015; Robins 2016; Van Kampen 2012	+/- Test-retest ICC 0.61 - 0.92 SEM 1.18 pts; 2.2 - 10.0% $MDC_{90:}$ 6.2%; MDC_{95} 3.3 pts (27.5%)		No floor/No ceiling effect	
ASES				
Cook 2002; Kocher 2005; Moser 2012; Robins 2016; Yahia 2011; Piitulainen 2014	+/- Test-retest ICC 0.75 - 0.96 LoA (bias) 9.5 (0.7)		No floor/No ceiling effect	
Surgical				
Constant				AUC 0.84 Constant vs 0.79 DASH (Christie 2011) ES ASES 0.61 vs. Constant 0.57 vs. 0.50 SST SRM 0.77 ASES vs. 0.58 Constant vs. SST 0.47 (Oh 2009) SRM 2.24 Constant vs.2.17 SST (Ge 2013)
Christie 2009; Rocourt 2008; Christie 2011; Oh 2009; Ge 2013		MCID 16.6 PASS 42.0 - 44.0	+ AUC 0.84 No floor/No ceiling effect	
DASH				
Christie 2009; Schmitt 2004; Christie 2011	+ Test-retest ICC 0.91 SEM 5.2; MDC 12.2	MCID 10.1 - 10.2; PASS 42.9 - 43.0 No floor/ceiling effect	+ AUC 0.79 Change correlation: 0.66	
ASES				
Beaton 1998; Oh 2009; Ge 2013; Cook 2002	+ Test-retest ICC 0.91 - 0.96			
SST				
Oh 2009 Beaton 1998	+ Test-retest ICC 0.99			

Measurement properties established on a diversified shoulder conditions sample (continued)				
	Reliability and measurement error	**Interpretability**	**Responsiveness**	**Direct comparison**
Mixed surgical/non-surgical				
DASH				ICC 0.86 SST vs. 0.85 QuickDASH vs. 0.83 DASH (Van Kampen 2013)
Van Kampen 2013; Diniz Lopes 2009; Beaton 2005; Beaton 2011; Fayad 2008a	+/- Test-retest ICC 0.83 - 0.95 LoA (bias) -7.2 - 13.2 (3)	MCID 3.9 - 15.0*; MDC 16.3; MCII: 12.4 * depends on method		
QuickDASH				
Van Kampen 2013; Beaton 2005	- Test-retest ICC 0.85 MDC 17.1	MCII 13.4		
SST				
Roddey 2000; Van Kampen 2013	- Test-retest ICC 0.86 SEM 11.65%; MDC 2.8 pts (23.3%);	MCII 2.2 pts (18.3%)		
ASES				
Celik 2013; Michener 2002; Vroutsou 2016; Cook 2003; Sallay 2003	+/- Test-retest ICC 0.84 - 0.96 SEM 6.7; MDC 9.94;	MCID 6.4 No floor/No ceiling effect	+/- AUC 0.74 - 0.82	

The PROMs' measurement properties were frequently defined based on samples including various shoulder pathologies. Among them, 18 studies reported measurement properties in non-surgical treatments, eight following surgery and nine in a mixed sample of non-surgical and surgical treatments.

5.3.1.2.1. PROMs, DCS non-surgical treatment (NSu)

5.3.1.2.1.1. Constant (NSu-DCS)

Three studies investigated the properties of the Constant Score (Blonna et al., 2012; Celik, 2016; Conboy et al., 1996) (results sub-section 5.3.1.2. "PROMs measurement properties in a diversified shoulder conditions sample", p. 196). Blonna et al. found important differences between expert (ICC = 0.93 for absolute Constant and 0.94 for relative Constant) and non-expert users (ICC = 0.80 and 0.81) for intra-rater reliability, demonstrating that the level of experience influences the reliability of the score (2012). They also demonstrated that the degree of standardisation of the procedure influences the reliability, especially for the non-expert users (standardized expert ICC = 0.97; non-expert 0.95).

The inter-rater reliability was found to be lower than the intra-rater reliability (Celik, 2016; Conboy et al., 1996; Blonna et al., 2012). Blonna et al.'s results for reliability and LoA were more favourable when experts performed the test, and less favourable when non-experts undertook the task.

The intra- and inter-rater LoAs were within the acceptable ≤ ± 10% threshold only for intra-rater reliability, when experts performed the evaluation using the revised guidelines of Constant et al. for measurement standardisation (Constant et al., 2008; Blonna et al., 2012; Celik, 2016).

The SEM (6 points) and MDC (16.4 points) were the only interpretability aspects to be reported and the absence of floor and ceiling effects is the only available result related to the responsiveness (Conboy et al., 1996; Celik, 2016).

5.3.1.2.1.2.DASH and QuickDASH (NSu-DCS)

Four articles addressed the measurement properties of the DASH (results sub-section 5.3.1.2. “PROMs measurement properties in a diversified shoulder conditions sample, p. 196). The test-retest reliability (ICC = 0.86) was slightly under the expected standard of 0.90, but higher than that of the ASES (ICC = 0.75) (Moser et al., 2012). Real change might have interfered with these results as all patients, including those who changed their shoulder status in-between the test and the retest session one week later, were considered in the calculations of the ICCs. The MCID (11.7 points) was the only reported interpretability aspect (Lundquist et al., 2014).

Three studies reported convergent results concerning the responsiveness of the DASH, with an acceptable AUC (AUC ≥ 0.70) (De vet 2011) for the detection of improved patients (AUC = 0.76 – 0.77) and a moderate change correlation ranging from r = -0.52 – -0.59 above the defined threshold for adequacy (r ≥ 0.50) (Lundquist et al., 2014; Negahban et al., 2015). No floor or ceiling effects were detected (Lundquist et al., 2014).

Test-retest reliability of the QuickDASH was also adequate (ICC = 0.90 – 0.94) (Fayad et al., 2009; Mintken et al., 2009). The LoA were slightly above the ± 10% threshold (- 8.4 – 15.2), and a 3.4% bias was stated between measurements (Fayad et al., 2009). The SEM (4.8%), MDC (11.2%) and MCII (8%) but not the PASS were determined (Mintken et al., 2009).

No study directly compared the responsiveness of the DASH with its simplified version, the QuickDASH. However, comparisons between studies showed comparable change correlation and AUC for the detection of improved patients between these PROMs (Lundquist et al., 2014; Negahban et al., 2015; Fayad et al., 2009; Mintken et al., 2009).

5.3.1.2.1.3.SST (NSu-DCS)

The results of re-retest reliability were conflicting between the four studies that addressed this aspect (Ebrahimzadeh et al., 2016; Neto et al., 2013; Membrilla-Mesa et al., 2015b; van Kampen et al., 2012) (results sub-section 5.3.1.2. “PROMs measurement properties in a diversified shoulder conditions sample, p. 196). Two studies (Membrilla-Mesa et al., 2015b; van Kampen et al., 2012) found adequate

reliability (ICC = 0.91 – 0.92), but a recall effect cannot be excluded in the first one (Membrilla-Mesa et al., 2015b), as the time interval between measurements was 48h only. This shorter interval might also have favourably influenced the calculation of the SEM (2.21% vs. 10% in Van Kampen et al. 2012). MDCs could not be compared between studies as MDC_{90} was used in one study (Membrilla-Mesa et al., 2015b), while the more current MDC_{95} was used in the other one (van Kampen et al., 2012).

Two studies found test-retest reliability ≥ 0.90 at a one-week interval (Ebrahimzadeh et al., 2016; Neto et al., 2013). However, no procedure was apparently implemented in these studies to ensure that patients' performance remained stable between time points. Therefore, part of the test-retest variability might be due to real change in patients. One study found no floor or ceiling effect. Nevertheless, 1.8% of patients reported the worst possible score and 13.6% the best possible score (van Kampen et al., 2012). Another study reported 7.5% floor and 7.5% ceiling effect, defined as the percentage of minimum and maximum scores, respectively (Robins et al., 2017). So, floor and ceiling effect were considered as being present or absent by the authors, as a function of the criteria used to define them, but would have been classified as being absent in all studies when the recommended criteria, defined as < 15% of patients reaching the minimum or maximum scores, was used (Terwee et al., 2007).

5.3.1.2.1.4. ASES (NSu-DCS)

Five studies investigated the measurement properties of the ASES (Cook et al., 2002; Kocher et al., 2005; Moser et al., 2012; Robins et al., 2017; Yahia et al., 2011a) (results sub-section 5.3.1.2. "PROMs measurement properties in a diversified shoulder conditions sample, p. 196). The ASES reliability results were incongruent between studies, with the ICC ranging from 0.96 to 0.75 across studies. The highest reliability was obtained in a study in which the time interval was 1-3 days only and the patients were questioned by two different raters (Yahia et al., 2011b). The lowest one was obtained in a study were all patients, and not only those who reported as unchanged, were included in the analysis (Moser et al., 2012). Real change may have negatively interfered with this result, but it was nevertheless lower than that of the DASH, which had been tested in the same conditions in this study (0.75 vs.0.86).

Floor and ceiling effects are the only other reported measurement properties, with no floor and no ceiling effects, though 2.5% of patients reached the maximum score (Robins et al., 2017).

Levels of evidence of PROMs measurement properties in samples including diversified conditions non-surgically treated are summarised in Table 5.3.

Table 5.3: Summary table for the level of evidence of PROMs measurement properties in samples including diversified conditions non-surgically treated *

	Reliability (ICC and LoAs)	**Responsiveness (AUC, ES, SRM, change correlation, floor/ceiling effect)**	**Interpretability aspects (SEM, MDC, MCII/MCID, PASS)**
Constant	Inter-: - moderate Intra- : +/- low	?	?
Relative Constant	?	?	?
DASH	- low	+ moderate	?
QuickDASH	+ moderate	+ moderate	+/- low
SST	+/- low	?	+/- low
ASES	+/- low	?	+/- low
Comparisons	+/- low	DASH superior to ASES: low	

Legend: Inter-: inter-rater; Intra-: intra-rater; DASH: Disabilities of the Arm, Shoulder and Hand score SST: Simple Shoulder Test; ASES: American Shoulder and Elbow Surgeons score; ICC: intraclass correlation coefficient; LoA: Limits of Agreement; AUC: Area Under the receiver operating characteristic Curve; ES: Effect Size; SRM: Standardised Response Mean; SEM: Standard Error of Measurement; MDC: Minimal Detectable Change; MCII/MCID Minimal Clinically Important Improvement/Difference; PASS: Patient Acceptable Symptom State.
+ sufficient, +/- undetermined, - insufficient, ? non-investigated measurement properties
* Evidence graded using the modified GRADE approach for grading the quality of evidence of measurement properties (Prinsen 2018)

5.3.1.2.1. PROMs DSC surgical treatment (Su)

5.3.1.2.1.1. Constant (Su-DSC)

The five studies that investigated the measurement properties of the Constant Score had assessed differing clinimetric properties for the Constant Score, except a shared evaluation of the SRM.

No relevant information on the Constant Score's reliability was available. One study found excellent intra- and inter-tester Pearson correlations (≥ 0.90) (Rocourt et al., 2008). However, this result was not taken into consideration in the table because ICC would have been the recommended statistics as it integrates systematic error in its calculation (Weir, 2005). MCID (17 points) and PASS (42 or 44 points according to the method) were determined (Christie et al., 2011), but not the LoA, SEM and MDC.

Considering responsiveness, the AUC (AUC = 0.84) for improved patients detection was above the cut-off considered as sufficient (≥ 0.70) and superior to that of the DASH (AUC = 0.79) (Christie et al., 2011; De Vet et al., 2011c). SRM at 6 months was found to be much higher in the study by Ge et al. than in Oh et al.'s (SRM = 2.24 vs. 0.58) (Oh et al., 2009; Ge et al., 2013), possibly because these researchers had included samples of patients with surgical interventions that cannot be compared across studies. These values were thus higher than that of the ASES in Ge et al. (SRM = 2.17) and lower than in Oh et al. (SRM = 0.77). No floor and ceiling effect were detected (Christie et al., 2009).

5.3.1.2.1.2. DASH (Su-DSC)

The three studies assessing the DASH had each investigated a large panel of measurement properties but, other than estimates for MCID, had no properties in common. The single-measurement reliability was above the expected threshold (ICC = 0.91) when measured using sequenced test-retest trials (Schmitt and Di Fabio, 2004). All the interpretability aspects were determined: SEM 5.22, MDC 12.2 and MCID 10.1 – 10.2 and PASS 42.9 – 43 (Christie et al., 2009; Schmitt and Di Fabio, 2004). The MCID and PASS values were consistent across the two studies that reported them (Christie et al., 2009; Schmitt and Di Fabio, 2004).

The correlation coefficient between change score of the DASH and a global rating of change scale was moderate (r = 0.66) and the AUC for improvement detection (AUC = 0.79) was acceptable (AUC ≥ 0.70) but inferior to that of the Constant (AUC = 0.84) (Christie et al., 2011). No floor or ceiling affect was detected (Christie et al., 2009).

5.3.1.2.1.3. ASES (Su-DSC)

Four studies investigated the measurement properties of the ASES. Three of them directly compared it to other PROMs, which is convenient to determine the respective reliability or the responsiveness of several scores in the same population (Ge et al., 2013; Beaton and Richards, 1998; Oh et al., 2009). However, no interpretability aspects were investigated in these studies.

Beaton and Cook (Beaton 1996, Cook 2002) both found an adequate reliability (ICC ≥ 0.90), with 0.96 and 0.91, respectively. This was slightly lower than for the SST (ICC = 0.99) (Beaton and Richards, 1998).

The SRM varied considerably between studies (0.77 – 2.17). No aggregation can be made from these results as the sample composition varied across studies (Beaton and Richards, 1998; Oh et al., 2009; Ge et al., 2013).

The comparison of the responsiveness of the SST, Constant and ASES showed a lower ES and SRM of the SST (ES = 0.50, SRM = 0.47) and Constant (ES = 0.57, SRM = 0.58) at 6 months compared to the ASES (ES = 0.61, SRM = 0.77) (Oh et al., 2009). The SRM of the ASES was also slightly superior to that of the SST (SRM = 0.93 vs. 0.87) in another study (Beaton 1998). Conversely, the Constant showed a higher SRM in another study (SRM = 2.24 vs. 2.17 for the ASES) (Ge et al., 2013).

5.3.1.2.1.4. SST (Su-DSC)

One study found an excellent test-retest reliability at a one-week interval (ICC = 0.99) (Beaton and Richards, 1998).

One study that compared the responsiveness of the SST to the ASES, found a lower ES and SRM for the SST at 6 months for the latter (ES = 0.50, SRM = 0.47 for the SST vs. ES = 0.61, SRM = 0.77 for the ASES) (Oh et al., 2009). Beaton found a higher

SRM at 6 months (SRM = 0.87) (Beaton and Richards, 1998), but this result cannot be compared with Oh et al. due to the differences in sample composition.

The levels of evidence of PROMs measurement properties in samples including diversified conditions surgically treated are summarised in Table 5.4.

Table 5.4: Summary table for the level of evidence of PROMs measurement properties in samples including diversified conditions surgically treated *

	Reliability (ICC and LoAs)	**Responsiveness (AUC, ES, SRM, change correlation, floor/ceiling effect)**	**Interpretability aspects (SEM, MDC, MCII/MCID, PASS)**
Constant	?	+ low	?
Relative Constant	?	?	?
DASH	+ low	+ low	?
QuickDASH	?	?	?
SST	+ low	?	?
ASES	+ moderate	?	?
Comparisons	-	DASH superior to Constant, low ASES superior to Constant, low ASES superior to SST, low Constant superior to SST, low	-

Legend: DASH: Disabilities of the Arm, Shoulder and Hand score SST: Simple Shoulder Test; ASES: American Shoulder and Elbow Surgeons score; ICC: intraclass correlation coefficient; LoA: Limits of Agreement; AUC: Area Under the receiver operating characteristic Curve; ES: Effect Size; SRM: Standardised Response Mean; SEM: Standard Error of Measurement; MDC: Minimal Detectable Change; MCII/MCID Minimal Clinically Important Improvement/Difference; PASS: Patient Acceptable Symptom State.
+ sufficient, +/- undetermined, - insufficient, ? non-investigated measurement properties
* Evidence graded using the modified GRADE approach for grading the quality of evidence for measurement properties

5.3.1.2.1. PROMs measurement properties in diversified shoulder conditions mixed surgical/non-surgical (Mi-DSC)

5.3.1.2.1.1. Constant (Mi-DSC)

No study investigated the Constant Score's properties in a diversified sample, including surgically and non-surgically treated patients.

5.3.1.2.1.2. DASH and QuickDASH (Mi-DSC)

Four studies investigated the measurement properties of the DASH. Two of them tested the test-retest reliability and found it to be either lower than required (0.83) (van Kampen et al., 2013) or adequate (0.95) (Fayad et al., 2008a).

One study found test-retest LoAs (bias) of -7.2 - 10.3 points (3 points), which is within requires clinimetric standards (Fayad et al., 2008a). One study had determined the value of the MDC as 16.3 and MCII as 12.4 (van Kampen et al., 2013). The fact that the MDC was higher than the MCII implies that the latter cannot be considered as valid because when an individual patient records an apparent change score equivalent to the quoted MCII, this score remains within the 95% confidence limits for the estimation of random measurement error (MDC), and as such, error and the potential effects of an intervention cannot be differentiated with confidence (van der Linde et al., 2017; De Vet et al., 2011a). Another study found a MCID value of 11.5 using an anchor-based method (Beaton et al., 2005) approaching the MCII value found by Van Kampen, though MCII considers the change only for patients who improved, while MCID considers patients who improved or deteriorated. Interestingly, Beaton compared the results obtained from various options for MCID calculation and found values ranging from 3.9 to 15, and only moderate agreement between approaches (Kappa = 0.47). This reinforces the controversy raised by other authors about MCID calculation (Tubach et al., 2005c).

The responsiveness has been investigated in two studies using ES and SRM. The SRM at 3 months was quite different between studies (SRM = 0.85 for Diniz-Lopez

and 1.13 for Beaton), but the sample composition was not similar between studies (Diniz Lopes et al., 2009; Beaton et al., 2005).

Two studies investigated the properties of the QuickDASH with the purpose to compare its properties to that of the DASH. Van Kampen found values close to that of the DASH for single-measurement reliability based on sequenced test-retest trials (ICC = 0.83 for the DASH and 0.85 for the QuickDASH), and comparable MDC (17.1 vs. 16.3 points for the DASH) and MCII (13.4 vs. 12.4 points for the DASH) (van Kampen et al., 2013). The ICC value was nevertheless under the ≥ 0.90 threshold. Similarly, Beaton found approaching values between the DASH and the QuickDASH for the SRM (SRM = 1.08 vs 1.13) (Beaton et al., 2005).

5.3.1.2.1.3.SST (Mi-DSC)

One study investigated the reliability of the SST and found it to be lower than required (ICC = 0.86), though comparable to that of the DASH and QuickDASH (van Kampen et al., 2013). The SEM (11.65%), MDC (2.8 points; 23.3%) and MCII (2.2 points; 18.3%) were also determined (Roddey et al., 2000; van Kampen et al., 2013). No study investigated the responsiveness of the SST.

5.3.1.2.1.4.ASES (Mi-DSC)

Two studies found the test-retest reliability of the ASES to be meeting expected standards (0.94-0.96) (Sallay and Reed, 2003; Celik et al., 2013) and one to be lower (0.84) (Michener et al., 2002). These results cannot be directly compared, as the sample composition was different in each study. The interpretability aspects SEM (6.7 points), MDC (9.4 points) and MCID (6.4 points) were determined in one study (Michener et al., 2002).

Variations between studies were observed for the ES (0.80 vs. 1.35) and SRM (0.75 vs. 1.54), probably due to variations in sample composition and variable timeframes (Michener et al., 2002; Vrotsou et al., 2016). The discriminative power for improvement determined by the AUC was adequate (0.74 in Cook and 0.82 in Michener et al. 2002) in two studies (Michener et al., 2002; Cook et al., 2003). The percentages of patient reaching the maximum or the minimum scores were both 8%,

which was considered as no floor/ceiling effects as these values are lower than the 15% considered for threshold (Celik et al., 2013).

Levels of evidence of PROMs measurement properties in samples including diversified conditions either surgically or non-surgically treated are summarised in Table 5.5.

Table 5.5: Summary table for the level of evidence of PROMs measurement properties in samples including diversified conditions either surgically or non-surgically treated *

	Reliability (ICC and LoAs)	**Responsiveness (AUC, ES, SRM, change correlation, floor/ceiling effect)**	**Interpretability aspects (SEM, MDC, MCII/MCID, PASS)**
Constant	?	?	?
Relative Constant	?	?	?
DASH	+/- low	?	+/- low
QuickDASH	- low	?	+/- low
SST	- low	?	+/- low
ASES	+/- moderate	+ low	+/- low
Comparisons	DASH, QuickDASH and SST have comparable ICCs < 0.90	-	-

Legend: DASH: Disabilities of the Arm, Shoulder and Hand score SST: Simple Shoulder Test; ASES: American Shoulder and Elbow Surgeons score; ICC: intraclass correlation coefficient; LoA: Limits of Agreement; AUC: Area Under the receiver operating characteristic Curve; ES: Effect Size; SRM: Standardised Response Mean; SEM: Standard Error of Measurement; MDC: Minimal Detectable Change; MCII/MCID Minimal Clinically Important Improvement/Difference; PASS: Patient Acceptable Symptom State.
+ sufficient, +/- undetermined, - insufficient, ? non-investigated measurement properties
* Evidence graded using the modified GRADE approach for grading the quality of evidence for measurement properties

5.3.1.3. PROMs measurement properties in a rotator cuff conditions sample

	Reliability and measurement error	Interpretability aspects	Responsiveness	Direct comparison
Non-surgical				
Constant				Change correlation with WORC: Constant 0.61 vs. 0.84 DASH ES Constant 0.89 vs. 0.61 DASH SRM Constant 1.16 vs. 0.68 DASH (De Witte 2012) No floor effect ASES vs. 21% SST (Beckman 2015)
Henseler 2015; Holmgren 2014; Moeller 2014;De Witte 2012	+ Intra-rater ICC 0.93 - 0.95 + Inter-rater ICC 0.94 SEM 4.1 - 8.0 pts (relative Constant 10%) MDC 11.2 - 23 pts (28% relative Constant) LoA intra-rat. ± 11.3 - ± 12.6 pts LoA inter-rat. ± 11.6 pts	MCII 15 - 19 pts No floor/ceiling effect for absolute score No floor/17% ceiling effect relative score	Change correlation with WORC 0.61 ES 0.89 SRM 1.16	
DASH				
Haldorsen 2014; Mehta 2015; Michener 2013; Rysstad 2017; De Witte 2012	+/- ICC 0.86 - 0.91 SEM 4.3 - 4.7 LoA -11.9 - 14.1	MDC 11.8 - 13.1 MCII 4.4 No floor/ceiling effect	+ AUC 0.77 Change correlation: 0.61 Change correlation with WORC 0.84	
SST				
Beckman 2015; Naghdi 2015; Tashjian 2010	+ Test-retest 0.94 SEM 0.7 pts/5.5% MDC 3.7 pts/15.3%	MCID 2.05 pts Floor effect: no to 21%/ No ceiling effect		
ASES				
Beckman 2015; Tashjian 2010		MCID 12.01 - 16.72 No floor/ceiling effect		

Measurement properties established on a rotator cuff conditions sample (continued)				
	Reliability and measurement error	**Interpretability aspects**	**Responsiveness**	**Direct comparison**
Surgical				
Constant				SRM 3 months DASH 0.50 vs. 0.51 QuickDASH SRM 6 months DASH 0.75 vs. 0.78 QuickDASH (MacDermid 2015) SRM SST 1.79 vs. 1.63 DASH (MacDermid 2006) SRM absolute Constant 1.38 vs. relative Constant 1.34 vs. ASES 0.94 (Holtby 2005)
Christiansen 2015; Holtby 2005; Kukkonen 2013; O'Connor 1999		MCID 9.9 - 11.0 pts	+ AUC 0.85 for Constant and 0.78 for relative Constant Change correlation 0.32 - 0.78 Change correlation with WORC 0.77	
DASH				
Macdermid 2015; MacDermid 2006				
QuickDASH				
Macdermid 2015				
SST				
MacDermid 2006; Godfrey 2007	+ ICC Test-retest 0.97	No floor/ceiling effect		
ASES				
Holtby 2005; Kocher 2005		No floor/ceiling effec	Change correlation with WORC 0.85t	

5.3.1.3.1. PROMs measurement properties in non-surgical rotator cuff conditions (RCC) samples (NSu-RCC)

Among the instruments selected in this review, the Constant, DASH, ASES and SST, but not the QuickDASH, properties have been investigated in non-surgically treated rotator cuff samples.

5.3.1.3.1.1. Constant (NSu-RCC)

Four studies investigated the measurement properties of the Constant Score in this population (non-surgically treated rotator cuff samples) (Henseler et al., 2015; Holmgren et al., 2014; Moeller et al., 2014; de Witte et al., 2012).

The intra- rater (0.93 – 0.95) and inter-rater (0.94) reliability was found to be adequate (ICC ≥ 0.90) in this population (Portney and Watkins, 2015; Moeller et al., 2014)(Portney and Watkins, 2015, Moeller et al., 2014)(Portney and Watkins, 2015, Moeller et al., 2014). The SEM (8 points; 10% for relative Constant) and MDC (28 points, 23% for relative Constant) were determined in a sample including patients with various diagnoses related to disorders of the rotator cuff, with some variations in subgroups (Henseler et al., 2015). The latter indices of clinimetric performance were found to be lower in another study, with SEM ranging from 4.1 to 4.7 points and intra-rater MDC from 11.2 to 13.1 points according to rater, and inter-rater SEM reaching 11.6 points (Moeller et al., 2014). These differences between studies were observed despite the fact that both had used the revised guidelines for the Constant Score use. However the study that got the most favourable clinimetric values used a fixed isometric dynamometer for strength measurements (Moeller et al., 2014), while a hand-held dynamometer was used in the other one (Henseler 2015). The intra-rater LoA ranged from ± 11.2 – ± 13.1 points and the inter-rater LoA were ± 11.6 points, i.e. larger than the ± 10% defined threshold in the study that used a fixed isometric dynamometer (Moeller et al., 2014).

The responsiveness of the Constant Score had been evaluated using ES (0.89), SRM (-1.16) and change correlation with the WORC (r = 0.61) (de Witte et al., 2012). These values were higher than those of the DASH were (ES = -0.61; SRM = -0.68), while

the WORC had slightly higher ES (-0.96) and lower SRM (-0.91). No ceiling effect was observed for the absolute Constant Score, while it was 17% for the relative Constant Score, with important variations in subgroups (53% maximum scores for impingement, 7% for tear and 15% for massive tear). A floor effect was absent in all circumstances (Henseler et al., 2015). No AUC has been reported for unimportant/important change discrimination.

The MCII was 17 points, with some variations according to the rotator cuff integrity (intact: 19 points; torn 15 points) (Holmgren 2014). Similar variations were observed for the specificity (intact: 76%; torn 91%; overall 91%) and the sensitivity (intact: 97%; torn 82%; overall 79%) for unimportant/important change discrimination (Holmgren et al., 2014).

5.3.1.3.1.2. DASH (NSu-RCC)

Five studies investigated the properties of the DASH in this population. The three studies that investigated the test-retest reliability found ICC values either slightly over or under the 0.90 expected threshold (ICC = 0.86 – 0.91) (Haldorsen et al., 2014; Mehta et al., 2015; Rysstad et al., 2017).

All the interpretability aspects were determined except the PASS. The studies that investigated the SEM (4.7 – 4.3 points) and the MDC (11.8 – 13.1 points) found concordant results (Haldorsen et al., 2014; Mehta et al., 2015; Rysstad et al., 2017). The LoA were -11.9 to 14.1 points, but the bias was not reported (Haldorsen et al., 2014). The reported MCII was 4.4 points, but this value cannot be considered valid, as it is smaller than the MDC.

The two studies that calculated the ES and SRM found results of very different magnitudes (ES 2.2 vs. 0.61; SRM 6.1 vs. 0.68) for comparable timeframes between measurements (Mehta 2015; de Witte 2012). The comparison with the Constant Score showed lower responsiveness for the DASH (ES: 0.61 vs. -0.89; SRM: 0.68 vs. -1.16) (de Witte et al., 2012). The change score was moderately correlated with perceived recovery (r = -0.61) (Rysstad et al., 2017) and strongly with the WORC (r = -0.84) (de Witte et al., 2012; Hinkle et al., 2003). The AUC for improved/unimproved discrimination (AUC = 0.71) was just above the threshold considered as acceptable

(De Vet et al., 2011c), with sensitivity of 0.77% and specificity of 0.69% (Rysstad et al., 2017).

All the studies investigated the floor and ceiling effects and none of them found one of these effects.

5.3.1.3.1.3.SST (NSu-RCC)

Three studies investigated the properties of the SST in this population. The test-retest reliability was found to be adequate (ICC = 0.94) in the only article that had reported this property (Naghdi et al., 2015).

The SEM (0.7 points; 5.5%), MDC (3.7 points; 15.3%) and the MCID (2.05 points, corresponding to 17% or 2.33 points, corresponding to 19%, for four- and fifteen-point anchor method) were reported (Naghdi et al., 2015; Tashjian et al., 2010), but not the PASS and the LoA.

Only the floor and ceiling effects were investigated among responsiveness-related properties. One study detected no such effects (Naghdi et al., 2015), while the other found that 6.1% of patients reached the maximum score, which was considered as no ceiling effect and 21% had reached the minimum score, which was above the defined 15% threshold for a floor effect (Beckmann et al., 2015). Comparatively, no floor and ceiling effects were found for the ASES in the same study.

5.3.1.3.1.4.ASES (NSu-RCC)

Two studies investigated the properties of the ASES in this population (Tashjian et al., 2010; Beckmann et al., 2015) but none of them investigated its reliability. Only the MCID was reported among interpretability aspects. Values of 12 and 17 points were found for shoulder function MCID using a four- or a fifteen-point anchor, respectively (Tashjian et al., 2010). The percentages of patients reaching the maximum and the minimum score effects were both 2.3%, which was lower than the 15% defined as the threshold for a floor or a ceiling effect. Conversely, the SST showed a floor effect that had exceeded the threshold, in the same study (Beckmann et al., 2015). No other responsiveness properties have been investigated for the ASES.

Levels of evidence of PROMs measurement properties in samples including rotator cuff conditions non-surgically treated are summarised in Table 5.6.

Table 5.6: Summary table for the level of evidence of PROMs measurement properties in samples including rotator cuff conditions non-surgically treated*

	Reliability (ICC and LoAs)	**Responsiveness (AUC, ES, SRM, change correlation, floor/ceiling effect)**	**Interpretability aspects (SEM, MDC, MCII/MCID, PASS)**
Constant	Inter-: + low Intra-: + low	+/- low	+/- low
Relative Constant	?	?	?
DASH	+/- low	+ low	+/- low
QuickDASH	?	?	?
SST	+ low	?	+/- low
ASES	?	?	+/- low
Comparisons		Higher ES and SRM for Constant than DASH: low Floor effect for the SST vs. none for the ASES: low Ceiling effect for the relative Constant vs. no for the Constant	

Legend: Intra-: Intra-rater; Inter-: Inter-rater; DASH: Disabilities of the Arm, Shoulder and Hand score SST: Simple Shoulder Test; ASES: American Shoulder and Elbow Surgeons score; ICC: intraclass correlation coefficient; LoA: Limits of Agreement; AUC: Area Under the receiver operating Curve; ES: Effect Size; SRM: Standardised Response Mean; SEM: Standard Error of Measurement; MDC: Minimal Detectable Change; MCII/MCID Minimal Clinically Important Improvement/Difference; PASS: Patient Acceptable Symptom State.
+ sufficient, +/- undetermined, - insufficient, ? non-investigated measurement properties
* Evidence graded using the modified GRADE approach for grading the quality of evidence of measurement properties (Prinsen 2018)

5.3.1.3.1. PROMs measurement properties in surgical rotator cuff conditions (Su-RCC)

Among the instruments selected in this review, the Constant, DASH, QuickDASH, SST and ASES properties have been investigated in surgically treated rotator cuff samples.

5.3.1.3.1.1. Constant (Su-RCC)

Four studies investigated the measurement properties of the Constant Score following rotator cuff surgery (Christiansen, 2015; Holtby and Razmjou, 2005; Kukkonen et al., 2013; O'Connor et al., 1999). No usable information is available on the Constant Score's reliability in this population. One study found excellent intra-rater Spearman correlations (r = 0.96) and good to excellent inter-rater reliability (r = 0.91 and 0.89) (Livain et al., 2007). However, this result was not taken into consideration in the table because ICC would have been the recommended statistics as it integrates systematic error in its calculation (Weir, 2005).

The responsiveness was evaluated in three studies. The ES were not comparable as one study determined it in all patients (O'Connor et al., 1999) and the other one in improved patients only (Christiansen, 2015). Similarly, the SRM was not comparable between studies as the type of surgery was different (decompression vs. mix of different surgeries related to rotator cuff), which probably led to difference in the magnitude of the measured effects (O'Connor et al., 1999; Holtby and Razmjou, 2005). The change correlations (0.32 – 0.78) were also varying amongst studies as they had used different reference tools and time points for comparison (Holtby and Razmjou, 2005; Christiansen et al., 2015; O'Connor et al., 1999). The AUC for improved/unimproved discrimination was above the required threshold (AUC ≥ 0.70) (Terwee et al., 2007; De Vet et al., 2011c) for the absolute (AUC =0.85) and relative (AUC = 0.78) Constant Score. Concerning responsiveness, the Constant Score displayed higher SRM than the ASES (SRM = 1.38 for absolute score and 1.34 for relative score vs. 0.94 for ASES) (Holtby and Razmjou, 2005).

The MCID is the only interpretability aspect that was investigated. Two studies found that its value was considerably influenced by the calculation method (Christiansen et al., 2015; Kukkonen et al., 2013). Using the common anchor-based approach, the

values were close between studies (10.4 vs. 11 points) (Christiansen et al., 2015; Kukkonen et al., 2013).

5.3.1.3.1.2. DASH and QuickDASH (Su-RCC)

The only investigated measurement property of the DASH and of the QuickDASH was the responsiveness. The SRM values differ considerably between studies, but are not comparable because one study considered all patients following cuff repair (SRM = 0.50 at 3 months; 0.75 at 6 months) (Macdermid et al., 2015), while the other one calculated the specific SRM of subgroups with different patterns of progress (positive, equivocal, negative) (MacDermid et al., 2006). The comparison of the DASH's and QuickDASH's SRMs showed close values between these scores (SRM = 0.75 and 0.78, respectively) (Macdermid et al., 2015).

5.3.1.3.1.3. SST (Su-RCC)

Two studies investigated the measurement properties of the SST in patients undergoing surgery. The test-retest reliability was adequate (ICC = 0.97) (Godfrey et al., 2007). Two studies evaluated the responsiveness using the SRM, but they were not comparable because one study considered all patients following cuff repair post-surgery without defining a precise timeframe (SRM 1.01; ES 1.08) (Godfrey et al., 2007), while the other one calculated the specific SRM of subgroups with different patterns of progress [positive (SRM = 1.79), equivocal (SRM = 0.17), negative (SRM = -0.73)] (MacDermid et al., 2006). The floor (2.1%) and ceiling (5.1%) effects were under the 15% defined threshold (Godfrey et al., 2007). No interpretability aspect was investigated.

5.3.1.3.1.4. ASES (Su-RCC)

Two studies investigated the responsiveness of the ASES in this population, while the other measurement properties were not investigated. One study reported and adequate change correlation with the WORC (r = 0.85) (Holtby and Razmjou, 2005), while the second one reported no floor or ceiling effects (Kocher et al., 2005).

Levels of evidence of PROMs measurement properties in samples including rotator cuff conditions surgically treated are summarised in Table 5.7.

Table 5.7: Summary table for the level of evidence of PROMs measurement properties in samples including rotator cuff conditions surgically treated*

	Reliability (ICC and LoAs)	**Responsiveness (AUC, ES, SRM, change correlation, floor/ceiling effect)**	**Interpretability aspects (SEM, MDC, MCII/MCID, PASS)**
Constant	?	+ low	+/- low
Relative Constant	?	?	?
DASH	?	?	?
QuickDASH	?	?	?
SST	+ low	?	?
ASES	?	+ low	?
Comparisons		Comparable SRM between DASH and QuickDASH: low SRM of DASH and SST are comparable: low Constant and relative Constant have higher SRM than ASES	

Legend: DASH: Disabilities of the Arm, Shoulder and Hand score SST: Simple Shoulder Test; ASES: American Shoulder and Elbow Surgeons score; ICC: intraclass correlation coefficient; LoA: Limits of Agreement; AUC: Area Under the receiver operating Curve; ES: Effect Size; SRM: Standardised Response Mean; SEM: Standard Error of Measurement; MDC: Minimal Detectable Change; MCII/MCID Minimal Clinically Important Improvement/Difference; PASS: Patient Acceptable Symptom State.
+ sufficient, +/- undetermined, - insufficient, ? non-investigated measurement properties
* Evidence graded using the modified GRADE approach for grading the quality of evidence of measurement properties (Prinsen 2018)

5.3.1.4. PROMs Measurement properties in an osteoarthritis shoulder condition sample

	Reliability	Interpretability aspects	Responsiveness	Direct comparison
Non-surgical				
DASH				ES DASH 1.07 vs. WORC 1.33 SRM DASH 0.90 vs. WORC 1.11.
Corona 2016		No floor/No ceiling effect		
Surgical				
Constant				
Angst 2004; Angst 2008; Torrens 2016; Sciascia 2017		MCID 8 pts PASS 78 pts No floor/ceiling effect	+ AUC 0.77 - 0.85	ES Constant 2.23 vs. ASES 2.13 vs. DASH 1.19 SRM Constant 1.99 vs. ASES 1.81 vs. DASH 1.22 AUC Constant 0.77 vs. ASES 0.76 vs. DASH 0.71 (Angst 2008) SRM 3 months QuickDASH 0.84 vs. DASH 0.82 SRM 6 months QuickDASH 1.06 vs. DASH 1.07 (Macdermid 2015) ES 1.26 QuickDASH vs. 1.17 DASH (Angst 2009) ES SST 2.23 vs. DASH 1.41 SRM SST 1.73 vs. DASH 1.76 (Roy 2010) ES Constant 2.9 vs. ASES 2.5 SRM Constant 2.4 vs. ASES 2.2 Ceiling effect: none Constant vs. 21% ASES (Sciascia 2017)
DASH				
Macdermid 2015; Angst 2004; Angst 2008; Roy 2010		No floor/No ceiling effect	+ AUC 0.71 Change correlation with SST 0.50	
QuickDASH				
Angst 2009; Macdermid 2015				
SST				
Roy 2010; Tashjian 2017		MCID 2.4 - 3 pts	+/- AUC 0.66 Change correlation with DASH 0.50	
ASES				
Angst 2004; Angst 2008; Sciascia 2017; Goldhahn 2008; Kocher 2005; Tashjian 2017; Werner 2016	+ ICC Test-retest 0.93	MCID 13.5 - 21 PASS 73 No floor/No ceiling effect	+/- AUC 0.76 - 0.88	
Measurement properties established on an osteoarthritis shoulder condition sample (continued)				
Mix of surgical and non-surgical				
Constant and ASES				SRM Constant 1.21 vs. ASES 1.29 (Lo 2001)
Lo 2001				

5.3.1.4.1. PROMs measurement properties in non-surgical osteoarthritis (NSu-OA)

The data were very scarce for non-surgical treatment as only one study addresses the measurement properties of outcome measures in this population (Corona et al., 2016). Only the responsiveness of the DASH score has been partially investigated for this approach as the effect size, SRM, floor and ceiling effects have been calculated at six months interval (ES= 1.07; SRM = 0.90), which were lower than those of the Western Ontario Rotator Cuff index (WORC), a condition-specific PROM. No floor and ceiling effects were detected.

5.3.1.4.2. PROMs measurement properties in surgical osteoarthritis PROMs (Su-OA)

The measurement properties of all the selected PROMs have been investigated for surgically treated osteoarthritis, with the exception of the relative Constant.

5.3.1.4.2.1. Constant (Su-OA)

The Constant measurement tool's reliability has not been tested for surgically treated OA. The MCID and PASS are the only interpretability aspects that have been established (MCID 8 points; PASS 78 points) for the Constant Score (Sciascia et al., 2017; Torrens et al., 2016).

The responsiveness of the Constant has been found to be superior to that of the DASH based on ES and SRM, and slightly above that of the ASES (Constant ES = 2.23 vs. 2.13 for ASES vs. 1.19 for DASH; SRM Constant = 1.99 vs. 1.81 for ASES vs. 1.22 for DASH) (Angst et al., 2008). Another study also found a slightly better ES and SRM for the Constant compared to the ASES (Constant ES = 2.9 vs. 2.5 for ASES; Constant SRM = 2.4 vs. 2.2 for ASES) (Roy et al., 2010). Discriminative power between improved/unimproved patients was adequate (AUC = 0.77 - 0.85), which was close to that of the ASES (AUC = 0.76) and slightly superior to that of the DASH (AUC = 0.71) (Angst et al., 2008; Sciascia et al., 2017). No ceiling or floor effects for the Constant was shown five to six years after surgery (Angst et al., 2004).

5.3.1.4.2.2. DASH and QuickDASH (Su-OA)

The reliably and the interpretability aspects of the DASH and QuickDASH have not been evaluated.

The responsiveness of the DASH was acceptable (AUC = 0.71) but slightly inferior to that of the Constant and ASES based on AUCs. The responsiveness of the DASH was also inferior to that of the Constant and ASES based on ES and SRM (see previous sub-section) (Angst et al., 2008). The comparison of the DASH and the QuickDASH showed comparable SRM (DASH SRM = 0.82 vs QuickDASH 0.84) and ES (DASH ES = 1.07 vs QuickDASH 1.06) (Angst et al., 2009; Macdermid et al., 2015). The DASH had inferior ES and similar SRM to the SST (ES = 1.41 vs. 2.23; SRM = 1.76 vs 1.71) (Roy et al., 2010). Its change correlation with the SST was moderate, precisely at the r = 0.50 threshold (Roy et al., 2010). No ceiling or floor effects were reported for the DASH five to six years after surgery (Angst et al., 2004).

5.3.1.4.2.3. SST (Su-OA)

The reliability of the SST has not been tested and MCID is the only interpretability aspect that has been determined (2.4 – 3 points) (Roy et al., 2010). Though the SST had a superior ES and equivalent SRM to the DASH (ES = 2.23 vs. 1.41; SRM = 1.71 vs. 1.76), it showed a discriminative power lower than the ≥ 0.70 threshold (AUC = 0.66).

5.3.1.4.2.4. ASES (Su-OA)

The ASES is the only PROM for which the test-retest reliability, which was excellent (ICC = 0.93), and has been evaluated in this population (Goldhahn et al., 2008). The PASS has been determined to be 73 points (Sciascia et al., 2017) and the MCID 13.5 – 21 points (Tashjian et al., 2017; Werner et al., 2016).

The ASES (ES = 2.5, SRM = 2.2, AUC = 0.88) and Constant (ES = 2.9, SRM = 2.4, AUC = 0.85) showed similar levels of responsiveness, and discriminative power between satisfied and unsatisfied patients (Sciascia et al., 2017).

The same study of Sciascia et al. (2017) showed a ceiling effect for the ASES at two years following arthroplasty, while Angst et al. found no ceiling or floor effects for the

Constant, DASH and ASES five to six years after the same surgery (Angst et al., 2004; Sciascia et al., 2017).

Levels of evidence for measurement properties of PROMs in samples including patients with osteoarthritis surgically treated are summarised in Table 5.8.

Table 5.8: Summary table for the level of evidence for measurement properties of PROMs in samples including patients with osteoarthritis surgically treated*

	Reliability (ICC and LoAs)	**Responsiveness (AUC, ES, SRM, change correlation, floor/ceiling effect)**	**Interpretability aspects (SEM, MDC, MCII/MCID, PASS)**
Constant	?	+ moderate	+/- low
Relative Constant	?	?	?
DASH	?	+ low	?
QuickDASH	?		+ low
SST	?	+/- low	+/- low
ASES	+ low	+/- moderate	+/- low
Comparisons	+/- low	Constant superior to ASES: moderate ASES superior to DASH: low SST comparable to DASH: low	-

Legend: DASH: Disabilities of the Arm, Shoulder and Hand score SST: Simple Shoulder Test; ASES: American Shoulder and Elbow Surgeons score; ICC: intraclass correlation coefficient; LoA: Limits of Agreement; AUC: Area Under the receiver operating Curve; ES: Effect Size; SRM: Standardised Response Mean; SEM: Standard Error of Measurement; MDC: Minimal Detectable Change; MCII/MCID Minimal Clinically Important Improvement/Difference; PASS: Patient Acceptable Symptom State.
+ sufficient, +/- undetermined, - insufficient, ? non-investigated measurement properties

* Evidence graded using the modified GRADE approach for grading the quality of evidence of measurement properties (Prinsen 2018)

5.3.1.4.3. PROMs measurement properties in mixed surgical/non-surgical osteoarthritis (Mi-OA)

Only one study investigated this type of composition within a population. It found closely matched SRMs for the Constant and the ASES in a mixed sample of non-surgically and surgically treated patients (SRM = 1.21 vs. 1.29) (Lo et al., 2001). Due to the scarce literature, no table is provided to summarize the levels of evidence for measurement properties in this subpopulation. No information on the reliability and interpretability aspects is available. Concerning the responsiveness, only the low level of evidence for the equivalence between the Constant and ASES can be stated.

5.3.1.5. PROMs Measurement properties in a shoulder instability sample

Measurement properties established in s shoulder instability sample				
	Reliability and measurement error	**Interpretability aspects**	**Responsiveness**	**Direct comparison**
Non-surgical				
Constant				
Dawson 1999; Kirkley 1998			ES 0.2	SRM 3 months WOSI 0.93 vs. DASH 0.71 vs. Constant 0.59 vs. ASES 0.54 AUC WOSI 0.90 vs. Constant 0.76 (Kirkley 1998) ES WOSI 1.57 vs. DASH 1.47 SRM WOSI 1.94 vs. DASH 1.43 (Cacchio 2012)
DASH				
Cacchio 2012; Kirkley 1998		MCID 22 pts No floor/ceiling effect	+ AUC 0.76	
SST				
Godfrey 2007;	+ Test-retest ICC 1.00	No floor/ceiling effect		
ASES				
Kirkley 1998				
WOSI				
Cacchio 2012; Hatta 2011; Hofstaetter 2010; Kirkley 1998; Skare 2013; Van der Linde 2014; Wiertsema 2014; Basar 2017	+ Test-retest ICC 0.91 - 0.98 SEM 71 pts (3.4%) - 130 (6.2%) MDC 196 pts (9.3%) - 483 (23.0%) LoA 333.9 - 344.8 pts (15.9% - 16.4%) - LoA ± 400 pts (19%)	MCID 400 pts (19%) No floor/No ceiling effect	+ AUC 0.90	

Measurement properties established in a shoulder instability sample (continued)				
	Reliability and measurement error	**Interpretability aspects**	**Responsiveness**	**Direct comparison**
Surgical				
WOSI				ICC test-retest WOSI 0.84 vs. QuickDASH 0.75 SRM WOSI 1.55 vs. QuickDASH 0.87 (Gaudelli 2014) ES 0.62 vs Rowe 0.46 SRM WOSI 0.65 vs. Rowe 0.34 (Oh 2009)
Gaudelli 2014; Salomonsson 2009; Oh 2009	+/- Test-retest ICC 0.84 - 0.94	No floor/No ceiling effect		
QuickDASH				
Gaudelli 2014				
Mixed surgical non-surgical				
DASH				Change correlation with WOSI: DASH 0.75 vs. SST 0.69 (van der Linde 2017)
van der Linde 2017				
SST (van der Linde 2017)				
van der Linde 2017				
ASES				
Kocher 2005		No floor/No ceiling effect		
WOSI				
Yuguero 2016; van der Linde 2017	+ Test-retest ICC 0.95	MCID 294 pts (14%) No floor/No ceiling effect	+ AUC 0.82 Change correlation WOSI 0.64	

5.3.1.5.1. PROMs measurement properties in non-surgical shoulder instability (NSu-SI)

Ten studies on the measurement properties of the Constant, DASH, SST and WOSI were found, but not on the QuickDASH and the ASES.

5.3.1.5.1.1. Constant (NSu-SI)

Two studies investigated the responsiveness of the Constant using the SRM. Limited information can be drawn from these results, as the SRMs were not compared to that of other outcome measures. One study found a lower value (SRM = 0.2) (Dawson et al., 1999) than the other one (SRM = 0.59) (Kirkley et al., 1998). Based on the global rating of change scale that was used in both studies, this difference is influenced by the higher proportion of patients who improved in the second study.

5.3.1.5.1.2. DASH (NSu-SI)

Two studies investigated the responsiveness of the DASH, among which one also defined the MCID (22 points). The SRM was much higher in one study (SRM = 1.43) (Cacchio et al., 2012) than the other (SRM = 0.71) (Kirkley et al., 1998), possibly because the patients' change was influenced by the structured rehabilitation treatment applied between measurements that had occurred within the first study only. In both studies, the responsiveness was lower than that of the WOSI (SRM = 1.94 and 0.93, respectively). The AUC was adequate (AUC = 0.76) but lower than that of the WOSI (AUC = 0.91-0.98) and mainly due to its lack of sensitivity (61%) (Cacchio et al., 2012). The proportion of patients who reported the lowest possible score was 1.5% at baseline, which was under the 15% threshold to consider that a floor effect had been present, but appeared to be nevertheless higher than for the WOSI (0%). No ceiling effect was detected (Cacchio et al., 2012).

5.3.1.5.1.3. SST (NSu-SI)

One studies investigated the measurement properties of the SST (Godfrey et al., 2007). The reliability was perfect (ICC = 1.00). The SRM was calculated (SRM = 0.63) but not compared to that of other PROMs. The proportion of patients reaching the

minimal and maximal score was 2.0% and a 9.3%, respectively, which was under the 15% threshold to consider that a floor or ceiling effect is present.

5.3.1.5.1.4. ASES (NSu-SI)

One studies investigated the measurement properties of the ASES, in a larger study that considered several PROMs. The responsiveness was compared to that of the other investigated PROMs. Based on the SRM values (SRM = 0.54), it was lower than that of the Constant (SRM = 0.59), DASH (SRM = 0.71) and WOSI (SRM = 0.93).

5.3.1.5.1.5. WOSI (NSu-SI)

The WOSI was the most frequently investigated PROM for the evaluation shoulder function in instability, as eight studies addressed its measurement properties compare to maximum two for other PROMs. All studies found a reliability above the ICC ≥ 0.90 threshold (ICC = 0.91 – 0.98).

Among measurement error issues, the SEM and MDC were evaluated in three studies. The reported SEMs ranged from 71 – 171 points (3.4% – 8.3%) and the MDC from 196 – 483 points (9.3% – 23%) (Cacchio et al., 2012; Wiertsema et al., 2014; van der Linde et al., 2014). The examination of methods used in the studies did not provide possible explanations for these discrepancies, except that the SEM tended to be smaller when the test-retest interval was shorter, which might have been induced by recall effects. The LoAs were investigated in two studies. One found them to be -333.0 – 344 points (-15.9% – 16.4%) (Skare et al., 2013), while the estimation based on the graphs presented in the second one was around ± 400 points (19%) (Wiertsema et al., 2014). Thus, LoAs were larger than the 10% threshold in any case and indicating potentially excessive errors based on single measurements when using the SST.

The SRM was investigated in two studies, with a larger value (SRM = 1.57) in the study that calculated it before and after a structured rehabilitation (Cacchio et al., 2012), than in the one without specific rehabilitation (SRM = 0.93) (Kirkley et al., 1998). In direct comparisons, the responsiveness of the WOSI was higher than that of the DASH (SRM = 1.94 vs. 1.43 and ES 1.57 vs. 1.47 in the first mentioned study, and SRM = 0.93 vs. 0.71 in the second one) and of the Constant and ASES (SRM 0.93 vs. 0.59 for Constant and 0.54 for SST). The AUC was excellent (AUC = 0.90)

and superior to that of the Constant (AUC = 0.76), with better sensitivity (92%) than specificity (83%) (Cacchio et al., 2012). No floor or ceiling effects were detected in the five studies that investigated this aspect using the 15% minimal/maximal score criteria (Cacchio et al., 2012; Hofstaetter et al., 2010; van der Linde et al., 2014; Wiertsema et al., 2014; Basar et al., 2017). However, when the criteria for floor and ceiling effect was based on the patients who performed at the maximum score minus MDC or the minimum score plus MDC, respectively, 17% had a score < MDC (floor effect) and 5% has a score > 100-MDC (ceiling effect) (van der Linde et al., 2014). Levels of evidence for measurement properties of outcome measures in samples including patients with shoulder instability non-surgically treated are summarised in Table 5.9.

Table 5.9: Summary for the level of evidence of PROMS measurement properties in samples including patients with shoulder instability non-surgically treated *

	Reliability (ICC and LoAs)	**Responsiveness (AUC, ES, SRM, change correlation, floor/ceiling effect)**	**Interpretability aspects (SEM, MDC, MCII/MCID, PASS)**
Constant	?	- low	?
Relative Constant	?	?	?
DASH	?	- low	+/- low
QuickDASH	?	?	?
SST	+ low	?	?
ASES	?	- low	+/- low
WOSI	+ high - LoA > 10% : moderate	+ high	+ moderate

Legend: Disabilities of the Arm, Shoulder and Hand score SST: Simple Shoulder Test; ASES: American Shoulder and Elbow Surgeons score; WOSI: Western Ontario Shoulder Disability Index; ICC: intraclass correlation coefficient; LoA: Limits of Agreement; AUC: Area Under the receiver operating Curve; ES: Effect Size; SRM: Standardised Response Mean; SEM: Standard Error of Measurement; MDC: Minimal Detectable Change; MCII/MCID Minimal Clinically Important Improvement/Difference; PASS: Patient Acceptable Symptom State.
+ sufficient, +/- undetermined, - insufficient, ? non-investigated measurement properties
* Evidence graded using the modified GRADE approach for grading the quality of evidence of measurement properties (Prinsen 2018)

5.3.1.5.1. PROMs measurement properties in surgical shoulder instability (Su-SI)

Only four studies investigated the measurement properties of the PROMS in surgical shoulder instability treatment. Only the DASH and the WOSI outcome measures were concerned by these studies.

5.3.1.5.1.1. QuickDASH (Su-SI)

One study investigated the properties of the DASH. The test-retest reliability's ICC was under the 0.90 threshold (ICC = 0.75). The SRM (SRM = 0.87) was considerably lower than for the WOSI (SRM = 1.55) at one year (Gaudelli et al., 2014).

5.3.1.5.1.2. WOSI (Su-SI)

Three studies investigated the measurement properties of the WOSI. Concerning the reliability, one found an ICC above the defined ≥ 0.90 threshold (ICC = 0.94) (Salomonsson et al., 2009), while the other one found a lower value (ICC = 0.84), which was nevertheless higher than that of the QuickDASH (ICC = 0.75) (Gaudelli et al., 2014). The questionnaire was administrated by means of telephone conversation in the study in which lower values were found. No interpretability aspects were determined for surgical treatment.

Concerning the WOSI measurement tool's responsiveness, one study found much lower SRM at 6 months (SRM = 0.65) (Oh 2009) than the two other ones that investigated this parameter (SRM = 1.40 at 6 months and 1.55 at 1 year) (Salomonsson et al., 2009; Gaudelli et al., 2014), respectively. The SRM and ES (SRM = 0.66; ES = 0.62) were nevertheless higher than those of the concurrent Rowe score for shoulder instability (SRM = 0.34; ES 0.46) (Oh et al., 2009) and of the QuickDASH (SRM = 0.87).

Levels of evidence for measurement properties of PROMs in samples including patients with surgically treated shoulder instability are summarised in Table 5.10.

Table 5.10: Summary table for the level of evidence for measurement properties of PROMS outcome measures in samples including patients with shoulder instability surgically treated*

	Reliability (ICC and LoAs)	**Responsiveness (AUC, ES, SRM, change correlation, floor/ceiling effect)**	**Interpretability aspects (SEM, MDC, MCII/MCID, PASS)**
Constant	?	?	?
Relative Constant	?	?	?
DASH	?	?	?
QuickDASH	- low	- low	?
SST	?	?	?
ASES	?	?	?
WOSI	+ low (- low over phone)	+ low	?
Comparisons	-	WOSI superior to QuickDASH: low WOSI superior to Rowe: low	-

Legend: DASH: Disabilities of the Arm, Shoulder and Hand score SST: Simple Shoulder Test; ASES: American Shoulder and Elbow Surgeons score; WOSI: Western Ontario Shoulder Disability Index; ICC: intraclass correlation coefficient; LoA: Limits of Agreement; AUC: Area Under the receiver operating characteristic Curve; ES: Effect Size; SRM: Standardised Response Mean; SEM: Standard Error of Measurement; MDC: Minimal Detectable Change; MCII/MCID Minimal Clinically Important Improvement/Difference; PASS: Patient Acceptable Symptom State.

+ sufficient, +/- undetermined, - insufficient, ? non-investigated measurement properties

* Evidence graded using the modified GRADE approach for grading the quality of evidence of measurement properties (Prinsen 2018)

5.3.1.5.1. PROMs measurement properties in mixed surgical/non-surgical shoulder instability (Mi-SI)

Three studies investigated the measurement properties in mixed samples of patients with shoulder instability treated either non-surgically or surgically, using the WOSI, SST, DASH or ASES (van der Linde et al., 2017; Kocher et al., 2005; Yuguero et al., 2016).

5.3.1.5.1.1. DASH and SST (Mi-SI)

The DASH and the SST were investigated in the same study (van der Linde et al., 2017). Their moderate to high change correlations with the global rating of change scale [SST (r = 0.69) and DASH (r = 0.75)] are indicative of adequate responsiveness.

5.3.1.5.1.1. ASES (Mi-SI)

Only the responsiveness of the ASES measurement tool was investigated. However, little can be inferred from the ES (ES = 0.86) and SRM (SRM = 0.93) in the absence of comparison with another outcome measure. No floor effect was found and 1.3% of patients reached the maximal score, which is under the 15% defined threshold for a ceiling effect (Kocher et al., 2005).

5.3.1.5.1.2. WOSI (Mi-SI)

The reliability of the WOSI was adequate (0.95) (Yuguero et al., 2016). The SRM (SRM = 0.61) and ES (ES = 0.25) were determined but not compared to other outcome measures. The moderate to high change correlations with the global rating of change scale (r = 0.64), SST (r = 0.69) and DASH (r = 0.75) are indicative of adequate responsiveness. The discriminative power for improved/unimproved discrimination was adequate (AUC = 0.82) (van der Linde et al., 2017). The MCID (14%) was the only determined interpretability aspect (van der Linde et al., 2017).

5.3.1.6. PROMs Measurement properties in a non-surgically treated capsulitis sample

	Reliability and measurement error	Interpretability aspects	Responsiveness	Direct comparison
DASH				
Staples 2010			+ Active treatment AUC 0.71 + Improved patients AUC 0.82 + Markedly improved patients AUC 0.86 Change correlations 0.66	

5.3.1.6.1. PROMs in non-surgically treated capsulitis (NSu-C)

Only the DASH Score's measurement properties have been partially evaluated, following arthrographic joint distension or oral prednisolone (Staples et al., 2010). No data are available for test-retest reliability and no interpretability aspects have been defined.

The responsiveness has been differentiated between the "receiving treatment of known efficacy", the "improved" and the "markedly improved" groups. The AUC was higher than the 0.70 threshold in all of these groups (AUC = 0.71 – 0.86). The possibility of interpreting the magnitude of the ES and SRM results is nevertheless limited as no comparison with other PROMs was performed. Due to the scarce literature, no table is provided to summarize the levels of evidence for measurement properties in this subpopulation. It could only be stated that the DASH display adequate responsiveness, with a low level of evidence.

5.3.1.7. PROMs Measurement properties in a shoulder fracture sample

	Reliability and measurement error	Interpretability aspects	Responsiveness	Direct comparison
Non-surgical				
Constant				Test-retest ICC Constant 0.91 vs. 0.87 DASH Change correlation with SSV: Constant 0.66 vs. DASH - 0.68 (Van de Water 2014) ES Constant 0.31 vs. DASH 0.44 (Van de Water 2016)
Van de Water 2014; Van de Water 2016	+ Test-retest ICC 0.91 SEM 4.5 LoA ~ ± 10 pts (bias ~5%) (from graph)	MCID 5.1 - 11.4	Change correlation with SSV 0.66; with DASH - 0.72	
DASH				
Fayad 2008b; Van de Water 2014; Van de Water 2016	- Test-retest ICC 0.87 SEM 6.5 LoA ~ ± 15 pts (bias ~5%) (from graph)	MCID -8.1 - -13.0	Change correlation with SSV 0.68 ; with Constant -0.72 Change correlation with handicap scale 0.33	
Mixed surgical/non-surgical				
Constant				
Mahabier 2016	SEM 6.4 MDC 17.7	MCID 6.1 No floor/ceiling effect	- AUC 0.59 Change correlation with DASH -0.60	AUC DASH 0.66 vs. Constant 0.59 ES Constant 1.71 vs. DASH -1.55 SRM DASH -1.63 vs. Constant 1.60 Ceiling effect DASH 31.1% vs. none for Constant (Mahabier 2016)
DASH				
Slobogean 2010; Mahabier 2016	+ Test-retest ICC 0.93 SEM 6.9 MDC 19.0 LoA 15.2 - 15.9 (bias 0.4)	MCID 6.7 (95% CI, 5.0-15.8) No floor/31.1% ceiling effect	- AUC 0.66 Change correlation with Constant -0.60	

5.3.1.7.1. PROMs measurement properties in non-surgical humerus fracture (NSu-F)

Three studies investigated the measurement properties of the DASH and Constant PROMs for the assessment of function in patients with fractures of the shoulder (van de Water et al., 2014; van de Water et al., 2016a; Fayad et al., 2008b). Among them, two compared the DASH and Constant outcome measures (van de Water et al., 2014; van de Water et al., 2016a) .

The reliability of these outcome measures were found to be within the specified requirement for the Constant (ICC = 0.91), but slightly under it for the DASH (ICC = 0.87) (van de Water et al., 2014).

Concerning interpretability aspects, the SEM, MCID and LoAs were reported for both the Constant and DASH outcome measures, but not their MDC and PASS measurement characteristics (van de Water et al., 2014). The SEM was lower for the Constant (4.5 points) than for the DASH (6.5 points). For both outcome measures the MCID magnitude was quite different if the anchor-based (Constant 11.4 points; DASH -13.0 points) or the distribution-based method (Constant 5.1 points; DASH -8.1 points) was used. The LoAs were not numerically reported but could be estimated from graphical inspection. The LoAs (~ 10%) were at the limits defined as acceptable in this review for the Constant and larger for the DASH (~15%). A ~ 5% test-retest bias was visible for the two PROM-derived assessment tools, which it at the limit defined as acceptable in this review.

Two studies investigated the responsiveness of the Constant and the DASH. Both were correlated to the change score of the SSV, as well as to each other's change score, with a similar strength (r = 0.68 – 0.72) (van de Water et al., 2014). The effect size of the Constant (ES = 0.31) was somewhat lower than that of the DASH (ES = 0.44) (van de Water et al., 2016a). One research study reported a considerably higher ES (ES = 1.2), at a stage of recovery when progress is expected to be more marked, but did not compare it to that of the Constant (Fayad et al., 2008b).

5.3.1.7.1. PROMs measurement properties in surgical humerus fracture (Su-F)

No PROM has been evaluated in this population.

5.3.1.7.2. PROMs measurement properties in mixed surgical/non-surgical humerus fracture (Mi-F)

Two studies investigated the measurement properties of PROMs in patients with fractures, among which, one had compared the DASH's and Constant's measurement properties.

5.3.1.7.2.1. Constant and DASH (Mi-F)

As most of the results originate from one study that investigated the measurement properties of the Constant in this population, the measurement properties of the two Score are presented together in the same sub-section to avoid repetitions (Mahabier et al., 2017). The reliability of the outcome measures has not been evaluated in this study, but has been evaluated for the DASH only in another study (ICC = 0.93). The SEM (Constant 6.4 points; DASH 6.9 points), MDC (Constant 17.7 points; DASH 19.0 points) and MCID (Constant 6.1 points; DASH 6.7 points), but not the LoAs and PASS have been evaluated among the interpretability aspects. However, the MCID value found in this study cannot be considered as a valid threshold for the determination of the change that matters to the patient, because it was smaller than the MDC (van der Linde et al., 2017; De Vet et al., 2011a).

For the Constant, the specificity (58%) and the sensitivity (61%) were only fair, so that the area under the curve (AUC = 0.59) was lower than the defined threshold (≥ 0.70). For the DASH, the sensitivity (45%) was lower than the specificity (81%), so that the area under the curve (AUC = 0.68) was slightly lower than the defined threshold. However, the criterion was the discrimination between patients who scored "a little better" and patients who did not change, which was a more stringent criterion that the most frequently used "improved/unimproved" discrimination. The Constant and DASH SRM (Constant SRM = 1.60; DASH SRM = -1.63) and ES (Constant ES = 1.71; DASH

SRM = -1.55) were evaluated six and 12 months after the injury. The change correlation between the Constant and DASH was r = -0.60 (Mahabier et al., 2017). No floor and ceiling effects were detected for the Constant, which contrasted to the DASH, for which 31.1% ceiling effect was detected 12 months after the fracture.

5.3.2. Movement analysis-based outcome measures results

5.3.2.1. Presentation of movement analysis-based outcome measures

Presentation of movement analysis-based outcome measures				
Article	**Pathology**	**Measurement method**	**Convergent validity**	**NOTES**
Duc 2014	Rotator cuff non-surgical	Three scores : Duration of muscular activation (T_{EMG}), duration of arm movement (T_{mov}) and $T_{emg/mov}$ (relative electromyography (EMG) time over movement time) measured with inertial sensor system and EMG	Absolute correlation between the DASH, SST and Constant and the T_{mov}, upper trapezius T_{EMG}, and $T_{emg/mov}$ ranged 0.45 - 0.79 in laboratory setting and non-significant in other cases. Best correlations with the DASH, SST and Constant in daily condition was found for the $T_{emg/mov}$ of upper trapezius (0.56 - 0.62); lower and mostly non-significant in other cases	More scores have been explored in this study: only those which showed a significant difference between patient and control groups are reported here
Jolles 2011	Mixed sample of patients with rotator cuff conditions and osteoarthritis, surgically treated	Three scores: Between-sides balance for power-related metric (Power score), range of angular velocity (RAV Score) and moment (Moment Score) measured with an inertial sensor system during a series of seven movements at a self-selected speed	Absolute correlation with DASH, SST, ASES and Constant Power score 0.69 - 0.80 RAV Score 0.67 - 0.76 Moment Score 0.61 - 0.70	Power Score (equivalent to P Score) is the parent score of the B-B Score

Presentation of movement analysis-based outcome measures (continued)				
Article	**Pathology**	**Measurement method**	**Convergent validity**	**NOTES**
Korver 2014a Korver2014b	Rotator cuff	Two Scores : ARS: angular rate signal (equivalent to above mentioned RAV score) COMP: combination of angular rate signal and acceleration signal (equivalent to above mentioned P Score) Measured with inertial sensor system during two selected movements (hand to the back and to the ceiling) at self-selected speed	Weakly correlated with functional score DASH and SST < 0.25 ⇨ captures a different aspect of shoulder function than PROMs	Completion time < 5min Two movements : 'arm to the back' and 'arm behind the head' Mean of 3 repetitions No between-sides asymmetry related to hand dominance
Pichonnaz 2015a *	Rotator cuff Capsulitis Humerus fracture Shoulder instability (all non-surgical)	Smartphone B-B Score : between side balance of for power-related metric measured with a smartphone inertial sensor system during 2 selected movements (hand to the back and to the ceiling) at self-selected speed	Absolute correlations with Constant, relative Constant, SST, QuickDASH and WOSI: Rotator cuff: 0.55–0.84 Humerus fracture: 0.66 – 0.70, no correlation with QuickDASH Capsulitis: 0.64 – 0.76 Instability: 0.46 – 0.58	Completion time 2 - 3 minutes Mean of 3 repetitions of the two movements used for the score calculation
Pichonnaz 2015b **	Rotator cuff surgical	Pathological arm underuse percentage in everyday life environment compared to population with the same hand dominance, measured with an inertial sensor system	Non-significant correlation with clinical scores except with Constant 3 months 0.46	

* Publication based on the MSc dissertation of the thesis's author, in which the B-B Score conception was developed
** Although the author of the thesis is the author of this publication, it was related to a specific project that was not part of the PhD

Presentation of movement analysis-based outcome measures (continued)				
Article	**Pathology**	**Measurement method**	**Convergent validity**	**NOTES**
Pichonnaz 2015c	Mixed sample of patients with rotator cuff conditions and osteoarthritis, surgically treated	Study on the conception of the B-B Score (see Pichonnaz 2015a for description of the B-B Score	Absolute correlations with DASH, SST and Constant Score ranged from 0.51 to 0.77	
Pichonnaz 2017	Diversified, non-surgical	B-B Score (please see above) Inertial sensor system B-B Score (please see above)	Smartphone and inertial sensor system equivalent	Study demonstrated the equivalency of smartphone and inertial sensor system for B-B Score measurement
Yang 2014	Capsulitis	Shoulder physical activity (SPA): accelerometer net vector magnitude data counts; higher counts represent more complex strategies caused by pain and discomfort	Correlation with Flexilevel scale of shoulder function 0.47	

5.3.2.2. Measurement properties of movement analysis-based outcome measures

Measurement properties of movement analysis-based outcome measures					
Score	**Pathology**	**Reliability and measurement error**	**Interpretability aspects**	**Responsiveness**	**Normal performance**
Power Score RAV Score Moment Score (Jolles 2011)	Mixed sample of patients with rotator cuff conditions and osteoarthritis, surgically treated			Effect size between patients with abnormal and normal pain Power Score -1.91 RAV Score -1.90 Moment Score -1.72 (vs. 1.01 DASH; -0.96 ASES; -1.09 Constant; -1.13 SST)	Power score mean (SD) 91% (7%) RAV score 92% (5%) Moment score 84% (10%)
Tmov Temg upper trapezius Temg/mov (Duc 2014)	Rotator cuff non-surgical	Test re-test: T_{mov}: ICC 0.74 T_{emg} upper trapezius: ICC 0.83 $T_{emg/mov}$ upper trapezius: ICC 0.81			Please see original publication for details of each score
ARS COMP (Korver 2014; Korver2014b)	Rotator cuff, non-surgical	Intra-rater ARS: ICC 0.94 COMP: ICC 0.95 Inter-tester ARS: ICC 0.90 COMP: ICC 0.91 Inter-tester DASH ICC 0.63 Inter-tester SST ICC 0.70		Discriminative power patient/healthy: Specificity: ARS asymmetry 81.0% COMP asymmetry 85.0% Sensitivity: ARS asymmetry 98.0% COMP asymmetry 84.0% Floor effect: no	Asymmetry between shoulder 14.6% for COMP and 9.6 for ARS Healthy/pathological cut-off 27% difference between sides for COMP and 16% for ARS
Vector magnitude data counts (Yang 2014)	Capsulitis			AUC 0.83	

Measurement properties of movement analysis-based outcome measures (continued)					
Score	**Pathology**	**Reliability and measurement error**	**Interpretability aspects**	**Responsiveness**	**Normal performance**
Arm underuse (Pichonnaz 2015b)	Rotator cuff surgical			Correlation change with Constant 0.47, DASH, 0.49, SST no correlation (3-6 months, NS at other stages)	Mean (SD) of dominance arm use: Right handed: 61.2% (6.6%) Left handed: 54.3% (6.7%)
B-B Score (Pichonnaz 2015a; Pichonnaz 2015c; Pichonnaz 2017)	Mixed sample of patients with various shoulder conditions non-surgically treated (n = 65)	Intra-rater ICC Smartphone 0.92 Reference System 0.92 Inter-rater ICC Smartphone 0.92 Reference System 0.93 Inter-devices ICC: 0.97 SEM: Intra-rater: Smartphone 6.6%; Inertial sensor system 6.6% Inter-rater: Smartphone 6.6%; Inertial sensor system 6.4% LoA (bias): Intra-rater: Smartphone -17.4 - 20.3% (1.5%) Inertial sensor system -19.3 - 19.6% (0.1%) Inter-rater: Smartphone - 16.9% - 20.0% (1.5%); Reference System - 18.1 - 20.0% (1.0%) SEM 6.4% – 6.6% MDC 18.1%	MCII 25.2% PASS 77.6%	AUC (patients-controls): all patients 0.88; indicated pathologies 0.96 AUC (improved/unimproved): all patients 0.73; indicated pathologies 0.70 ES/SRM for all patients: 0.90/0.90 ES/SRM for indicated pathologies: 0.81/1.18 Change correlation with PROMs for all patients: 0.55 – 0.71 Change correlation with PROMs for indicated pathologies: 0.47 – 0.69 No floor/ceiling effect	Healthy/pathological cut-off: > 82.1

Measurement properties of movement analysis-based outcome measures (continued)					
Score	**Pathology**	**Reliability and measurement error**	**Interpretability aspects**	**Responsiveness**	**Normal performance**
B-B Score (continued)	Healthy (n=20)				Healthy mean Score 95% (mean of baseline and 6 months measurements)
	Rotator cuff (n=20)		No floor/ceiling effect	AUC (patients-controls): 0.90 Healthy/pathological cut-off: > 83.6 ES/SRM: 0.69/1.98 (> Constant, QuickDASH, SST) Change correlations with PROMs: no – 0.55	
	Humerus fracture (n=23)		No floor/ceiling effect	AUC (patients-controls): 0.98 Healthy/pathological cut-off: > 71.6 ES/SRM: 1.94/1.98 (< Constant; > QuickDASH, SST) Change correlation with PROMs: 0.56 – 0.75	
	Capsulitis (n=22)		No floor/ceiling effect	AUC (patients-controls): 0.99 Healthy/pathological cut-off: > 82.1 ES/SRM:1.16/1.68 (> Constant, QuickDASH, SST) Change correlation with PROMs: no – 0.47	
	Shoulder instability (n=23)		No floor/ 22% ceiling effect	AUC (patients-controls): 0.67 Healthy/pathological cut-off: > 81.6 ES/SRM: 0.10/0.13 (< WOSI and Constant; > QuickDASH and SST) Change correlation with PROMs: no – 0.50	

5.3.2.2.1. Normal performance for MAB outcome measures

Normative values have been defined systematically for all selected MAB outcome measures, except for the shoulder physical activity determined by net vector magnitude data count (Yang et al., 2014). However, the samples were small in all studies (≤ 100 participants), which prevented any sample' stratification. The influence of age, sex amongst factors that could potentially influence the outcome is thus presently unknown.

Normative values and healthy-pathological cut-off values were determined for the ARS and COMP scores (Korver et al., 2014a), arm underuse score (Pichonnaz et al., 2015b), EMG muscular activity duration (Duc et al., 2014), Power Score, RAV Score, Moment Score (Jolles et al., 2011) and B-B Score for inertial sensor system and smartphone measurement (Pichonnaz et al., 2015a)

5.3.2.2.2. MAB outcome measures in diversified shoulder conditions mixed surgical/non-surgical (NSu- and Su-DSC)

Three studies on kinematic shoulder function outcome measures relied on a patients of patient with diversified pathologies (Jolles et al., 2011; Pichonnaz et al., 2017; Pichonnaz et al., 2015a). One study included patients operated on for rotator cuff repair or total shoulder arthroplasty (Jolles et al., 2011) and two other ones non-surgically treated patients with rotator cuff, adhesive capsulitis, humerus fracture or shoulder instability (Pichonnaz et al., 2015a; Pichonnaz et al., 2017).

5.3.2.2.2.1.P, RAV and Moment Scores (Su-DSC)

Jolles et al. investigated the properties of three MAB outcome measures (Power score, RAV score and Moment score) that displayed a close similarity amongst measurement properties (Jolles et al., 2011). The criterion-based validity indicated that the MAB outcome measures were actually indicative of shoulder function for the P, RAV and M Scores (r = 0.61 – 0.80 with the DASH, SST, ASES and Constant

PROMs). The effect sizes for the difference between the patients with abnormal and normal pain at follow-up was in favour of the three MAB outcome measures (absolute ES = 1.72 – 1.91), compared to the DASH (ES = 1.01), SST (ES = - 1.13), Constant (ES = 1.09) and ASES (ES = - 0.96) shoulder function PROMs (absolute ES 0.96 – 1.13). The authors also highlighted that the MAB outcome measures were able to detect treatment failures at an earlier stage than PROMs. It can be considered that the three investigated MAB outcome measures reflected shoulder function as they were moderately to highly related to PROMs pursuing the same purpose (r = 0.61 – 0.80).

5.3.2.2.2.2.B-B Score (NSu-DSC)

Previous research has shown that the B-B Score had convergent validity with the DASH, SST and Constant (r = 0.51 – 0.77) in a sample with diversified pathologies (Pichonnaz et al., 2015c).

The thesis Phase 2 and 3 studies provided material for the comparison of the measurement properties of outcomes measures investigated in a sample including diversified pathologies. This subsequent literature review was therefore an opportunity to conduct a comparative analysis of the measurement properties of the B-B score. In the Phase 2 study that compared the measurement properties of the MAB shoulder function B-B Score, measured with an inertial sensor system or a smartphone, the reliability was above the required cut-off (ICC ≥ 0.90), with intra-rater ICCs reaching 0.92 and inter-tester ICCs 0.92-0.93, regardless of device (Pichonnaz et al., 2017). The SEM ranged from 6.4% – 6.6% for intra- and inter-tester measurements regardless of device. The intra- and inter-tester agreements were comparable, with LoAs ranging from ± 18.8% to ± 19.5%, i.e. higher than the ± 10% threshold. Separate analyses were conducted for each pathological subgroup in the Phase 3 study, but some statistics were also calculated for the whole patient group including all pathologies (“All patients” group) or for the whole patient group excluding patients with shoulder instabilities, for whom the B-B Score is known to be inadequate (“Indicated pathologies group”) (see sub-section 4.2.2 Analysis, within Chapter four, p. 133 - 134). It was determined that the AUCs were adequate for the discrimination between controls and patients (AUCs = 0.88 “All patients” and 0.96 “Indicated pathologies”) (see Table 4.4 in sub-section 4.3.2 “Discriminative power”, within

Chapter four, p. 139). The ES (ES = 0.81 “All patients” and 1.21 “Indicated pathologies”) and SRM (SRM = 0.90 “All patients” and 1.26 “Indicated pathologies”) were comparable to those of the Constant and relative Constant and superior to those of the SST and QuickDASH (see sub-section 4.3.4 “Responsiveness”, within Chapter 4, Table 4.6 p. 142 and 4.7 p. 143). The change score correlations with the Constant, relative Constant, and SST were adequate (r = 0.65 – 0.71), but below the required threshold for the QUICKDASH in the “Indicated pathologies” subgroup only (“All patients r = 0.55; “Indicated pathologies r = 0.45) (same sub-section, Table 4.8 p. 144). The AUCs for the discrimination between improved and unimproved patients were adequate (AUC = 0.73 “All patients” and 0.70 “Indicated pathologies”), but lower than those of the Constant, relative Constant, DASH, SST, QuickDASH (AUC = 0.78 – 0.83 “All patients” and 0.73 – 0.83 “Indicated pathologies”) (same sub-section, Table 4.9 p. 145). The interpretability aspects of the B-B Score were 18.1% for the MDC, 25.2% for the MCII and 77.6% for the PASS). No issues related to floor and ceiling effects was detected (sub-section 4.4.2 “Interpretability aspects”, within Chapter four, p. 147) (Pichonnaz et al., 2015a).

Levels of evidence of MAB outcome measures measurement properties in samples including diversified conditions are summarised in Table 5.11.

Table 5.11: Summary table for the level of evidence of MAB outcome measures measurement properties in samples including diversified conditions *

	Reliability (ICC and LoAs)	Responsiveness (AUC, ES, SRM, change correlation, floor/ceiling effect)	Interpretability aspects (SEM, MDC, MCII/MCID, PASS)	Correlation to PROMs
B-B Score (NSu)	+ low	+ low	+ low	+ low
P Score (Su)	?	+ low	?	+ low
RAV Score (Su)	?	+ low	?	+ low
Moment Score (Su)	?	+ low	?	+ low
Comparisons	-	ES of P, RAV and M superior to Constant, DASH, ASES and SST ES and SRM of the B-B Score comparable to Constant and relative Constant, and superior to QuickDASH and SST	-	-

Legend: P Score: Power Score; RAV Score Range of Angular Velocity Score; M Score Moment Score; ICC: intraclass correlation coefficient; Disabilities of the Arm, Shoulder and Hand score SST: Simple Shoulder Test; ASES: American Shoulder and Elbow Surgeons score; LoA: Limits of Agreement; AUC: Area Under the receiver operating Curve; ES: Effect Size; SRM: Standardised Response Mean; SEM: Standard Error of Measurement; MDC: Minimal Detectable Change; MCII/MCID Minimal Clinically Important Improvement/Difference; PASS: Patient Acceptable Symptom State.

+ sufficient, +/- undetermined, - insufficient, ? non-investigated measurement properties

* Evidence graded using the modified GRADE approach for grading the quality of evidence for measurement properties

Su: surgical treatment; NSu: non-surgical treatment

5.3.2.2.3. MAB outcome measures measurement properties in surgical and non-surgical rotator cuff conditions (NSu- and Su-RCC)

Four studies investigated the measurement properties of MAB outcome measures for rotator cuff conditions, among which three had focused on non-surgical treatment (Korver et al., 2014a; Pichonnaz et al., 2015a; Duc et al., 2014), and one on surgical treatment (Pichonnaz et al., 2015b).

5.3.2.2.3.1. ARS and COMP Scores (NSu-RCC)

Korver investigated two scores, the ARS (peak-to-peak difference in the angular rate signal for the three axes) and the COMP (area described by combining the angular rate signal and acceleration signal), measured during two basic shoulder movements (arm to the back and arm behind the head). Both scores had adequate intra- and inter-rater reliability (ICC ≥ 0.90) (ARS ICC = 0.94 and 0.90; COMP ICC = 0.95 and 0.91, respectively), which was better than the DASH's (ICC = 0.63) and SST's (ICC = 0.70) reliability reported in this study. No interpretability aspect was reported for these scores. The specificity (ARS 81%; COMP 84%) and sensitivity (ARS 98%; COMP 84%) were high for the discrimination between patients and healthy controls. However, such results were indicative of discriminative power between patients and controls rather than responsiveness of the outcome measure, as the specificity and sensitivity did not address the score's improved/unimproved discrimination power. The negligible correlation ($r < 0.25$) (Hinkle 2003) with the DASH and SST indicated that the ARS and COMP scores had limited convergent validity, as they did not capture the same dimension of shoulder function as these PROMs. Interpretability aspects and responsiveness were not reported for these scores, except for no floor effects.

5.3.2.2.3.2. T_{mov}, T_{mov} and $T_{EMG/mov}$ Scores (NSu-RCC)

A study investigated shoulder muscle activation during active movements, using a combination of inertial sensors and EMG (Duc et al., 2014). Several alternative scores were tested among which the duration of arm movement (T_{mov}), duration of the upper

trapezius activation ($T_{EMG]}$) and the upper trapezius percentage of muscular activation time over movement time ($T_{EMG/mov[\%]}$) were able to show a significant difference in performance between healthy controls and patients. The reliability of these three scores (ICC = 0.74-0.81) was lower than the threshold defined in this review (ICC ≥ 0.90). The absolute correlations with the DASH, SST and Constant PROMs ranged from $r = 0.46 – 0.79$ in laboratory settings and from ($r = 0.56 – 0.62$) for the $T_{EMG/mov[\%}$ for measurements undertaken in everyday conditions. The responsiveness and interpretability aspects were not determined.

5.3.2.2.3.3. Arm underuse (NSu-RCC)

One study investigated some measurement properties of a score that quantified arm underuse following rotator cuff surgery (Pichonnaz et al., 2015b). This score significantly differentiated the patients from the healthy controls three months after surgery and showed the recovery pattern of arm usage over time. However, no correlation was found with the DASH and the SST at several post-surgical stages and a significant correlation with the Constant was found only three month after surgery ($r = 0.49$), indicating that the arm underuse score did not capture the same dimension of shoulder function as these PROMs. Correlations between change scores were found only between three and six months for the DASH ($r = 0.49$) and the Constant ($r = 0.47$).

The change score correlation was the only measurement property that was investigated in this study. It showed no correlation between change scores for the SST and low correlations for the Constant ($r = 0.47$) and the DASH ($r = 0.49$).

5.3.2.2.3.4. B-B Score (NSu-RCC)

The Phase 3 study and its related article allowed for the investigation of the B-B Score measurement properties in a non-surgical sample (Pichonnaz et al., 2015a). The moderate to high correlations with the Constant ($r = 0.82$), relative Constant ($r = 0.84$), SST ($r = 0.63$) and QuickDASH ($r = -0.55$) indicated that the B-B Score measures a dimension close to the shoulder function PROMs (see sub-section 4.3.3 “Convergent validity”, Table 4.5, within Chapter four p. 141). The reliability and interpretability aspects were not specifically determined for rotator cuff conditions (Pichonnaz et al., 2017). The ES/SRM six months after baseline measurement were SRM = 0.69/ ES =

0.69, which was superior to the Constant (SRM = 0.54/0.58), relative Constant (SRM = 0.50/ ES = 0.57), SST (SRM =0.52/ ES = 0.48) and QuickDASH (SRM = 0.35/ ES = 0.47) (see sub-section 4.3.4 “Responsiveness”, Table 4.6 and 4.7, within Chapter 4 p. 142 - 143). The AUC (AUC = 0.90) was excellent for the discrimination between patients and healthy controls, with an affected-non affected cut-off at 83.6. The specificity (90%) and sensitivity (90%) were high for rotator cuff conditions. Conversely, the AUC for the improved/unimproved discrimination was not determined specifically for rotator cuff conditions (see same sub-section, Table 4.9, p. 145). The change correlation was moderate with the Constant (r = 0.51) and relative Constant (r = 0.55) but non-significant with the SST and QuickDASH (see same sub-section, Table 4.8, p. 144).

Levels of evidence for measurement properties of outcome measures in samples including patients with non-surgical rotator cuff conditions are summarised in Table 5.12.

Table 5.12: Summary table for the level of evidence of MAB outcome measures measurement properties in samples including non-surgical rotator cuff conditions *

	Reliability (ICC and LoAs)	**Responsiveness (AUC, ES, SRM, change correlation, floor/ceiling effect)**	**Interpretability aspects (SEM, MDC, MCII/MCID, PASS)**	**Correlation to PROMs**
B-B Score	?	+ low	?	+ low
RAV and COMP Scores	+ low	+ low	?	- low
T_{mov}, T_{EMG} and $T_{EMG/mov}$ Scores	?	?	?	- low
Arm underuse	?	+ low	?	- low
Comparisons	-	-	-	-

5.3.2.2.1. Measurement properties of MAB outcome measures in osteoarthritis (OA)

No MAB outcome measure has been validated specifically for OA. Thus, no table is provided to summarize levels of evidence for measurement properties in this subpopulation.

5.3.2.2.1. Measurement properties of MAB outcome measures in non-surgical shoulder instability (NSu-SI) (OA)

One study evaluated the measurement properties of a MAB outcome measure (B-B Score) for non-surgical shoulder instability evaluation (Pichonnaz et al., 2015a). The specific interpretability aspects were not reported in this study.

The B-B Score was correlated with the WOSI (r = 0.58), the QuickDASH (r = -0.57), the SST (r = 0.52), the Constant (r = 0.46) and the relative Constant (r = 0.43). This indicates adequate convergent validity for the evaluation of shoulder function (see sub-section 4.3.4 “Convergent validity”, within Chapter 4, Table 4.5 p. 141 - 142).

The ES and SRM for a change of baseline until three months was small (ES = 0.10; SRM = 0.13). When directly compared to PROMs, these value were lower than the ES and SRM of the WOSI (ES = 0.47; SRM = 0.41) and, to a lesser extent, lower than that of the relative Constant (ES = 0.27; SRM = 0.22) and Constant (ES = 0.21; SRM = 0.19), but equivalent to the ES of the SST (ES = 0.10; SRM = 0.08) and superior to that of the QuickDASH (ES = 0.01; SRM = 0.01). The AUC for the discrimination of patients with shoulder instability from healthy controls was under the 0.70 threshold (AUC = 0.67). The specificity was excellent (98%) but the sensitivity was low (48%). No floor and ceiling effects were detected (Pichonnaz et al., 2015a) (see sub-section 4.3.4 “Responsiveness”, within Chapter 4, Table 4.6 p. 142, Table 4.7 p. 143 and 4.8 p. 144).

Table 5.13: Summary table for the level of evidence for measurement properties of MAB outcome measures in samples including patients with shoulder instability non-surgically treated *

	Reliability (ICC and LoAs)	**Responsiveness (AUC, ES, SRM, change correlation, floor/ceiling effect)**	**Interpretability aspects (SEM, MDC, MCII/MCID, PASS)**
B-B Score	?	- low	-
Comparisons	-	WOSI superior to generic shoulder PROMs Constant, DASH, ASES: high WOSI and to a lower extent Constant superior to B-B Score: low	-

Legend: DASH Disabilities of the Arm, Shoulder and Hand score ASES: American Shoulder and Elbow Surgeons score; WOSI: Western Ontario Shoulder Disability Index; ICC: intraclass correlation coefficient; LoA: Limits of Agreement; AUC: Area Under the receiver operating Curve; ES: Effect Size; SRM: Standardised Response Mean; SEM: Standard Error of Measurement; MDC: Minimal Detectable Change; MCII/MCID Minimal Clinically Important Improvement/Difference; PASS: Patient Acceptable Symptom State.
+ sufficient, +/- undetermined, - insufficient, ? non-investigated measurement properties
* Evidence graded using the modified GRADE approach for grading the quality of evidence of measurement properties (Prinsen 2018)

5.3.2.2.1. Measurement properties of MAB outcome measures in capsulitis (NSu-C)

One study evaluated the measurement of a MAB outcome measure (kinematic B-B Score) for the non-surgical treatment of a capsulitis (Pichonnaz et al., 2015a) (results sub-section 5.3.2.2 “Measurement properties of movement analysis-based outcome measures”, p. 239). The correlation strength was moderate to high with the SST (r = 0.76), relative Constant (r = 0.69), Constant (r = 0.68) and QuickDASH (r = -0.64) (Hinkle et al., 2003), indicating adequate convergent validity for shoulder function evaluation. No specific reliability and interpretability aspects for capsulitis were determined in this study (see sub-section 4.3.4 “Convergent validity”, within Chapter 4, Table 4.5 p. 141).

Concerning responsiveness preceding investigations conducted within this thesis (Phase 3 study) showed that ES and SRM were ES = 1.16 and SRM = 1.68 for the baseline until three months change. When directly compared to PROMs, the B-B Score ES and SRM were higher than those of the SST were (ES = 0.86; SRM = 1.24) and QuickDASH (ES = 0.55; SRM = 1.07), which were used for comparison. Concerning the Constant, the ES and SRM provided divergent results concerning the superiority or the inferiority of one outcome measure over the other, with higher ES for the B-B Score and higher SRM for the Constant (ES = 1.05; SRM = 1.98) and relative Constant (ES = 1.04; SRM = 2.02). The AUC was calculated for the discrimination between patients with or without a capsulitis. Its value (AUC = 0.99) was excellent, largely above the 0.70 threshold, with an affected-non-affected side cut-off value at 82.1%. The specificity was nominally perfect (100%) and the sensitivity was excellent (95%). No floor and ceiling effects were detected (see sub-section 4.3.4 “Responsiveness”, within Chapter 4, Table 4.6 p. 142, Table 4.7 p. 143 and 4.8 p. 144). Due to the scarce literature, no table is provided to summarize the levels of evidence for measurement properties in this subpopulation. It can only be stated that the B-B Score displays adequate responsiveness, with a low level of evidence, and that its comparison with PROMs shows lower responsiveness than the Constant and better responsiveness than the SST and QuickDASH, with a low level of evidence.

5.3.2.2.1. Measurement properties of MAB outcome measures in mixed surgical/non-surgical humerus fractures (Mi-F)

The Phase 3 study was the only measurement properties study on MAB outcome measures following proximal humeral fracture. In this study, the measurement properties of the B-B Score have been partially investigated, in a mixed sample of surgically and conservatively treated patients. The reliability and interpretability aspects of this outcome measure have not been reported specifically in this population. Concerning the convergent validity with PROMs, the B-B Score correlation was moderate with the Constant (r = 0.70), relative Constant (r = 0.69), SST (r = 0.66) and low with the QuickDASH (r = - 0.40).

Concerning responsiveness, the ES (ES = 1.94) and SRM (SRM = 1.98) were slightly inferior to that of the Constant (Constant ES = 2.09 and SRM = 2.02; relative Constant

ES = 2.10 and SRM = 2.09) but superior to that of the SST (ES =1.65; SRM = 1.70) and the QuickDASH (ES = 1.25; SRM = 1.07). The discriminative power between affected and non-affected participants was excellent (AUC = 0.98), with an affected-non-affected cut-off at 71.6%, due to high specificity (96%) and perfect sensitivity (100%) No floor effect, defined as 0 + MDC, was found. The absolute change correlations were moderate to high ($r = 0.61 – 0.78$). No ceiling effect was detected, as no patient reached 100% in this subgroup (see sub-section 4.3.4 “Responsiveness”, within Chapter 4, Table 4.6 p. 142, Table 4.7 p. 143 and 4.8 p. 144).

5.4. Discussion

5.4.1. Overview of the literature review process

This review collated and compared the measurement properties of currently used patient-reported and MAB outcome measures of function in frequent shoulder pathologies. It aimed therefore at determining if an approach has advantages over the other one, considering their respective measurement properties.

More specifically to this thesis, the literature review aimed at challenging the measurement properties of the B-B Score reported in Phase 2 and Phase 3 of the thesis with the measurement properties of alternative outcome measures, considering both PROMs and MAB outcome measures. This comparison of the B-B Score clinimetric performances from Phase 2 and 3 with those of other outcome measures pursuing the same purpose has aimed at laying the foundation for circumstantiated recommendations on its use in various in various clinical contexts.

The investigation of the specific measurement properties of several PROMs and MAB outcome measures for several treatment approaches in several pathologies was necessary to avoid the inappropriate aggregation of data that were produced in obviously different testing conditions. Although this detailed approach may *in fine* increase the specificity of recommendations and allow a thorough and fair comparison of the Phase 2 and 3 results of the B-B Score with alternative outcome measures, it implied that the results should be reported and hereafter discussed with a considerable level of details.

The results were reported separately for each one of the selected common shoulder pathologies (rotator cuff condition, capsulitis, osteoarthritis, proximal humerus fracture and shoulder instability) and for studies including samples with diversified shoulder pathologies. The results of each one of these groups were reported separately for surgical samples, non-surgical samples and mixed surgical/non surgical samples according to the treatment approach applied to the patients included in the study. This detailed reporting was required to account for the context-dependency of the measurement properties (Robertson et al., 2017; Riddle and Stratford, 2013; Collins and Roos, 2016; El Gaafary, 2016). It was implemented with the purpose of providing the foundations for targeted recommendations concerning the choice of outcome measures for the types of patients' scenarios commonly encountered in physiotherapy practice.

For feasibility reasons, the measurement properties were investigated only for the most commonly used PROMs, based on the published literature, while all MAB outcome measures were considered. The double-check that was conducted at all stages showed that the initial bibliographic search was near from exhaustive, considering the investigated databases. Only four additional articles were retrieved following this checking process.

5.4.2. Score selection

It was crucial to limit the number of PROMs to ensure the feasibility of the review and its adequacy in reflecting actual clinical practice in shoulder function measurement. The number of investigated outcome measures was thus limited to five frequently used and considered as valid PROMs (Makhni et al., 2015; Gartsman et al., 2015).

The PROMs' selection based on the frequency observed in our bibliographic search was in line with other reviews that had investigated the use of PROMs in the literature. Gartsman and al. (2015) found a close ranking for the frequency of use of PROMs for the articles published from 2004 to 2014 in The Journal of Bone & Joint Surgery, except for the DASH, which was found to be more used than the SST in the bibliographic investigations conducted for this review. The UCLA was frequently used according to Gartsman and al., but these authors had estimated that this score could not be considered as a validated. Makhni et al. also found a similar PROM-use' ranking specifically for rotator cuff evaluation in a review that encompassed six major

journals publishing articles on shoulder issues (Makhni et al., 2015). They reported that the Constant, ASES, UCLA, SST were the most commonly used, while the DASH only ranked 9th for rotator cuff conditions in this publication. The minor differences in selection between these reviews and the present one are explainable by the fact that the latter's scope was larger, as the frequency of use was considered for four pathologies in all Medline/Pubmed indexed journals.

The inclusion of a condition-specific tool for instability (WOSI) was necessary because it is recognised that generic tools perform lower than specific tools for this condition. This had been reported for the Constant (Conboy et al., 1996; Kemp et al., 2012; Dawson et al., 1999; Oh et al., 2009), UCLA (Romeo et al., 1996; Oh et al., 2009), ASES (Kemp et al., 2012; Romeo et al., 1996; Goldhahn et al., 2008; Oh et al., 2009), SST (Oh et al., 2009) and QuickDASH (Gaudelli et al., 2014). Though frequently used in shoulder instabilities, the ROWE was not retained because it had not previously undergone a full validation process (Rouleau et al., 2010; Fayad et al., 2004; Kirkley et al., 2003; Gartsman et al., 2015).

Several versions of the DASH, the Constant and the ASES were available. The DASH and the QuickDASH PROMs were both considered and were compared to help users to make an informed choice between these two very similar instruments. As the burden is lighter using the QuickDASH, equivalent measurement properties were considered as advantageous for this PROM (Kolber et al., 2013; Institute for Work & Health).

For the Constant Score, the relative Constant approach, which compares the patient's performance to a sex and age matched group, has been developed to overcome the gender dependency and the decline with increasing age that were observed using the original approach of the Constant Score (Constant, 1986; Yian et al., 2005; Katolik et al., 2005; Fialka et al., 2005). Both approaches were included in this review due to their frequent use and because both the absolute and the relative performance to a matched group are of interest. Age and sex dependency of the outcome have also been reported for the DASH and QuickDASH, but no relative score has been found for these PROMS (Aasheim and Finsen, 2014; Hunsaker et al., 2002).

Although that several versions of the ASES have been developed (Fayad et al., 2004; Angst et al., 2011), with different measurement properties (Beaton and Richards,

1998), the specific version that had been used was not always specified in the articles. This was an important limitation in the interpretation of the results and might explain controversial measurement properties for this score.

5.4.3. Overview of the retrieved literature

The number of retrieved articles was roughly comparable for PROMs (1800) and MAB outcome measures (1642). Conversely, a much lower number of articles (9 vs. 85) could be included for the latter category. Most of the research on movement analysis was focused on phenomena' analyses, and very little on the development and validation of interpretable outcome measures for clinicians. In addition, numerous articles stated solely a difference between a patient group and a healthy group for a given kinematic or kinetic parameter, but did not report a more extensive validation process. These articles were not retained within this review, as the mere statement of a difference between a patient and a healthy control groups is not sufficient to allow the monitoring of patients' progress.

Several factors limited the ability to compare the measurement properties more rigorously in this review. First, few researchers compared directly several outcome measures in the same population and context. The performances of outcome measures were not compared between articles in this review because it was hardly possible to determine if the variations in properties were caused by the measurement conditions or actually by the clinimetric performances of the outcome measures. More specifically, only three original studies comparing directly PROMs and MAB approaches were identified (Korver et al., 2014a; Jolles et al., 2011; Pichonnaz et al., 2015a), though such studies would be of great interest to contrast their measurement properties in strictly identical conditions.

In addition, the possibility to proceed to comparisons between studies was limited by the heterogeneity of the methodological approaches used to determine measurement properties. For example, eleven methods, each leading to a different result, have been listed to calculate the MCID in shoulder function (Beaton et al., 2011). Similarly, considering the responsiveness, the AUCs were calculated at different follow-up times and according to varying reference criteria (e.g. improvement, important improvement, satisfaction, perceived handicap). Likewise, heterogeneous reference instruments were used for calculation of correlations associated with changes in

performance over time. Conversely, the ES and SRM were commonly reported estimators of responsiveness throughout studies, though they are not recommended by the authors of the COSMIN checklist (Mokkink et al., 2010b). The ESs or SRMs of were nevertheless compared amongst outcome measures, provided that they had been calculated in the same study under the same conditions. Between-studies comparisons of ESs or SRMs were not made, due to the heterogeneity of testing conditions between studies. All these examples illustrate the lack of consensus that surrounds the methods for measurement properties' determination.

Additionally, the descriptions of the tested populations were frequently imprecise. This is an important limitation to the application of the research results to a population of interest, which had previously been reported by the EQUATOR Network for quality and transparency of health research (Yamato et al., 2016). A considerable number of articles investigated the measurement properties in a sample with various shoulder problems, of which respective proportions were not reported. These articles were nevertheless retained in this review because this approach is sustainable to run exploratory investigations at the initial stage of an outcome measure development. However, the use of a diversified sample limits the possibility of applying the results to a patient, who has by definition, a specific and not a generic shoulder condition.

Although 82 articles were included for PROMs, it appeared that some conditions, such as capsulitis, surgically treated fractures or conservatively treated osteoarthritis, had been scarcely investigated or not investigated at all. Considering these literature limitations, studies with large sample sizes, specific populations and comparative use of several tools by independent researchers would be required to improve the ability to compare outcome measures and to increase the precision of estimation of their clinimetric capabilities.

5.4.4. Interpretation of the results

5.4.4.1. Normal performance definition

The determination of a given aspect of performance capability in a healthy population is of importance for determining from which level of performance it should be considered that a shoulder condition has a functional impact, or for ascertaining whether a patient's shoulder completely recovered or not at the end of a treatment.

Norms have to be evaluated for each outcome measure, as it cannot be taken for granted that all healthy people are capable of reaching the maximum score, especially when accounting for the loss of physiological function due to aging. Additionally, the determination of a cut-off value for the discrimination between a normal a pathological performance is useful to account for the variably of the performance in healthy people..

It might seem relatively straightforward to determine an outcome measure normative value, but the literature review showed that the normal performance varies across subpopulations, according at least to age and sex. Thus, the norm can be subject of controversies, as it may vary with regard to the subpopulation in which it was determined (Yian et al., 2005). It would be recommended to derive normative values from large populations to allow stratified analyses, at least according to gender and age. As only the normative values of the Constant Score, DASH and QuickDASH were based on a stratified analysis among all the outcome measures investigated within the context of this review, the influence of the age and sex is unknown for the other outcome measures. The stratified analyses has led to the development of a version of the outcome measure accounting for age and sex for the Constant only (i.e. the relative Constant), but not for the DASH and the QuickDASH. The determination of cut-off values is useful to define the value of the outcome measure that differentiates a pathological from a healthy performance, accounting for the variability of normal performance within the healthy population.

The determination of a normal score has been frequently performed for MAB outcome measures at a developmental stage, but never in large populations. Thus, the reported values might be relatively imprecise and do not account for the influence of age and sex, due to the small size of the investigated samples.

The definition of the normal performance and of the healthy-pathological cut-off values of the B-B Score have been defined in the Phase 2 and 3 studies of this thesis based on a small sample size (n = 20) (Pichonnaz et al., 2015a). Though the side-to-side symmetry of the power-related metric measured in the Score was not likely to be age and sex-dependant, further investigations in larger samples are needed to test this hypothesis and to provide a precise estimation of the normal performance and of the healthy-pathological cut-off values.

5.4.4.2. Outcome measures for shoulder function in diversified condition populations

5.4.4.2.1. PROMs for non-surgical treatment of diversified condition populations (NSu-DCS)

The PROMs' measurement properties were frequently defined based on samples including various shoulder pathologies. As the validity of a measurement method is relevant only for the population in which it has been tested, this raises questions about the possibility of applying these results to a specific population that might be outside of the scope of the populations that were included in studies. As each pathology potentially affects shoulder function in a different way, it may be more appropriate to validate measurement properties in a sample with a single diagnosis, although this might complicate the validation process for practical reasons related to the recruitment of a precise target population.

Moreover, the possibilities to compare studies between them remains limited because the sample composition may differ from one study to the other, the main common point being merely the heterogeneity of shoulder conditions.

The formulation of recommendations associated with non-surgical treatments of shoulder disorders that could be based on the direct comparison between PROMs remains limited and indirect evidence can only be extrapolated cautiously from studies investigating diversified shoulder condition samples.

Concerning reliability, the Constant, SST and ASES had a mix of ICCs above and under the expected threshold, while the QuickDASH had ICCs above the ≥ 0.90 threshold. The only direct comparison for test-retest reliability favours the DASH over the ASES score (Moser et al., 2012), but the DASH's ICC remains nevertheless under the required ≥ 0.90 ICC value when single measurement reliability is considered (Portney and Watkins, 2015).

The complete set of interpretability aspects has not been determined for any PROM. However, only the PASS was missing for the QuickDASH. The only study comparing

the floor and ceiling effects of the SST and ASES was in favour of the ASES for which no patients obtained the maximum or the minimum score, though the percentage of patients reaching maximum/minimum scores was lower than the defined <15% threshold for both outcome measures (Robins et al., 2017).

In case the Constant Score is chosen for use in clinical practice, it should be considered that precise adherence to the recommended procedure for its use and previous training of the assessor are prerequisites to the tool's application under optimal conditions, as its measurement properties have been shown to be better under these conditions (Blonna et al., 2012). This PROM is probably more sensitive than the other ones to these aspects, as the completion of some items (e.g. strength and range of motion) requires clinical skills.

As the comparison of the DASH and the QuickDASH showed no disadvantage for the QuickDASH in non-surgical treatment, the latter one should be preferred in this context for its convenience. Although no firm recommendation can be formulated based on the literature, the use of this score is justifiable as it showed adequate ICC and AUC characteristics, and has had most of its interpretability aspects determined.

5.4.4.2.2. PROMS for surgical treatment of diversified condition populations (Su-DSC)

No information was found on the properties of the QuickDASH. The reliability was adequate for all tested PROMs. Concerning, the ASES performed better than the SST based on ES/SRM characteristics, and also better than the Constant but to a lesser extent. The Constant showed adequate and slightly superior AUC compared to the DASH. The DASH was the only one for which all interpretability aspects were available (SEM, MCID, MDC, PASS), which can facilitate the clinical interpretation of results.

Thus, while investigated properties had reached required thresholds, no clear recommendation can be formulated based on measurement properties in surgically-treated but diversified populations. The latter was due to missing information about

some PROMs and some conflicting results concerning the few direct comparisons that had been performed.

5.4.4.2.3. PROMS for mixed surgical/non-surgical treatment of diversified conditions populations (Mi-DSC)

The DASH, QuickDASH, SST and ASES, but not the Constant have been investigated in mixed samples of surgically or non-surgically treated patients.

When compared in the same study, the DASH, QuickDASH and SST had comparable test-retest ICCs ranging from 0.83 to 0.86, which is lower than the 0.90 required threshold (van Kampen et al., 2013). In contrast the DASH's test-retest reliability, was found to be adequate (ICC = 0.95) in a study in which it was not compared to other PROMs (Fayad et al., 2008a). For the ASES, one study found also insufficient test-retest reliability (ICC = 0.84), (Michener 2002) while another other ones found adequate reliability (Sallay and Reed, 2003; Celik et al., 2013). Bias due to real change that induced systematic variability' intrusion cannot be excluded as no study clearly reported that only stable patients were included in the test-retest evaluation.

The PASS had never been calculated, and the SEM had been calculated only for the SST (Roddey et al., 2000). The MDC and the MCID or MCII had been calculated for the DASH, QuickDASH, SST and ASES (van Kampen et al., 2013; Beaton et al., 2005; Michener et al., 2002). However, these reported MCID or MCII were of little interest as they were systematically lower than the MDC. This implies that when a clinician measures an improvement at the MCII/MCID level, he cannot be sure that the measured change is not due to measurement error (van der Linde et al., 2017). In these cases, the MDC should be considered as the threshold from which a clinically meaningful change happened.

Little can be said on the responsiveness, as no comparison was possible between studies using various similar methods for its evaluation. It could only be stated that the DASH was slightly more responsive than the QuickDASH at 3 months when both PROMs were compared in the same study (Beaton et al., 2005). It should thus be the first choice among them when the patient change is of concern.

Based on analysed results, no recommendation can be made about the choice of a PROM for the evaluation in a diversified sample including surgically and conservatively treated patients, beyond this specific point concerning the use of the DASH and QuickDASH.

5.4.4.2.4. MAB outcome measures for surgical and non-surgical diversified shoulder conditions (NSu- and Su-DSC)

To date, the research on the measurement properties of MAB outcome measure of shoulder function in diversified samples is scarce. Four outcome measures (P Score, RAV Score, M Score and B-B Score) were investigated in three studies, of which two are related to the Phase 2 and 3 of this thesis (Jolles et al., 2011; Pichonnaz et al., 2015a; Pichonnaz et al., 2017).

The reliability of the P Score, RAV Score and M Score was not reported in the literature, while the Phase 2 study had demonstrated adequate properties for intra- and inter-rater reliability.

Concerning responsiveness, the comparison of SRM and ES was globally in favour of the MAB outcome measures, except for the Constant and relative Constant that compared to the B-B Score. The AUC demonstrated adequate discriminative power between improved and unimproved patients, though they were lower than those of the PROMs to which the B-B Score was compared.

Based on Phase 2 results, the kinematic B-B Score's measurement properties were equivalent between an inertial sensor system and a smartphone, while this equivalency has not been demonstrated for the P, RAV and Moment Scores (Pichonnaz et al., 2017). This is advantageous for the latter that is cheaper and more accessible and is of interest for the accessibility of clinicians to the necessary technology. The LoAs were higher than the 10% threshold defined in this thesis, which

indicates that divergences between the outcome and the real performance are possible for single measurements.

Based on the few studies that have investigated the measurement properties of MAB outcome measures in diversified samples, this approach demonstrated convergent validity and adequate measurement properties. The B-B Score was the only MAB outcome measure that had undergone an extensive validation both previously within the literature and now having received further critical scrutiny with Phase 1, 2 and 3 studies within this thesis. However, the level of evidence about this score's measurement properties remains limited, as the results have not been replicated in other studies to date.

In the present state of the literature, it can be stated that the existing MAB outcome measures display adequate measurement properties, but the body of knowledge is insufficient to conclude that they might have superiority or inferiority compared to the analysed PROMs. Levels of evidence of MAB outcome measures measurement properties in samples including diversified conditions are summarised in Table 5.6.

5.4.4.2.5. Benchmarking of measurement properties of all outcome measures determined in diversified populations samples (NS-, S- and Mi-DSC)

The use of samples including patients with various pathologies is very frequent in validation studies. This is understandable due to the difficulty of recruiting samples of patients with precisely determined pathologies. The latter represents high administrative workloads to achieve the separate validation of an outcome measure in a variety of precisely defined shoulder conditions. However, this situation raises several concerns with regard to the reported results, all other things being equal.

There is first a conceptual problem about the definition of the target population. "Patients with shoulder conditions" cannot be considered as a homogenous population, as each separate pathology potentially impairs function in a different way and affects people having different characteristics, (e.g. when the pathology is related

to aging or lifestyle). For example, day-to-day variation and change over time will obviously influence test-retest reliability and responsiveness, respectively.

Moreover, the possibility of comparing studies including diversified pathologies in various proportions remains limited. In fact, the common point between the research samples is essentially that they are heterogeneous**.**

Thus, researchers including diversified samples within their studies should mainly have the goal of running exploratory studies to offer a first insight into measurement properties, unless they include large enough samples to conduct more targeted analyses of subgroups. Though this thesis' author is aware of this limitation, statistics including a mix of several pathologies were conducted to allow for the subsequent comparisons of the B-B Score measurement properties in this literature review.

No clear recommendation can be made for the choice of a shoulder function PROM in diversified populations involving non-surgically treated or surgically treated patients or a mixture of both. No PROM showed consistently superior reliability or responsiveness across studies. The DASH was the most extensively investigated outcome measure for surgically treated patients, and the QuickDASH for conservatively treated patients, which can facilitate the clinical interpretation of results in these populations when these outcome measures are used.

The information on MAB outcome measures was incomplete, which might restrict possibilities for the users to interpret the results of their measurements in some circumstances. However, the reliability of MAB outcome measures was consistently within required standards. MAB outcome measures' responsiveness was adequate considering AUC and change correlation values, and their ES was even superior to that of the PROMs for the Power, RAV and Moment Scores, as reported by Jolles et al. (2011). However, the data are not sufficient in scope yet to draw conclusions on the superiority or inferiority of their responsiveness compared to PROMs. The overall level of evidence remains low due to the lack of replication studies to date.

Despite research that is emerging from the literature and added to by the results of Phase 1, 2 and 3 studies in this thesis, it would seem that the MAB outcome measures represent a promising but still to be fully-investigated alternative to PROMs for shoulder function evaluation in samples, including those involving diversified

pathologies. Their superiority or inferiority over PROMs cannot presently be established due to the lack of data.

5.4.4.3. Outcome measures for shoulder function in rotator cuff conditions (RCC)

5.4.4.3.1. PROMs measurement properties in non-surgical rotator cuff (NSu-RCC)

The Constant showed adequate reliability. There is a controversy about the magnitude of the interpretability aspects, which is problematic for clinical interpretation of results, as the values considerably vary from one study to single-measurement another. It might be that the use of a fixed isometric dynamometer produces more stable values than a hand-held dynamometer but this should be confirmed by a dedicated study.

A ceiling effect was observed for the relative Constant only. This ceiling effect might be due to the fact that rotator cuff does not always induce an important functional loss and some undetected tears are also present in the general population (Sher et al., 1995). The distinction of the function of patients and supposedly healthy subjects may thus be difficult. This may explain why the Constant discriminative power is lower in rotator cuff conditions when the cuff is intact compared to when the cuff is torn (Holmgren et al., 2014).

The DASH score's reliability was below or above the threshold for acceptable reliability according to the findings of a specific study (ICC ≥ 0.90). Though reported, its MCID cannot be considered as valid as the MDC is considerably higher than the MCID. The DASH score's responsiveness was found to be adequate (AUC = 0.77) (De Vet et al., 2011c). No study has investigated yet if the QuickDASH could efficiently replace the DASH for the evaluation of conservative treatment of rotator cuff conditions.

As little information is available on the ASES clinimetric qualities, this PROM cannot be recommended for non-surgical rotator cuff evaluation. The SST showed adequate reliability and most of its interpretability aspects were investigated, but little is known about its responsiveness. Caution is warranted when interpreting the score of patients

performing low on the SST, as there is a controversy on the presence or absence of a floor effect.

The comparison between the SST and the ASES was advantageous for the ASES that showed no floor effect while the SST did when both scores were investigated in the same testing conditions (Beckmann et al., 2015; Tashjian et al., 2010). The study that compared the responsiveness of the Constant and the DASH showed lower responsiveness for the latter (de Witte et al., 2012).

The DASH and the Constant Score were the most extensively investigated outcome measures, while limited information was available on the SST's responsiveness and the ASES's reliability and responsiveness. The Constant should be preferred in situations where the responsiveness is paramount, and the DASH when the interpretation is to be based on consensual interpretability aspects and this approach is of prime importance.

5.4.4.3.2. PROMs measurement properties in surgical rotator cuff (Su-RCC)

The data on measurement properties were sparse following rotator cuff surgery, as none of the selected PROMs has been extensively investigated. The reliability has been investigated for the SST only, with an excellent result for this outcome measure. Only the MCID of the Constant Score has been investigated. Using the common anchor-based method, value around 10 – 11 points represents a clinically useful change for the patient using the Constant Score. However, the two studies that investigated the MCID emphasised more generally that the MCID value is highly dependent of the method used to determine it (Christiansen et al., 2015; Kukkonen et al., 2013).

The responsiveness was the most frequently investigated property. However, separate analysis of the studies provided little specific and usable information on ES and SRM magnitude, due to the variations in sample composition, timeframe and applied treatments. Similarly, change correlations were not comparable, as the change was correlated with scores from varying reference outcome measures across studies.

Direct comparison amongst PROMs within the same study showed that the QuickDASH and the DASH have similar responsiveness (Macdermid et al., 2015). This result was favourable for the latter score, which is simpler in its administration. The comparison of the DASH's and SST's responsiveness in improved, equivocal and negative response to treatment subgroups showed no clear advantage of one score over the other (MacDermid et al., 2006). The Constant Score showed adequate ability to discriminate patients who improved or not (Christiansen et al., 2015), but various change correlations (O'Connor et al., 1999). The comparison of the Constant and the ASES responsiveness was in favour of the Constant Score, when considering responses for either the absolute or for the relative Constant Score (Holtby and Razmjou, 2005).

No strong recommendation as to which might be the best PROM amongst those selected for the assessment of patients' shoulder function following rotator cuff surgery can be made based on the retrieved data, due to the inherent limitations of research within the literature. The Constant Score has a slight advantage over the other PROMs in the present state of knowledge, as it showed superior responsiveness to ASES in a direct comparison and its MCID has been determined in this population, which is useful for interpreting the meaning of a change in the function of the shoulder for the patient. However, its reliability has not been investigated in this population.

5.4.4.3.3. Measurement properties of MAB outcome measures in surgical rotator cuff conditions (NSu- and Su-RCC)

A limited number of studies investigated the measurement properties of MAB outcome measures for shoulder function evaluation in rotator cuff conditions. Two studies, assessing arm underuse and muscular activation time were essentially exploratory and provided limited information on measurement properties. In both studies, the scores from the MAB outcome measures were poorly correlated to those derived from shoulder function PROMs. Thus, further research is needed to determine more specifically the concepts that these outcome measures are measuring.

Several properties were investigated for the ARS and COMP. However, the convergent validity of these scores was limited, as shown by their low correlations with the shoulder function PROMs.

The discriminative power between patients and healthy controls was investigated for the ARS, the COMP and the B-B Score, which all showed good to excellent discrimination capacity. The B-B Score showed better specificity and lower sensitivity than the ARS and better specificity and sensitivity than the COMP (B-B Score: specificity 90%, sensitivity 90%; ARS: specificity 81%, sensitivity 98%; . COMP: specificity 85%; sensitivity 84%). In contrast to the other MAB outcome measures, the B-B Score was related to PROMs, and can thus be considered as a specific measurement of shoulder function.

The literature review showed that the B-B Score discriminative power and responsiveness had been extensively investigated in non-surgical rotator cuff conditions. In addition, it highlighted some limitations, as the ICCs for intra- and inter-rater reliability and the clinical values had not been specifically defined for this pathology, either to avoid overwhelming details in the reporting of results, or because the statistics required large samples to be conducted. Nevertheless, more information was available for the B-B Score, than for other MAB outcome measures. It was also shown to be the only MAB outcome measure that was consistently related to shoulder function PROMs, though the COMP Score rely on the same metric and almost the same movements (hand to the back + hand behind the head vs. hand to the back + hand to the ceiling as to change a bulb for the B-B Score). Both Score are simplified versions of the P Score, but the systematic approach that had been used at the conception stage of the B-B Score appears to have preserved the relationship to shoulder function during the simplification process (Coley et al., 2007a; Pichonnaz et al., 2015c; Korver et al., 2014a).

5.4.4.3.4. Benchmarking of measurement properties of outcomes measures in surgical rotator cuff conditions (NSu- and Su-RCC)

No general recommendation can be formulated for the informed choice of an outcome measure for the evaluation of patients with rotator cuff conditions, as none demonstrated superior measurement properties over the other ones. Users should refer to the above intermediate syntheses (PROMs non-surgical p. 265 - 266, PROMs surgical p. 266 - 267, MAB outcome measures p. 267 - 268 and Benchmarking of measurement properties of outcomes measures in rotator cuff outcome measures p. 269 - 270) to choose the best tool for their specific needs.

No PROM demonstrated globally superior measurement properties, either for non-surgical or for surgical treatment of patients. On the other hand, the research on MAB outcome measures validation in this field is still in its infancy. Few MAB outcome measures exist and their measurement properties have not been exhaustively investigated for this population. The thesis Phase 3 study was useful in this respect, as the B-B Score is presently the only MAB outcome measure that has demonstrated convergent validity with the PROMs and can therefore claims to assess shoulder function.

Most studies reported an adequate reliability of outcome measures, with comparable ICC values for MAB outcome measures and PROMs. When a direct comparison was made between PROMs and MAB outcome measure, the latter outcome measures (ARS and COMP) showed better reliability than the DASH and SST. No interpretability aspects have been specifically determined in surgically or conservatively treated rotator cuff populations for MAB outcome measures, which limits the possibility to interpret the results of clinical measurements. The B-B Score showed superior responsiveness to that of four currently used shoulder function PROMs, when this characteristic was assessed using ES and SRM. However, more research is needed before making conclusions about responsiveness, because the B-B Score change correlation was adequate with the Constant and relative Constant Scores only.

Globally, the lack of data prevents conclusions on the respective advantages/limitations of the PROMs or the MAB approach in rotator cuff pathologies. However, although the body of knowledge on them is still limited, the MAB outcome measures represent a promising path for further exploration as they displayed equivalent or superior properties when direct comparisons were performed.

5.4.4.4. Outcome measures for shoulder function in glenohumeral osteoarthritis (OA)

5.4.4.4.1. PROMs measurement properties in osteoarthritis (Su-, NS- and Mi-OA)

Conversely to the situation for non-surgical treatments, some properties have been calculated for all the selected outcome measures concerning the evaluation of shoulder functional capabilities after surgical treatments. Most of the studies proceeded to direct comparison between scores, which allows for comparison between tools' measurement properties tested under the same conditions. The responsiveness has been tested for all outcome measures, and five studies have performed a comparison between several of the selected outcome measures (Angst et al., 2004; Angst et al., 2008; Angst et al., 2009; Sciascia et al., 2017; Roy et al., 2010; Macdermid et al., 2015). Yet, little information was available for reliability and interpretability aspects following surgery, which limits the interpretability of the score value or score change in patients.

Although the responsiveness and discriminative power of the Constant and the ASES are closely matched, Sciascia et al. calculated a better relative efficiency for the Constant (0.8) (Sciascia et al., 2017). Comparable responsiveness was found between the DASH and the QuickDASH, which is advantageous for the QuickDASH, which is simpler to complete.

The DASH was less responsive than the SST, Constant and ASES when directly compared to them. These three PROMs constitute preferable options for the evaluation of shoulder function following surgery to address shoulder OA, in the present state of knowledge. However, these recommendations might need to be

refined based on future research to determine the reliability and interpretability aspects of these scores.

5.4.4.5. Outcome measures for shoulder function in shoulder instability (SI)

5.4.4.5.1. PROMs measurement properties in non-surgical shoulder instability (NSu-SI)

The WOSI was the most extensively validated PROMs for shoulder function in non-surgically treated shoulder instability, while the data were patchy or absent for the other PROMs. Its single-measurement reliability was unanimously found to be adequate. All the interpretability aspects have been investigated except the PASS. However, there is a controversy on the exact magnitude of the SEM and MDC, which is a limitation for the clinical interpretation of results.

c The comparisons were consistently favouring to the WOSI, which always displayed larger SRM than the other score outcome measures and significantly higher discriminative power between improved and unimproved patients. No floor or ceiling effects were detected using the reference 15% threshold for these aspects of measurement properties.

Therefore, the WOSI appears to be the first choice for the evaluation of non-surgical shoulder instability among the tested PROMs. The WOSI has adequate measurement properties, though LoAs should warrant caution in outcome interpretation when analysing the performance of a patient on one occasion. The latter controversies about interpretability aspects may also render the need for clinical interpretations to be undertaken with caution. Another review should compare the WOSI measurement properties to that of other outcome measures that are specific to shoulder instability, in order to compare this PROMs to other measurement tools that were designed exactly for the same purpose (e.g. the Oxford Shoulder Instability Score (OSIS), Melbourne Instability Shoulder Score (MISS) and the Rowe instability score) (Plancher and Lipnick, 2009).

5.4.4.5.2. PROMs measurement properties in surgical shoulder instability (Su-SI)

Three studies investigated the measurement properties of PROMs for surgical shoulder instability treatment (Gaudelli et al., 2014; Salomonsson et al., 2009; Oh et al., 2009). Only the WOSI and QuickDASH were involved. It appears from the study that investigated their measurement properties that the WOSI was more responsive and more reliable than the QuickDASH, though its ICC was lower than required when the questionnaire was administered and completed remotely by telephone.

The clinical interpretation' possibilities are limited as no interpretability aspect was determined in this context. Based on these limited results, the only conclusion that can be stated was that the WOSI offers advantages over the QuickDASH for the evaluation of surgical shoulder instability treatments and that the outcome measure completion over the phone has insufficient reliability.

5.4.4.5.3. PROMs measurement properties in mixed surgical/non-surgical shoulder instability (Mi-SI)

Little information was available, so that no direct comparison can be made between outcome measures. In this suboptimal situation to propose recommendation, the WOSI showed adequate reliability (at a low level of evidence) and responsiveness (at a low level of evidence), but no interpretability aspect is available for the clinical interpretation of the score in this population, except for the MCID. Due to the scarce literature, no table is provided to summarize the levels of evidence for measurement properties in this subpopulation

5.4.4.5.4. Measurement properties of MAB outcome measures in non-surgical shoulder instability (NSu-SI)

Little attention has been given to the validation of MAB outcome measures for the evaluation of shoulder instability, as only one study had evaluated the measurement

properties of a kinematic outcome measure for non-surgical shoulder instability (Pichonnaz et al., 2015a). The Phase 3 study that investigated the B-B Score measurement properties specifically in this population was innovative, in the sense that it was the first study that intended to develop and investigate a score specifically for these pathologies, according to the literature retrieved on MAB outcome scores and to the best of the author's knowledge.

The strength of its correlations with PROMs showed that the B-B Score had a moderate relationship to the other outcome measures of shoulder function, when instability was considered. The direct comparison between the B-B Score and several PROMs showed that it was considerably less responsive than the WOSI, and to a lesser extent the Constant Score for shoulder instability evaluation. Moreover, its discriminative power between patients and controls was insufficient, essentially because of a lack of specificity. Therefore, the score was not efficient for identifying correctly the patients with shoulder instability, because of an excessive proportion of false positive results.

In summary, the WOSI was superior to the B-B Score for shoulder instability evaluation in a sample of conservatively treated patients and no other MAB outcome measure was available to date for shoulder instability evaluation.

The Phase 2 study highlighted the poor measurement properties of the B-B Score for conservatively treated shoulder instability. The aim of this thesis (i.e. validate the simplest possible kinematic shoulder function scoring procedure applicable in clinical practice and research) was therefore not reached specifically for this condition, as the B-B Score clinimetric weaknesses prevent its application for shoulder instabilities. By contrast, this result highlight that the detailed analysis of pathological subgroups provided in this thesis was required to offer a realistic picture of the B-B Score measurement properties in each investigated shoulder pathology. Analyses of the complete sample of patients were useful for the comparison between the B-B Score clinimetric performance and its alternative outcome measures. They made it possible to use the abundant literature that relies on study samples including diversified pathologies, but would have been insufficient to highlight the contrasted results between pathologies.

5.4.4.5.5. Benchmarking of measurement properties of outcome measures in shoulder instability (Su-, NS- and M-SI)

The WOSI was the most frequently evaluated PROM in nonsurgical, surgical and mixed nonsurgical/surgical samples. It consistently showed higher clinimetric performance when compared to concurrent alternatives, whether it is PROMs or MAB outcome measures. This result highlights that a specific tool for shoulder function evaluation in instability is better performing than any generic tool for shoulder function. Several authors had previously put forward that the WOSI is the most rigorously validated instability outcome measure and that it has thus become the most used in recent years (Rouleau et al., 2010; Angst et al., 2011; Wylie et al., 2014).

Thus, the WOSI is presently the first choice for shoulder instability among selected PROMs and MAB outcome measures. The WOSI has however, some limitations related to the difficulty to interpret the clinical meaning of results due to some controversial or missing interpretability aspects. An interpretation of results based on a single measurement may also be compromised by the variability of measurement, as indicated by the relatively inflated magnitude of the LoA.

Very little attention has been put on the development of a valid MAB outcome measure for instability evaluation. The only the properties of the B-B Score have been partially investigated, with diminished results for this particular pathology (shoulder instability). This is contrary to the other shoulder pathologies for which the B-B Score had been tested (Pichonnaz et al., 2015a).

The results of the researches conducted in this thesis were further reinforced by the subsequent literature review. This might indicate that a specific approach to shoulder instability should be used to develop an outcome measure able to capture the function-related movement alterations in this condition. Shoulder instability is actually mainly characterised by apprehension of movements at risk of dislocation, while the other shoulder pathologies retained in this review are essentially characterised by the association of pain, stiffness and weakness, in various proportions. A MAB outcome measure for shoulder instability should thus ideally be able to challenge the patient's

range of motion that causes apprehension during shoulder movements, but which does so without compromising the patient's safety.

5.4.4.6. Outcome measures for shoulder function in capsulitis (C) (frozen shoulder)

Measurement properties of outcome measures in capsulitis (NSu-C)

Little focus has been put on the evaluation of PROMs and MAB outcome measures for assessing function capabilities in capsulitis. Only one study was found for each approach.

The DASH demonstrated an adequate discriminative power between improved/unimproved patients and the B-B Score demonstrated an excellent discriminative power between affected/healthy controls. These values cannot be compared as the discrimination criteria were not the same.

The B-B Score's convergent validity was adequate considering its correlation with PROMs. Conversely, the change correlations were lower, indicating that the evaluation of change had limited relationship to that measured using PROMs.

The BB Score's ES compared favourably, and its SRM equivalently, with those of PROMs, when a direct comparison was performed in similar conditions.

Based on the limited evidence available, it was stated that the B-B Score displayed adequate responsiveness. Nevertheless, the direct comparison of responsiveness of the B-B Score and that derived from PROMs favoured the B-B Score considering ES' responses, and equivalent when considering the SRM's responses.

Provided that the other measurement properties are adequate, these results suggest that a MAB outcome measures may have the potential to challenge the PROMs for effective functional assessment of this pathology. However, relevant data were scarce and more research is needed to complete the knowledge on both approaches for shoulder function evaluation in capsulitis.

5.4.4.7. Outcome measures for shoulder function in humerus fracture (F)

5.4.4.7.1. PROMs measurement properties in non-surgical humerus fracture (NSu-F)

Limited evidence was available on the measurement properties of PROMs in non-surgical treatment of proximal humerus fracture. From the available results, it appears that the DASH is slightly less reliable, but also slightly more responsive than the Constant. Thus, the Constant has an advantage as an outcome measure when a measure at a given time point is required, due to its superior reliability, while the DASH is more efficient for detecting the patient's change amongst several time points. Due to the scarce literature, no table is provided to summarize the levels of evidence for measurement properties in this subpopulation.

5.4.4.7.2. PROMs measurement properties in mixed surgical/non-surgical humerus fracture (Mi-F)

There is limited evidence on the measurement properties of the PROMs in mixed samples of non-surgically and surgically treated patients. Only two outcome measures were partially investigated and compared, i.e. the DASH and the Constant.

No information was available on the reliability of the Constant, while the DASH's reliability was adequate. The defined interpretability aspects (SEM, MDC, MCID) were closely matched amongst them. The LoAs were also evaluated for the DASH only, but they were larger than the ± 10% threshold used in this review.

Concerning responsiveness, the change scores of both PROMs were moderately correlated. The ES and SRM were of comparable magnitude between scores, but the DASH showed a better ability to discriminate the patient who improved by a small amount better from those who did not improve. However, it showed a marked ceiling effect at 12 months that was not present for the Constant.

The DASH should be preferred to the Constant at an early stage of recovery, because it demonstrated a higher ability to detect improvement and more interpretability aspects were determined for this score than for the Constant. However, it should not be used at a late stage of recovery, due to its consequent ceiling effect.

The reliability still needs to be compared between these scores to be able to formulate more informed recommendations for the preferred use of one or other PROM.

5.4.4.7.3. Measurement properties of MAB outcome measures in humerus fracture (NS- and Mi-F)

Little research has been conducted on MAB outcome measures in patients with proximal humeral fracture, as only the study related to this thesis was found in a mixed sample of surgically and non-surgically treated patients (Pichonnaz et al., 2015a).

The strength of the correlations with the selected PROMs showed that the B-B Score evaluated a shoulder function concept that is related to those underpinning the Constant and SST, and to a lesser extend the QuickDASH. The reliability and critical the outcome measure in this population were not determined. Thus, no comparison with PROMs can be made on these aspects. The responsiveness was adequate, as only the Constant showed a slightly higher ES and SRM. Moreover the change correlations showed a good relationship with the selected PROMs for the evaluation of shoulder function' change.

5.4.4.7.4. Benchmarking of measurement properties of outcome measures in humerus fractures (NSu- and Mi-F)

Little research has been conducted to investigate the clinimetric properties of outcome measures for shoulder function assessment following humerus fracture. However, some of the researches compared several outcome measures, so that conclusions can be drawn on some issues.

The Constant and the DASH were the only scores that were investigated and compared, in non-surgically treated and in mixed of surgically/non-surgically treated samples. Their measurement properties were globally comparable, with a better reliability for the Constant in non-surgically treated shoulders and a better responsiveness for the DASH at an early stage, which revealed an advantage clinimetrically but was later diminished by a consequent ceiling effect at one year following fracture.

Only one MAB outcome measure, the B-B Score was investigated and compared to PROMs in mixed samples of surgically/non-surgically treated patients. This score was correlated to shoulder function PROMs and can be thus considered as a shoulder function focused outcome measure. Its responsiveness was slightly lower than that of the Constant and higher than that of the QuickDASH, and to a lower extent, the SST based on ES and SRM. Its change correlations with PROMs showed that the evaluation of change using this score is comparable to that using PROMs.

Based on these statements and on the compared measurement properties, it would appear that the Constant and B-B Score perform comparably concerning the responsiveness following shoulder fracture, but that the literature was insufficient to draw firm conclusions.

5.4.4.8. Synthesis on the measurement properties of PROMs and MAB outcome measures in shoulder disorders, with emphasis on the thesis achievements

The synthesis of the results of the literature review provides an opportunity to address general considerations on methods for measuring function of the shoulder. It also allows a certain distance to be taken with the results of the studies carried out in the thesis, which is why it was carried out at the end of the work, and not upstream, as is traditionally the case.

The corpus of research on PROMs and MAB outcome measures was substantial and approximately equivalent in scope considering the number of publications. However, the research on MAB outcome measures has rarely led to a basis from which an

effective scoring system can be developed, and even less so in the pursuit of a tool that is applicable within clinical practice.

Based on this statement from the literature, the developments undertaken in this thesis can be considered as innovative in that they rely on MAB methods, while taking advantage of technological opportunities (use of a smartphone) and possibilities to simplify the measurement procedure (use of the B-B Score) to propose an approach that can be applied for routine clinical assessment. The interdisciplinary collaboration between physiotherapists, medical doctors and engineers has helped to reduce the gap between the conception of an efficient movement analysis method and its clinical application.

Measurement properties were frequently determined in diversified patient samples and, to a lower extent, within rotator cuff populations. Conversely, little research had addressed the measurement properties in fractures and even less in capsulitis. This is problematic for the clinical transference of useful information, as the measurement properties are context-dependent (Riddle and Stratford, 2013; Robertson et al., 2017; El Gaafary, 2016; Collins and Roos, 2016) and the patient, who presents with a defined pathology and with specific consequences to shoulder function, is not always comparable to the population that has been the target for research. The thesis' data were gathered and analysed separately for various common shoulder disorders to allow a circumstantiated interpretation of the investigated measurement properties.

The Constant and the DASH/QuickDASH were the most extensively investigated PROMs. Based on the results of this review, no PROM can considered as globally superior to the other ones, and thus none can be recommended as a generic standard for shoulder function' measurement. These results were in line with previously published systematic reviews on PROMs measurement properties standard (Fayad et al., 2005; Oh et al., 2009; Placzek et al., 2004; Roy et al., 2009).

Nevertheless, some PROMs might have demonstrated an equivalency or an advantage compared to the others when a direct comparison was performed in a given target population. The WOSI is the only one PROM that demonstrated overall superior measurement properties to the other and in comparison to MAB outcome measures as well, when shoulder instability was considered. The good measurement properties of the WOSI had previously been reported by other authors (Angst et al.,

2011; Salomonsson et al., 2009). As the WOSI was the only condition-specific outcome measure amongst the selected ones, this finding raises the question of whether specific outcome measures should be developed to improve the quality of shoulder function assessment, although this option would further increase the already high number of shoulder function outcome measures.

Despite the substantial body of literature on shoulder movement analysis, few MAB outcome measures exist. Furthermore, the development and the clinical application of most of them has not been continued beyond the initial studies. Conversely, this thesis' work was oriented toward clinical applicability. Much communication, marketing and technical work would still be necessary for its routine application to become a reality, but the smartphone B-B Score has nevertheless been designed to make it technically feasible.

Some of MAB outcome measures (B-B Score and P Score) are correlated to PROMs and can thus be considered to investigate the same concept of shoulder function. As a consequence, they might concur with the PROMs for shoulder function' evaluation. Other ones (ARS, COMP, arm underuse, $T_{emg/mov}$), that were hardly correlated with PROMs, investigate different concepts of shoulder function compared to PROMs. More research is needed to understand better what encompasses the concepts captured by these MAB outcome measures. A large variety of biomechanical parameters can be measured using MAB methods. However, they cannot be *a priori* considered to reflect shoulder function until their convergent validity has been demonstrated by an adequate correlation between them and recognised shoulder function measurement tools (de los Reyes-Guzman et al., 2014).

The B-B Score relationship to PROMs was expected, because the measured power-related parameter [(deg/s)*(m/s^2)] had precisely been selected due to its relationship to shoulder function from the conception of the P Score (the parent score from which the B-B Score was derived) and of the B-B Score (Coley et al., 2007a; Pichonnaz et al., 2015c). The Phase 3 study confirmed the adequacy ($r \geq 0.50$) of the B-B Score for shoulder function evaluation of patients' populations with rotator cuff condition, humerus fracture or capsulitis, but not for shoulder instability. Again, these results, which are differentiated according to pathologies, highlight the context-dependency of the measurement properties of outcome measures.

More research is needed to investigate exhaustively the measurement properties of MAB outcome measures in various populations with shoulder conditions. The thesis aimed at an extensive validation of the B-B Score and, consequently, it came up as the most extensively validated MAB outcome measures to date. For recall, its normal performance, its reliability and interpretability aspects in a diversified sample and its convergent validity and responsiveness for the pathologies selected in this review have been investigated in Phase 3 study and published (Pichonnaz et al., 2015a). Additionally, it had been demonstrated in Phase 2 study and its related article that the B-B Score measurement can be performed using a smartphone, with similar properties to a dedicated IMU device (Pichonnaz et al., 2017).

It would be worth developing more the exploration and validation of MAB outcome measures, as well as their transfer into clinical practice, provided that they display sound measurement properties at the initial stage of testing. On this latter point, the systematic step-by-step validation approach has been effective, since it demonstrated that the use of a smartphone did not lead to a degradation of the measurement properties of the B-B Score, compared to an inertial measurement system, which is the tool used for all other selected MAB outcome measure. Though the practicality was not a formally investigated aspect of the literature review, it appears that the B-B Score is the only MAB Score that was designed to be measured with a cheap and accessible device. However, to date no score, including the B-B Score, has been exhaustively tested, including reliability and interpretability aspects for specific shoulder conditions. The literature review highlighted additional patients' population for which validation studies would be useful to extent the knowledge about the B-B Score measurement properties.

It was striking to state in this review, that the development of the few existing MAB outcome measures have very rarely been followed by applications in treatment outcome studies, as could be observed during the inspection of titles and abstracts of articles. This is a limitation to the acquisition of experience on these outcome measures. This situation highlights the lack of focus on clinical applicability at the development stage of MAB outcome measures, as well as the shortcomings of the knowledge transfer from research into professional practice. Though the B-B Score was designed to be easily applicable in clinical practice and research, mainly in the study of Phase 3 that tested the measurement capacities of the smartphone, the lack

of actual clinical applications also apply to this MAB outcome measure to date. Its actual use in the future, if possible in conjunction with other outcome measures to allow for comparison, will be necessary to gain more knowledge on the B-B Score clinical capabilities.

The Phase 2 and 3 studies demonstrated that, except for shoulder instability, the B-B Score measurement properties were appropriate, to the exception of the LoAs that were larger than the ≤ 10% arbitrary defined threshold. The literature review confirmed that the B-B Score compared equally and sometimes favourably to PROMs in direct and indirect comparisons with alternative outcome measures to the exception of LoAs that were larger. Although LoAs ≥ 10% were frequently reported in this review, this appeared as a shortcoming of the B-B Score. Future work could address this issue for example by modifying the testing instructions in order to increase the movement repeatability (e.g. by using targeted movement or setting a pace) or increasing the number of replications. Three replications had been defined in Phase 1 study as an optimal number to contain measurement variability while limiting measurement constraints. Nevertheless, further investigations on the use of a higher number of replications might be conducted with the aim to improve the B-B Score reliability for single measurements.

More generally, it was stated that the measurement properties of MAB outcome measures generally and adequately complied with requirements. In the long run, they might thus represent a viable alternative to overcome the controversies surrounding shoulder function evaluation with PROMs for most current shoulder pathologies, provided that more research is conducted to extensively validate and improve the existing MAB outcome measures, and that greater emphasis is placed on clinical applicability and knowledge transfer.

5.4.5. Study limitations

For clinical interest and feasibility reasons, this review was conducted on the most frequently used PROMs in the most frequent shoulder pathologies. Thus, the results do not apply to all PROMs and are not transferable to other less frequent shoulder pathologies. The fact that the investigated PROMs are frequently used reflects the present practice, but does not imply that the selected outcome measures are necessarily the ones with the best measurement properties. New PROMs that have

been recently developed may have been developed based on presently recommended standards but scarcely diffused to date. The fact that the WOSI Score stood out for shoulder instability raise the question if conditions-specific PROMs would have obtained better measurement properties that generic shoulder PROMs in this review.

As explained in the introduction, the conditions were not met to proceed to a formal quality analysis of the literature. Thus, it was not possible to present a hierarchical analysis of the quality of the articles. Other authors, who have conducted literature reviews where the COSMIN checklist could potentially be used, have chosen either not to rely on this checklist, or to use it without indicating how the difficulties were overcome, or to adapt the assessment in a transparent manner (Andreopoulou et al. 2018; Zanudin et al. 2017). However, a selective qualitative analysis of the methodology was undertaken when conflicting results were found in order to determine the factors explaining the discrepancies. The inability to make a quantitative assessment of the quality of the studies prevented a comparison between the studies conducted in this thesis and those found in the literature.

Differences in results were frequently induced by a lack of consensus about the methods to be used. This was for example, the case for the AUC criteria for responsiveness' evaluation, the reference scores used for change correlations and the MCID/MCII determination methods. No meta-analysis could be conducted because of the heterogeneity of the methods, timeframes and sample composition. Moreover, no well-established quantitative meta-analysis methods were found to aggregate the data of some measurement properties. The qualitative synthesis of the results did not allow robust conclusion to be drawn from statistical inferences concerning the differences in measurement properties of the selected outcome measures. Similarly, previous reviews that had addressed this topic had also proposed a qualitative synthesis of the data (Kirkley et al., 2003; Oh et al., 2009; Huang et al., 2015; Harvie et al., 2005; Fayad et al., 2005; Placzek et al., 2004; Roy et al., 2009). The inability to perform a meta-analysis prevented drawing generalizable conclusions on the clinical performance of the B-B score compared to alternative outcome measures.

The languages of the included articles were limited to English and French. Some validation studies of PROMs in a translated language might therefore have been

ignored. Validation studies of translations published in the two aforementioned languages were nevertheless included.

In some cases, several versions of testing instructions exist for a PROM. Whenever possible, observed discrepancies between studies were checked to assess whether they might have originated because of the use of different versions or testing procedures. However, this was not always possible because the used version was not systematically reported.

The validation studies that were only presented in congresses or in academic works were not retrieved in this review, because they had not undergone a full peer-reviewed process. However, this most likely would have had a marginal impact of the results, as no scientific communication of this nature was identified during the manual search in the articles' references and in the websites that compile outcome measures' properties. Globally, the search conducted within the retained databases was near to exhaustive, as the manual search elicited only a small number of additional articles.

This review investigated the measurement properties of the measurements tools. Validity issues were not exhaustively investigated, except for the convergent validity between MAB outcome measures and PROMs, and for the floor and ceiling effects, which are part of content validity. This aspect was essential to address in order to determine whether a MAB outcome measures should be considered as an indicator of shoulder function or not. Other validity issues were not investigated in order to avoid adding to the complexity of this work. Relevant information about the validity of measurement tools for the assessment of shoulder functional capabilities can be found in other reviews (Roe et al., 2013; Makhni et al., 2015; De Baets et al., 2017; Oh et al., 2009; Bot et al., 2004; Fayad et al., 2004), and this should clearly be taken into account also when choosing a measurement tool.

A strong point of this review was that it differentiated the measurement properties in several selected populations of patients and for surgical and non-surgical treatment. It addressed separately the issues about measurement properties known to be context-dependent (Robertson et al., 2017; Riddle and Stratford, 2013; Collins and Roos, 2016), which had not been addressed in most previous systematic reviews. The analysis accounting for the specificity of each pathological population added to the complexity of the work, but contributed to the precision and relevance of the

results. However, some limitations inherent in the literature have been identified in the implementation of this approach. The population pathologies were often defined succinctly, compromising the possibilities to apply the results to a single patient during the clinical encounter. This was especially the case when a sample with diversified shoulder pathologies was enrolled in a study. The latter studies were nevertheless, not excluded from analysis, as they represented a considerable part of the body of knowledge on the subject.

Considering the substantial body of knowledge on shoulder movement analysis, it had been unexpected that only seven articles on the measurement properties of MAB outcome measures would have been retrieved. This paucity limited the possible comparisons with PROMS, particularly for OA, humerus fracture and capsulitis. Moreover, six of these articles originated from the same laboratory, with which this thesis' author had collaborated. The origin of the developed MAB outcome measures was expected to be much more diversified at the initiation of this review. Although actions have been taken to ensure a fair analysis based on previously defined objective criteria, the author's methodological background and experience might have influenced the results' interpretation against his intentions.

5.5. Conclusion

This systematic review allowed for the comparison of the measurement properties of the B-B Score with alternative outcome measures. It provided therefore an opportunity to challenge its clinimetric performance investigated in the Phase 2 and 3 studies with those of outcome measures using a questionnaires-based approach (PROMs) and those using also a MAB approach.

More generally, it added to the body of knowledge on the outcome measures of shoulder function, as it was the first literature review that compared measurement properties of frequently-used shoulder function PROMs (Constant, DASH, SST, ASES and WOSI) and MAB outcome measures to the best of the author's knowledge. It reported the outcome measures' respective measurement properties separately for current shoulder conditions, to account for the context dependency of measurement properties.

Similarly to previous systematic reviews, it stated that no PROM was globally superior to the other ones for shoulder function evaluation, except for the WOSI that performed

better than generic shoulder function PROMs for shoulder instability' evaluation. In other shoulder conditions, a PROM may merely display particular advantages over the other ones only for a given set of conditions of evaluation. Thus, the choice of a PROM should be oriented by its specific measurement properties for the target population, and not based on general considerations.

Concerning the retrieved body of literature, it was stated that despite the considerable amount of literature on PROMs, little information about the clinimetric performance of outcome measures was found for capsulitis and fracture evaluation. The Constant and DASH/QuickDASH were the most extensively investigated PROMs. Although they cannot be considered as superior to concurrent outcome measures in all aspects, they nevertheless possess a more consistent body of knowledge about their clinimetric characteristics to better orientate the potential user's choice.

The review of MAB outcome measures showed that despite the consistency of research on shoulder movement analysis, few investigations had resulted in the development of an outcome measure for shoulder function evaluation. All MAB outcome measures, with the exception of the B-B Score, had had their measurement properties investigated in one sample of population only at the development stage of the measurement tools. It can thus be considered that the development of MAB outcome measures for shoulder function assessment is still in its infancy.

Nevertheless, the investigated properties of MAB outcome measures were generally adequate for their intended purposes. Also, they compared equally and sometimes favourably to PROMs in direct comparisons within pathologies. The B-B Score was the most extensively investigated MAB outcome measures to date, though its reliability and interpretability aspects have still to be defined in specific populations.

This literature review allowed for an extended benchmarking for the measurement properties that had been previously investigated in Phase 2 and 3 studies of the thesis. It showed that the shortcomings of the B-B Score concerned specifically the clinimetric performances for the assessment of function in shoulder instability and the variability of single measurements highlighted by large LoAs. All other measurement properties were comparable to those of concurrent scores, with slight nuances for each testing conditions, and complied with the established standards for adequate measurement. The literature review highlighted that further researches on the B-B Score should primarily investigate the influence of modified measurement procedures

on the variability of single measurements, so that it performs comparably to alternative outcome measures with regard to this specific shortcoming. Concerning shoulder instability, a condition-specific approach, which differs considerably from the B-B Score is probably needed to asses shoulder function.

Based on the results of this review, it appears that MAB shoulder function evaluation is still an emerging field. The results are presently too limited to be conclusive on their superiority or inferiority over PROMs. Nevertheless, studies on measurement properties conducted to date showed that they constitute, including for the B-B Score, a sustainable alternative or complement to frequently used PROMs. However, it would be worth investigating if devices that are more accessible can substitute inertial sensor systems to facilitate the widespread application of MAB outcome measures in clinical conditions, as was done for the B-B Score in Phase 2 study. Future researches are needed to investigate exhaustively the measurement properties of existing MAB outcome measures and optimise their testing procedures, as well as attempting to develop ones that are more efficient. The clinical applicability and the knowledge transfer toward clinical users are aspects that need to be considered in these future developments.

CHAPTER SIX

GENERAL DISCUSSION AND CONCLUSIONS

6.1. General achievements

6.1.1. Conception of a founded measurement method

This thesis has endeavoured to explore, based on contemporary clinical need, an alternative path for shoulder function evaluation, in order to provide, if possible, the clinicians with a valid, cheap and straightforward shoulder outcome measure. The research underpinning these ambitions took place in a context where there is an ongoing controversy about the best shoulder PROM to use and where transfer of laboratory-based movement analysis into clinical practice has remained scarce due to its technical complexity (Kirkley et al., 2003; Oh et al., 2009; Huang et al., 2015; Harvie et al., 2005; Aminian and Najafi, 2004; Clark et al., 2017). Concurrently, technological progress and the widespread diffusion of sensors within daily-life' objects has revealed pathways for the exploration of new opportunities to improve the efficiency of shoulder function' evaluation in clinical practice (Ciuti, 2015).

Based on these statements and considering that access to suitable devices, time constraints and familiarity with technology are important barriers that could potentially be overcome, it was decided to explore to what extend a very simple movement analysis-based outcome measure using a smartphone, met the requirements of a valid/efficient measurement tool of shoulder function. During the initial phases of the PhD research programme (Phase 1, 2 and 3 studies) involving 'proof of this concept', everything was focused towards keeping the measurement procedure and instrumentation to their simplest expressions, while preserving measurement properties, in order to develop an efficient outcome measure.

The kinematic B-B Score was chosen for focused explorations because it had been designed to capture shoulder function using only two essential movements, i.e. "hand to the back" + "hand to the ceiling as to change a bulb". Initial research suggested that this score had potentially sound measurement properties, though this remained to be established with more precision in specific conservatively treated shoulder conditions (Pichonnaz et al., 2015c).

Initial research conducted by the author for the purpose of his MSc dissertation (Pichonnaz, 2010; Pichonnaz et al., 2015c) suggested that this score had potentially sound measurement properties, though the measurement process remained to be

optimised, the practicalities and the measurement properties had to be further investigated and specified for common conservatively treated shoulder conditions. Having addressed these issues, it was then necessary to compare the clinimetric performance of the optimised version of the B-B Score to the alternative outcome measures for shoulder function evaluation, in order to provide a substantiated insight into its contribution to the measurement of shoulder function.

Incremental steps from the first to the third Phase of the thesis addressed these issues. The Phase 1 study, which explored various possible alternatives for the delivery and calculation of the B-B Score, resulted in the definition of optimal testing and calculation procedures, amongst tested possibilities. The Phase 2 study, which investigated the influence of the use of a dedicated IMU system or a smartphone on the measured values, established that the B-B Score could be acquired with greater simplicity without deterioration of the measurement properties using a smartphone (Pichonnaz et al., 2017). The Phase 3 study, in which the detailed measurement properties of the smartphone B-B Score using the optimised procedure were investigated, resulted in their specific determination in several common shoulder pathologies (Pichonnaz et al., 2015a). Following these studies, it was necessary to consider the defined measurement properties of the B-B Score within a larger scope, which was done by means of a systematic literature review that included a wide range of alternative PROMs and MAB outcome measures for shoulder function evaluation

6.1.2. Scoring method optimisation

6.1.2.1. Achievements of the B-B score optimisation study (Phase 1)

Phase 1 provided the foundations to underpin the choice of a justifiable calculation method for the B-B Score among several theoretically relevant ones. A summary of the clinimetric performance for the measurement properties of the B-B Score investigated in the Phase 1 study is available in Table 6.1.

Table 6.1: Summary of the clinimetric performance for the measurement properties of the B-B Score investigated in the Phase 1 study

Measurement property	Clinimetric performance
Discriminative power	Significant difference between patient and control groups $p < 0.01$ Large ES for difference between healthy and groups (Cohen's d 1.60 – 1.70)
Stability between replications	Non-significant 1.8% increase over replications (p = 0.06 in patient group, 0.16 in control group) ICC between replications 0.90
Intra-rater reliability	ICC: 1st measurement 0.93 2nd measurement 0.97 Bias (LoA): 1st measurement 1.2% (± 12.7%) 2nd measurement 2.3% (± 16.7%)
Inter-rater reliability	ICC: 1st rater 0.94 2nd rater 0.96 Bias (LoA): 1st rater - 0.9% (± 13.3%) 2nd rater - 2.6% (± 16.6%)

Legend: ES: effect size; LoA: limits of agreement; ICC: intraclass coefficient of correlation

These results were sufficiently encouraging to support further research on the measurement properties of the B-B Score for measuring shoulder function in conservatively treated patients, who represent much larger populations than the surgically treated ones (Colvin et al., 2012). The adequacy of these first results with current standards for clinimetric performance was of importance for the good continuation the thesis' project. As the Phase 1 study was the first investigation of the B-B Score in a population that had not been surgically treated, it was not *a priori* obvious that the measuring properties would be adequate.

The exploration of the relevance of the "area" computation method as an alternative to the original "range" method did not allow improving the B-B Score measurement properties. The investigations rather confirmed that the results using the original "range" method was not importantly influenced by possible peak measurements. This reinforces the results on the P Score developments published by Coley et al. (Coley et al. 2007a)

The results of the Phase 1 study complied with the standards for adequate measurement properties (please see Table 5.2, within sub-section 5.2.6 “Interpretation delimitation”, within Chapter five, p. 185), except for the LoAs that were ≥ ± 10%. Although this threshold is not widely recognised, as it has been defined based on clinical considerations for the needs of the subsequent literature review, LoAs ranging from ± 12.7% to ± 16.7% were indicative of a level of variability that this might affect the precision of single measurements.

At the Phase 1 advancement stage of the thesis, the magnitude of the LoAs were attributed to the inexperience of the raters in the B-B Score delivery, as close data inspection had revealed the influence of a limited number of divergent values on the LoAs (please see sub-section 2.3.3.3 “B-B Score determined by mean or median of replications”, Figure 2.5 and 2.6 Bland and Altman graphs for 1 replication, within Chapter two, p. 86 - 87). However, this assumption could be *a posteriori* invalidated, because the LoAs of the Phase two were found to be larger than that of the Phase 1 (Smartphone intra-rater LoAs ± 18.8%; inter-rater LoA ± 18.5%), though the data had been collected by trained users. The origin of the variability lies probably more in the difficulty for participants to perform exactly the same movement several times than in the inaccuracy in the placement of the sensors or the imprecision of the sensors themselves.

As the intra- and inter-rater magnitudes of the LoAs stated in Phase 2 study represents the main shortcoming of the B-B Score for the shoulder function measurement in rotator cuff conditions, capsulitis and proximal humerus fracture, it is questionable whether the choice of only three replications in the subsequent thesis’ phases was adequate. Taking a larger number of replications into consideration for the calculation of the B-B Score should mathematically decrease the LoAs, but would be of interest only if this modified procedure does not induce any carry-over effects (like warm-up or fatigue effect) (Mercer, 2002).

In addition, the chosen measurement procedure was optimal only within the investigated alternatives. Other modifications could be explored, to determine whether they improve the ability of the measured persons to execute the score’s movements in a consistent manner. Amongst other factors, the use of targeted movements, the specification of a speed, the wearing of a light weight or the provision of a cadence might influence the consistency of executed movements, and therefore

the magnitudes of the LoAs. These alternatives were not tested in the context of this thesis, because it was aimed to keep the scoring procedure at its simplest expression.

The results from the Phase 1 study also provided information to ensure the feasibility of the research protocol and the implementation of an efficient recruitment procedure. The information was precious for the subsequent studies, as a more efficient recruitment procedure could be implemented and the experience acquired by the raters ensured the collection of proper data in the next research phases. The fact that a pilot study had been undertaken also proved to be an advantage when applying for the Swiss National Science Foundation funding.

It could thus be considered that the issues related to the definition of optimal testing and calculations procedures, amongst tested possibilities, were addressed in Phase 1, but that issues related to the instrumentation and to the B-B Score's measurement properties were still to be investigated, which was addressed in the Phase 2 and 3 studies.

6.1.3. Development and testing of a smartphone approach

6.1.3.1. Achievements of the smartphone evaluation study (Phase 2)

Phase 2 dealt with the issues related to the simplification of the B-B Score's instrumentation. It compared the respective measurement properties of a middle-segment smartphone and a dedicated movement analysis IMU system, used as a reference for the B-B Score measurement.

A summary of the clinimetric performance for the measurement properties of the B-B Score investigated in the Phase 2 study is available in Table 6.2.

Table 6.2: Summary of the clinimetric performance for the measurement properties of the B-B Score investigated in the Phase 2 study

Measurement property	Clinimetric performance
Discriminative power	Significant difference between control and patient groups $p < 0.01$
Intra-devices reliability	ICC: 0.97 Bias (LoA): - 0.6 (± 12.6) ME: 0.7 SEM: 4.0
Intra-rater reliability	ICC: reference device 0.92 smartphone 0.92 Bias (LoA): reference device 0.1 (± 19.4) smartphone 1.5 (± 18.8) ME: reference device: 0.8 smartphone: 0.7 SEM: reference device: 6.6 smartphone: 6.6
Inter-rater reliability	ICC: reference device: 0.92 Smartphone: 0.93 Bias (LoA): reference device: 1.5 (± 19.0) smartphone: 1.0 (± 18.4) ME: reference device: 0.7 smartphone: 0.7 SEM: reference device: 6.4 smartphone: 6.6

Legend: ES: effect size; LoA: limits of agreement; ICC: intraclass coefficient of correlation; ME measurement error; SEM standard error of measurement.

The Phase 2 study investigated essentially the influence of the measurement device on the quality of the measurement and provided an insight into the measurement properties of the B-B Score using a sample that included various shoulder conditions (rotator cuff conditions, proximal humerus fracture and capsulitis).

The results of the device comparison highlighted that the IMU system and the smartphone were interchangeable for group measurements, but that the magnitude of the LoA might preclude the devices' routine exchange when measurements concern individual participants. This makes the smartphone a possible substitute to

inertial sensor systems that can be used with confidence for the group evaluation of shoulder function using the B-B Score. Previous research had already shown that smartphone measurements are adequate for shoulder ROM evaluation, but no study had investigated the validity of smartphone measurement for shoulder function evaluation up to now (Cuesta-Vargas, 2016; Johnson et al., 2015; Shin et al. 2012). This result was of importance to overcome the tendency for movement analysis-based methods to be confined to laboratory settings, as was highlighted by the subsequent literature review on the measurement properties of outcome measures of shoulder function.

This result was also important with regard to the aims of the thesis, which intended to validate the simplest possible kinematic shoulder function scoring procedure applicable in clinical practice and research. Following the Phase 2 study, the process of simplification of the testing procedure could be considered as successfully achieved, as the combination of a score that includes only essential movements and a device whose use has entered into daily life reduces the testing procedure to its simplest expression. Conversely to other previously tested simple procedures for shoulder function evaluation, the B-B Score was related to alternative outcome measures of shoulder function, which demonstrated its convergent validity (Korver et al. 2014a; Korver et al. 2014b).

However, caution is warranted when interpreting the measured outcome of a single measurement that concern an individual patient. It should be considered that the typical error is ± 6.6% based on the SEM, and that errors of up to ± 18.6 % may occasionally occur based on the limits of agreement. Individual measurements at regular intervals can be used to overcome the disadvantages associated with the variability of a single measure, as the follow-up curve of the patient's performance will be correct, due to the random distribution of errors. Due to the lack of comparable investigations available in the literature, it was not possible to determine if this degree of variability was specific to the B-B Score, or more generally related to ability of participants to perform consistently shoulder movements over measurements.

The Phase two results confirmed the measurements properties that had been previously explored in the Phase 1 study, concerning the discriminative power and the reliability of the B-B Score. The results were slightly more favourable in Phase 1 than in Phase 2 (intra- and inter-rater ICC range 0.94 – 0.96 vs. 0.92 – 0.92 in Phase

2, LoA range ± 12.7% – ± 16.7% vs. 18.8% – ± 19.5%), but the Phase 2 results should be considered as the reference, because they were derived from data acquired by experienced raters in considerably larger samples of patients and controls. The intra- and inter-rater measurement properties were comparable, indicating that the B-B Score measurement had negligible dependency on the person performing the measurement. Previous studies had already shown the adequate reliability of smartphones for shoulder ROM evaluation, but none had previous investigated their reliability for shoulder function evaluation (Cuesta-Vargas, 2016; Johnson et al., 2015; Lim, 2015).

The development of an accessible and quickly delivered measurement method was required to allow for the implementation of movement analysis in routine clinical practice. It was nevertheless not enough to meet all the requirements of measurement in professional practice, as the use of an outcome measure is only warranted to the extent that the user is assured of the quality of its measurement properties. Therefore, the subsequent investigations undertaken in the Phase 3 study of the thesis aimed at an in-depth assessment of the measurement properties of the smartphone B-B Score, in order to be able to provide the necessary information to users.

6.1.4. Extensive investigation of measurement properties

6.1.4.1. Achievements of the study on the measurement properties of the smartphone B-B Score (Phase 3)

The Phase 3 study was aimed at the investigation of the measurement properties of the B-B Score in four current shoulder conditions. The measurement properties were analysed by the yardstick of established references for the quality of outcome measures and compared to those of frequently used PROMs that are considered as current standards for shoulder function evaluation.

A summary of the clinimetric performance for the measurement properties of the B-B Score investigated in the Phase 3 study is available in Table 6.3**.**

Table 6.3: Summary of the clinimetric performance for the measurement properties of the B-B Score investigated in the Phase 3 study

Measurement property	Clinimetric performance for rotator cuff conditions, proximal humerus fracture and capsulitis*
Convergent validity	$r \geq 0.50$, except for the QuickDASH for humerus fractures ($r = -0.40$).
Discriminative power	Significant difference between patient and control group ($p < 0.01$) Significant difference between baseline and 6 months stage ($p < 0.01$) AUC for patients vs. control discrimination: 0.90 - 0.96
Responsiveness	ES and SRM: B-B Score, Constant Score and relative Constant Score show close responsiveness, and superior responsiveness to QuickDASH and SST Change correlation with Constant, relative Constant, QuickDASH and SST: - Humerus fractures, "Indicated pathologies" and "All patients" group: $r > 0.50$ - Rotator cuff Constant and relative Constant $r > 0.50$; QuickDASH and SST: no correlation - Capsulitis: $r < 0.50$ AUC for improved vs. unimproved discrimination: - 0.73 All patients - 0.70 Indicated pathologies vs. AUC 0.73 – 0.83 for Constant and relative Constant QuickDASH and SST
Interpretability aspects	MDC: - Rotator cuff:15.7% - Humerus fracture: 17.5% - Capsulitis: 14.6% MCII: 25.2% PASS: 77.6%

* The measurement properties of the B-B Score for shoulder instability evaluation were demonstrated to be inadequate in this study: the B-B Score should not be used for the evaluation of shoulder function in this patients' population

Legend: QuickDASH: Quick Disabilities of the Arm, Shoulder and Hand score; AUC Area Under the operator receiving Curve; ES: Effect Size; SRM: Standardised Response Mean; SST: Simple Shoulder Test; MDC: Minimal Detectable Change; PASS: Patient Acceptable Symptoms State.

The Phase three study allowed for the extensive determination of the measurement properties of the B-B Score in four common shoulder pathologies. The measurement

properties were found to be in line with the established standards for clinimetric performance for the assessment of patients with rotator cuff conditions, capsulitis and proximal humerus fractures, but not for patients with shoulder instabilities. As no alternative kinematic shoulder function score exists, to the best of our knowledge, it is not possible to determine is this is weakness is related to an intrinsic shortcoming of MAB methods or is specific to the B-B Score. The subsequent literature review highlighted that following the Phase 2 and 3 studies, the B-B Score was the MAB outcome measure of shoulder function with the highest number of properties investigated.

The B-B Score clearly demonstrated adequate discriminative power as it was able to differentiate groups, stages and types of study participants. The interpretation of the results with reference to the responsiveness was less straightforward. Although the clinimetric performance globally complied with the standards for adequate responsiveness, the comparison of the smartphone B-B Score's responsiveness with that of PROMs provided mixed results, depending on the shoulder condition and the methods used (ES, SRM, correlations between change scores, AUC). Thus, neither the superiority nor inferiority of a shoulder function evaluation method over the others could be established in this phase of the PhD's research programme. This result was expected for PROMS, as several previous reviews had reached the same conclusion, but is new concerning the current equivalency of MAB outcome measures and PROMS, as no previous review comparing them had been conducted so far, to the best of our knowledge (Fayad et al., 2005; Oh et al., 2009; Placzek et al., 2004; Roy et al., 2009). Based on the findings of the Phase 3 study of the thesis, the only conclusion that can be drawn is that the B-B Score performs equivalently to other established measurement methods, as long as patients with shoulder instability are not the target population.

Although the subgroups' sample size was sufficient to provide an insight into the B-B Score's measurement properties within specific shoulder conditions, ultimately studies with larger homogeneous groups will be needed to establish with more precision, the measurement properties of the smartphone B-B Score in various shoulder conditions and to compare them more definitively with current standards. Studies about the relevant parameters and testing procedures to evaluate shoulder function in shoulder instability will also be required.

Interpretability aspects that are important for results' interpretation were also defined in this Phase 3 study. This study offered novel information that was of importance to provide the users with the necessary information to determine if a performance is normal, if a difference is real, if it is meaningful for a patient and if the patient's present state if acceptable for him/her.

Although all B-B Score' measurement properties could not be established with precision for all investigated shoulder pathologies within the Phase 3 study, nevertheless, a novel and quite extensive validation process has been conducted, laying the important foundations for a sound interpretation of results by clinicians and researchers. The literature review showed that, despite the limitation stated above for subgroup analysis, the measurement properties of the B-B Score were established on a larger sample than the alternatives scores, of which none had been tested on large samples following the score development study (Duc et al., 2014; Coley et al., 2007a; Korver et al., 2014a; Korver et al.; Yang et al., 2014).

Further research using larger samples should be undertaken to increase the precision of the results and to establish the B-B Score measurement properties in other patients populations (e.g. osteoarthritis, shoulder arthroplasty or, rotator cuff tears repair). It should also explore the potential of MAB outcome measures for the assessment of function in shoulder instability, as the B-B Score clinimetric performances were clearly insufficient for the assessment of shoulder function in this pathology. A MAB outcome measure for shoulder instability should ideally be able to challenge the patient's range of motion that causes apprehension during shoulder movements, but without compromising the patient's safety. No such investigation has been conducted to date on this issue, to the best of our knowledge.

After the establishment of the B-B Score's measurement properties accumulated by means of Phase 1, 2 and 3 studies, it was possible to offer a culmination to the thesis, which focused on an up-to-date contextualising the B-B Score's performance from a broader perspective and critically-evaluating it against presently-used, concurrent PROMs and against any other alternative movement analysis-based outcome measure. This aspiration was achieved by means of a review of literature that was conducted following the Phase 3 study.

6.1.5. Benchmarking of the measurement properties of the smartphone B-B Score with concurrent methods

6.1.5.1. Achievements of the systematic literature review comparing the properties of PROMs and MAB outcome measures

In order to appraise the soundness of the research orientations taken in this thesis, it was of importance to determine globally, to what extent might movement analysis represent a viable alternative approach to the contemporary reliance on PROMs, in attempting to overcome the issues of assessing changes to functional capacity of the shoulder, and more specifically, to critically-appraise the strengths and weaknesses of the B-B Score compared to all other approaches.

A summary of the key points of the literature review comparing the measurement properties of PROMs and MAD outcome measures is available in Table 6.4

Table 6.4: Summary of the key points of the literature review comparing the measurement properties of PROMs and MAD outcome measures

- First literature review comparing the measurement properties of PROMs
- Lack of a tool for quantitative evaluation of the literature that would be adapted to the review purpose
- Consequent but heterogeneous body of literature on PROMs measurement properties prevents meta-analysis
- Scarce body of knowledge on MAB outcome measures for shoulder function assessment
- No retrieved PROM or MAB outcome measure superior to any other, except for the WOSI for non-surgically treated shoulder instability
- Investigated properties of MAB outcome measures were generally adequate for their intended purposes.
- MAB outcome measures, including the B-B Score compared equally to PROMs in direct comparisons within pathologies.
- B-B Score was the most extensively investigated MAB outcome measure to date

Specifically considering the benchmarking of the outcome measure developed in this thesis, the B-B Score appeared to be the most extensively validated MAB outcome measure to date. It is the only MAB Score for which the reliability, measurement error, interpretability and responsiveness have all been evaluated (Duc et al., 2014; Coley et al., 2007a; Korver et al., 2014a; Korver et al.; Yang et al., 2014). This statement does not imply that its measurement properties are superior to those of current - or yet to be developed - MAB outcome measures, but allows potential users to rely on the information available in the literature to use the B-B Score and interpret its outcome. As stated in the Phase 3 study's research findings, it was confirmed that its measurement properties were comparable to those of currently used PROMs, except for function evaluation in shoulder instabilities for which the WOSI score was superior to the other investigated outcome measures. This finding reinforces those of previous literature reviews that reported that the measurement properties of the WOSI were superior to those of generic shoulder function outcome measures for shoulder instability (Cacchio et al., 2012; Kirkley et al., 1998; Kirkley et al., 1998).

Concerning practicalities and accessibility, the B-B Score was the only one that had proven to be possibly measured using a smartphone, with similar properties to a dedicated IMU device. Despite the need for future possible improvements, the B-B Score appeared to be well positioned among MAB outcome measures, as it has undergone an extensive validation process in four shoulder pathologies and has practical advantages on its alternative MAB outcome measures. Due to its moderate to high correlation with currently used PROMs and comparable measurement properties, it also appears to be a viable alternative to shoulder function PROMs for the evaluation of rotator cuff conditions, proximal humerus fractures and capsulitis, but not for shoulder instability. This good convergent validity was also found for the P Score, if which the B-B Score is extracted, but not for the accelerometer net vector magnitude data counts, arm underuse percentage, COMP Score, TEMG, Tmov and Temg/mov (Duc et al., 2014; Coley et al., 2007a; Korver et al., 2014a; Korver et al.; Pichonnaz et al. 2015a ; Yang et al., 2014). This highlights the need for MAB outcome measures to respond to certain features to capture shoulder function rather mere movement alterations.

The realisation of the review was a complex issue, due to the need to retrieve a large range of measurement properties for several outcome measures in four pathologies.

This detailed classification was thought to be necessary for the sake of precision and comprehensiveness, as measurement properties are valid only in the context in which they were measured. However, it is questionable to what extent it is necessary to detail the properties of measurement properties for each possible context, as this may make it impossible to synthesize the results and may ultimately calls into question the possibility of generalizing the results.

The difficulties encountered in carrying out the literature review highlighted some inherent limitations to the subject that could not be overcome using current approaches. No quantitative rating of the articles could be performed, due to the lack of an instrument that would have allowed doing so for all investigated measurement properties. The comparison between tools could not be performed based on precise objective criteria, due to the nature of the literature, as only a small proportion of articles directly compared the properties of several PROMS within the same study, and even less directly compared the properties of PROMs and MAB methods. Importantly, the heterogeneity amongst the studied populations that had been investigated, follow-up times and the methods used to calculate measurement properties, had prevented the aggregation of results into a meta-analysis. Therefore, the quality of the body of literature and the results concerning measurement properties could be qualitatively but not quantitatively discussed and critically evaluated. Other authors who had previously addressed the topic also renounced proceeding to a meta-analysis and reported the heterogeneity of the literature (Kirkley et al., 2003; Oh et al., 2009; Huang et al., 2015; Harvie et al., 2005; Fayad et al., 2005; Placzek et al., 2004; Roy et al., 2009). Most of them renounced to conduct a meta-analysis, while Roy et al. have undertaken a weighting of measurement properties across studies. However, this approach was not adopted in this thesis' review, as it does not allow overcoming the issues related to data heterogeneity.

When viewed collectively, all these issues explain why the controversy surrounding the evaluation of shoulder function using PROMs continues to exist despite decades of research on the topic. The characteristics of the current body of literature and the shortcomings of the literature evaluation methods for validation studies make it difficult to provide a clear synthesis on the respective strength and weaknesses of the outcome measures for shoulder function evaluation. This situation is problematic for users, who lack easily interpretable information to make a well-grounded and

informed decision on the choice of a shoulder function outcome measure adapted to their specific needs.

This review nevertheless highlighted some useful issues for the orientation of future research and measurement practice. Despite the substantial number of publications addressing movement analysis, very few have led to the development, let alone the clinical validation of a MAB outcome measure for shoulder function assessment. The body of knowledge remains thus limited in this area, so that the development of shoulder function MAB outcome measures can still be considered as an emerging field. A previous literature review had previously concluded that more research was needed to develop more MAB outcome measures that related to shoulder function (De Baets et at., 2017).

An overall illustration of the thesis accomplishments, which highlights the extent of the thesis achievements compared to the initial process illustrated in Figure 1.4, “Structure of the thesis process”, p. 54, is available in Figure 6.1” Achievements of the thesis process”

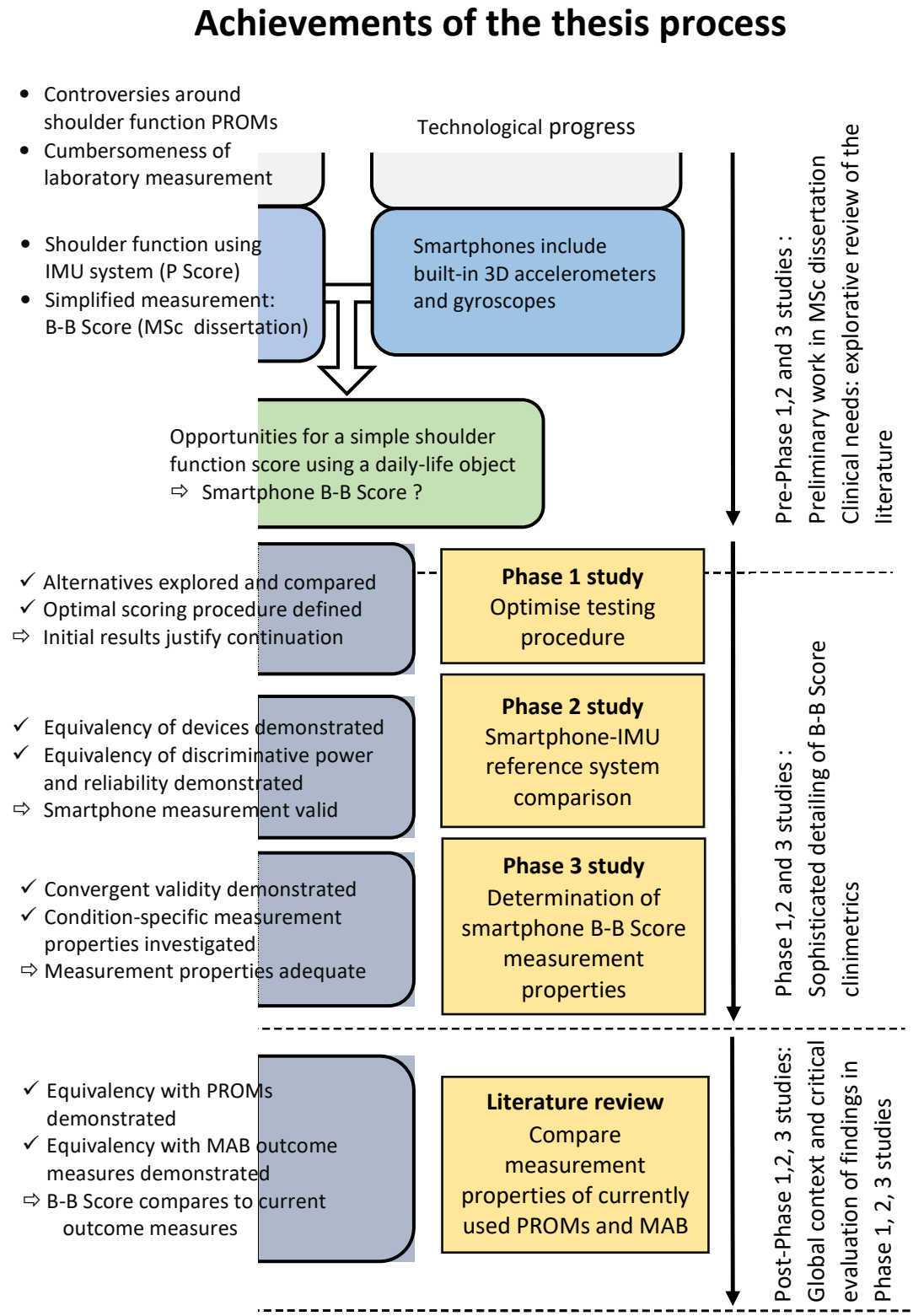


Figure 6.1: Achievement of the thesis process, to compare with Figure 1.4: "Overview of the planned thesis process" within sub-section 1.1.4.4 "Implication of practical issues for the thesis, p. 54.

6.2. Implications of the thesis' findings for clinics and research

6.2.1. Scope of application of the B-B Score

The main outcome of this thesis centres on in the demonstration that the use of a cheap, accessible, valid and straightforward MAB outcome measure is justifiable for clinical and research purposes in rotator cuff conditions, proximal humerus fracture and capsulitis. Previous research had already shown that smartphone measurements are adequate for shoulder ROM evaluation, but this thesis was the first to demonstrate the validity of smartphone measurement for shoulder function evaluation up to now (Cuesta-Vargas, 2016; Johnson et al., 2015; Shin et al. 2012).

The main limitation of the B-B Score for use with these pathologies was the rather large LoAs. As errors were randomly distributed, this might affect the precision of single measurements, but not necessarily of groups' measurements or of the mean outcome of repeated measurements. Due to the lack of data in the literature, it is not possible to interpret if this is a specific shortcoming of the B-B Score, or a more general characteristic of MAB shoulder function outcome measures. Thus, the most suitable applications for the B-B Score concern group-based measurements and the definition of a recovery trend in the follow-up of patients.

As the B-B Score is the only computerised MAB outcome measure applicable for routine measurement, it is of particular interest in situations where laboratory measurements are not possible or where the 'paper and pencil' approach shows limitations (i.e. whenever language, item interpretation, or data communication issues are involved). Thus, the smartphone B-B Score could for example be especially suited for multicentre studies over several countries.

6.2.2. Decision making about shoulder function evaluation

The literature review confirmed the value of computerised MAB outcome measures for shoulder function evaluation, but also highlighted that little research had addressed the conception of clinically usable scores to date. The conclusions of this review might underpin further research in this field. Concerning shoulder function PROMs, the review confirmed the current impossibility to make strong recommendations for the use of one tool over another in a given situation, due to the dispersion of results caused by the heterogeneity of research and calculation methods. The same limitation was reported by previous authors that had conducted

literature reviews on the topic (Kirkley et al., 2003; Oh et al., 2009; Huang et al., 2015; Harvie et al., 2005; Fayad et al., 2005; Placzek et al., 2004; Roy et al., 2009).

This thesis did not bring new solutions on this point. This outcome had been somewhat expected and confirms one of the fundamental issues raised within the thesis' introduction i.e. that the investigation of new pathways is needed to overcome the controversy surrounding shoulder function PROMs. This situation remains problematic for users, who have to rely on inconclusive findings to underpin the choice of an outcome measure. The tables of measurement properties elaborated and detailed within the review, may help them to make an informed choice when selecting a tool for shoulder function evaluation in rotator cuff conditions, humerus fracture, capsulitis, instability and glenohumeral osteoarthritis.

In the present situation, the association of several outcome measures represent the most robust approach. Although any chosen recommendation on this topic may be debated, the use in conjunction of the DASH (extensively validated subjective score), Constant (extensively validated mix of clinical measurements and subjective questions) and B-B Score (objective MAB score) might represent a justifiable approach for research purposes in rotator cuff conditions, proximal humerus fractures and capsulitis. The complementary nature of these outcome measures may provide an effective large-scale overview of shoulder function. For instability, the WOSI was superior to all the other outcome measures evaluated in this thesis. However, further research should still challenge this score with concurrent shoulder instability scores like the Rowe score, Melbourne Instability Shoulder Score (MISS) or Oxford Shoulder Instability Score (OSIS) (Plancher and Lipnick, 2009).

6.3. Suggestions for practice and future research work

6.3.1. Reconsideration of initial assumptions

The basic assumptions underlying this thesis were based on a limited number of available studies, most of which reported results from a small sample. Though these studies constituted the best available evidence at the time of the thesis' conception, they can now be partly reconsidered in the light of the results produced during this work.

The P Score had been taken as a reference for the conception of the B-B Score. This approach was sustainable, as this score had demonstrated promising measurement properties and was the most advanced in its development at the time of the thesis' inception (Coley et al., 2007a; Jolles et al., 2011; Pichonnaz et al., 2015c). The investigations run in the thesis confirmed that the approach of the P Score - using a power-related metric for the evaluation of shoulder function - was sound, as the measurement properties of the B-B score were within expected standards for rotator cuff condition, capsulitis and humerus fracture. As the sample size of the thesis was considerably larger and included conservatively treated patients, contrary to the studies that aimed at the development of the P and B-B Score, the results of the thesis reinforced and extended the basis for the use of a power-related metric for the evaluation of shoulder function.

The soundness of the choice of the two B-B Score movements (hand to the back and hand to the ceiling) and their weighting based on principal component analysis and multiple regressions was also confirmed. Actually, Korver et al. studies, which used the same metric but slightly different movements (hand behind the head instead of hand to the ceiling) and did not weight them, found weak correlations with shoulder function PROMs. This indicated limited ability of this score to capture shoulder function, conversely to the B-B Score (Korver et al., 2014a; Korver et al., 2014b).

This approach had also inherent limitations in that, when taking the P Score as a reference for the conception of the B-B Score, it was at best possible to design a score that closely matched the P Score, but not a score that would have superior measurement properties. Some suggestions for the improvement of the B-B Score will be made in the next subsection. However, it might also be of interest to challenge the background of the B-B Score conception, in order to overcome the limitations of the basic assumptions on which it relies.

Coley's work found that a power-related metric was a better indicator of shoulder function than ROM, especially when the patient is able to reach full ROM but with difficulty (Coley, 2007). However, it would be interesting to investigate if a combination of a power-related metric and ROM would further increase the relationship with other outcome measures of shoulder function.

The B-B Score movements are representative of the difficulties that are reported by patients suffering shoulder function loss (van der Windt et al., 1995; Magermans et al., 2005). Nevertheless, the two movements do not cover all possible shoulder movements. A recent approach that has been investigated consists in taking the hand reachable space to capture shoulder function. This approach has been used either considering the active ROM in various directions based on clinical observation (Riley and al. 2018) or based on computerized movement analysis (LMAM-EPFL, 2018). The observation-based approach showed a relationship with the SPADI shoulder function PROM, while no publication is yet available for the movement analysis based approach.

Instead of making two movements interrupted by a pause in the rest position, it might also be possible to link the two movement in a row and calculate the B-B Score on a single large movement that goes from hand to the back to hand to the ceiling. This would make the test even simpler to perform.

Another possible simplification that deserves investigations could be to perform the movements holding the smartphone in the hand instead of attaching it to the arm with an armband. It should then be checked whether the elbow and wrist movements interfere substantially with the evaluation of shoulder function, but this would save the time used for fixing the armband, which represent approximately half the time required to perform the B-B Score. This would also facilitate the self-evaluation of the patient without supervision.

6.3.2. B-B Score improvement

Further researches may address shortcomings of the B-B Score reported in this thesis, i.e. the magnitude of the LoAs and the lack of validity for shoulder instability measurement.

Concerning the first issue, several investigations could be conducted to contain the extent of measurement variability. The simplest approach could investigate the influence of the number of replications on the B-B Score variability. The variability should theoretically decrease with the square root of the repetitions' number (Mercer and Gleeson, 2002). Although the use of three replications had been retained based on Phase 1 study investigations, because most of the decrease in measurement-to-

measurement variability occurred within this number of replications, it might retrospectively still be of interest to use a higher number of replications.

The optimal number of replications to contain the single measurement variability within the level previously defined as acceptable for clinical measurements (LoA ≤ ± 10%) cannot be inferred from the data collected in the studies conducted for the purpose of the thesis. Based on the above-mentioned theoretical considerations, the single measurement variability should decrease with each added replication, but in a progressively lower proportion for each added replication. Although no carry-over effect between replications was stated in the Phase 1 study (non-significant progressive mean increase of measured values reaching 1.8% between the first and the fifth replication), such effect cannot be excluded using more replications, making that the results might not be in line with the theoretical expectations. For example, it cannot be excluded that a warm-up or a fatigue effect interferes with the results of the measurements when more than five replications are used. The influence of a selectively-used number of the replications should also be explored, like for example discounting the highest and lowest values, taking the mean of the three central values out of a higher number of replications, or omitting the first replication, followed by a reasonable (yet to be determined) number of replicates reflecting random variability.

With regard to the second shortcoming of the B-B Score, investigations concerning the evaluation of shoulder function in shoulder instability would imply to reconsider the Score conception. While the B-B Score was conceived to detect shoulder function alterations at a self-chosen speed in essential movements, these movements are obviously not challenging enough to induce apprehension of dislocation, which is pathognomonic of shoulder instability. The so-called ‘squaring of the circle’ would be to find a solution to generate apprehension without putting the patient at risk of dislocation. The examination of the end of active range of motion in the shoulder instability position (typically combination of flexion, abduction and lateral rotation), e.g. when throwing a ball, could be of particular interest in this situation. This difficulty may explain why no other MAB shoulder function outcome measure has been proposed to date, to the best of our knowledge.

Other Score’s refinements of the B-B Score may be possible using IMU systems but not smartphone-based measurements. The addition of an IMU module on the acromion in order to capture the scapula movement might allow a more precise

location of the source of shoulder dysfunction, because the B-B Score as it conceived currently, captures only the resultant of all involved body segments. This might increase the clinical relevance of measurements as it would allow targeting more precisely the treatment goals according to the degree of involvement of the scapula (De Baets et al., 2017). The reliability of this approach is yet to be demonstrated, because the scapular motion capture using IMU remains a complex issue, especially at the end range of humerus elevation (Coley, 2007; Lempereur et al., 2014). The addition of an IMU module on the trunk might also be of interest to record the movements of the trunk interfering with the shoulder function measurement, and thus be able to discard them from the analyses in order to obtain a purer outcome (Duc et al., 2013).

The study samples have been designed in order to determine the measurement properties of the B-B Score in different shoulder pathologies, but were not large enough to investigate the possible influence of subgroup characteristics on the score. The performance of the healthy population has thus been determined based on a 20-participant sample, which was sufficient to reach the thesis' aims, but not to investigate the possible specificities of subgroups.

Typically, age, sex and dominance might potentially influence the results. As an illustration, such influences have been stated for the Constant, the DASH and the QuickDASH scores, concerning age and sex (Constant, 1986; Yian et al., 2005; Katolik et al., 2005; Aasheim and Finsen, 2014; Hunsaker et al., 2002).

The influence of these two characteristics is less likely for the B-B Score, which compares the performance of the two shoulders. It is theoretically not likely that age or sex might influence mean population symmetry between sides in one of these subpopulation, as aging and sex affects both shoulders in a similar way.

However, the variability in asymmetry might potentially be different within one of the subpopulations, which would affect the range of the score that is considered as normal. For example, the variability of the symmetry might potentially be larger in an older population, due to a possible interaction between age and dominance, which could hypothetically lead to a different age-related performance decline according to shoulder side. This could have an impact on the discriminative power of the score.

Also, the effect on dominance has been considered as negligible in this thesis, due to the absence of a significant difference between sides and to the limited magnitude of the between-side difference observed in the healthy group. However, the establishment of a more precise norm based on a large sample would be of interest to increase the accuracy of the evaluation, as possible population-related factors could be taken into account in the performance assessment. This investigation could rather easily be conducted due to the practicality of the B-B Score.

6.3.3. Possible future research pathways

In the context of this research, movement variability was considered as a drawback that negatively influenced the B-B Score precision. However, movement variability might also be investigated as a parameter of interest for shoulder evaluation. A parallel can be made with the investigations on gait variability that revealed that it was indicative of fall risk (Hausdorff, 2005). Concerning the shoulder, the precise meaning of movement variability still needs to be clarified, though some recent studies have already investigated its relationship with pain and motor control strategies (Mehler et al., 2017; Major et al., 2014; Lopez-Pascual et al., 2017b).

Another pathway could concern the development of MAB outcome measures of shoulder function. As stated in the literature review, only a small fraction of the studies on the shoulder movement was extended by further works to lead to the development of a scoring system. Future researches in this direction may investigate the relevance of either measurements over a short span of time in controlled conditions or several hours' measurements in a free-living environment. For instance, these investigations might be a continuation of the previous works on the area of functional reach, the used arm position in daily life or the shoulder muscular activity in daily life (Hurd et al., 2014; Clement et al., 2018; Duc et al., 2013; Duc et al., 2014; Coley et al., 2008a; Coley et al., 2009).

A recent literature review that included the B-B Score suggested several possible research path for MAB outcome measures, of which certain appear to be relevant to the author of the thesis (De Baets et al., 2017). It was proposed to investigate also 'movement smoothness', 'movement path' and 'trajectory length' to represent the functional status of a joint. It would actually be of interest to investigate if these parameters are mere indicators of movement alterations or more largely indicators of

shoulder function. The review's authors also suggest that taking the thoraco-humeral movement into consideration is not sufficient to capture shoulder function. This statement is questionable, as the relationship of the B-B Score, and of the P Score as well with shoulder function PROMS, could be demonstrated. This was expected as the thoraco-humeral movements is the resultant of the whole shoulder joint complex. The addition of scapulo-humeral movement into the algorithm would surely be of interest to locate more precisely the location of shoulder function alteration, though as the expense of an increased complexity of the testing and analysis procedures. Moreover, this can be more conveniently achieved based on clinical evaluation of each joint of the shoulder complex, as scapulo-humeral movement is itself the resultant of several joints (scaplulo-thoracic, acromio-clavicular, sterno- costo-clavicular joints) that can hardly be analysed separately using computerised movement analysis.

Future researches may also address the on-going controversy on the validity and measurement properties of shoulder function PROMs, though no simple solution is likely to solve the problem. A first useful step would consist of the elaboration of a consensual definition of shoulder function (Roe et al., 2013). A larger consensus should also be reached concerning the recognised methods for establishing measurement properties (Mokkink et al., 2010). To ensure objectivity and admissibility, these consensuses should be developed by large panels of experts under the patronage of an independent organisation. These panels should also recommend orientations for future research on PROMS. Nowadays, it is not clear if the most promising path consists of the improvement of present PROMs or the development of new ones based on a still to be elaborated consensual approach. The degree of specificity of the new PROMs to be developed should also be stipulated, knowing that tools that are more specific will tend to be more valid for the evaluated condition, but that their development would most likely lead eventually to a plethora of shoulder function outcome measures (Michener and Leggin, 2001; Longo et al., 2011; Beaton et al., 2002).

6.3.4. Possible future development pathways

Some technological development projects can also be envisioned. The B-B Score software application stores the data on the smartphone only, in order to prevent all undesired outcome' communication. The results can be communicated only using an email address, typically that of the patient or of a stakeholder that the patient has given permission to be contacted. This simple approach was chosen to prevent the contravention of data protection and professional confidentiality issues during the application's development. Provided that these issues are correctly handled and only with the patient's agreement, further development of the application may improve the possibilities to communicate and centralise the results. For example, it might be possible to inform concerned stakeholders or construct a centralised database for the establishment of norms (e.g. the expected recovery trend or final outcome for various shoulder conditions) or for benchmarking.

Another possible technological development may address the construction of a more comprehensively featured smartphone application oriented toward the health professionals' needs, in which numerous useful and well-validated applications would be accessible in a coordinated interface. The development of smartphone health applications is a growing field, so that it becomes arduous for the user to select the most reliable ones and to find in a timely manner, the ones they have downloaded on their smartphone. Such an application should allow the straightforward activation of meaningful applications for various pathologies, outcomes and body regions, and an easy transfer toward the patient's file. In this context, the B-B Score would be one of the possible applications integrated within the section for shoulder evaluation. Importantly, the conception of such an application should be guided by public health and professional needs considerations.

6.4. Final conclusion

The situation concerning shoulder function evaluation is currently not optimal for clinicians and researchers. On one hand, PROMS have intrinsic limitations and none of them has demonstrated its superiority over the others. On the other hand, no easily accessible MAB outcome measure is available for routine assessment. Thus, this thesis was initiated with the aim to develop and assess the simplest possible MAB shoulder function scoring procedure for clinical measurement.

Following optimisation of the testing procedure and extensive investigations of the measurement properties of the B-B Score, the research' findings allowed the conclusion that this score met the current standards for adequate measurement properties for the evaluation of rotator cuff, shoulder capsulitis or humerus fracture function, either using a dedicated inertial sensor system or a smartphone for measurement. It was concluded from the benchmarking of the B-B Score with the other existing measurement methods that its measurement properties are globally comparable to those of alternative shoulder function measurement methods.

The shortcomings of the B-B Score concerned specifically the clinimetric performances for the assessment of function in shoulder instability and the variability of single measurements. Further research should investigate these issues, either by further refining the B-B Score or by investigating alternatives using simple testing procedures for the evaluation of shoulder function.

Though it can still be improved, the B-B Score already represents a sustainable measurement method for the evaluation of shoulder function in rotator cuff, shoulder capsulitis or humerus fracture. These thesis' results thus demonstrated that a valid MAB shoulder function evaluation can be achieved using a simple procedure and an accessible device. This constitutes a useful contribution to facilitating routine objective evaluation of shoulder function in clinical and research conditions, which is one of the cornerstones of adequate decision-making in patients care.

.

REFERENCES

AASHEIM, T. & FINSEN, V. 2014. The DASH and the QuickDASH instruments. Normative values in the general population in Norway. *J Hand Surg Eur Vol,* 39**,** 140-4.

AMASAY, T., ZODROW, K., KINCL, L., HESS, J. & KARDUNA, A. 2009. Validation of tri-axial accelerometer for the calculation of elevation angles. *International Journal of Industrial Ergonomics,* 39**,** 783-789.

AMERICAN ACADEMY OF ORTHOPAEDIC SURGEONS. 2009. *The DASH Outcome Measure* [Online]. Available: http://www.dash.iwh.on.ca/ (archived on 12 May 2015 at http://www.webcitation.org/6ZEN143eU) [Accessed 23 August 2018].

AMINIAN, K. & NAJAFI, B. 2004. Capturing human motion using body-fixed sensors: outdoor measurement and clinical applications. *Computer Animation and Virtual Worlds,* 15**,** 79-94.

ANAES 2000. Recommandations pratiques pour le diagnostic de la maladie d'Alzheimer. *In:* SERVICE DES RECOMMANDATIONS ET RÉFÉRENCES PROFESSIONNELLES (ed.). Agence National d'Accréditation et d'Evaluation en Santé.

ANDREOPOULOU, G., MERCER, T. H. & VAN DER LINDEN, M. L. 2018. Walking measures to evaluate assistive technology for foot drop in multiple sclerosis: A systematic review of psychometric properties. *Gait Posture,* 61**,** 55-66

ANGST, F. 2011. The new COSMIN guidelines confront traditional concepts of responsiveness. *BMC Med Res Methodol,* 11**,** 152; author reply 152.

ANGST, F., GOLDHAHN, J., DRERUP, S., AESCHLIMANN, A., SCHWYZER, H. K. & SIMMEN, B. R. 2008. Responsiveness of six outcome assessment instruments in total shoulder arthroplasty. *Arthritis Rheum,* 59**,** 391-8.

ANGST, F., GOLDHAHN, J., DRERUP, S., FLURY, M., SCHWYZER, H. K. & SIMMEN, B. R. 2009. How sharp is the short QuickDASH? A refined content and validity analysis of the short form of the disabilities of the shoulder, arm and hand questionnaire in the strata of symptoms and function and specific joint conditions. *Qual Life Res,* 18**,** 1043-51.

ANGST, F., PAP, G., MANNION, A. F., HERREN, D. B., AESCHLIMANN, A., SCHWYZER, H. K. & SIMMEN, B. R. 2004. Comprehensive assessment of clinical outcome and quality of life after total shoulder arthroplasty: usefulness and validity of subjective outcome measures. *Arthritis Rheum,* 51**,** 819-28.

ANGST, F., SCHWYZER, H. K., AESCHLIMANN, A., SIMMEN, B. R. & GOLDHAHN, J. 2011. Measures of adult shoulder function: Disabilities of the Arm, Shoulder, and Hand Questionnaire (DASH) and its short version (QuickDASH), Shoulder Pain and Disability Index (SPADI), American Shoulder and Elbow Surgeons

(ASES) Society standardized shoulder assessment form, Constant (Murley) Score (CS), Simple Shoulder Test (SST), Oxford Shoulder Score (OSS), Shoulder Disability Questionnaire (SDQ), and Western Ontario Shoulder Instability Index (WOSI). *Arthritis Care Res (Hoboken),* 63 Suppl 11**,** S174-88.

BAGULEY, T. 2009. Standardized or simple effect size: what should be reported? *Br J Psychol,* 100**,** 603-17.

BASAR, S., GUNAYDIN, G., HAZAR KANIK, Z., SOZLU, U., ALKAN, Z. B., PALA, O. O., CITAKER, S. & KANATLI, U. 2017. Western Ontario Shoulder Instability Index: cross-cultural adaptation and validation of the Turkish version. *Rheumatol Int,* 37**,** 1559-1565.

BEATON, D. & RICHARDS, R. R. 1998. Assessing the reliability and responsiveness of 5 shoulder questionnaires. *J Shoulder Elbow Surg,* 7**,** 565-72.

BEATON, D. E., BOERS, M. & WELLS, G. A. 2002. Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. *Curr Opin Rheumatol,* 14**,** 109-14.

BEATON, D. E., BOMBARDIER, C., KATZ, J. N., WRIGHT, J. G., WELLS, G., BOERS, M., STRAND, V. & SHEA, B. 2001a. Looking for important change/differences in studies of responsiveness. OMERACT MCID Working Group. Outcome Measures in Rheumatology. Minimal Clinically Important Difference. *J Rheumatol,* 28**,** 400-5.

BEATON, D. E., HOGG-JOHNSON, S. & BOMBARDIER, C. 1997. Evaluating changes in health status: reliability and responsiveness of five generic health status measures in workers with musculoskeletal disorders. *J Clin Epidemiol,* 50**,** 79-93.

BEATON, D. E., KATZ, J. N., FOSSEL, A. H., WRIGHT, J. G., TARASUK, V. & BOMBARDIER, C. 2001b. Measuring the whole or the parts? Validity, reliability, and responsiveness of the Disabilities of the Arm, Shoulder and Hand outcome measure in different regions of the upper extremity. *J Hand Ther,* 14**,** 128-46.

BEATON, D. E. & RICHARDS, R. R. 1996. Measuring function of the shoulder. A cross-sectional comparison of five questionnaires. *J Bone Joint Surg Am,* 78**,** 882-90.

BEATON, D. E., VAN EERD, D., SMITH, P., VAN DER VELDE, G., CULLEN, K., KENNEDY, C. A. & HOGG-JOHNSON, S. 2011. Minimal change is sensitive, less specific to recovery: a diagnostic testing approach to interpretability. *J Clin Epidemiol,* 64**,** 487-96.

BEATON, D. E., WRIGHT, J. G., KATZ, J. N. & UPPER EXTREMITY COLLABORATIVE, G. 2005. Development of the QuickDASH: comparison of three item-reduction approaches. *J Bone Joint Surg Am,* 87**,** 1038-46.

BECKMANN, J. T., HUNG, M., BOUNSANGA, J., WYLIE, J. D., GRANGER, E. K. & TASHJIAN, R. Z. 2015. Psychometric evaluation of the PROMIS Physical Function Computerized Adaptive Test in comparison to the American

Shoulder and Elbow Surgeons score and Simple Shoulder Test in patients with rotator cuff disease. *J Shoulder Elbow Surg,* 24**,** 1961-7.

BÉTHOUX, F. C., PAUL 2003. *Guide des outils de mesure et d'évaluation en Médecine physique et Réadaptation,* Paris, Frison Roche

BLAND, J. & ALTMAN, D. 1986a. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet,* 1**,** 307 - 310.

BLAND, J. M. & ALTMAN, D. G. 1986b. Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement. *Lancet,* 1**,** 307-310.

BLONNA, D., BELLATO, E., CARANZANO, F., BONASIA, D. E., MARMOTTI, A., ROSSI, R. & CASTOLDI, F. 2014. Validity and reliability of the SPORTS score for shoulder instability. *Joints,* 2**,** 59-65.

BLONNA, D., SCELSI, M., MARINI, E., BELLATO, E., TELLINI, A., ROSSI, R., BONASIA, D. E. & CASTOLDI, F. 2012. Can we improve the reliability of the Constant-Murley score? *J Shoulder Elbow Surg,* 21**,** 4-12.

BORSTAD, J. D. & LUDEWIG, P. M. 2002. Comparison of scapular kinematics between elevation and lowering of the arm in the scapular plane. *Clinical Biomechanics,* 17**,** 650-659.

BOT, S. D., TERWEE, C. B., VAN DER WINDT, D. A., BOUTER, L. M., DEKKER, J. & DE VET, H. C. 2004. Clinimetric evaluation of shoulder disability questionnaires: a systematic review of the literature. *Ann Rheum Dis,* 63**,** 335-41.

BROPHY, R. H., BEAUVAIS, R. L., JONES, E. C., CORDASCO, F. A. & MARX, R. G. 2005. Measurement of shoulder activity level. *Clin Orthop Relat Res,* 439**,** 101-8.

CACCHIO, A., NECOZIONE, S., MACDERMID, J. C., ROMPE, J. D., MAFFULLI, N., DI ORIO, F., SANTILLI, V. & PAOLONI, M. 2012. Cross-cultural adaptation and measurement properties of the italian version of the Patient-Rated Tennis Elbow Evaluation (PRTEE) questionnaire. *Phys Ther,* 92**,** 1036-45.

CARRASCO, J. L. & JOVER, L. 2003. Estimating the generalized concordance correlation coefficient through variance components. *Biometrics,* 59**,** 849-858.

CELIK, D. 2016. Turkish version of the modified Constant-Murley score and standardized test protocol: reliability and validity. *Acta Orthop Traumatol Turc,* 50**,** 69-75.

CELIK, D., ATALAR, A. C., DEMIRHAN, M. & DIRICAN, A. 2013. Translation, cultural adaptation, validity and reliability of the Turkish ASES questionnaire. *Knee Surg Sports Traumatol Arthrosc,* 21**,** 2184-9.

CENTER FOR DISEASE CONTROL. 2012. *Anthropometric Reference Data for Children and Adults: United States, 2007–2010* [Online]. Available: https://www.cdc.gov/nchs/data/series/sr_11/sr11_252.pdf (archived at

http://www.webcitation.org/728YVwr8g on 2 September 2018) [Accessed 2 September 2018].

CHANG, A. 2014. *StatsToDo : Sample Size for Receiver Operator Characteristics (ROCs) Program* [Online]. Available: https://www.statstodo.com/SSizROCs_Pgm.php (archived at https://www.statstodo.com/SSizROCs_Pgm.php on 2 September 2018).

CHRISTE, G. 2017. Validité d'un test diagnostique : utilité clinique de la sensibilité, spécificité et rapports de vraisemblance. *Mains Libres***,** 47-52.

CHRISTIANSEN, D. H., FROST, P., FALLA, D., HAAHR, J. P., FRICH, L. H. & SVENDSEN, S. W. 2015. Responsiveness and Minimal Clinically Important Change: A Comparison Between 2 Shoulder Outcome Measures. *J Orthop Sports Phys Ther,* 45**,** 620-5.

CHRISTIANSEN, D. H. F., POUL; FALLA, DEBORAH; HAAHR, JENS PEDER; FRICH, LARS HENRIK; SVENDSEN SUSANNE WULFF 2015. Responsiveness and Minimal Clinically Important Change: A Comparison Between Two Shoulder Outcome Measures. *Journal of Orthopaedic & Sports Physical Therapy,* 45**,** 1-19.

CHRISTIE, A., DAGFINRUD, H., GARRATT, A. M., RINGEN OSNES, H. & HAGEN, K. B. 2011. Identification of shoulder-specific patient acceptable symptom state in patients with rheumatic diseases undergoing shoulder surgery. *J Hand Ther,* 24**,** 53-60; quiz 61.

CHRISTIE, A., HAGEN, K. B., MOWINCKEL, P. & DAGFINRUD, H. 2009. Methodological properties of six shoulder disability measures in patients with rheumatic diseases referred for shoulder surgery. *J Shoulder Elbow Surg,* 18**,** 89-95.

CLARK, C. C. T., BARNES, C. M., STRATTON, G., MCNARRY, M. A., MACKINTOSH, K. A. & SUMMERS, H. D. 2017. A Review of Emerging Analytical Techniques for Objective Physical Activity Measurement in Humans. *Sports Medicine,* 47**,** 439-447.

CLEMENT, J., RAISON, M. & ROULEAU, D. M. 2018. Reproducibility analysis of upper limbs reachable workspace, and effects of acquisition protocol, sex and hand dominancy. *J Biomech,* 68**,** 58-64.

CLINE, M. G., MEREDITH, K. E., BOYER, J. T. & BURROWS, B. 1989. Decline of height with age in adults in a general population sample: estimating maximum height and distinguishing birth cohort effects from actual loss of stature with aging. *Hum Biol,* 61**,** 415-25.

COHEN, J. 1988. *Statistical power analysis for the behavioral sciences,* Hillsdale, Lawrence Earlbaum Associates.

COLEY, B. 2007. *Shoulder function and outcome evaluation after surgery using 3D inertial sensors.* Doctorate ès Sciences, Swiss Institute of Technology.

COLEY, B., JOLLES, B. M., FARRON, A. & AMINIAN, K. 2008a. Arm position during daily activity. *Gait Posture,* 28**,** 581-7.

COLEY, B., JOLLES, B. M., FARRON, A. & AMINIAN, K. 2009. Detection of the movement of the humerus during daily activity. *Med Biol Eng Comput,* 47**,** 467-74.

COLEY, B., JOLLES, B. M., FARRON, A., BOURGEOIS, A., NUSSBAUMER, F., PICHONNAZ, C. & AMINIAN, K. 2007a. Outcome evaluation in shoulder surgery using 3D kinematics sensors. *Gait Posture,* 25**,** 523-32.

COLEY, B., JOLLES, B. M., FARRON, A., PICHONNAZ, C., BASSIN, J. P. & AMINIAN, K. 2008b. Estimating dominant upper-limb segments during daily activity. *Gait Posture,* 27**,** 368-75.

COLLINS, N. J. & ROOS, E. M. 2016. PROMs for Osteoarthritis. *In:* MIEDANY, Y. E. (ed.) *Patient Reported Outcome Measures in Rheumatic Diseases.* Springer.

COLVIN, A. C., EGOROVA, N., HARRISON, A. K., MOSKOWITZ, A. & FLATOW, E. L. 2012. National trends in rotator cuff repair. *J Bone Joint Surg Am,* 94**,** 227-33.

CONBOY, V. B., MORRIS, R. W., KISS, J. & CARR, A. J. 1996. An evaluation of the Constant-Murley shoulder assessment. *J Bone Joint Surg Br,* 78**,** 229-32.

CONSTANT, C. 1986. Age related recovery of shoulder function after surgery [thesis]. *Cork, Ireland: University College,* 1986.

CONSTANT, C. R., GERBER, C., EMERY, R. J., SOJBJERG, J. O., GOHLKE, F. & BOILEAU, P. 2008. A review of the Constant score: modifications and guidelines for its use. *J Shoulder Elbow Surg,* 17**,** 355-61.

CONSTANT, C. R. & MURLEY, A. H. 1987. A clinical method of functional assessment of the shoulder. *Clin Orthop Relat Res,* 214**,** 160-4.

COOK, K. F., RODDEY, T. S., GARTSMAN, G. M. & OLSON, S. L. 2003. Development and psychometric evaluation of the Flexilevel Scale of Shoulder Function. *Med Care,* 41**,** 823-35.

COOK, K. F., RODDEY, T. S., OLSON, S. L., GARTSMAN, G. M., VALENZUELA, F. F. & HANTEN, W. P. 2002. Reliability by surgical status of self-reported outcomes in patients who have shoulder pathologies. *J Orthop Sports Phys Ther,* 32**,** 336-46.

CORONA, K., CERCIELLO, S., MORRIS, B. J., VISONA, E., MEROLLA, G. & PORCELLINI, G. 2016. Cross-cultural adaptation and validation of the Italian version of the Western Ontario Osteoarthritis of the Shoulder index (WOOS). *J Orthop Traumatol*.

CORTINA, J. M. 1993. What is coefficient alpha? An examination of theory and applications. *Journal of applied psychology,* 78**,** 98.

COSMIN. 2010. *COnsensus-based Standards for the selection of health Measurement INstruments* [Online]. Available: http://www.cosmin.nl/ [Accessed 25 October 2017].

COSMIN. 2018a. *COSMIN Risk of Bias checklist* [Online]. Available: https://www.cosmin.nl/wp-content/uploads/COSMIN-RoB-checklist-V2-0-v17_rev3.pdf (archived on 31 August 2018 at http://www.webcitation.org/7250gnbvO) [Accessed 31 August 2018].

COSMIN. 2018b. *COSMIN methodology for assessing the content validity of PROMs* [Online]. Available: https://cosmin.nl/wp-content/uploads/COSMIN-methodology-for-content-validity-user-manual-v1.pdf (archived on 31 August 2018 at http://www.webcitation.org/72lmhfi94) [Accessed 31 August 2018].

COSTA, L. O., LIN, C. W., GROSSI, D. B., MANCINI, M. C., SWISHER, A. K., COOK, C., VAUGHN, D., ELKINS, M. R., SHEIKH, U., MOORE, A., JULL, G., CRAIK, R. L., MAHER, C. G., GUIRRO, R. R., MARQUES, A. P., HARMS, M., BROOKS, D., SIMONEAU, G. G. & STRUPSTAD, J. H. 2012. Clinical trial registration in physiotherapy journals: recommendations from the International Society of Physiotherapy Journal Editors. *J Physiother,* 58**,** 211-3.

COURT-BROWN, C. M. & CAESAR, B. 2006. Epidemiology of adult fractures: A review. *Injury,* 37**,** 691-7.

CUESTA-VARGAS, A. I. & ROLDAN-JIMENEZ, C. 2016. Validity and reliability of arm abduction angle measured on smartphone: a cross-sectional study. *BMC Musculoskelet Disord,* 17**,** 93.

CULHAM, E. & PEAT, M. 1993. Functional anatomy of the shoulder complex. *J Orthop Sports Phys Ther,* 18**,** 342-50.

CUTTI, A. G., GIOVANARDI, A., ROCCHI, L., DAVALLI, A. & SACCHETTI, R. 2008. Ambulatory measurement of shoulder and elbow kinematics through inertial and magnetic sensors. *Medical & Biological Engineering & Computing,* 46**,** 169-178.

CUTTI, A. G., PAREL, I., RAGGI, M., PETRACCI, E., PELLEGRINI, A., ACCARDO, A. P., SACCHETTI, R. & PORCELLINI, G. 2014. Prediction bands and intervals for the scapulo-humeral coordination based on the Bootstrap and two Gaussian methods. *J Biomech,* 47**,** 1035-44.

CIUTI, G., RICOTTI, L., MENCIASSI, A. & DARIO, P. 2015. MEMS sensor technologies for human centred applications in healthcare, physical activities, safety and environmental sensing: a review on research activities in Italy. Sensors (Basel), 15, 6441-68.

DAWSON, J., FITZPATRICK, R. & CARR, A. 1999. The assessment of shoulder instability. The development and validation of a questionnaire. *J Bone Joint Surg Br,* 81**,** 420-6.

DE BAETS, L., VAN DER STRAATEN, R., MATHEVE, T. & TIMMERMANS, A. 2017. Shoulder assessment according to the international classification of

functioning by means of inertial sensor technologies: A systematic review. *Gait Posture,* 57**,** 278-294.

DE BAETS, L., VAN DEUN, S., DESLOOVERE, K. & JASPERS, E. 2013. Dynamic scapular movement analysis: is it feasible and reliable in stroke patients during arm elevation? *PLoS One,* 8**,** e79046.

DE LOS REYES-GUZMAN, A., DIMBWADYO-TERRER, I., TRINCADO-ALONSO, F., MONASTERIO-HUELIN, F., TORRICELLI, D. & GIL-AGUDO, A. 2014. Quantitative assessment based on kinematic measures of functional impairments during upper extremity movements: A review. *Clin Biomech (Bristol, Avon),* 29**,** 719-27.

DE VET, H., TERWEE, C., OSTELO, R., BECKERMAN, H., KNOL, D. & BOUTER, L. 2006a. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. *Health and Quality of Life Outcomes,* 4**,** 54.

DE VET, H. C., TERWEE, C. B. & BOUTER, L. M. 2003. Current challenges in clinimetrics. *J Clin Epidemiol,* 56**,** 1137-41.

DE VET, H. C., TERWEE, C. B., KNOL, D. L. & BOUTER, L. M. 2006b. When to use agreement versus reliability measures. *J Clin Epidemiol,* 59**,** 1033-9.

DE VET, H. C., TERWEE, C. B., MOKKINK, L. B. & KNOL, D. L. 2011a. Interpretability. *In:* TERWEE, C. B., KNOL, D. L., DE VET, H. C. W. & MOKKINK, L. B. (eds.) *Measurement in Medicine: A Practical Guide.* Cambridge: Cambridge University Press.

DE VET, H. C., TERWEE, C. B., MOKKINK, L. B. & KNOL, D. L. 2011b. Reliability. *In:* TERWEE, C. B., KNOL, D. L., DE VET, H. C. W. & MOKKINK, L. B. (eds.) *Measurement in Medicine: A Practical Guide.* Cambridge: Cambridge University Press.

DE VET, H. C., TERWEE, C. B., MOKKINK, L. B. & KNOL, D. L. 2011c. Responsiveness. *In:* TERWEE, C. B., KNOL, D. L., DE VET, H. C. W. & MOKKINK, L. B. (eds.) *Measurement in Medicine: A Practical Guide.* Cambridge: Cambridge University Press.

DE VET, H. C., TERWEE, C. B., MOKKINK, L. B. & KNOL, D. L. 2011d. Systematic reviews of measurement properties. *In:* TERWEE, C. B., KNOL, D. L., DE VET, H. C. W. & MOKKINK, L. B. (eds.) *Measurement in Medicine: A Practical Guide.* Cambridge: Cambridge University Press.

DE VET, H. C., TERWEE, C. B., MOKKINK, L. B. & KNOL, D. L. 2011e. Validity. *In:* TERWEE, C. B., KNOL, D. L., DE VET, H. C. W. & MOKKINK, L. B. (eds.) *Measurement in Medicine: A Practical Guide.* Cambridge: Cambridge University Press.

DE VRIES, W. H., VEEGER, H. E., BATEN, C. T. & VAN DER HELM, F. C. 2016. Can shoulder joint reaction forces be estimated by neural networks? *J Biomech,* 49**,** 73-9.

DE WITTE, P. B., HENSELER, J. F., NAGELS, J., VLIET VLIELAND, T. P. & NELISSEN, R. G. 2012. The Western Ontario rotator cuff index in rotator cuff disease patients: a comprehensive reliability and responsiveness validation study. *Am J Sports Med,* 40**,** 1611-9.

DINIZ LOPES, A., CICONELLI, R. M., CARRERA, E. F., GRIFFIN, S., FALOPPA, F. & BALDY DOS REIS, F. 2009. Comparison of the responsiveness of the Brazilian version of the Western Ontario Rotator Cuff Index (WORC) with DASH, UCLA and SF-36 in patients with rotator cuff disorders. *Clin Exp Rheumatol,* 27**,** 758-64.

DUC, C. 2013. *Objective outcome evaluation of the shoulder and cervical function after surgery using body-fixed sensors.* Doctorate ès Sciences, Swiss Institute of Technology.

DUC, C., FARRON, A., PICHONNAZ, C., JOLLES, B. M., BASSIN, J. P. & AMINIAN, K. 2013. Distribution of arm velocity and frequency of arm usage during daily activity: objective outcome evaluation after shoulder surgery. *Gait Posture,* 38**,** 247-52.

DUC, C., PICHONNAZ, C., BASSIN, J. P., FARRON, A., JOLLES, B. M. & AMINIAN, K. 2014. Evaluation of muscular activity duration in shoulders with rotator cuff tears using inertial sensors and electromyography. *Physiol Meas,* 35**,** 2389-400.

EBRAHIMZADEH, M. H., VAHEDI, E., BARADARAN, A., BIRJANDINEJAD, A., SEYYED-HOSEINIAN, S. H., BAGHERI, F. & KACHOOEI, A. R. 2016. Psychometric Properties of the Persian Version of the Simple Shoulder Test (SST) Questionnaire. *Arch Bone Jt Surg,* 4**,** 387-392.

EL GAAFARY, M. 2016. A Guide to PROMs Methodology and Selection Criteria. *In:* EL MIEDANY, Y. (ed.) *Patient Reported Outcome Measures in Rheumatic Diseases.* Cham: Springer International Publishing.

ENCYCLOPÆDIA BRITANNICA ONLINE. *Kinematics* [Online]. Available: https://www.britannica.com/science/kinematics [Accessed 31 November 2017.

ENCYCLOPÆDIA BRITANNICA ONLINE. *Kinetics* [Online]. Available: https://www.britannica.com/science/kinetics [Accessed 31 November 2017.

EREMENCO, S., PEASE, S., MANN, S., BERRY, P. & ON BEHALF OF THE, P. R. O. C. S. P. S. 2017. Patient-Reported Outcome (PRO) Consortium translation process: consensus development of updated best practices. *Journal of Patient-Reported Outcomes,* 2**,** 12.

EUROQOL, G. 2018. *EQ-5D* [Online]. Available: http://www.webcitation.org/6ytibh0Hk [Accessed 23 April 2018 2018].

FAYAD, F., LEFEVRE-COLAU, M. M., GAUTHERON, V., MACE, Y., FERMANIAN, J., MAYOUX-BENHAMOU, A., ROREN, A., RANNOU, F., ROBY-BRAMI, A., REVEL, M. & POIRAUDEAU, S. 2009. Reliability, validity and responsiveness

of the French version of the questionnaire Quick Disability of the Arm, Shoulder and Hand in shoulder disorders. *Man Ther,* 14**,** 206-12.

FAYAD, F., LEFEVRE-COLAU, M. M., MACE, Y., FERMANIAN, J., MAYOUX-BENHAMOU, A., ROREN, A., RANNOU, F., ROBY-BRAMI, A., GAUTHERON, V., REVEL, M. & POIRAUDEAU, S. 2008a. Validation of the French version of the Disability of the Arm, Shoulder and Hand questionnaire (F-DASH). *Joint Bone Spine,* 75**,** 195-200.

FAYAD, F., LEFEVRE-COLAU, M. M., MACE, Y., GAUTHERON, V., FERMANIAN, J., ROREN, A., ROBY-BRAMI, A., REVEL, M. & POIRAUDEAU, S. 2008b. Responsiveness of the French version of the Disability of the Arm, Shoulder and Hand questionnaire (F-DASH) in patients with orthopaedic and medical shoulder disorders. *Joint Bone Spine,* 75**,** 579-84.

FAYAD, F., MACE, Y. & LEFEVRE-COLAU, M. M. 2005. [Shoulder disability questionnaires: a systematic review]. *Ann Readapt Med Phys,* 48**,** 298-306.

FAYAD, F., MACE, Y., LEFEVRE-COLAU, M. M., POIRAUDEAU, S., RANNOU, F. & REVEL, M. 2004. [Measurement of shoulder disability in the athlete: a systematic review]. *Ann Readapt Med Phys,* 47**,** 389-95.

FEINSTEIN, A. R. 1983. An additional basic science for clinical medicine: IV. The development of clinimetrics. *Ann Intern Med,* 99**,** 843-8.

FENG, D., SVETNIK, V., COIMBRA, A. & BAUMGARTNER, R. 2014. A comparison of confidence interval methods for the concordance correlation coefficient and intraclass correlation coefficient with small number of raters. *J Biopharm Stat,* 24**,** 272-93.

FIALKA, C., OBERLEITNER, G., STAMPFL, P., BRANNATH, W., HEXEL, M. & VECSEI, V. 2005. Modification of the Constant-Murley shoulder score-introduction of the individual relative Constant score Individual shoulder assessment. *Injury,* 36**,** 1159-65.

FOUGEYROLLAS, P., CLOUTIER, R., BERGERON, H., ST-MICHEL, G., CÔTÉ, J., CÔTÉ, M., BOUCHER, N., ROY, K. & RÉMILLARD, M.-B. 1998. *Classification québécoise: Processus de production du handicap*, Québec RIPPH/SCCIDIH.

GAIT UP. 2018. *Hands Up shoulder testing App is now available* [Online]. Available: http://www.webcitation.org/6ytgJhY2N [Accessed 23 April 2018].

GARTSMAN, G. M., MORRIS, B. J., UNGER, R. Z., LAUGHLIN, M. S., ELKOUSY, H. A. & EDWARDS, T. B. 2015. Characteristics of clinical shoulder research over the last decade: a review of shoulder articles in The Journal of Bone & Joint Surgery from 2004 to 2014. *J Bone Joint Surg Am,* 97**,** e26.

GAUDELLI, C., BALG, F., GODBOUT, V., PELET, S., DJAHANGIRI, A., GRIFFIN, S. & ROULEAU, D. M. 2014. Validity, reliability and responsiveness of the French language translation of the Western Ontario Shoulder Instability Index (WOSI). *Orthop Traumatol Surg Res,* 100**,** 99-103.

GE, Y., CHEN, S., CHEN, J., HUA, Y. & LI, Y. 2013. The development and evaluation of a new shoulder scoring system based on the view of patients and physicians: the Fudan University shoulder score. *Arthroscopy,* 29**,** 613-22.

GIAVARINA, D. 2015. Understanding Bland Altman analysis. *Biochem Med (Zagreb),* 25**,** 141-51.

GLEESON, N. P. & MERCER, T. H. 1996. The utility of isokinetic dynamometry in the assessment of human muscle function. *Sports Med,* 2**1,** 18-34.

GODFREY, J., HAMMAN, R., LOWENSTEIN, S., BRIGGS, K. & KOCHER, M. 2007. Reliability, validity, and responsiveness of the simple shoulder test: psychometric properties by age and injury type. *J Shoulder Elbow Surg,* 16**,** 260-7.

GOLDHAHN, J., ANGST, F., DRERUP, S., PAP, G., SIMMEN, B. R. & MANNION, A. F. 2008. Lessons learned during the cross-cultural adaptation of the American Shoulder and Elbow Surgeons shoulder form into German. *J Shoulder Elbow Surg,* 17**,** 248-54.

GRADE HANDBOOK. 2013. *Handbook for grading the quality of evidence and the strength of recommendations using the GRADE approach.* [Online]. Available: https://gdt.gradepro.org/app/handbook/handbook.html (archived on 31 August 2018 at http://www.webcitation.org/72ImfW6fe) [Accessed 7 September 2018].

GREEN, S., BUCHBINDER, R. & HETRICK, S. 2003. Physiotherapy interventions for shoulder pain. *Cochrane Database Syst Rev***,** CD004258.

HALDORSEN, B., SVEGE, I., ROE, Y. & BERGLAND, A. 2014. Reliability and validity of the Norwegian version of the Disabilities of the Arm, Shoulder and Hand questionnaire in patients with shoulder impingement syndrome. *BMC Musculoskelet Disord,* 15**,** 78.

HANDOLL, H. H., ALMAIYAH, M. A. & RANGAN, A. 2004. Surgical versus non-surgical treatment for acute anterior shoulder dislocation. *Cochrane Database Syst Rev***,** CD004325.

HANDOLL, H. H., OLLIVERE, B. J. & ROLLINS, K. E. 2012. Interventions for treating proximal humeral fractures in adults. *Cochrane Database Syst Rev,* 12**,** CD000434.

HANLEY, J. A. & MCNEIL, B. J. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology,* 143**,** 29-36.

HARVIE, P., POLLARD, T. C. B., CHENNAGIRI, R. J. & CARR, A. J. 2005. The use of outcome scores in surgery of the shoulder. *Journal of Bone and Joint Surgery-British Volume,* 87b**,** 151-154.

HAUSDORFF, J. M. 2005. Gait variability: methods, modeling and meaning. *Journal of NeuroEngineering and Rehabilitation,* 2**,** 19-19.

HEFFORD, C., ABBOTT, J. H., ARNOLD, R. & BAXTER, G. D. 2012. The patient-specific functional scale: validity, reliability, and responsiveness in patients with upper extremity musculoskeletal problems. *J Orthop Sports Phys Ther,* 42**,** 56-65.

HENSELER, J. F., KOLK, A., VAN DER ZWAAL, P., NAGELS, J., VLIET VLIELAND, T. P. & NELISSEN, R. G. 2015. The minimal detectable change of the Constant score in impingement, full-thickness tears, and massive rotator cuff tears. *J Shoulder Elbow Surg,* 24**,** 376-81.

HETTRICH CM, D. W., KUHN JE. 2007. Measurement of Shoulder Outcomes. *In:* IANNOTTI, J. P. & WILLIAMS, G. R. (eds.) *Disorders of the shoulder: diagnosis & management.* 2nd ed. Philadelphia, USA: Lippincott Williams & Wilkins.

HINKLE, D. E., WIERSMA, W. & JURS, S. G. 2003. *Applied statistics for the behavioral sciences,* Boston, Houghton Mifflin

HOFSTAETTER, J. G., HANSLIK-SCHNABEL, B., HOFSTAETTER, S. G., WURNIG, C. & HUBER, W. 2010. Cross-cultural adaptation and validation of the German version of the Western Ontario Shoulder Instability index. *Arch Orthop Trauma Surg,* 130**,** 787-96.

HOLMGREN, T., OBERG, B., ADOLFSSON, L., BJORNSSON HALLGREN, H. & JOHANSSON, K. 2014. Minimal important changes in the Constant-Murley score in patients with subacromial pain. *J Shoulder Elbow Surg,* 23**,** 1083-90.

HOLTBY, R. & RAZMJOU, H. 2005. Measurement properties of the Western Ontario rotator cuff outcome measure: a preliminary report. *J Shoulder Elbow Surg,* 14**,** 506-10.

HORN, K. K., JENNINGS, S., RICHARDSON, G., VLIET, D. V., HEFFORD, C. & ABBOTT, J. H. 2012. The patient-specific functional scale: psychometrics, clinimetrics, and application as a clinical outcome measure. *J Orthop Sports Phys Ther,* 42**,** 30-42.

HUANG, H., GRANT, J. A., MILLER, B. S., MIRZA, F. M. & GAGNIER, J. J. 2015. A Systematic Review of the Psychometric Properties of Patient-Reported Outcome Instruments for Use in Patients With Rotator Cuff Disease. *Am J Sports Med,* 43**,** 2572-82.

HUDAK, P. L., AMADIO, P. C. & BOMBARDIER, C. 1996. Development of an upper extremity outcome measure: the DASH (disabilities of the arm, shoulder and hand) [corrected]. The Upper Extremity Collaborative Group (UECG). *Am J Ind Med,* 29**,** 602-8.

HUNSAKER, F. G., CIOFFI, D. A., AMADIO, P. C., WRIGHT, J. G. & CAUGHLIN, B. 2002. The American academy of orthopaedic surgeons outcomes instruments: normative values from the general population. *J Bone Joint Surg Am,* 84-A**,** 208-15.

HURD, W. J., MORROW, M. M., MILLER, E. J., ADAMS, R. A., SPERLING, J. W. & KAUFMAN, K. R. 2014. Novel approaches to objectively assess shoulder function. *J Shoulder Elbow Surg,* 23**,** e251-5.

HURD, W. J., MORROW, M. M., MILLER, E. J., ADAMS, R. A., SPERLING, J. W. & KAUFMAN, K. R. 2017. Patient-Reported and Objectively Measured Function Before and After Reverse Shoulder Arthroplasty. *J Geriatr Phys Ther*.

HUSTED, J. A., COOK, R. J., FAREWELL, V. T. & GLADMAN, D. D. 2000. Methods for assessing responsiveness: a critical review and recommendations. *J Clin Epidemiol,* 53**,** 459-68.

INSTITUTE FOR WORK & HEALTH. *About the QuickDASH* [Online]. Available: http://www.dash.iwh.on.ca/about-quickdash [Accessed 27 October 2017].

JENSEN, K. U., BONGAERTS, G., BRUHN, R. & SCHNEIDER, S. 2009. Not all Rowe scores are the same! Which Rowe score do you use? *J Shoulder Elbow Surg,* 18**,** 511-4.

JIMERSON, S. R. B., MATTHEW K; VANDERHEYDEN, AMANDA 2007. *Handbook of response to intervention: The science and practice of assessment and intervention*, Springer Science & Business Media.

JOHNSON, L. B., SUMNER, S., DUONG, T., YAN, P., BAJCSY, R., ABRESCH, R. T., DE BIE, E. & HAN, J. J. 2015. Validity and reliability of smartphone magnetometer-based goniometer evaluation of shoulder abduction--A pilot study. *Man Ther,* 20**,** 777-82.

JOLLES, B., AMINIAN, K., COLEY, B., DUC, C., PICHONNAZ, C., BASSIN, J. & FARRON, A. 2010. Mission: observer l'épaule. *Forum Med Suisse,* 10**,** 48.

JOLLES, B. M., DUC, C., COLEY, B., AMINIAN, K., PICHONNAZ, C., BASSIN, J. P. & FARRON, A. 2011. Objective evaluation of shoulder function using body-fixed sensors: a new way to detect early treatment failures? *J Shoulder Elbow Surg,* 20**,** 1074-81.

KAPANDJI, I. 1971. The Physiology of the Joints, Volume I, Upper Limb. *American Journal of Physical Medicine & Rehabilitation,* 50**,** 96.

KATOLIK, L. I., ROMEO, A. A., COLE, B. J., VERMA, N. N., HAYDEN, J. K. & BACH, B. R. 2005. Normalization of the Constant score. *J Shoulder Elbow Surg,* 14**,** 279-85.

KELLEY, M. J., SHAFFER, M. A., KUHN, J. E., MICHENER, L. A., SEITZ, A. L., UHL, T. L., GODGES, J. J. & MCCLURE, P. W. 2013. Shoulder pain and mobility deficits: adhesive capsulitis. *J Orthop Sports Phys Ther,* 43**,** A1-31.

KEMP, K. A., SHEPS, D. M., BEAUPRE, L. A., STYLES-TRIPP, F., LUCIAK-COREA, C. & BALYK, R. 2012. An evaluation of the responsiveness and discriminant validity of shoulder questionnaires among patients receiving surgical correction of shoulder instability. *ScientificWorldJournal,* 2012**,** 410125.

KENNEDY, C. A., BEATON, D. E., SMITH, P., VAN EERD, D., TANG, K., INRIG, T., HOGG-JOHNSON, S., LINTON, D. & COUBAN, R. 2013. Measurement properties of the QuickDASH (disabilities of the arm, shoulder and hand) outcome measure and cross-cultural adaptations of the QuickDASH: a systematic review. Qual Life Res, 22, 2509-47.

KHADILKAR, L., MACDERMID, J. C., SINDEN, K. E., JENKYN, T. R., BIRMINGHAM, T. B. & ATHWAL, G. S. 2014. An analysis of functional shoulder movements during task performance using Dartfish movement analysis software. *Int J Shoulder Surg,* 8**,** 1-9.

KIBLER, W. B., LUDEWIG, P. M., MCCLURE, P., UHL, T. L. & SCIASCIA, A. 2009. Scapular Summit 2009: introduction. July 16, 2009, Lexington, Kentucky. *J Orthop Sports Phys Ther,* 39**,** A1-A13.

KIBLER, W. B., LUDEWIG, P. M., MCCLURE, P. W., MICHENER, L. A., BAK, K. & SCIASCIA, A. D. 2013. Clinical implications of scapular dyskinesis in shoulder injury: the 2013 consensus statement from the 'Scapular Summit'. *Br J Sports Med,* 47**,** 877-85.

KING, M. T. 2011. A point of minimal important difference (MID): a critique of terminology and methods. *Expert Rev Pharmacoecon Outcomes Res,* 11**,** 171-84.

KIRKLEY, A., GRIFFIN, S. & DAINTY, K. 2003. Scoring systems for the functional assessment of the shoulder. *Arthroscopy,* 19**,** 1109-20.

KIRKLEY, A., GRIFFIN, S., MCLINTOCK, H. & NG, L. 1998. The development and evaluation of a disease-specific quality of life measurement tool for shoulder instability. The Western Ontario Shoulder Instability Index (WOSI). *Am J Sports Med,* 26**,** 764-72.

KOCHER, M. S., HORAN, M. P., BRIGGS, K. K., RICHARDSON, T. R., O'HOLLERAN, J. & HAWKINS, R. J. 2005. Reliability, validity, and responsiveness of the American Shoulder and Elbow Surgeons subjective shoulder scale in patients with shoulder instability, rotator cuff disease, and glenohumeral arthritis. *J Bone Joint Surg Am,* 87**,** 2006-11.

KOLBER, M. J., SALAMH, P. A., HANNEY, W. J. & SAMUEL CHENG, M. 2013. Clinimetric evaluation of the disabilities of the arm, shoulder, and hand (DASH) andQuickDASH questionnaires for patients with shoulder disorders. *Physical Therapy Reviews,* 19**,** 163-173.

KOO, T. K. & LI, M. Y. 2016. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine,* 15**,** 155-163.

KORVER, R. J., HEYLIGERS, I. C., SAMIJO, S. K. & GRIMM, B. 2014a. Inertia based functional scoring of the shoulder in clinical practice. *Physiol Meas,* 35**,** 167-76.

KORVER, R. J., SENDEN, R., HEYLIGERS, I. C. & GRIMM, B. 2014b. Objective outcome evaluation using inertial sensors in subacromial impingement syndrome: a five-year follow-up study. *Physiol Meas,* 35**,** 677-86.

KOTTNER, J., AUDIGE, L., BRORSON, S., DONNER, A., GAJEWSKI, B. J., HROBJARTSSON, A., ROBERTS, C., SHOUKRI, M. & STREINER, D. L. 2011. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *J Clin Epidemiol,* 64**,** 96-106.

KRUEGER, D., KRAUS, N., PAULY, S., CHEN, J. & SCHEIBEL, M. 2011. Subjective and objective outcome after revision arthroscopic stabilization for recurrent anterior instability versus initial shoulder stabilization. *Am J Sports Med,* 39**,** 71-7.

KUKKONEN, J., KAUKO, T., VAHLBERG, T., JOUKAINEN, A. & AARIMAA, V. 2013. Investigating minimal clinically important difference for Constant score in patients undergoing rotator cuff surgery. *J Shoulder Elbow Surg,* 22**,** 1650-5.

KVIEN, T. K., HEIBERG, T. & HAGEN, K. B. 2007. Minimal clinically important improvement/difference (MCII/MCID) and patient acceptable symptom state (PASS): what do these concepts mean? *Ann Rheum Dis,* 66 Suppl 3**,** iii40-1.

LABORATORY OF MOVEMENT ANALYSIS AND MEASUREMENT—SWISS INSTITUTE OF TECHNOLOGY OF LAUSANNE. 2016. *Smartphone App iShould* [Online]. Available: http://lmam.epfl.ch/smartphone/ishould (archived at URL http://www.webcitation.org/6yj35kZhv on 16 April 2018) [Accessed 16 April 2018].

LMAM-EPFL - LABORATORY OF MOVEMENT ANALYSIS AND MEASUREMENT—SWISS INSTITUTE OF TECHNOLOGY OF LAUSANNE. 2018. *Shoulder test* [Online]. Available: https://lmam.epfl.ch/page-125471-en-html/page-125544-en-html/page-125555-en-html/ (archived at URL http://www.webcitation.org/6yj35kZhv on 24 July 2019) [Accessed July 24 2019]

LAMAL 1994. Art. 32 de la loi fédérale sur l'assurance-maladie du 18 mars 1994 (= LAMal ; RS 832.10).

LANDIS, J. R. & KOCH, G. G. 1977. The measurement of observer agreement for categorical data. *Biometrics,* 33**,** 159-74.

LEFÈVRE-COLAU, M.-M., NGUYEN, C., PALAZZO, C., SROUR, F., PARIS, G., VUILLEMIN, V., POIRAUDEAU, S., ROBY-BRAMI, A. & ROREN, A. 2017. Recent advances in kinematics of the shoulder complex in healthy people. *Annals of Physical and Rehabilitation Medicine*.

LEMPEREUR, M., BROCHARD, S., LEBOEUF, F. & REMY-NERIS, O. 2014. Validity and reliability of 3D marker based scapular motion analysis: a systematic review. *J Biomech,* 47**,** 2219-30.

LENTH, R. V. 2010. *Java applets for power and sample size* [Online]. Available: http://www.webcitation.org/6ZEMrvmpu [Accessed Archived on 12 May 2015 2010].

LIAVAAG, S., SVENNINGSEN, S., REIKERAS, O., ENGER, M., FJALESTAD, T., PRIPP, A. H. & BROX, J. I. 2011. The epidemiology of shoulder dislocations in Oslo. *Scand J Med Sci Sports,* 21**,** e334-40.

LIM, J. Y., KIM, T. H. & LEE, J. S. 2015. Reliability of measuring the passive range of shoulder horizontal adduction using a smartphone in the supine versus the side-lying position. *J Phys Ther Sci,* 27**,** 3119-22

LINKEL, A., RAUDONYTE, I., SHIPPEN, J., MAY, B., DAUNORAVICIENE, K., SAWICKI, A. & GRISKEVICIUS, J. 2017. Intrapersonal and interpersonal evaluation of upper extremity kinematics. *Technol Health Care,* 25**,** 939-948.

LIPPITT, S. B. H., D. T.; MATSEN, F. A. 1993. A practical tool for evaluating function: the Simple Shoulder Test. *In:* MATSEN PA, F., FH. ; HAWKINS, RJ (ed.) *The shoulder: a balance of mobility and stability.* Rosemont: American Academy of Orthopaedic Surgery.

LITTLEWOOD, C. & COOLS, A. M. J. 2017. Scapular dyskinesis and shoulder pain: the devil is in the detail. *British Journal of Sports Medicine*.

LIVAIN, T., PICHON, H., VERMEULEN, J., VAILLANT, J., SARAGAGLIA, D., POISSON, M. F. & MONNET, S. 2007. [Intra- and interobserver reproducibility of the French version of the Constant-Murley shoulder assessment during rehabilitation after rotator cuff surgery]. *Rev Chir Orthop Reparatrice Appar Mot,* 93**,** 142-9.

LO, I. K., GRIFFIN, S. & KIRKLEY, A. 2001. The development of a disease-specific quality of life measurement tool for osteoarthritis of the shoulder: The Western Ontario Osteoarthritis of the Shoulder (WOOS) index. *Osteoarthritis Cartilage,* 9**,** 771-8.

LONGO, U. G., VASTA, S., MAFFULLI, N. & DENARO, V. 2011. Scoring systems for the functional assessment of patients with rotator cuff pathology. *Sports Med Arthrosc,* 19**,** 310-20.

LOPES, A. D., TIMMONS, M. K., GROVER, M., CICONELLI, R. M. & MICHENER, L. A. 2015. Visual scapular dyskinesis: kinematics and muscle activity alterations in patients with subacromial impingement syndrome. *Arch Phys Med Rehabil,* 96**,** 298-306.

LOPEZ-PASCUAL, J., PAGE, A. & SERRA-ANO, P. 2017a. Dynamic thoracohumeral kinematics are dependent upon the etiology of the shoulder injury. *PLoS One,* 12**,** e0183954.

LOPEZ-PASCUAL, J., PAGE, A. & SERRA-ANO, P. 2017b. Movement Variability Increases With Shoulder Pain When Compensatory Strategies of the Upper Body Are Constrained. *J Mot Behav***,** 1-7.

LUDEWIG, P. M. & COOK, T. M. 2000. Alterations in shoulder kinematics and associated muscle activity in people with symptoms of shoulder impingement. *Phys Ther,* 80**,** 276-91.

LUDEWIG, P. M., PHADKE, V., BRAMAN, J. P., HASSETT, D. R., CIEMINSKI, C. J. & LAPRADE, R. F. 2009. Motion of the shoulder complex during multiplanar humeral elevation. *J Bone Joint Surg Am,* 91**,** 378-89.

LUDEWIG, P. M. & REYNOLDS, J. F. 2009. The association of scapular kinematics and glenohumeral joint pathologies. *J Orthop Sports Phys Ther,* 39**,** 90-104.

LUINGE, H. J. & VELTINK, P. H. 2005. Measuring orientation of human body segments using miniature gyroscopes and accelerometers. *Medical & Biological Engineering & Computing,* 43**,** 273-282.

LUINGE, H. J., VELTINK, P. H. & BATEN, C. T. 2007. Ambulatory measurement of arm orientation. *J Biomech,* 40**,** 78-85.

LUNDQUIST, C. B., DOSSING, K. & CHRISTIANSEN, D. H. 2014. Responsiveness of a Danish version of the Disabilities of the Arm, Shoulder and Hand (DASH) questionnaire. *Dan Med J,* 61**,** A4813.

MACDERMID, J. C., DROSDOWECH, D. & FABER, K. 2006. Responsiveness of self-report scales in patients recovering from rotator cuff surgery. *J Shoulder Elbow Surg,* 15**,** 407-14.

MACDERMID, J. C., KHADILKAR, L., BIRMINGHAM, T. B. & ATHWAL, G. S. 2015. Validity of the QuickDASH in patients with shoulder-related disorders undergoing surgery. *J Orthop Sports Phys Ther,* 45**,** 25-36.

MAGERMANS, D. J., CHADWICK, E. K., VEEGER, H. E. & VAN DER HELM, F. C. 2005. Requirements for upper extremity motions during activities of daily living. *Clin Biomech (Bristol, Avon),* 20**,** 591-9.

MAHABIER, K. C., DEN HARTOG, D., THEYSKENS, N., VERHOFSTAD, M. H. J., VAN LIESHOUT, E. M. M. & INVESTIGATORS, H. T. 2017. Reliability, validity, responsiveness, and minimal important change of the Disabilities of the Arm, Shoulder and Hand and Constant-Murley scores in patients with a humeral shaft fracture. *J Shoulder Elbow Surg,* 26**,** e1-e12.

MAJOR, M. J., STINE, R. L., HECKATHORNE, C. W., FATONE, S. & GARD, S. A. 2014. Comparison of range-of-motion and variability in upper body movements between transradial prosthesis users and able-bodied controls when executing goal-oriented tasks. *J Neuroeng Rehabil,* 11**,** 132.

MAKHNI, E. C., STEINHAUS, M. E., MORROW, Z. S., JOBIN, C. M., VERMA, N. N., COLE, B. J. & BACH, B. R., JR. 2015. Outcomes assessment in rotator cuff pathology: what are we measuring? *J Shoulder Elbow Surg,* 24**,** 2008-15.

MARK, D. N., JACK; LAMARCHE, JEFF 2011. *Beginning iOS 5 Development: Exploring the iOS SDK*, Apress.

MATSEN, F. A., 3RD, TANG, A., RUSS, S. M. & HSU, J. E. 2017. Relationship Between Patient-Reported Assessment of Shoulder Function and Objective Range-of-Motion Measurements. *J Bone Joint Surg Am,* 99**,** 417-426.

MATSUI, K., SHIMADA, K. & ANDREW, P. D. 2006. Deviation of skin marker from bone target during movement of the scapula. *J Orthop Sci,* 11**,** 180-4.

MCDOWELL, I. 2006. *Measuring health: a guide to rating scales and questionnaires*, Oxford University Press.

MCHORNEY, C. A. & TARLOV, A. R. 1995. Individual-patient monitoring in clinical practice: are available health status surveys adequate? *Qual Life Res,* 4**,** 293-307.

MEHLER, D. M. A., REICHENBACH, A., KLEIN, J. & DIEDRICHSEN, J. 2017. Minimizing endpoint variability through reinforcement learning during reaching movements involving shoulder, elbow and wrist. *PLoS One,* 12**,** e0180803.

MEHTA, S. P., TIRUTTANI, R., KAUR, M. N., MACDERMID, J. & KARIM, R. 2015. Psychometric Properties of the Hindi Version of the Disabilities of Arm, Shoulder, and Hand: A Pilot Study. *Rehabil Res Pract,* 2015**,** 482378.

MEMBRILLA-MESA, M. D., CUESTA-VARGAS, A. I., POZUELO-CALVO, R., TEJERO-FERNANDEZ, V., MARTIN-MARTIN, L. & ARROYO-MORALES, M. 2015a. Shoulder pain and disability index: cross cultural validation and evaluation of psychometric properties of the Spanish version. *Health Qual Life Outcomes,* 13**,** 200.

MEMBRILLA-MESA, M. D., TEJERO-FERNANDEZ, V., CUESTA-VARGAS, A. I. & ARROYO-MORALES, M. 2015b. Validation and reliability of a Spanish version of Simple Shoulder Test (SST-Sp). *Qual Life Res,* 24**,** 411-6.

MERCER, T. H. & GLEESON, N. P. 2002. The efficacy of measurement and evaluation in evidence-based clinical practice. *Physical Therapy in Sport,* 3**,** 27-36.

MICHENER, L. A. 2011. Patient- and clinician-rated outcome measures for clinical decision making in rehabilitation. *J Sport Rehabil,* 20**,** 37-45.

MICHENER, L. A. & LEGGIN, B. G. 2001. A review of self-report scales for the assessment of functional limitation and disability of the shoulder. *J Hand Ther,* 14**,** 68-76.

MICHENER, L. A., MCCLURE, P. W. & SENNETT, B. J. 2002. American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form, patient self-report section: reliability, validity, and responsiveness. *J Shoulder Elbow Surg,* 11**,** 587-94.

MINTKEN, P. E., GLYNN, P. & CLELAND, J. A. 2009. Psychometric properties of the shortened disabilities of the Arm, Shoulder, and Hand Questionnaire (QuickDASH) and Numeric Pain Rating Scale in patients with shoulder pain. *J Shoulder Elbow Surg,* 18**,** 920-6.

MITCHELL, C., ADEBAJO, A., HAY, E. & CARR, A. 2005. Shoulder pain: diagnosis and management in primary care. *BMJ,* 331**,** 1124-8.

MITCHELL, K., GUTIERREZ, S. B., SUTTON, S., MORTON, S. & MORGENTHALER, A. 2014. Reliability and validity of goniometric iPhone applications for the assessment of active shoulder external rotation. *Physiother Theory Pract,* 30**,** 521-5.

MOELLER, A. D., THORSEN, R. R., TORABI, T. P., BJOERKMAN, A. S., CHRISTENSEN, E. H., MARIBO, T. & CHRISTIANSEN, D. H. 2014. The Danish version of the modified Constant-Murley shoulder score: reliability, agreement, and construct validity. *J Orthop Sports Phys Ther,* 44**,** 336-40.

MOKKINK, L. B., DE VET, H. C. W., PRINSEN, C. A. C., PATRICK, D. L., ALONSO, J., BOUTER, L. M. & TERWEE, C. B. 2018. COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. *Quality of Life Research,* 27**,** 1171-1179.

MOKKINK, L. B., TERWEE, C. B., GIBBONS, E., STRATFORD, P. W., ALONSO, J., PATRICK, D. L., KNOL, D. L., BOUTER, L. M. & DE VET, H. C. 2010a. Inter-rater agreement and reliability of the COSMIN (COnsensus-based Standards for the selection of health status Measurement Instruments) Checklist. *BMC Medical Research Methodology,* 10**,** 82.

MOKKINK, L. B., TERWEE, C. B., KNOL, D. L., STRATFORD, P. W., ALONSO, J., PATRICK, D. L., BOUTER, L. M. & DE VET, H. C. W. 2010b. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. *BMC Med Res Methodol,* 10.

MOKKINK, L. B., TERWEE, C. B., PATRICK, D. L., ALONSO, J., STRATFORD, P. W., KNOL, D. L., BOUTER, L. M. & DE VET, H. C. 2010c. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Qual Life Res,* 19**,** 539-49.

MOKKINK, L. B., TERWEE, C. B., PATRICK, D. L., ALONSO, J., STRATFORD, P. W., KNOL, D. L., BOUTER, L. M. & DE VET, H. C. 2010d. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol,* 63**,** 737-45.

MOKKINK, L. B., TERWEE, C. B., PATRICK, D. L., ALONSO, J., STRATFORD, P. W., KNOL, D. L., BOUTER, L. M. & DE VET, H. C. 2010e. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol,* 63.

MOOSMAYER, S., SMITH, H. J., TARIQ, R. & LARMO, A. 2009. Prevalence and characteristics of asymptomatic tears of the rotator cuff: an ultrasonographic and clinical study. *J Bone Joint Surg Br,* 91**,** 196-200.

MOSER, A. D., KNAUT, L. A., ZOTZ, T. G. & SCHARAN, K. O. 2012. Validity and reliability of the Portuguese version of the American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form. *Rev Bras Reumatol,* 52**,** 348-56.

MOUSTGAARD, H., BELLO, S., MILLER, F. G. & HROBJARTSSON, A. 2014. Subjective and objective outcomes in randomized clinical trials: definitions differed in methods publications and were often absent from trial reports. *J Clin Epidemiol,* 67**,** 1327-34.

MUNRO, B. H. 2005. *Statistical methods for health care research,* Philadelphia, Lippincott Williams & Wilkins.

NAGHDI, S., NAKHOSTIN ANSARI, N., RUSTAIE, N., AKBARI, M., EBADI, S., SENOBARI, M. & HASSON, S. 2015. Simple shoulder test and Oxford Shoulder Score: Persian translation and cross-cultural validation. *Arch Orthop Trauma Surg,* 135**,** 1707-18.

NEGAHBAN, H., BEHTASH, Z., SOHANI, S. M. & SALEHI, R. 2015. Responsiveness of two Persian-versions of shoulder outcome measures following physiotherapy intervention in patients with shoulder disorders. *Disabil Rehabil,* 37**,** 2300-4.

NENDAZ, M. R. & PERRIER, A. 2004. [Sensitivity, specificity, positive and negative predictive value of a diagnostic test]. *Rev Mal Respir,* 21**,** 390-3.

NETO, J. O., GESSER, R. L., STEGLICH, V., BONILAURI FERREIRA, A. P., GANDHI, M., VISSOCI, J. R. & PIETROBON, R. 2013. Validation of the Simple Shoulder Test in a Portuguese-Brazilian population. Is the latent variable structure and validation of the Simple Shoulder Test Stable across cultures? *PLoS One,* 8**,** e62890.

NORDIN, M. & FRANKEL, V. H. 2001. *Basic biomechanics of the musculoskeletal system*, Lippincott Williams & Wilkins.

O'CONNOR, D. A., CHIPCHASE, L. S., TOMLINSON, J. & KRISHNAN, J. 1999. Arthroscopic subacromial decompression: responsiveness of disease-specific and health-related quality of life outcome measures. *Arthroscopy,* 15**,** 836-40.

O'KANE, J. W. & TORESDAHL, B. G. 2014. The evidenced-based shoulder evaluation. *Curr Sports Med Rep,* 13**,** 307-13.

OH, J. H., JO, K. H., KIM, W. S., GONG, H. S., HAN, S. G. & KIM, Y. H. 2009. Comparative evaluation of the measurement properties of various shoulder outcome instruments. *Am J Sports Med,* 37**,** 1161-8.

OÏHÉNART, L., DUC, C. & AMINIAN, K. 2012. iShould: Functional evaluation of the shoulder using a Smartphone. *Gait & Posture,* 36**,** S61-S62.

OLLEY, L. M. & CARR, A. J. 2008. The use of a patient-based questionnaire (the Oxford Shoulder Score) to assess outcome after rotator cuff repair. *Ann R Coll Surg Engl,* 90**,** 326-31.

OWENS, B. D., DUFFEY, M. L., NELSON, B. J., DEBERARDINO, T. M., TAYLOR, D. C. & MOUNTCASTLE, S. B. 2007. The incidence and characteristics of shoulder instability at the United States Military Academy. *Am J Sports Med,* 35**,** 1168-73.

PAGE, M. J., MCKENZIE, J. E., GREEN, S. E., BEATON, D. E., JAIN, N. B., LENZA, M., VERHAGEN, A. P., SURACE, S., DEITCH, J. & BUCHBINDER, R. 2015. Core domain and outcome measurement sets for shoulder pain trials are needed: systematic review of physical therapy trials. *J Clin Epidemiol,* 68**,** 1270-81.

PANDYAN, A. P. V. W., FREDERIKE; JOHNSON, GARTH R.; GREENFIELD, TONY; TONY, GREENFIELD 2002. Instrumentation in experimentation. *Research methods for postgraduates.* London: Arnold.

PATEL, A. A., DONEGAN, D. & ALBERT, T. 2007. The 36-item short form. *J Am Acad Orthop Surg,* 15**,** 126-34.

PICAVET, H. S. & SCHOUTEN, J. S. 2003. Musculoskeletal pain in the Netherlands: prevalences, consequences and risk groups, the DMC(3)-study. *Pain,* 102**,** 167-78.

PICHONNAZ, C. 2009. *Clinical measurement of arm elevation with accelerometers.* MSc Module validation work, Queen Margaret University.

PICHONNAZ, C. 2010. *Development of a kinematic functional shoulder test including only essential movements.* Master of Sciences in physiotherapy MSc dissertation, Queen Margaret University, Edinburgh.

PICHONNAZ, C., AMINIAN, K., ANCEY, C., JACCARD, H., LECUREUX, E., DUC, C., FARRON, A., JOLLES, B. M. & GLEESON, N. 2017. Heightened clinical utility of smartphone versus body-worn inertial system for shoulder function B-B score. *PLoS One,* 12**,** e0174365.

PICHONNAZ, C., DUC, C., GLEESON, N., ANCEY, C., JACCARD, H., LECUREUX, E., FARRON, A., JOLLES, B. M. & AMINIAN, K. 2015a. Measurement properties of the smartphone-based B-B Score in current shoulder pathologies. *Sensors (Basel),* 15**,** 26801-17.

PICHONNAZ, C., DUC, C., JOLLES, B. M., AMINIAN, K., BASSIN, J. P. & FARRON, A. 2015b. Alteration and recovery of arm usage in daily activities after rotator cuff surgery. *J Shoulder Elbow Surg,* 24**,** 1346-52.

PICHONNAZ, C., LECUREUX, E., BASSIN, J. P., DUC, C., FARRON, A., AMINIAN, K., JOLLES, B. M. & GLEESON, N. 2015c. Enhancing clinically-relevant shoulder function assessment using only essential movements. *Physiol Meas,* 36**,** 547-60.

PINES, J. M., CARPENTER, C. R., RAJA, A. S. & SCHUUR, J. D. 2012. *Evidence-based emergency care: diagnostic testing and clinical decision rules,* Chichester, John Wiley & Sons.

PLACZEK, J. D., LUKENS, S. C., BADALANMENTI, S., ROUBAL, P. J., FREEMAN, D. C., WALLEMAN, K. M., PARROT, A. & WIATER, J. M. 2004. Shoulder outcome measures: a comparison of 6 functional tests. *Am J Sports Med,* 32**,** 1270-7.

PLANCHER, K. D. & LIPNICK, S. L. 2009. Analysis of evidence-based medicine for shoulder instability. *Arthroscopy,* 25**,** 897-908.

PORTNEY, L. G. & WATKINS, M. P. 2015. *Foundations of Clinical Research: Applications To Practice,* Philadelphia, F.A. Davis Company/Publishers.

PRINSEN, C. A. C., MOKKINK, L. B., BOUTER, L. M., ALONSO, J., PATRICK, D. L., DE VET, H. C. W. & TERWEE, C. B. 2018. COSMIN guideline for systematic reviews of patient-reported outcome measures. *Qual Life Res,* 27**,** 1147-1157.

RAGAB, A. A. 2003. Validity of self-assessment outcome questionnaires: patient-physician discrepancy in outcome interpretation. *Biomed Sci Instrum,* 39**,** 579-84.

RIDDLE, D. & STRATFORD, P. 2013. Acknowledging barriers and identifying barrier busting strategies. *Is this change real? Interpreting patient outcomes in physical therapy.* Philadelphia: FA Davis.

ROBERTSON, S., KREMER, P., AISBETT, B., TRAN, J. & CERIN, E. 2017. Consensus on measurement properties and feasibility of performance tests for the exercise and sport sciences: a Delphi study. *Sports Medicine - Open,* 3**,** 2.

ROBINS, R. J., ANDERSON, M. B., ZHANG, Y., PRESSON, A. P., BURKS, R. T. & GREIS, P. E. 2017. Convergent Validity of the Patient-Reported Outcomes Measurement Information System's Physical Function Computerized Adaptive Test for the Knee and Shoulder Injury Sports Medicine Patient Population. *Arthroscopy,* 33**,** 608-616.

ROCOURT, M. H., RADLINGER, L., KALBERER, F., SANAVI, S., SCHMID, N. S., LEUNIG, M. & HERTEL, R. 2008. Evaluation of intratester and intertester reliability of the Constant-Murley shoulder assessment. *J Shoulder Elbow Surg,* 17**,** 364-9.

RODDEY, T. S., OLSON, S. L., COOK, K. F., GARTSMAN, G. M. & HANTEN, W. 2000. Comparison of the University of California-Los Angeles Shoulder Scale and the Simple Shoulder Test with the shoulder pain and disability index: single-administration reliability and validity. *Phys Ther,* 80**,** 759-68.

ROE, Y., SOBERG, H. L., BAUTZ-HOLTER, E. & OSTENSJO, S. 2013. A systematic review of measures of shoulder pain and functioning using the International classification of functioning, disability and health (ICF). *Bmc Musculoskeletal Disorders,* 14**,** 73.

ROLDAN-JIMENEZ, C. & CUESTA-VARGAS, A. I. 2016. Age-related changes analyzing shoulder kinematics by means of inertial sensors. *Clin Biomech (Bristol, Avon),* 37**,** 70-6.

ROMEO, A. A., BACH, B. R., JR. & O'HALLORAN, K. L. 1996. Scoring systems for shoulder conditions. *Am J Sports Med,* 24**,** 472-6.

ROULEAU, D. M., FABER, K. & MACDERMID, J. C. 2010. Systematic review of patient-administered shoulder functional scores on instability. *J Shoulder Elbow Surg,* 19**,** 1121-8.

ROWE, P. J. 1999. Measurement systems. *In:* DURWARD, B. R. B., G. D.; ROWE, P. J. (ed.) *Functional human movement. Measurement and assessment.* Oxford: Butterworth Heinemann.

ROY, J. S., MACDERMID, J. C. & WOODHOUSE, L. J. 2009. Measuring shoulder function: a systematic review of four questionnaires. *Arthritis Rheum,* 61**,** 623-32.

ROY, J. S., MACDERMID, J. C. & WOODHOUSE, L. J. 2010. A systematic review of the psychometric properties of the Constant-Murley score. *J Shoulder Elbow Surg,* 19**,** 157-64.

RUNDQUIST, P. J., ANDERSON, D. D., GUANCHE, C. A. & LUDEWIG, P. M. 2003. Shoulder kinematics in subjects with frozen shoulder. *Arch Phys Med Rehabil,* 84**,** 1473-9.

RUNDQUIST, P. J. & LUDEWIG, P. M. 2004. Patterns of motion loss in subjects with idiopathic loss of shoulder range of motion. *Clin Biomech (Bristol, Avon),* 19**,** 810-8.

RUNDQUIST, P. J. & LUDEWIG, P. M. 2005. Correlation of 3-dimensional shoulder kinematics to function in subjects with idiopathic loss of shoulder range of motion. *Phys Ther,* 85**,** 636-47.

RUSSEK, L. 2004. Factors affecting interpretation of reliability coefficients. *J Orthop Sports Phys Ther,* 34**,** 341-9.

RYSSTAD, T., ROE, Y., HALDORSEN, B., SVEGE, I. & STRAND, L. I. 2017. Responsiveness and minimal important change of the Norwegian version of the Disabilities of the Arm, Shoulder and Hand questionnaire (DASH) in patients with subacromial pain syndrome. *BMC Musculoskelet Disord,* 18**,** 248.

SAHINOGLU, E., ERGIN, G. & UNVER, B. 2019. Psychometric properties of patient-reported outcome questionnaires for patients with musculoskeletal disorders of the shoulder. Knee Surg Sports Traumatol Arthrosc. [ahead of print

SALLAY, P. I. & REED, L. 2003. The measurement of normative American Shoulder and Elbow Surgeons scores. *J Shoulder Elbow Surg,* 12**,** 622-7.

SALOMONSSON, B., AHLSTROM, S., DALEN, N. & LILLKRONA, U. 2009. The Western Ontario Shoulder Instability Index (WOSI): validity, reliability, and responsiveness retested with a Swedish translation. *Acta Orthop,* 80**,** 233-8.

SCHMITT, J. S. & DI FABIO, R. P. 2004. Reliable change and minimum important difference (MID) proportions facilitated group responsiveness comparisons using individual threshold criteria. *J Clin Epidemiol,* 57**,** 1008-18.

SCHULLER, W., OSTELO, R. W., JANSSEN, R. & DE VET, H. C. 2014. The influence of study population and definition of improvement on the smallest detectable change and the minimal important change of the neck disability index. *Health Qual Life Outcomes,* 12**,** 53.

SCIASCIA, A. D., MORRIS, B. J., JACOBS, C. A. & EDWARDS, T. B. 2017. Responsiveness and Internal Validity of Common Patient-Reported Outcome Measures Following Total Shoulder Arthroplasty. *Orthopedics***,** 1-7.

SHER, J. S., URIBE, J. W., POSADA, A., MURPHY, B. J. & ZLATKIN, M. B. 1995. Abnormal findings on magnetic resonance images of asymptomatic shoulders. *J Bone Joint Surg Am,* 77**,** 10-5.

SHIN, S. H., RO DU, H., LEE, O. S., OH, J. H. & KIM, S. H. 2012. Within-day reliability of shoulder range of motion measurement with a smartphone. *Man Ther,* 17**,** 298-304.

SHROUT, P. E. & FLEISS, J. L. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull,* 86**,** 420-8.

SINGH, J. A., SPERLING, J., BUCHBINDER, R. & MCMAKEN, K. 2010. Surgery for shoulder osteoarthritis. *Cochrane Database Syst Rev***,** Cd008089.

SKARE, O., LIAVAAG, S., REIKERAS, O., MOWINCKEL, P. & BROX, J. I. 2013. Evaluation of Oxford instability shoulder score, Western Ontario shoulder instability index and Euroqol in patients with SLAP (superior labral anterior posterior) lesions or recurrent anterior dislocations of the shoulder. *BMC Res Notes,* 6**,** 273.

SLOBOGEAN, G. P., NOONAN, V. K., FAMUYIDE, A. & O'BRIEN, P. J. 2011. Does objective shoulder impairment explain patient-reported functional outcome? A study of proximal humerus fractures. *J Shoulder Elbow Surg,* 20**,** 267-72.

SLOBOGEAN, G. P. & SLOBOGEAN, B. L. 2011. Measuring shoulder injury function: common scales and checklists. *Injury,* 42**,** 248-52.

SOPER, D. S. 2004. *Statistics Calculators* [Online]. Available: http://www.webcitation.org/6ZEMd2NIS [Accessed Archived on 12 May 2015 2015].

STAPLES, M. P., FORBES, A., GREEN, S. & BUCHBINDER, R. 2010. Shoulder-specific disability measures showed acceptable construct validity and responsiveness. *J Clin Epidemiol,* 63**,** 163-70.

STRATFORD, P. W. & RIDDLE, D. L. 2005. Assessing sensitivity to change: choosing the appropriate change coefficient. *Health Qual Life Outcomes,* 3**,** 23.

STREINER, D. L. 2003. Clinimetrics vs. psychometrics: an unnecessary distinction. *J Clin Epidemiol,* 56**,** 1142-5; discussion 1146-9.

TASHJIAN, R. Z., DELOACH, J., GREEN, A., PORUCZNIK, C. A. & POWELL, A. P. 2010. Minimal clinically important differences in ASES and simple shoulder

test scores after nonoperative treatment of rotator cuff disease. *J Bone Joint Surg Am,* 92**,** 296-303.

TASHJIAN, R. Z., HUNG, M., KEENER, J. D., BOWEN, R. C., MCALLISTER, J., CHEN, W., EBERSOLE, G., GRANGER, E. K. & CHAMBERLAIN, A. M. 2017. Determining the minimal clinically important difference for the American Shoulder and Elbow Surgeons score, Simple Shoulder Test, and visual analog scale (VAS) measuring pain after shoulder arthroplasty. *J Shoulder Elbow Surg,* 26**,** 144-148.

TEECE, R. M., LUNDEN, J. B., LLOYD, A. S., KAISER, A. P., CIEMINSKI, C. J. & LUDEWIG, P. M. 2008. Three-dimensional acromioclavicular joint motions during elevation of the arm. *J Orthop Sports Phys Ther,* 38**,** 181-90.

TERWEE, C. B., BOT, S. D., DE BOER, M. R., VAN DER WINDT, D. A., KNOL, D. L., DEKKER, J., BOUTER, L. M. & DE VET, H. C. 2007. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol,* 60**,** 34-42.

TERWEE, C. B., DEKKER, F. W., WIERSINGA, W. M., PRUMMEL, M. F. & BOSSUYT, P. M. 2003. On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. *Qual Life Res,* 12**,** 349-62.

THABANE, L., MA, J., CHU, R., CHENG, J., ISMAILA, A., RIOS, L. P., ROBSON, R., THABANE, M., GIANGREGORIO, L. & GOLDSMITH, C. H. 2010. A tutorial on pilot studies: the what, why and how. *BMC Med Res Methodol,* 10**,** 1.

THOOMES-DE GRAAF, M., SCHOLTEN-PEETERS, G. G., SCHELLINGERHOUT, J. M., BOURNE, A. M., BUCHBINDER, R., KOEHORST, M., TERWEE, C. B. & VERHAGEN, A. P. 2016. Evaluation of measurement properties of self-administered PROMs aimed at patients with non-specific shoulder pain and "activity limitations": a systematic review. *Qual Life Res,* 25**,** 2141-60.

TORRENS, C., GUIRRO, P. & SANTANA, F. 2016. The minimal clinically important difference for function and strength in patients undergoing reverse shoulder arthroplasty. *J Shoulder Elbow Surg,* 25**,** 262-8.

TORRENS, C., SANCHEZ, J. F., ISART, A. & SANTANA, F. 2015. Does fracture of the dominant shoulder have any effect on functional and quality of life outcome compared with the nondominant shoulder? *J Shoulder Elbow Surg,* 24**,** 677-81.

TROFA, D., RAJAEE, S. S. & SMITH, E. L. 2014. Nationwide trends in total shoulder arthroplasty and hemiarthroplasty for osteoarthritis. *Am J Orthop (Belle Mead NJ),* 43**,** 166-72.

TUBACH, F., RAVAUD, P., BARON, G., FALISSARD, B., LOGEART, I., BELLAMY, N., BOMBARDIER, C., FELSON, D., HOCHBERG, M., VAN DER HEIJDE, D. & DOUGADOS, M. 2005a. Evaluation of clinically relevant changes in patient

reported outcomes in knee and hip osteoarthritis: the minimal clinically important improvement. *Ann Rheum Dis,* 64**,** 29-33.

TUBACH, F., RAVAUD, P., BARON, G., FALISSARD, B., LOGEART, I., BELLAMY, N., BOMBARDIER, C., FELSON, D., HOCHBERG, M., VAN DER HEIJDE, D. & DOUGADOS, M. 2005b. Evaluation of clinically relevant states in patient reported outcomes in knee and hip osteoarthritis: the patient acceptable symptom state. *Ann Rheum Dis,* 64**,** 34-7.

TUBACH, F., RAVAUD, P., BEATON, D., BOERS, M., BOMBARDIER, C., FELSON, D. T., VAN DER HEIJDE, D., WELLS, G. & DOUGADOS, M. 2007. Minimal clinically important improvement and patient acceptable symptom state for subjective outcome measures in rheumatic disorders. *J Rheumatol,* 34**,** 1188-93.

TUBACH, F., RAVAUD, P., MARTIN-MOLA, E., AWADA, H., BELLAMY, N., BOMBARDIER, C., FELSON, D. T., HAJJAJ-HASSOUNI, N., HOCHBERG, M., LOGEART, I., MATUCCI-CERINIC, M., VAN DE LAAR, M., VAN DER HEIJDE, D. & DOUGADOS, M. 2012. Minimum clinically important improvement and patient acceptable symptom state in pain and function in rheumatoid arthritis, ankylosing spondylitis, chronic back pain, hand osteoarthritis, and hip and knee osteoarthritis: Results from a prospective multinational study. *Arthritis Care Res (Hoboken),* 64**,** 1699-707.

TUBACH, F., WELLS, G. A., RAVAUD, P. & DOUGADOS, M. 2005c. Minimal clinically important difference, low disease activity state, and patient acceptable symptom state: methodological issues. *J Rheumatol,* 32**,** 2025-9.

VALDERAS, J. M., FERRER, M., MENDIVIL, J., GARIN, O., RAJMIL, L., HERDMAN, M. & ALONSO, J. 2008. Development of EMPRO: A tool for the standardized assessment of patient-reported outcome measures. *Value Health,* 11.

VAN ALPHEN, A., HALFENS, R., HASMAN, A. & IMBOS, T. 1994. Likert or Rasch? Nothing is more applicable than good theory. *J Adv Nurs,* 20**,** 196-201.

VAN DE WATER, A. T., DAVIDSON, M., SHIELDS, N., EVANS, M. C. & TAYLOR, N. F. 2016a. The Shoulder Function Index (SFInX): evaluation of its measurement properties in people recovering from a proximal humeral fracture. *BMC Musculoskelet Disord,* 17**,** 295.

VAN DE WATER, A. T., SHIELDS, N., DAVIDSON, M., EVANS, M. & TAYLOR, N. F. 2014. Reliability and validity of shoulder function outcome measures in people with a proximal humeral fracture. *Disabil Rehabil,* 36**,** 1072-9.

VAN DE WATER, A. T. M., DAVIDSON, M., SHIELDS, N., EVANS, M. C. & TAYLOR, N. F. 2016b. The Shoulder Function Index (SFInX): evaluation of its measurement properties in people recovering from a proximal humeral fracture. *BMC Musculoskeletal Disorders,* 17**,** 295.

VAN DEN NOORT, J. C., WIERTSEMA, S. H., HEKMAN, K. M., SCHONHUTH, C. P., DEKKER, J. & HARLAAR, J. 2014. Reliability and precision of 3D wireless measurement of scapular kinematics. *Med Biol Eng Comput,* 52**,** 921-31.

VAN DER LINDE, J. A., VAN KAMPEN, D. A., VAN BEERS, L., VAN DEURZEN, D. F. P., SARIS, D. B. F. & TERWEE, C. B. 2017. The Responsiveness and Minimal Important Change of the Western Ontario Shoulder Instability Index and Oxford Shoulder Instability Score. *J Orthop Sports Phys Ther,* 47**,** 402-410.

VAN DER LINDE, J. A., VAN KAMPEN, D. A., VAN BEERS, L. W., VAN DEURZEN, D. F., TERWEE, C. B. & WILLEMS, W. J. 2015. The Oxford Shoulder Instability Score; validation in Dutch and first-time assessment of its smallest detectable change. *J Orthop Surg Res,* 10**,** 146.

VAN DER LINDE, J. A., WILLEMS, W. J., VAN KAMPEN, D. A., VAN BEERS, L. W., VAN DEURZEN, D. F. & TERWEE, C. B. 2014. Measurement properties of the Western Ontario Shoulder Instability index in Dutch patients with shoulder instability. *BMC Musculoskelet Disord,* 15**,** 211.

VAN DER WINDT, D. A., KOES, B. W., BOEKE, A. J., DEVILLE, W., DE JONG, B. A. & BOUTER, L. M. 1996. Shoulder disorders in general practice: prognostic indicators of outcome. *Br J Gen Pract,* 46**,** 519-23.

VAN DER WINDT, D. A., KOES, B. W., DE JONG, B. A. & BOUTER, L. M. 1995. Shoulder disorders in general practice: incidence, patient characteristics, and management. *Ann Rheum Dis,* 54**,** 959-64.

VAN KAMPEN, D. A., VAN BEERS, L. W., SCHOLTES, V. A., TERWEE, C. B. & WILLEMS, W. J. 2012. Validation of the Dutch version of the Simple Shoulder Test. *J Shoulder Elbow Surg,* 21**,** 808-14.

VAN KAMPEN, D. A., WILLEMS, W. J., VAN BEERS, L. W., CASTELEIN, R. M., SCHOLTES, V. A. & TERWEE, C. B. 2013. Determination and comparison of the smallest detectable change (SDC) and the minimal important change (MIC) of four-shoulder patient-reported outcome measures (PROMs). *J Orthop Surg Res,* 8**,** 40.

VEEGER, H. E. & VAN DER HELM, F. C. 2007. Shoulder function: the perfect compromise between mobility and stability. *J Biomech,* 40**,** 2119-29.

VROTSOU, K., CUELLAR, R., SILIO, F., RODRIGUEZ, M. A., GARAY, D., BUSTO, G., TRANCHO, Z. & ESCOBAR, A. 2016. Patient self-report section of the ASES questionnaire: a Spanish validation study using classical test theory and the Rasch model. *Health Qual Life Outcomes,* 14**,** 147.

WALTER, S. D., ELIASZIW, M. & DONNER, A. 1998. Sample size and optimal designs for reliability studies. *Stat Med,* 17**,** 101-10.

WEIR, J. P. 2005. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J Strength Cond Res,* 19**,** 231-40.

WERNER, B. C., CHANG, B., NGUYEN, J. T., DINES, D. M. & GULOTTA, L. V. 2016. What Change in American Shoulder and Elbow Surgeons Score Represents a Clinically Important Change After Shoulder Arthroplasty? *Clin Orthop Relat Res,* 474**,** 2672-2681.

WERNER, B. C., HOLZGREFE, R. E., GRIFFIN, J. W., LYONS, M. L., COSGROVE, C. T., HART, J. M. & BROCKMEIER, S. F. 2014. Validation of an innovative method of shoulder range-of-motion measurement using a smartphone clinometer application. *J Shoulder Elbow Surg,* 23**,** e275-82.

WICKHAM, J., PIZZARI, T., STANSFELD, K., BURNSIDE, A. & WATSON, L. 2010. Quantifying 'normal' shoulder muscle activity during abduction. *J Electromyogr Kinesiol,* 20**,** 212-22.

WIERTSEMA, S. H., DE WITTE, P. B., RIETBERG, M. B., HEKMAN, K. M., SCHOTHORST, M., STEULTJENS, M. P. & DEKKER, J. 2014. Measurement properties of the Dutch version of the Western Ontario Shoulder Instability Index (WOSI). *J Orthop Sci,* 19**,** 242-9.

WILD, D., GROVE, A., MARTIN, M., EREMENCO, S., MCELROY, S., VERJEE-LORENZ, A., ERIKSON, P., TRANSLATION, I. T. F. F. & CULTURAL, A. 2005. Principles of Good Practice for the Translation and Cultural Adaptation Process for Patient-Reported Outcomes (PRO) Measures: report of the ISPOR Task Force for Translation and Cultural Adaptation. *Value Health,* 8**,** 94-104.

WINER, B., ; BROWN, D. ; MICHELS, K. 1991. *Statistical principles in experimental design.,* New York, McGraw-Hill.

WONG, W. Y., WONG, M. S. & LO, K. H. 2007. Clinical applications of sensors for human posture and movement analysis: a review. *Prosthet Orthot Int,* 31**,** 62-75.

WORLD HEALTH ORGANIZATION. 2001. *International Classification of Functioning, Disability and Health (ICF)* [Online]. Geneva: World Health Organization. Available: http://www.who.int/classifications/icf/en/ [Accessed 21 November 2017].

WRIGHT, R. W. & BAUMGARTEN, K. M. 2010. Shoulder outcomes measures. *J Am Acad Orthop Surg,* 18**,** 436-44.

WU, G., VAN DER HELM, F. C. T., VEEGER, H. E. J., MAKHSOUS, M., VAN ROY, P., ANGLIN, C., NAGELS, J., KARDUNA, A. R., MCQUADE, K., WANG, X. G., WERNER, F. W. & BUCHHOLZ, B. 2005. ISB recommendation on definitions of joint coordinate systems of various joints for the reporting of human joint motion - Part II: shoulder, elbow, wrist and hand. *Journal of Biomechanics,* 38**,** 981-992.

WYLIE, J. D., BECKMANN, J. T., GRANGER, E. & TASHJIAN, R. Z. 2014. Functional outcomes assessment in shoulder surgery. *World J Orthop,* 5**,** 623-33.

WYLIE, J. D., SUTER, T., POTTER, M. Q., GRANGER, E. K. & TASHJIAN, R. Z. 2016. Mental Health Has a Stronger Association with Patient-Reported Shoulder Pain and Function Than Tear Size in Patients with Full-Thickness Rotator Cuff Tears. *J Bone Joint Surg Am,* 98**,** 251-6.

YAHIA, A., GUERMAZI, M., KHMEKHEM, M., GHROUBI, S., AYEDI, K. & ELLEUCH, M. H. 2011a. Translation into Arabic and validation of the ASES index in

assessment of shoulder disabilities. *Annals of Physical and Rehabilitation Medicine,* 54**,** 59-72.

YAHIA, A., GUERMAZI, M., KHMEKHEM, M., GHROUBI, S., AYEDI, K. & ELLEUCH, M. H. 2011b. Translation into Arabic and validation of the ASES index in assessment of shoulder disabilities. *Ann Phys Rehabil Med,* 54**,** 59-72.

YAMAGUCHI, K., DITSIOS, K., MIDDLETON, W. D., HILDEBOLT, C. F., GALATZ, L. M. & TEEFEY, S. A. 2006. The demographic and morphological features of rotator cuff disease. A comparison of asymptomatic and symptomatic shoulders. *J Bone Joint Surg Am,* 88**,** 1699-704.

YAMAMOTO, A., TAKAGISHI, K., OSAWA, T., YANAGAWA, T., NAKAJIMA, D., SHITARA, H. & KOBAYASHI, T. 2010. Prevalence and risk factors of a rotator cuff tear in the general population. *J Shoulder Elbow Surg,* 19**,** 116-20.

YAMATO, T., MAHER, C., SARAGIOTTO, B., MOSELEY, A., HOFFMANN, T., ELKINS, M. & JETTE, A. 2016. The TIDieR Checklist Will Benefit the Physical Therapy Profession. *Phys Ther,* 96**,** 930-1.

YANG, J. L., LIN, J. J., HUANG, H. Y., HUANG, T. S. & CHAO, Y. W. 2014. Shoulder physical activity, functional disability and task difficulties in patients with stiff shoulders: interpretation from RT3 accelerator. *Man Ther,* 19**,** 349-54.

YAP, B. W. & SIM, C. H. 2011. Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation,* 81**,** 2141-2155.

YIAN, E. H., RAMAPPA, A. J., ARNEBERG, O. & GERBER, C. 2005. The Constant score in normal shoulders. *J Shoulder Elbow Surg,* 14**,** 128-33.

YUGUERO, M., HUGUET, J., GRIFFIN, S., SIRVENT, E., MARCANO, F., BALAGUER, M. & TORNER, P. 2016. Transcultural adaptation, validation and assessment of the psychometric properties of the spanish version of the Western Ontario Shoulder Instability Index questionnaire. *Rev Esp Cir Ortop Traumatol,* 60**,** 335-345.

ZANUDIN, A., MERCER, T. H., JAGADAMMA, K. C. & VAN DER LINDEN, M. L. 2017. Psychometric properties of measures of gait quality and walking performance in young people with Cerebral Palsy: A systematic review. *Gait Posture,* 58**,** 30-40.

Appendix I

MSc related published article

IOP Publishing | Institute of Physics and Engineering in Medicine Physiological Measurement

Physiol. Meas. **36** (2015) 547–560 doi:10.1088/0967-3334/36/3/547

Enhancing clinically-relevant shoulder function assessment using only essential movements

C Pichonnaz[1,2], E Lécureux[3], J-P Bassin[1], C Duc[4], A Farron[2], K Aminian[4], B M Jolles[2] and N Gleeson[5]

[1] Haute Ecole de Santé Vaud, HES-SO // University of Applied Sciences Western Switzerland, Physiotherapy Department, Av. de Beaumont 21, 1011 Lausanne Switzerland
[2] CHUV-UNIL Orthopedics and Traumatology Department, Av. Pierre-Decker 4, 1011 Lausanne, Switzerland
[3] CHUV-UNIL, direction médicale, Rue du Bugnon 46, 1011 Lausanne, Switzerland
[4] Laboratory of Movement Analysis and Measurement, Ecole Polytechnique Fédérale de Lausanne (EPFL), ELH 135 (Bâtiment ELH), Station 11, 1015 Lausanne, Switzerland
[5] Queen Margaret University, School of Health Sciences, Queen Margaret University, Edinburgh EH21 6UU, UK

E-mail: claude.pichonnaz@hesav.ch, Estelle.Lecureux@chuv.ch, jean-philippe.bassin@hesav.ch, cyntia.duc@epfl.ch, Alain.Farron@chuv.ch, kamiar.aminian@epfl.ch, Brigitte.Jolles-Haeberli@chuv.ch and ngleeson@qmu.ac.uk

Received 7 April 2014, revised 10 December 2014
Accepted for publication 16 January 2015
Published 18 February 2015

Abstract

Kinematic functional evaluation with body-worn sensors provides discriminative and responsive scores after shoulder surgery, but the optimal movements' combination has not yet been scientifically investigated. The aim of this study was the development of a simplified shoulder function kinematic score including only essential movements. The P Score, a seven-movement kinematic score developed on 31 healthy participants and 35 patients before surgery and at 3, 6 and 12 months after shoulder surgery, served as a reference.

Principal component analysis and multiple regression were used to create simplified scoring models. The candidate models were compared to the reference score. ROC curve for shoulder pathology detection and correlations with clinical questionnaires were calculated.

The B–B Score (hand to the Back and hand upwards as to change a Bulb) showed no difference to the P Score in time*score interaction ($P > .05$) and its relation with the reference score was highly linear ($R^2 > .97$). Absolute

value of correlations with clinical questionnaires ranged from 0.51 to 0.77. Sensitivity was 97% and specificity 94%.

The B–B and reference scores are equivalent for the measurement of group responses. The validated simplified scoring model presents practical advantages that facilitate the objective evaluation of shoulder function in clinical practice.

Keywords: shoulder, outcome treatment, body-worn sensors, biomechanics, validation studies as topic

(Some figures may appear in colour only in the online journal)

1. Introduction

The assessment of functional outcome of shoulder treatments remains a controversial issue. Although many questionnaires exist, none has been universally recognized as a standard to date (Placzek *et al* 2004, Fayad *et al* 2005, Wilcox *et al* 2005, Oh *et al* 2009). Alternatively, the effectiveness of embedded kinematic measurement to assess shoulder function has not yet been extensively explored. Measurements based on body-worn sensors may potentially represent a well-balanced compromise between the practicality of questionnaires and the measurement precision and reliability of laboratory-based movement analysis (Pandyan *et al* 2002).

However, the most efficient testing procedure for the evaluation of shoulder function has not yet been defined. An approach to assessment that captures the essence of the complex patterns of movement comprising shoulder function may offer further progress towards an effective clinical tool. A simplified scoring procedure involving only essential movements would facilitate the use of movement analysis for outcome evaluation. Thus, this study focused on the development of an efficient and simple assessment model that should demonstrate content validity, relationship to shoulder function and ease of application.

Body-worn inertial sensors have been applied with promising results to measure shoulder movement in various conditions (Zhou *et al* 2006, Coley *et al* 2007, Luinge *et al* 2007, Wong *et al* 2007, Coley *et al* 2008, Teece *et al* 2008, Duc *et al* 2013). Their results are highly correlated to laboratory measurements and display adequate accuracy (Bernmark and Wiktorin 2002, Luinge and Veltink 2005, Zhou *et al* 2006, Coley *et al* 2007, Cutti *et al*, 2008). Among these authors, Coley *et al* (2007) proposed a shoulder kinematic score based on the P Score, which compares injured versus healthy arm power measured by accelerometers and gyroscopes. A power-related metric [(deg/s) * ($m\,s^{-2}$)] was used as it demonstrated more discrimination than angle measurements for shoulder outcome evaluation (Coley *et al* 2007). The clinical inference was that the ability of the patient to deliver energy and useful work in a timely manner during arm movements is typically reduced in shoulder pathologies (Murrell and Walton 2001, Bunker 2002).

The P Score procedure relied on a sequence of seven movements extracted from the Simple Shoulder Test (SST) and therefore included movements representative of daily life activities (Lippitt *et al* 1993, Coley *et al* 2007). The testing procedure requires around 20 min for completion. This approach demonstrated clinical relevance as the P Score discriminated healthy from pathological subjects, identified early treatment failure, was correlated to clinical scores and displayed good responsiveness after shoulder surgery (Coley *et al* 2007, Coley 2007, Jolles *et al* 2011).

Table 1. Sample characteristics.

	Patient group	Control group
Sample	35 participants	31 participants
Gender	25 male /10 female	15 male/16 female
Age mean (SD)	58 (9.6) years old	33.2 (8.1) years old
Weight mean (SD)	79.6 (14.7) kg	68.8 (10.4) kg
Height mean (SD)	1.70 (0.1) m	1.72 (0.1) m
BMI mean (SD)	27.2 (3.8) $kg\,m^{-2}$	23.2 (3.1) $kg\,m^{-2}$
Dominance	33 right-handed//2 left-handed	24 right-handed/7 left-handed
Surgery side dominant/ side	23 dominant /12 non dominant	—
Surgery intervention	28 rotator cuff /7 arthroplasty	—

Körver *et al* (2014b,2014a) proposed a kinematic score including only the movements 'arm to the back' and 'arm behind the head'. This simplified approach improved clinical applicability by reducing measurement time to less than 5 min. It demonstrated high intra- and inter-evaluator reliability, diagnostic sensitivity and specificity, but weak correlations with the DASH and SST clinical scores (Lippitt *et al* 1993, Jester *et al* 2005). Conversely to the P Score, its validity for shoulder function evaluation was thus limited.

It is of interest to explore if a simplification of the P Score procedure, based on a systematic approach, would ensure that measurement properties observed for the P Score are not compromised by the simplification process. The primary aim of this study was to design a simplified kinematic shoulder scoring model based on a selection of essential movements among the seven movements of daily life comprised in the reference P Score. It was hypothesized that the number of movements could be reduced based on components identified by principal component analysis (PCA). Multivariate regression (MR) was then used to combine the defined principal components into a simplified scoring model.

The secondary aim was to compare the results of the new simplified scoring model with those of the reference P score. It was hypothesized that the results of the simplified score would be comparable to those of the reference score in terms of descriptive statistics, linear relation, evolution pattern and agreement. The strength of the relationship with the shoulder function questionnaires was evaluated for the reference and the new kinematic scores. This evaluation aimed at estimating their concurrent validity relative to commonly used clinical questionnaires but not at validating the kinematic scores against a gold standard.

2. Methods

2.1. Reference score

This study was based on a secondary analysis of data gathered for the development of the P Score, which is detailed above (Coley *et al* 2007).

The sample (table 1) was made of participants from a prospective cohort study between 2005 and 2008 at the Department of Traumatology and Orthopaedic Surgery of the University Hospital of Lausanne. Ethical approval was granted by the local ethical board Human Research Ethics Committee of the Canton of Vaud (CER-VD). Patients gave their informed and signed consent for the secondary use of data for research purposes.

The included patients were adults with rotator cuff disease involving a supraspinatus rupture of at least 1 cm^2, as determined by an MRI, or with a gleno-humeral osteoarthritis

stage II or III according to the radiologic criteria published by Koss *et al* (1997). The criteria considered for rotator cuff surgery were significant pain or dysfunction affecting quality of life (American Academy of Orthopaedic Surgeons, 2010). Exclusion criteria were previous shoulder surgery or arthroscopy, intra-articular injection in the last six months, contralateral painful shoulder or malignant disorder. All patients were operated on by the same surgeon. The healthy participants, measured for normal usage characterization, were people without a history of shoulder condition/pain and were purposefully younger than the patients to avoid bias related to the high prevalence of asymptomatic rotator cuff tear above 40 years of age (Sher *et al* 1995).

Patients were measured before surgery and at 3, 6 and 12 months after surgery. The participants were asked to perform the following movements as shown by the evaluator:

(1) —Back: place hand to the back
(2) —Head: reach the back of the head with the hand
(3) —Flexion: lift the arm upwards to reach 90° flexion
(4) —Abduction: lift the arm on the side to reach 90° abduction
(5) —Shoulder: touch the opposite shoulder with the hand
(6) —Bulb: lift the hand upwards as to change a bulb
(7) —Rotation: rotate the arm laterally with a 90°elbow flexion

The participants were instructed before the test to perform the movements at their natural speed in the pain free range of motion. They were told that they should stop the movement should pain occur. The initial position was standing in front of the evaluator, with arm along the body in a relaxed position. The participants performed each of the 7 movements and got back to the initial position, as demonstrated by the evaluator. The movements were performed at 20 s intervals. The movements were performed on the affected side first for the patients and on the dominant side first for the healthy participants.

The movements' performance was assessed using the P Score, a metric related to the power of movement computed as the product of accelerations by angular velocities.

To measure this score, participants were equipped with two inertial sensors including a triaxial accelerometer and a triaxial gyroscope, placed on each humerus, 3 cm above the midpoint of the line connecting the lateral epicondyle (EL) and medial epicondyle (EM). The sensor's axes were aligned to the humerus anatomical frame following the ISB recommendations (Wu *et al* 2005): Yh on the line connecting the gleno-humeral (GH) joint and the midpoint of EL and EM, pointing to GH; Xh on the line perpendicular to the plane formed by EL, EM and GH, pointing forward; Zh on the line perpendicular to Xh and Yh, pointing to the right (figure 1).

Accelerations and angular velocities were amplified and low-pass filtered (cutoff frequency: 17 Hz) to remove noise (Mathie *et al* 2004, Aminian *et al* 2006) before being recorded by a data-logger (Physilog®, Gait Up, CH), at 200 Hz. A power-related parameter was extracted from the recorded signals: the range of acceleration was multiplied by the range of angular velocity, with a measurement unit of (deg/s) * ($m\,s^{-2}$), for each movement (figure 2). This parameter was calculated for each axis and then averaged, separately for each side. The P Score was then computed as the ratio of the performance of the affected side relative to the healthy side, expressed in percentage (Coley *et al* 2007). For example, while a typical healthy person performs near to 100%, the average patient might reach e.g. 46% before surgery, 67% at 3 months and 71% at 6 months. For healthy subjects, the P Score reflects the performance of the dominant side compared to the non dominant side.

Participants completed questionnaires to establish the relationship between P Score and respectively, pain assessed though Visual Analog Scale (VAS) and function estimated by

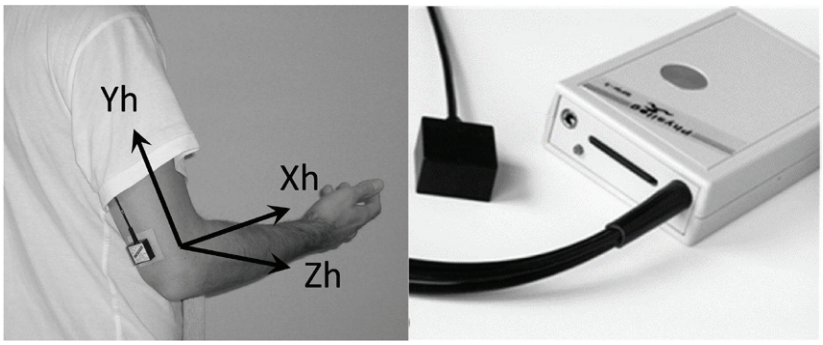


Figure 1. Arm sensors placement during measurement. Adapted with permission from Coley *et al* 2007, copyright 2007 Elsevier.

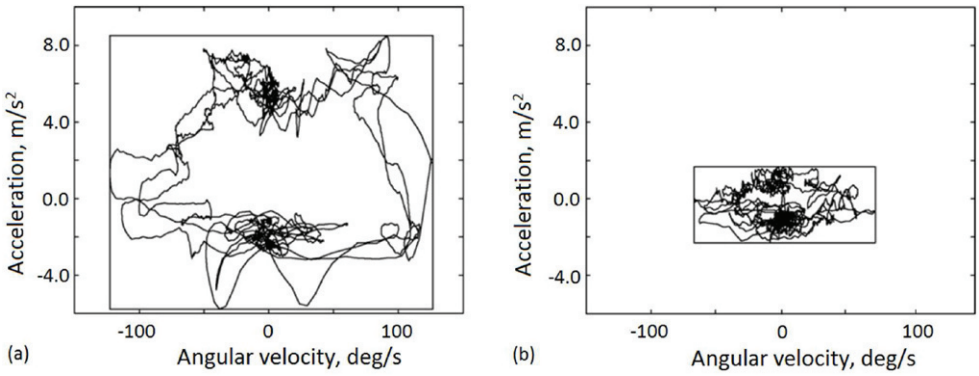


Figure 2. Humerus acceleration as a function of its angular velocity for a patient. (*a*) The trace represents the humerus acceleration versus angular velocity for the healthy side. (*b*) The trace represents the humerus acceleration versus angular velocity for the painful side. The rectangle, which circumscribes the curve corresponds to the product of the acceleration range by the angular velocity range (Pr). Reprinted with permission from Coley *et al* 2007, copyright 2007 Elsevier.

clinical questionnaires: Constant score, SST, Disability of Arm, Shoulder and Hand (DASH) (Constant and Murley 1987, Lippitt *et al* 1993, Fayad *et al* 2008).

2.2. Statistical analysis approach

The statistical analysis was conducted with PASW Statistics 18 (SPSS Inc, Chicago, Illinois). First, the development of a simplified scoring model based on selected movements was conducted, initially using data at 3 months after surgery. As rehabilitation is most intensive at this stage, it was of primary importance that the simplified score would be efficient. The scoring model was then applied to data at all stages to investigate its relevance over time. Finally, the reference and the simplified scores were compared.

PCA was used to identify components that explain most of the variance associated with the reference score. Among movements loading on a component, one was retained for each respective component for inclusion in a MR analysis. This planned linkage between multiple regression and antecedent PCA prevents multicollinearity problems that could cause

Table 2. Mean and SD of the reference score (P Score) and details for all performed movements for patients at 3 months. Unit of scores are % representing the performance of the pathological side compared to the healthy side.

		Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7
	P Score	Hand in back	Hand behind the head	Reach object ahead	Carry 4 kg in abduction	Touch opposite shoulder	Change a bulb	Move hand laterally
Mean (SD)	61.4 (19.4)	53.2 (26.4)	60.9 (22.6)	74.3 (25.0)	49.4 (27.2)	68.2 (21.0)	61.4 (27.8)	62.8 (26.2)

erratic changes in the regression coefficients (Jolliffe 2002, Portney and Watkins 2009). Several simplified scoring models were created based on MR results at 3 months. The models were then applied to data at baseline, 6 and 12 months to calculate their outcomes over all stages. The latter outcomes were compared with the reference score to assess the extent of congruency.

Then the progression pattern over time, for the reference and simplified scores, were compared using separate factorial (model [reference; simplified] *x* time [baseline, 3, 6 and 12 months post-surgery]) two-ways ANOVAs on both factors. 'Model' was used as factor and 'time' as covariate. Assumptions of normality and homoscedacity were verified using, respectively, the Shapiro–Wilk and the Levene's test. Results of the reference and simplified scores were reported using descriptive statistics (mean, standard deviation, 95% confidence interval, standard error, difference with reference score). The effect of arm dominance on score outcome was evaluated using a one-sample *t*-test against a test value of 100, indicating perfect symmetry. Simple linear regressions and Bland and Altman's limits of agreement (LOA) were performed at each stage of rehabilitation for the score which displayed the closest pattern of congruency to the reference score. The relationships between reference and simplified scores with clinical questionnaires were investigated using Spearman correlations. Type I error rates were set at $P < .05$, where applicable. Diagnostic power for shoulder pathology detection at baseline was calculated using receiver operating characteristic (ROC) analysis.

3. Results

3.1. Development of simplified scores

3.1.1. P Score and score for each movement. Mean and SD of the P Score at 3 months for each movement included in the PCA are presented in table 2.

3.1.2. Principal component analysis. The PCA highlighted two components. Arm movements related to the first component represented a dimension of 'elevation' while those related to the second component represented 'rotation'. This model was constant over all stages of rehabilitation. The movement 1 (Back) was systematically related to the rotation component. The movements 2 to 6 (Head, Flexion, Abduction, Shoulder, Bulb) were systematically related to the elevation component, with varying strength over stages. The movement 3 was excluded from the model at 3 months, as it was a complex variable i.e. correlated above .40 with several components. The movement 7 (Rotation) was related to one or the other component according to the stage.

The factor loadings and explained variance are presented in table 3.

Table 3. For each stage, the two PCA components are described with the eigenvalue, the factor loading to each movement, and the explained variance identified for the P score data, per component and for the cumulative variance (last column).

	PCA component	Eigenvalue	Factor loading per movement							Explained variance	
			1	2	3	4	5	6	7		
Baseline	1st elevation	3.0	—	0.72	0.62	0.83	0.72	0.80	—	43%	62%
	2nd rotation	1.3	0.83	—	—	—	—	—	0.82	19%	
3 months	1st elevation	3.4	—	0.74	—	0.85	0.81	0.85	0.69	57%	73%
	2nd rotation	1.0	0.96	—	—	—	—	—	—	16%	
6 months	1st elevation	3.3	—	0.88	—	0.73	0.86	0.82	0.63	56%	72%
	2nd rotation	1.0	0.97	—	—	—	—	—	—	16%	
12 months	1st elevation	2.1	—	—	0.80	0.90	—	—	—	51%	76%
	2nd rotation	1.0	0.76	—	—	—	—	—	0.91	25%	

Legend: 1. Hand to the back, 2. Hand behind the head, 3. Reach object ahead (90° flexion), 4. Carry 4 kg in abduction (90° abduction with load), 5. Touch opposite shoulder with hand, 6. Change a bulb (elevation), 7. Move hand laterally keeping elbow against the body (external rotation). Loadings are presented into brackets. Movements which do not appear in the table are complex movements (related to several components) that were therefore excluded from analysis.

3.1.3. Multiple regression at 3 months. Based on the two components identified by the PCA at 3 months, the MR included scores from two movements as independent variables. The first variable was the movement 1 that represented the rotation component. The movement 2, 4, 5 and 6 were alternatively included as a second variable in MR to represent the elevation component. The movement 7 was excluded from the candidate MRs as its relation to a component was erratic over time. Thus, four MRs were conducted with pairs of isolated movements as predictive variables: Back-Head (Movements 1 and 2), Back-Abduction (Movements 1 and 4), Back-Shoulder (Movements 1 and 5) and Back-Bulb (Movements 1 and 6) (table 4).

The regression equations from these four potential simplified scoring models were applied to data at all stages of rehabilitation. Shapiro–Wilk tests confirmed normality for all scores except for the Back-Abduction scoring model at baseline ($P < .05$) and this candidate was thus excluded from further analyses.

3.2. Selection of the simplified scoring model

The simplified scoring models developed at 3 months were applied to data at baseline, 6 and 12 months. A separate factorial ANOVA with measures on both factors was used to compare patterns of each remaining candidate scoring model (Back-Head, Back-Shoulder and Back-Bulb) with that of the reference score across stages of rehabilitation. Assumption of normality and homoscedacity were met for all scores at all stages ($P > .05$ for the Shapiro–Wilk and for the Levene's test).

The ANOVAs showed that the time-model interactions were significant for the Back-Head [$F(3, 90) = 7.0$; $P < .01$], the Back-Shoulder [$F(3, 78) = 3.0$; $P < 0.05$], but not for the Back-Bulb score, indicating that the latter model (B–B Score) offered better congruency between reference and simplified scoring models over the period of rehabilitation and should be selected for further comparative analysis on this basis (figure 3). Congruency was also confirmed by ANOVA that showed no significant difference in the model comparison between the reference and the simplified B–B Score.

Table 4. Details and results on the four candidate scoring models computed by the multiple regressions analysis (MR) at 3 months.

Predictive variables of reference score	Regression equations[a]	Standard error of estimate	Coefficient of determination (R^2)
Back-head	17.71 + (0.27 × back) + (0.47 × head)	10.5	70
Back-abduction	21.94 + (0.37 × back) + (0.40 × abduction)	9.5	76
Back-shoulder	7.38 + (0.19 × back) + (0.63 × head)	8.1	81
Back-bulb	16.71 + (0.32 × back) + (0.45 × bulb)	8.1	82

[a]All coefficients are significant at $P < .01$.

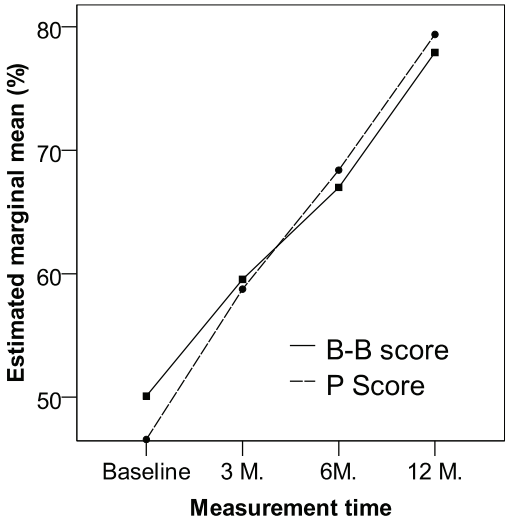


Figure 3. Graphs comparing evolution pattern over time of P Score and B–B Score (hand to the Back and hand upwards as to change a Bulb).

3.3. Comparison of the simplified score outcome with the reference score outcome

The mean, standard deviation, 95% confidence interval and standard error for the reference P Score and the simplified B–B score were calculated at each stage of rehabilitation (table 5). The difference between the mean reference and the mean B–B Score ranged respectively from −2.4 to − 0.2% according to stage in the patient group and was 5.3% in the control group.

The one sample *t*-test showed that the healthy group B–B Score of 102.9 was not significantly different from a 100% score indicating perfect symmetry between arms ($P = 0.28$). Conversely, the 108.2 P score showed a significant difference ($P < 0.01$).

The coefficients of the linear regressions between the reference and the B–B Score were significant at $P < .01$ at all stages. The slope coefficient was 1.03 at baseline, 1.01 at 3 months, 1.04 at 6 months and 1.01 at 12 months. Coefficient of determination was ≥.97 at all stages.

Bias and LOA between reference and B–B Scores were −3.1% ± 15.8, −0.7% ± 13.3, 1.8% ± 21.6 and 1.6% ± 19.6, respectively at baseline, 3, 6 and 12 months (figure 4).

Table 5. Descriptive statistics of the simplified scoring model (B–B Score) and the reference score (P Score) for patients at all stages and for healthy participants. Unit of scores are % representing the performance of the pathological side compared to the healthy side.

Stage	Scores	Mean	SD	95% CI Mean	Std Error	Difference in mean
Baseline	P Score	51.3	20.8	7.2	3.5	–0.6
	B–B	50.7	15.8	5.8	2.8	
3 months	P Score	62	19.2	6.7	3.3	–0.2
	B–B	61.8	16.8	6	3	
6 months	P Score	71.4	19.9	6.8	3.3	–2.4
	B–B	69	15.9	6.5	2.7	
12 months	P Score	81.5	22	7.6	3.7	–1
	B–B	80.5	21	7.6	3.7	
Controls	P Score	108.2	15.2	5.7	2.8	5.3
	B–B	102.9	14.5	5.4	2.6	

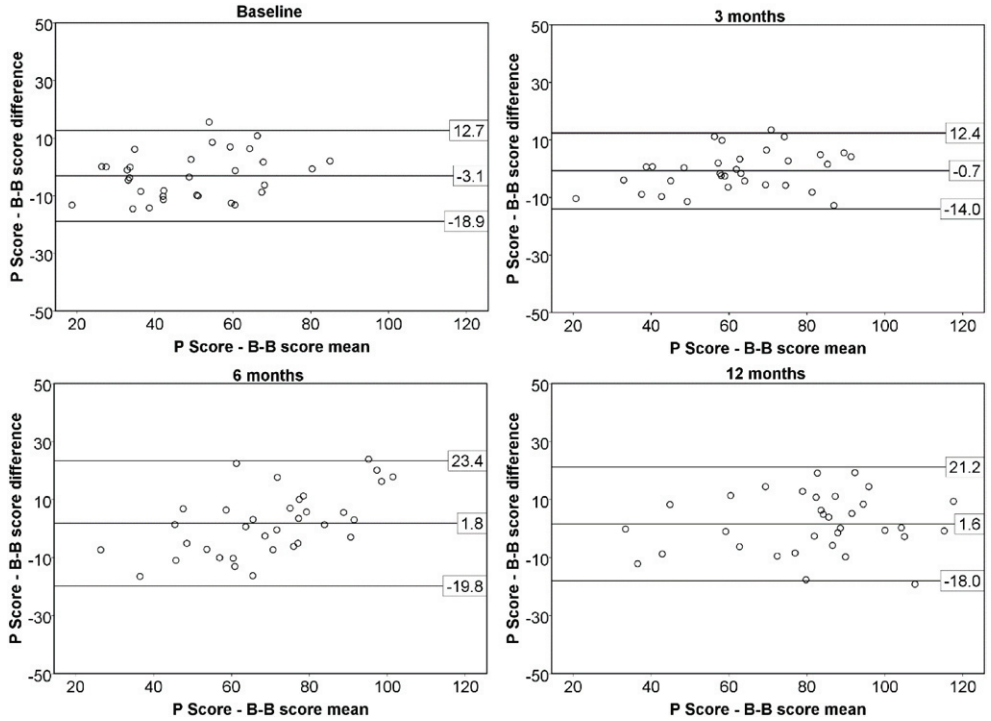


Figure 4. Bland and Altman plots for the agreement between P Score and B–B Score at all stages.

3.4. Relationship with shoulder function questionnaires

The absolute values of correlations between B–B Score and shoulder function assessed by clinical questionnaires (DASH, SST and Constant) ranged from 0.51 to 0.77 across the period of rehabilitation. The relationship between B–B Score and VAS pain (visual analog scale of pain) ranged from 0.35 to 0.50. All correlations were significant ($P < .05$) (table 6).

Table 6. Spearman correlations of all clinical questionnaires with the B-B Score over all stages.

	Baseline	3 months	6 months	1 year
DASH	−54[b]	−60[b]	−64[b]	−51[b]
SST	62[b]	62[b]	65[b]	55[b]
CST	57[b]	77[b]	61[b]	54[b]
VAS pain	−50[b]	−35[a]	−48[b]	−38[a]

[a] $P < 0.05$ [b] $P < 0.01$
Legend: DASH: Disabilities of the Arm, Shoulder and Hand questionnaire, SST: Simple Shoulder Test, CST: Constant shoulder score, VAS: visual analogic scale.

3.5. Diagnostic power for shoulder condition detection

The area under the ROC curve was 0.99 [95% CI 0.98–1.00]. Using a threshold score of 78.7, the sensitivity for shoulder pathology detection was 97% and the specificity 94%.

4. Discussion

4.1. Main results

This study investigated the possibility of reducing the number of movements performed during a reference kinematic test (P Score) and the influence of this reduction on the measured outcome following shoulder surgery.

The first step aimed at the development of a simplified kinematic scoring model using PCA and MR. It showed that 82% of the variation provided in the reference score involving seven movements was accounted for by the B–B Score at 3 months. As loss of information was minor, it should be reasonable to substitute this two-movement score for the reference score. The PCA highlighted two components, the 'elevation' and the 'rotation', which correspond to clinically relevant features. It was consistent over the three different assessment occasions during rehabilitation. This provided a robust indication that the data structure was correctly identified by PCA.

The PCA and MR did not allow the definition of a unique movement association reflecting the 'elevation' and the 'rotation' components. Therefore, several combinations were tested and the best choice was made in the second step of the analysis, based on the comparison of outcome with reference score outcomes using ANOVA.

No significant difference was found between the reference score and the simplified models. However, the model involving 'hand to the Back' and 'hand upwards as to change a Bulb' movements (B–B Score) displayed the closest patterning of minimized differences between simplified and reference models across the stages of rehabilitation. This model was thus selected for further comparison.

The congruency between the simplified model and the reference model was further confirmed by the descriptive statistics that showed little difference between reference and B–B Scores outcomes. Similarly, the linear regressions between the reference score and the B–B Score showed a close relationship ($R^2 > .97$). However, standard errors of estimate ranged from 7.0 to 11.1% indicating that consistent errors may occur in individual prediction.

Correspondingly, the Bland and Altman method showed that group results for the simplified model were closer to the reference score than those for individuals. The systematic error was limited (bias $< -3.1\%$) but the limits of agreements between reference score and B–B Score were large (13.3 to 21.6%) regardless of the stage.

The correlation between the B–B Score and the clinical questionnaires demonstrated that the outcome of the simplified scoring model is representative of the shoulder function and pain, with a closer link to function.

The non-significant one-sample *t*-test for the B–B Score against a test value of 100, indicated that the arm dominance had little influence on the outcome. Therefore, no correction was needed to account for the subjects' dominance.

The area under the ROC curve was 0.99, indicating an excellent ability of the B–B Score to distinguish affected from non-affected subjects (Hanley and McNeil1982). Concordantly, the diagnostic sensibility (97%) and specificity (94%) were excellent. The sensibility was 1% lower and the specificity was 5% higher than for a measurement method using range of angular velocities in subacromial impingement syndrome (Körver *et al* 2014a).

In summary, despite the possible divergence for single measurements, the findings confirm the hypothesis that the reference score and a simplified scoring model (B–B Score) provide comparable results for group measurements. It can be inferred from these analyses that the B–B Score is a reasonable substitute for the reference score during group-based measurements and offers the aforementioned characteristics of an efficient model.

4.2. Clinical interpretation

Rotation and elevation, as identified by the PCA, are two essential components of shoulder function. Seventy-seven and 73% of patients report difficulty in moving to reach the back of the head (elevation and external rotation) and the lower back (internal rotation) respectively in commonly occurring shoulder conditions (Van der Windt *et al* 1995). Some daily activities like perineal care require a large internal rotation while combing hair requires a large elevation and external rotation (Magermans *et al* 2005). Therefore, the inclusion of internal rotation and elevation in kinematic scores is underpinned by a close relation to shoulder function and confirmed by the correlations with clinical questionnaires.

4.3. Contribution to clinical practice

The new model for scoring shoulder function is a contribution to the transfer of new technology into clinical practice. Together with progress of hardware technology, miniaturization, wireless transmission, a drop in electronic costs and the development of user-friendly software, the simplification of body-worn sensors measurement procedure might render this approach more accessible to health professionals (Aminian and Najafi 2004).

As it is related to shoulder function questionnaires, the B–B score can be considered as a valid measurement tool of shoulder function. Due to its excellent sensitivity and specificity, it may be used in clinics to diagnose shoulder function alteration caused by rotator cuff tear or shoulder arthritis. Nevertheless, though the B–B Score is able to detect pathologies, it is not able to discriminate them.

The development of a simplified kinematic score is a contribution to an objective evaluation of shoulder function. Further research will be necessary to better understand the complementarities of objective and subjective approaches in shoulder function evaluation.

4.4. Study strengths and limitations

The process of analysis in this study implied that the B–B Score can at best perform equivalently to the reference score in the assessment of kinematic shoulder performance. Due to

its consistent resemblance to the reference score over the period of rehabilitation, it can be expected that the simplified score displays comparable kinanthropometric measurement properties. The advantage of the B–B Score over the reference score mainly resides in its clinical practicality.

The simplicity of the B–B Score allows measurement repetition. As the variability and error in a measurement mean score decreases with the square root of the repetitions number (assuming a normal distribution of error), test replication and averaging over intra-individual trials (Winer *et al* 1981) may overcome the limitation linked to the possible model's discrepancy in individual measurements.

The kinematic scores would be biased toward an overestimation in case of bilateral symptomatic shoulder condition. Therefore, the absolute value of the score would not be indicative of the real shoulder function in this case. When the reference side is not healthy, the score can only be used to follow-up shoulder function evolution toward improvement or degradation, provided that the reference side is stable.

The mean age of the patient group was purposefully higher compared to the control group, to avoid the inclusion of subjects with asymptomatic rotator cuff as control. Thus, the effect of age on the B–B score could not be investigated in this study. Theoretically, age-related degradation should have no influence as the subject serves as his own control. It must be considered that the B–B score reflects the function of the pathological shoulder compared to the normal shoulder function of the subject accounting for physiological aging. Further research is needed to investigate the possible effect of age on the B–B Score.

Further studies are warranted to validate exhaustively the B–B Score for various shoulder pathologies, with particular consideration given to measurement reproducibility, responsiveness and concurrent validity.

4.5. Conclusion

The primary aim of this study was to design a simplified kinematic shoulder scoring model based on a selection of essential movements among the seven movements of daily life comprised in the reference P Score. The secondary aim was to compare the results of the new simplified scoring model with the reference P score.

PCA and multiple regression were used to create simplified scoring models. Separate factorial ANOVA with measures on both factors were used to select the model presenting the best congruency with the reference model. The relationship between the reference and the new scoring model was evaluated using linear regression. The limits of agreement between models were evaluated using the Bland and Altman method. The validity of the new scoring model was evaluated calculating the correlations with shoulder function validated questionnaires. Finally, diagnostic power for shoulder pathology detection was calculated using receiver operating characteristic (ROC) analysis.

'Elevation' and 'rotation' movements were identified as the essential components of shoulder function. This study has shown that a measurement procedure including only two essential movements can replace a more complex seven-movement score without any significant information loss. Among all relevant two-movement models, the B–B Score (hand to the Back and hand upwards as to change a Bulb) was the best substitute for the reference score, due to its congruent evolution pattern across the period of rehabilitation compared to those of the reference model and to its clinical relevance for shoulder function evaluation.

The B–B Score and the reference score produced comparable outcomes as far as group measurement is concerned, but as might be expected, they could produce differing results during the assessment of individual patients.

The new score is a valid measurement method of shoulder function for the study population. It is able to discriminate accurately healthy subjects from patients suffering from rotator cuff or arthritis and is correlated to clinical questionnaires. The B–B Score is a contribution to objective evaluation of the shoulder function and to its routine application in physiotherapy, surgery and rehabilitation.

The practicality of the B–B Score allows for completion of repeated measurements, which could prove useful in decreasing measurement variability and establishing requisite measurement precision for effective intra-subject evaluations. Application of this new model to shoulder conditions other than those considered in this study should be validated prior to use. Further studies are warranted for an extensive validation of the B–B Score.

References

American Academy Of Orthopaedic Surgeons 2010 *Optimizing the management of rotator cuff problems: guideline and evidence report* Available: (www.aaos.org/research/guidelines/RCP_guideline.pdf) [Accessed 28.09.2014]

Aminian K 2006 Human movement capture and their clinical applications *Computational Intelligence for Movement Sciences: Neural Networks, Support Vector Machines and other Emerging Techniques* Eds R K Begg and M Palaniswami (USA: Ideas group)

Aminian K and Najafi B 2004 Capturing human motion using body-fixed sensors: outdoor measurement and clinical applications *Comput. Animat. Virtual Worlds* **15** 79–94

Bernmark E and Wiktorin C 2002 A triaxial accelerometer for measuring arm movements *Appl. Ergon.* **33** 541–7

Bunker T 2002 Rotator cuff disease *Curr. Orthop.* **16** 223–33

Coley B 2007 *Shoulder function and outcome evaluation after surgery using 3D inertial sensors* (Doctorate ès Sciences, Swiss Institute of Technology)

Coley B, Jolles B M, Farron A, Bourgeois A, Nussbaumer F, Pichonnaz C and Aminian K 2007 Outcome evaluation in shoulder surgery using 3D kinematics sensors *Gait Posture* **25** 523–32

Coley B, Jolles B M, Farron A, Pichonnaz C, Bassin J P and Aminian K 2008 Estimating dominant upper-limb segments during daily activity *Gait Posture* **27** 368–75

Constant C R and Murley A G 1987 A clinical method of functional assessment of the shoulder *Clin. Orthop. Relat. Res.* **214** 160

Cutti A G, Giovanardi A, Rocchi L, Davalli A and Sacchetti R 2008 Ambulatory measurement of shoulder and elbow kinematics through inertial and magnetic sensors *Med. Biol. Eng. Comput.* **46** 169–78

Duc C, Farron A, Pichonnaz C, Jolles B M, Bassin J P and Aminian K 2013 Distribution of arm velocity and frequency of arm usage during daily activity: objective outcome evaluation after shoulder surgery *Gait Posture* **38** 247–52

Fayad F *et al* 2008 Validation of the french version of the disability of the arm, shoulder and hand questionnaire (F-DASH) *Joint Bone Spine* **75** 195–200

Fayad F, Mace Y and Lefevre-colau M M 2005 Les échelles d'incapacité fonctionnelle de l'épaule: revue systématique *Annales de Réadaptation et de Médecine Physique* **48** 298–306

Hanley J A and Mcneil B J 1982 The meaning and use of the area under a receiver operating characteristic (ROC) curve *Radiology* **143** 29–36

Jester A, Harth A, Wind G, Germann G and Sauerbier M 2005 Disabilities of the arm, shoulder and hand (DASH) questionnaire: determining functional activity profiles in patients with upper extremity disorders *J. Hand Surg. Br.* **30** 23–8

Jolles B M, Duc C, Coley B, Aminian K, Pichonnaz C, Bassin J-P and Farron A 2011 Objective evaluation of shoulder function using body-fixed sensors: a new way to detect early treatment failures? *J. Shoulder Elbow Surg.* **20** 1074–81

Jolliffe I T 2002 *Principal Component Analysis* (New York: Springer)

Körver R J, Heyligers I C, Samijo S K and Grimm B 2014a Inertia based functional scoring of the shoulder in clinical practice *Physiol. Meas.* **35** 167–76

Körver R J, Senden R, Heyligers I C and Grimm B 2014b Objective outcome evaluation using inertial sensors in subacromial impingement syndrome: a five-year follow-up study *Physiol. Meas.* **35** 677–86

Koss S, Richmond J C and Woodward J S 1997 Two-to five-year followup of arthroscopic Bankart reconstruction using a suture anchor technique *Am. J. Sports Med.* **25** 809

Lippitt S B, Harryman D T and Matsen F A 1993 A practical tool for evaluating function: the simple shoulder test *The Shoulder: a Balance of Mobility and Stability* ed PA Matsen *et al* (Rosemont: American Academy of Orthopaedic Surgery)

Luinge H, Veltink P and Baten C 2007 Ambulatory measurement of arm orientation *J. Biomech.* **40** 78–85

Luinge H J and Veltink P H 2005 Measuring orientation of human body segments using miniature gyroscopes and accelerometers *Med. Biol. Eng. Comput.* **43** 273–82

Magermans D J, Chadwick E K J, Veeger H E J and Van Der Helm F C T 2005 Requirements for upper extremity motions during activities of daily living *Clin. Biomech.* **20** 591–9

Mathie M, Coster A, Lovell N and Celler B 2004 Accelerometry: providing an integrated, practical method for long-term, ambulatory monitoring of human movement *Physiol Meas* **25** R1–20

Murrell G A C and Walton J R 2001 Diagnosis of rotator cuff tears *Lancet* **357** 769–70

Oh J H, Jo K H, Kim W S, Gong H S, Han S G and Kim Y H 2009 Comparative evaluation of the measurement properties of various shoulder outcome instruments *Am. J. Sports Med.* **37** 1161–8

Pandyan A P, Van Wijck F, Johnson G R, Greenfield T and Tony G 2002 Instrumentation in experimentation *Research Methods For Postgraduates* (London: Arnold)

Placzek J D, Lukens S C, Badalanmenti S, Roubal P J, Freeman D C, Walleman K M, Parrot A and Wiater J M 2004 Shoulder outcome measures *Am. J. Sports Med.* **32** 1270–7

Portney L G and Watkins M P 2009 *Foundations of Clinical Research: Applications to Practice* (Upper Saddle River NJ, USA: Prentice Hall Health)

Sher J S, Uribe J W, Posada A, Murphy B J and Zlatkin M B 1995 Abnormal findings on magnetic resonance images of asymptomatic shoulders *J. Bone Joint Surg. Am.* **77** 10–5

Teece R, Lunden J, Lloyd A, Kaiser A, Cieminski C and Ludewig P 2008 Three-dimensional acromioclavicular joint motions during elevation of the arm *J. Orthop. Sports Phys. Ther.* **38** 181

Van Der Windt D A, Koes B W, De Jong B A and Bouter L M 1995 Shoulder disorders in general practice: incidence, patient characteristics and management *Ann. Rheum. Dis.* **54** 959–64

Wilcox R Arslanian L and Millett P 2005 Rehabilitation following total shoulder arthroplasty *J. Orthop. Sports Phys. Ther.* **35** 821–36

Winer B, Brown D and Michels K 1981 *Statistical Principles In Experimental Design* (New York: McGraw-Hill)

Wong W Y, Wong M S and Lo K H 2007 Clinical applications of sensors for human posture and movement analysis: a review *Prosthet. Orthot. Int.* **31** 62–75

Wu G *et al* 2005 ISB recommendation on definitions of joint coordinate systems of various joints for the reporting of human joint motion--Part II: shoulder, elbow, wrist and hand *J. Biomech.* **38** 981–92

Zhou H, Hu H, Harris N and Hammerton J 2006 Applications of wearable inertial sensors in estimation of upper limb movements *Biomed. Signal Process. Control* **1** 22–32

Appendix II

Acceptation letter Ré-Sa-R found

Hes·so
Haute Ecole Spécialisée
de Suisse occidentale
Fachhochschule Westschweiz
University of Applied Sciences
Western Switzerland

SIEGE HES-SO / MISSIONS HES-SO
GAETANO MASSA / COLLABORATEUR SCIENTIFIQUE
RUE DE LA JEUNESSE 1 / CASE POSTALE 452 / 2800 DELÉMONT /
TÉL. : 032 424 49 12 / FAX 032 424 49 01 /
GAETANO.MASSA@HES-SO.CH

WWW.HES-SO.CH

Haute école cantonale vaudoise de la santé
HECVSanté
Monsieur Claude Pichonnaz
Avenue de Beaumont 21
1011 Lausanne

Delémont, le 8 décembre 2010 / ach

Requête Ra&D Ré-Sa-R 17-10 n° SAGE-X 24519 : « Validation d'un score cinématique de la fonction de l'épaule incluant uniquement les mouvements essentiels »

Monsieur,

Nous avons pris connaissance avec attention du projet susmentionné et avons le plaisir de vous informer que les conditions sont réunies pour qu'il obtienne le financement prévu pour sa préparation. Le versement de la subvention à hauteur de CHF 22'000.- sera effectué par le service financier du siège de la HES-SO sur présentation d'une facture avec BV-BVR, avec la mention « demande de transfert de subvention » et après avoir complété les écrans no 220 et 222 (charges et heures réelles) de l'outil Sagex.

Nous restons à votre disposition pour tout complément d'information et vous prions d'agréer, Monsieur, nos salutations distinguées.

Marc-André Berclaz
Président du Comité directeur

Copies :
- A la direction de l'école concernée
- Au service financier de l'école/établissement
- Au coordinateur du réseau concerné
- Au service financier du siège

Appendix III

Acceptation letter Swiss National Science Foundation

FONDS NATIONAL SUISSE
DE LA RECHERCHE SCIENTIFIQUE

www.snf.ch
Wildhainweg 3, case postale 8232, CH-3001 Berne

Division des sciences humaines et sociales
DORE – Instrument de promotion pour la recherche orientée vers la pratique

Tél. +41 31 308 22 22
Fax +41 31 305 29 75
E-mail dore@snf.ch

Professeur Claude Pichonnaz
Filière Physiothérapie
Haute école cantonale vaudoise de la santé
HES-SO
Avenue Beaumont 21
CH-1011 Lausanne

Berne, le 23 mars 2011

Décision
13DPD6_135061 / 1

Monsieur,

Nous avons le plaisir de vous informer que le Conseil de la recherche a décidé de vous allouer un subside de recherche de CHF 140'600.00 pour le projet "Validation d'un score cinématique de la fonction de l'épaule incluant uniquement les mouvements essentiels".

La répartition et les conditions de l'octroi ci-jointes font partie intégrante de la décision. De plus, il convient d'observer particulièrement les dispositions du "Règlement des subsides" et du "Règlement d'exécution général relatif au règlement des subsides". Vous trouverez ces documents sur le serveur du FNS (cf. "Documents juridiques" ci-dessous). Des exemplaires vous sont volontiers remis sur demande. Si votre requête a été déposée en commun avec d'autres personnes, nous vous prions d'observer l'obligation d'informer mentionnée aux articles 14 et 32 ss du "Règlement des subsides".

Nous vous prions de remplir et nous remettre par voie électronique le formulaire en ligne "Demande de déblocage du subside" (www.mysnf.ch).

En vous souhaitant plein succès dans la réalisation de votre projet, nous vous prions d'agréer, Monsieur, nos salutations distinguées.

Susanne Grossniklaus

Appendix IV

Acceptation letter of Ethical Commission of the Faculty of biology and medicine of the University of Lausanne

canton de vaud
LIBERTÉ ET PATRIE

Commission cantonale
d'éthique de la recherche
sur l'être humain
Rue du Bugnon 21
1011 Lausanne

Prof. M. Burnier
Président

Secrétariat central
Tél. 021 692 50 08
Fax 021 692 50 05

Sous-Commission II
Président Prof. R. Darioli
Tél. 021 692 50 95
Secrétariat: tél. 021 692 50 40
E-mail: catherine.corbaz@unil.ch

M. Claude Pichonnaz
Professeur HES
HECVSanté-filière physiothérapeutes
Av. de Beaumont 21
1011 Lausanne

Lausanne, le 18 octobre 2010
RD/cc

Avis de la Commission cantonale (VD) d'éthique de la recherche sur l'être humain

---

Monsieur et cher Collègue,

Lors de sa séance du **7 septembre 2010**, la Commission cantonale (VD) d'éthique de la recherche sur l'être humain, Sous-Commission II (composition détaillée en page 4) a procédé à une évaluation approfondie du projet de recherche désigné ci-après :

Protocole 205/10 : Etude pilote pour la validation d'un score cinématique de la fonction de l'épaule incluant uniquement les mouvements essentiels

Investigateur :

M. Claude Pichonnaz
Professeur HES
HECVSanté-filière physiothérapeutes
Av. de Beaumont 21
1011 Lausanne

Copie : Mme Anne-Sylvie Fontannaz, Pharmacien cantonal, Service de la santé publique, Rue Cité-Devant 11, 1014 Lausanne

La Commission cantonale (VD) d'éthique de la recherche sur l'être humain, Sous-Commission II, base son appréciation sur les documents soumis les 21 juillet 2010 et 18 octobre 2010, soit :

- Vos lettres des 19 juillet 2010 et 7 octobre 2010
- Formulaire de base du 21 juillet 2010
- Protocole
- Feuille d'information, version modifiée du 7 octobre 2010
- Formulaire de consentement, version modifiée du 7 octobre 2010
- Dossier du patient
- Attestation du CHUV du 19 juillet 2010
- CV des Prof. Pichonnaz, Bassin, Farron, Aminian, Dr. Jolles-Haeberli, Mme Duc

☒ procédure ordinaire ☐ procédure simplifiée ☐ évaluation ultérieure

La Commission d'Ethique arrête l'**avis** suivant :

☒ **A Avis positif**

☐ **B Avis positif assorti de recommandations** (v. page 3 et suiv.)

Information écrite à la Commission d'éthique suffisante ☐

☐ **C Avis conditionnel** (v. page 3 et suiv.)

Evaluation ultérieure par la Commission d'éthique nécessaire ☐

Information écrite à la Commission d'éthique suffisante ☐

☐ **D Avis négatif motivé (et explication pour réexamen)** (v. page 3 et suiv.)

☐ **E Avis justifié de ne pas entrer en matière** (v. page 3 et suiv.)

L'avis s'applique également aux autres investigateurs mentionnés dans la demande d'évaluation, travaillant dans des sites de recherche relevant du champ de compétence de la Commission cantonale (VD) d'éthique de la recherche sur l'être humain.

Pour mémoire : **Obligations de l'investigateur**

Les produits testés et de comparaison (médicaments et dispositifs médicaux) doivent être fabriqués, évalués et utilisés conformément aux règles de l'art visant à en garantir la qualité et la sécurité.

Devoir de signaler :
a) immédiatement tout événement indésirable grave (serious adverse events)
b) toute information devenant disponible au cours de l'essai et ayant des conséquences directes pour la sécurité des sujets et la poursuite de l'essai
c) Modification du protocole
d) Fin ou arrêt prématuré de l'essai

Rapport intermédiaire : une fois par année

Notification d'essais de médicaments auprès de Swissmedic et de dispositifs médicaux auprès de l'OFSP (en cas d'étude sponsorisée, cette tâche incombe au promoteur)

Rapport final

Emolument : CHF 200.—(TVA en sus)

La Commission cantonale (VD) d'éthique de la recherche sur l'être humain, Sous-Commission II :

Prof. Roger Darioli
Président de la Sous-Commission II

FORMULAIRE DE RAPPORT INTERMEDIAIRE/FINAL
pour La Commission cantonale (VD) d'éthique de la recherche sur l'être humain
Sous-Commission II

Protocole 205/10 : Etude pilote pour la validation d'un score cinématique de la fonction de l'épaule incluant uniquement les mouvements essentiels

Investigateur:
M. Claude Pichonnaz
Professeur HES
HECVSanté-filière physiothérapeutes
Av. de Beaumont 21
1011 Lausanne

Etude terminée: OUI ❑ NON ❑ **Etude en cours:** OUI ❑ NON ❑

Etude arrêtée: OUI ❑ NON ❑

Si OUI, pourquoi?

.....

.....

.....

Conclusions qui peuvent être tirées de l'étude à ce stade:

.....

.....

.....

.....

Y a-t-il eu des incidents pendant l'étude? OUI ❑ NON ❑

Si OUI, lesquels?

.....

.....

.....

Fin de l'étude prévue pour: .....

Date: ..... **Signature de l'investigateur:** .....

A retourner au: Prof. M. Burnier
Président de la Commission cantonale (VD) d'éthique de la recherche sur l'être humain
Rue du Bugnon 21, 1011 Lausanne
En cas de publication, prière d'adresser un tiré-à-part

Date d'acceptation par la Commission cantonale (VD) d'éthique : **18.10.2010**
Date limite pour le renvoi de ce formulaire : **18.10.2011**

Séance du 7 septembre 2010

Composition de la Commission cantonale (VD) d'éthique de la recherche sur l'être humain

Sous-Commission II

L'avis de la Commission d'Ethique ayant siégé dans sa composition détaillée ci-après est valable, le quorum étant atteint (art. 32 de l'Ordonnance sur les essais cliniques de produits thérapeutiques du 17 octobre 2001).

					participe à l'avis	
	Nom, prénom	**Profession, titre**	H	F	**oui**	**non**
Présidence	Darioli Roger	Professeur honoraire, Spécialiste en médecine interne	☒	☐	☒	☐
Membres	Bringolf Michel	Spécialiste FMH en néphrologie	☒	☐	☒	☐
	Caci Mirela	Juriste	☐	☒	☒	☐
	Gardaz Jean-Patrice	Spécialiste FMH en anesthésiologie et médecine intensive, Adjoint-scientifique	☒	☐	☐	☒
	Jérôme-Choudja Cécile	Chef de clinique / Ethicienne	☐	☒	☐	☒
	Maillard Marc	Pharmacien, Chef de projet de recherche	☒	☐	☒	☐
	Manson Jan-Anders	Professeur ordinaire EPFL	☒	☐	☐	☒
	Manuel Oriol	Spécialiste en maladie infectieuse	☒	☐	☒	☐
	Paillex Roland	Physiothérapeute chef du CHUV	☒	☐	☒	☐
	Rège Walther Myriam	Cheffe de projet de recherche, IUMSP	☐	☒	☒	☐
	Rochat Etienne	Pasteur	☒	☐	☒	☐
	Schaller Marie-Denise	Professeur associée, Médecin-cheffe SIA	☐	☒	☐	☒
	Schild Laurent	Professeur ordinaire, Pharmacologue	☒	☐	☒	☐

Appendix V

Acceptation letter of amendment of the Ethical Commission of the Faculty of biology and medicine of the University of Lausanne

canton de vaud
LIBERTÉ ET PATRIE

Commission cantonale
d'éthique de la recherche
sur l'être humain
Rue César-Roux 19
1005 Lausanne

Prof. R. Darioli
Président

Secrétariat
Tél. 021 314 5598/5601/8622
Fax 021 314 76 01
E-mail: secretariatcervd@unil.ch

Sous-Commission II
Président Prof. R. Darioli
Tél. 021 314 56 29

M. Claude Pichonnaz
Professeur HES
HECVSanté-filière physiothérapeutes
Av. de Beaumont 21
1011 Lausanne

Lausanne, le 6 juin 2011
RD/cc

Protocole 205/10 : Etude pilote pour la validation d'un score cinématique de la fonction de l'épaule incluant uniquement les mouvements essentiels

Monsieur,

Je vous remercie de m'avoir adressé votre courrier du 1er juin 2011 concernant un amendement au protocole susmentionné.

Je vous remercie également de m'avoir soumis les documents suivants:

- Enumération des modifications apportées au protocole
- Formulaire d'information, version modifiée du 1er juin 2011
- Formulaire de consentement, version modifiée du 1er juin 2011
- Dossier d'étude des patients, version modifiée du 1er juin 2011

Ces documents ne posant pas de problème éthique, la Commission les **accepte**.

Je vous présente, Monsieur, mes meilleures salutations.

Prof. Roger Darioli
Président de la Sous-Commission II

Appendix VI

ClinicalTrials registry receipt of registration

Phase 1

ClinicalTrials.gov Protocol Registration and Results System (PRS) Receipt
Release Date: April 13, 2011

ClinicalTrials.gov ID: NCT01281085

Study Identification

Unique Protocol ID:	SAGEX-24519
Brief Title:	Validation of a Score for Shoulder Function Evaluation Based on Movement Analysis
Official Title:	Validation of a Kinematic Functional Shoulder Score Including Only Essential Movements: a Pilot Study
Secondary IDs:	

Study Status

Record Verification:	April 2011
Overall Status:	Completed
Study Start:	October 2010 []
Primary Completion:	April 2011 [Actual]
Study Completion:	April 2011 [Actual]

Sponsor/Collaborators

Sponsor:	Haute Ecole Cantonale Vaudoise de Santé
Responsible Party:	
Collaborators:	Réseau d'études appliquées des pratiques de Santé de Réadapation (ré)insertion

Oversight

U.S. FDA-regulated Drug:	
U.S. FDA-regulated Device:	
U.S. FDA IND/IDE:	No
Human Subjects Review:	Board Status: Approved Approval Number: 205/10 Board Name: Commission cantonale (VD) d'éthique de la recherche sur l'être humain Board Affiliation: Commission cantonale (VD) d'éthique de la recherche sur l'être humain Phone: +41021 692 50 08 Email: Roger.Darioli@hospvd.ch Address:

	Commission cantonale (VD) d'éthique de la recherche sur l'être humain Secrétariat central Rue du Bugnon 21 CH-1011 Lausanne
Data Monitoring:	No
FDA Regulated Intervention:	No

Study Description

Brief Summary:	Questionnaires are frequently used to evaluate shoulder function in various diseases or after surgery. However, measurement of shoulder function is presently a controversial issue. Shoulder movement analysis based on embedded sensors could be a promising alternative to questionnaires. Some studies already demonstrated the relevance of this approach. It has also been demonstrated that a simple testing procedure including only two arm movements produces comparable results to more complicated testing procedures. However, more studies are needed to extensively establish if this simplified testing procedure provides a trustworthy evaluation of patient shoulder function and its evolution. This study is a preliminary study which aims to develop a precise testing procedure which will be used in a future study aiming to evaluate measurement properties of a simple shoulder function test based on movement sensors.
Detailed Description:	Background Measurement of shoulder function is a controversial issue. There is a great variety of measurement tools but none of them has been universally accepted. There is therefore a need to develop extensively validated and convenient measurement tools. Embedded computerized movement analysis can potentially meet these requirements for measurement of shoulder function. Ambulatory measurement devices allow application in various clinical conditions, display adequate precision and accuracy, and are considerably more straightforward than laboratory-based systems. Using a Physilog ® II embedded system, Coley (2007) developed a relatively simple score of shoulder function (P Score). The method is based on arm power measurement by three-dimensional accelerometers and gyroscopes during seven consecutive shoulder movements. It demonstrated reliability, responsiveness and criterion-based validity. However, additional knowledge and technological progress could now contribute to further simplification of the. A secondary analysis of Coley's study data based on principal component analysis and multiple regressions highlighted that a procedure including only two selected movements produces comparable results to P Score. Moreover, the development of wireless systems considerably simplifies set up. Consequently, simpler but equivalent measurement procedure can now be considered. However some important issues have to be clarified before an extensive validation study can be undertaken. The simplicity of the testing procedure allows test replication. However, the number of movement replications needed to obtain a reliable outcome is presently unknown. Relevance of testing procedure and study feasibility have also to be evaluated. A pilot study is needed to clarify these issues. Aim The aim of this pilot study is to determine the number of movement replications needed to obtain a reliable result using a simplified cinematic

shoulder measurement procedure as well as to evaluate testing procedure and study protocol.

Methods Measurement will be carried out with four groups of patients presenting with frequent shoulder conditions (rotator cuff condition, shoulder instability, diaphyseal or subcapital humerus fracture, frozen shoulder). Measurement procedure includes two consecutive measurements, alternatively conducted by two evaluators. Currently used functional questionnaires will be completed at both stages.

Statistical analysis will address outcome variability according to number of replications and reproducibility.

Conditions

Conditions: Shoulder Pain

Keywords: Shoulder
Outcome treatment
Validation Studies
Biomechanics

Study Design

Study Type: Observational

Observational Study Model: Other

Time Perspective: Prospective

Biospecimen Retention: None Retained

Biospecimen Description:

Enrollment: 16 [Actual]

Number of Groups/Cohorts: 1

Groups and Interventions

Groups/Cohorts	Interventions
No treatment Shoulder conditions including rotator cuff condition treated conservatively, shoulder instability treated conservatively, diaphyseal humerus fracture or subcapital humerus fracture treated surgically and frozen shoulder	

Outcome Measures

Primary Outcome Measure:

1. variability of kinematic simplified functional score
   This pilot study mainly aims at defining how many tests replications are needed to obtain a reliable shoulder function measurement

   [Time Frame: only one measurement session]

Secondary Outcome Measure:

2. intra- and inter-reproducibility of kinematic functional shoulder score
   As measurement are performed twice respectively by two evaluators, this pilot study will provide a first insight of test reliability

[Time Frame: only one measurement session]

Eligibility

Study Population:	Patients diagnosed with a shoudler condition, as stated during the medical examination performed at the specialized shoulder consultation of the hospital
Sampling Method:	Non-Probability Sample
Minimum Age:	18 Years
Maximum Age:	
Sex:	All
Gender Based:	
Accepts Healthy Volunteers:	No
Criteria:	Inclusion Criteria:

- Rotator cuff condition, conservative treatment indicated
- Shoulder instability, conservative treatment indicated
- Diaphyseal humerus fracture or subcapital humerus fracture treated surgically, at 6 weeks post surgery
- Frozen shoulder, conservative treatment indicated

Exclusion Criteria:

- Bilateral shoulder condition or other shoulder condition than the ones mentioned in inclusion criteria
- Any concomitant pain or condition involving upper limb
- Cervical spine condition involving upper limb pain or mobility restriction
- Insufficient French language level to understand patient information form, consent form or questionnaires
- Insufficient ability to give truly informed consent or to understand questionnaires. It will be proceeded to a Mini Mental State score in case of uncertainty, with exclusion criteria at 24 points/30 (ANAES 2000).
- Medical contraindication to execute movements required for score completion
- Tumor
- Neurological condition interfering with test

Contacts/Locations

Central Contact Person:	Claude A. Pichonnaz, PT MSc Telephone: +41213168126 Email: cpichonn@hecvsante.ch
Central Contact Backup:	Jean-Philippe Bassin, PT OMT Telephone: +41 21 31 68 133 Email: jbassin@hecvsante.ch
Study Officials:	Claude A. Pichonnaz, PT MSc Study Principal Investigator HECVSanté and CHUV-UNIL
Locations:	Switzerland Département de l'Appareil Locomoteur - CHUV Lausanne, Switzerland, 1005 Contact: Claude A Pichonnaz, PT MSc 0041 21 3168126 cpichonn@hecvsante.ch

Contact: Bassin Jean-Philippe, PT OMT 0041 21 3168133
jbassin@hecvsante.ch
Sub-Investigator: Jean-Philippe Bassin, PT OMT
Sub-Investigator: Guillaume Christe, PT

IPDSharing

Plan to Share IPD:

References

Citations: Coley B, Jolles BM, Farron A, Bourgeois A, Nussbaumer F, Pichonnaz C, Aminian K. Outcome evaluation in shoulder surgery using 3D kinematics sensors. Gait Posture. 2007 Apr;25(4):523-32. Epub 2006 Aug 28. PubMed 16934979

Links: URL: http://www.resar.ch/images/stories/Projets_esquisses/esquisse_Claude_Pichonnaz.pdf
Description Brief study summary in French published by the sponsor

Available IPD/Information:

Appendix VII

Baseline patient file

Date :

Code patient :

Etude de validation d'un score cinématique simplifié de la fonction l'épaule

Baseline

Physiothérapeute responsable de l'étude : Claude Pichonnaz

Médecin responsable : Prof. Alain Farron

Collaborateurs de recherche:
M. Jean-Philippe Bassin, Mme Cyntia Duc, Prof. Brigitte Haeberli-Jolles, Prof. Kamiar Aminian

Adresse de contact

CHUV-UNIL

Département de l'appareil locomoteur – service de physiothérapie

Claude Pichonnaz

4, Avenue Pierre Decker

1005 Lausanne

021 316 81 26

Email: cpichonn@hecvsante.ch

Prof. A. Farron **Claude Pichonnaz** **Prof. Kamiar Aminian**

Prof. B. Jolles **Jean-Philippe Bassin** **Cyntia Duc**

Formulaire d'information

Validation d'un score cinématique fonctionnel de l'épaule incluant uniquement les mouvements essentiels

Présentation

Les physiothérapeutes et médecins disposent de différents moyens pour contrôler l'efficacité des traitements proposés pour les affections de l'épaule. Parmi ces moyens, on trouve de nombreux questionnaires, que le patient remplit lui-même afin d'évaluer par exemple la douleur, la mobilité de l'épaule ou les répercussions sur les activités de la vie quotidienne. Ces données subjectives sont évidemment très importantes, mais il manque en association un élément objectif d'évaluation du résultat du traitement. De précédentes études ont montré qu'un test simple basé sur l'analyse de deux mouvements de l'épaule par des capteurs permet de suivre efficacement l'évolution du patient. Cette étude vise à évaluer la qualité des mesures obtenues lors de ce test, ainsi que la fiabilité et la précision des résultats. Deux instruments de mesures seront utilisés pour recueillir les données : un système de mesure du mouvement Physilog ®, ainsi qu'un iPod®. Outre sa fonction première (permettre l'écoute de fichiers musicaux), l'iPod est en effet munis de capteurs de mouvement que nous allons utiliser pour mesurer votre épaule.

Objectif de l'étude

L'étude vise à évaluer la qualité des mesures obtenues lors d'un test de la fonction de l'épaule basé sur l'analyse par capteurs de deux mouvements essentiels (mettre la main dans le dos et lever le bras).

Quelles personnes sont concernées par l'étude ?

Les patients présentant des pathologies de l'épaule.

Des participants sans problème d'épaule seront également mesurés afin d'établir le résultat de référence de la personne saine.

Qui sont les investigateurs de l'étude ?

Plusieurs institutions collaborent de manière interdisciplinaire à cette étude. Des physiothérapeutes de la Haute Ecole Cantonal Vaudoise de Santé (HECVSanté), des médecins et des physiothérapeutes du Département de l'Appareil Locomoteur du CHUV, et des ingénieurs du Laboratoire de Mesure et d'Analyse du Mouvement de l'EPFL participent au projet.

Quels sont les principes de l'étude ?

Un capteur qui enregistre les vitesses angulaires et les accélérations du bras est collé par système velcro respectivement du côté atteint, puis du côté sain (Fig. 1). L'analyse de mesures permet d'avoir une représentation de la manière dont la personne bouge le bras et de comparer le mouvement des deux côtés.

Figure 1: Exemple de capteur positionné sur le bras

Comment se déroulent les tests ?

Deux collaborateurs de l'étude prennent le volontaire en charge pour la réalisation du test. Chacun des collaborateurs effectuera le test à tour de rôle avec vous.

Le 1er collaborateur met en place sur votre bras le capteur Physilog, qui communique avec le boîtier récepteur, ainsi que l'iPod. Le volontaire effectue cinq répétitions de mouvements simples de l'épaule, qui sont enregistrés par le boîtier récepteur.

Ensuite, le 2ème collaborateur répétera avec vous la même procédure de test que son collègue.

Nous vous prierons aussi de remplir des questionnaires qui permettront de mettre en relation les résultats obtenus avec ce que vous vivez quotidiennement.

L'ensemble de la procédure dure environ 60 minutes.

Combien de séances sont nécessaires ?

Deux séances de mesure espacées de 6 mois sont nécessaires. Ceci permet ainsi d'évaluer votre niveau initial et votre évolution dans le temps. La première séance se déroulera tel que décrit ci-dessus. Lors de la 2ème séance, un seul évaluateur vous prendra en charge et les tests ne seront donc répétés que deux fois.

Où se déroulent les tests ?

Les séances se dérouleront au service de physiothérapie de l'Hôpital Orthopédique, dans la cité Hospitalière du CHUV.

Y a-t-il des risques pour l'épaule ?

Les tests effectués ne présentent pas un risque supérieur aux mouvements que vous effectuez dans la vie courante. Les mouvements seront effectués en-dessous du seuil de douleur afin de ne pas augmenter des douleurs préexistantes.

Que devez-vous encore savoir ?

Cette étude ne modifie pas le traitement dont vous bénéficierez, qui est identique pour tous les patients, qu'ils fassent partie ou non de l'étude.

L'étude est financée par le Fonds National Suisse de la Recherche Scientifique. Par conséquent, aucun frais lié à l'étude n'est facturé aux assurances ou aux participants.

Le CHUV répond des dommages éventuels que vous pourriez subir dans le cadre de cette étude. Si, pendant ou après l'étude clinique, vous souffrez de problèmes de santé ou d'autres dommages en relation avec l'étude, vous voudrez bien en faire part à M. Claude Pichonnaz, investigateur principal de l'étude, dont les coordonnées sont notées à la fin de cette information, qui prendra les mesures adaptées à votre cas.

La participation à l'étude est volontaire. Vous avez la possibilité de vous retirer de l'étude à tout moment sans avoir à vous justifier et sans préjudice d'aucune sorte.

Toutes les données récoltées seront traitées de façon confidentielle. Elles pourront être transmises à des personnes extérieures en relation directe avec le projet de recherche, sous une forme anonyme uniquement, ainsi qu'à la Commission d'Ethique de la Faculté de Biologie et de Médecine de l'Université de Lausanne et à Swissmedic pour des activités de contrôle. Elles pourront être conservées durant 5 ans au maximum.

Aucun médicament ne sera utilisé pendant l'étude.

Votre médecin traitant pourra être informé de votre participation à l'étude. Une copie de votre dossier d'étude lui sera envoyée si vous faites part de ce souhait au responsable de l'étude, dont vous trouvez les coordonnées ci-dessous.

Un défraiement vous sera accordé pour compenser les frais de déplacement occasionnés par votre participation à l'étude.

Responsable de l'étude :

Claude Pichonnaz, Professeur HES-S2, HECVSanté filière physiothérapie
Avenue de Beaumont 21
1011 LAUSANNE
021 316 81 26
cpichonn@hecvsante.ch

DATE D ADMINISTRATION: ___/___/ 201

Prof. A. Farron **Claude Pichonnaz** **Prof. Kamiar Aminian**
Prof. B. Jolles **Jean-Philippe Bassin**

FORMULAIRE DE CONSENTEMENT ECLAIRE

Je soussigné(e) certifie que le Docteur …… m'a proposé de participer à l'étude intitulée:

« Validation d'un score cinématique fonctionnel de l'épaule incluant uniquement les mouvements essentiels »

- J'ai été informé(e) des buts et du déroulement de l'étude ci-dessus.
- J'affirme avoir lu attentivement et compris les informations écrites fournies, informations à propos desquelles j'ai pu solliciter toutes les explications nécessaires à la prise de ma décision.
- Je certifie avoir été informé(e) des avantages et des risques éventuels liés à cette étude et des obligations qui m'incombent pour la participation à l'étude.
- Je confirme notamment que j'ai eu suffisamment de temps pour réfléchir à ma participation.
- J'ai été informé(e) du fait que je pouvais interrompre à tout instant ma participation à cette étude sans avoir à me justifier et sans préjudice d'aucune sorte.
- Je consens à ce que les données recueillies pendant l'étude puissent être transmises à des personnes extérieures en relation avec le projet de recherche sous une forme anonyme, ainsi qu'à la Commission d'Ethique de la Faculté de Biologie et de Médecine de l'Université de Lausanne et à Swissmedic pour des activités de contrôle.
- Je consens à ce que mon médecin traitant soit informé de ma participation à cette étude.

J'accepte donc de participer à l'étude mentionnée dans l'en-tête.

Nom, prénom du patient / de la patiente :.....

Date : .....Signature du patient : .....

Nom du responsable de l'étude : .....

Date : .....Signature du responsable: .....

DATE D ADMINISTRATION: ___/___/ 201

A remplir par le clinicien

Diagnostic (entourez) :

Pathologie de la coiffe des rotateurs *Fracture* *Instabilité* *Capsulite rétractile*

Aucun problème d'épaule

1er test effectué par : _____ Taille __. ____ mètre(s)

2ème test effectué par : _____ Poids ____ kilogrammes

3ème test effectué par : _____

4ème test effectué par : _____

Intensité des douleurs et de la raideur dans l'épaule dominante (= la droite si vous êtes droitier, la gauche si vous êtes gaucher):

VAS 1. Quelle fut l'intensité des **douleurs** dans votre épaule au cours de la semaine passée ?

Réglette EVA en mm.: _____

VAS 2. Quelle fut l'intensité de la **raideur** dans votre épaule au cours de la semaine passée?

Réglette EVA en mm.: _____

Intensité des douleurs et de la raideur dans l'épaule non dominante (= la gauche si vous êtes droitier, la droite si vous êtes gaucher):

VAS 3. Quelle fut l'intensité des **douleurs** dans votre épaule au cours de la semaine passée?

Réglette EVA en mm.: _____

VAS 4. Quelle fut l'intensité de la **raideur** dans votre épaule au cours de la semaine passée?

Réglette EVA en mm.: _____

Remarques (particularité de la situation, déroulement des tests, biais éventuel, imprévus, pathologies associées interférant avec les tests…) :

_____

_____

_____

_____

DATE D ADMINISTRATION: ___/___/ 201

Score de Constant

		EPAULE **DROITE**	EPAULE **GAUCHE**
SUBJECTIF 35 POINTS / 100	DOULEUR / **15 points** Evaluation d'après échelle EVA (p. précédente) calcul : 15 - (valeur EVA 100mm x 1,5/10) (arrondir au point entier le plus proche. Si .5, arrondir au point supérieur)		
	NIVEAU D'ACTIVITÉ / **20 points** 1. Handicap professionnel ou occupationnel Evaluation d'après échelle EVA sur 4 points, zone dans laquelle se trouve le curseur (sévère = 0 → aucun = 4) 2. Handicap dans les activités de loisirs Evaluation d'après échelle EVA sur 4 points (sévère = 0 → aucun = 4) 3. La gêne dans le sommeil (oui = 0 pt ; parfois = 1 pt non = 2 pts) 4. Le niveau de travail confortable avec la main (10 pts) Taille 2 pts · Xyphoïde 4 pt · Cou 6 pts · Tête 8 pts · Au-dessus 10 pts SUB-TOTAL / 20 points		
OBJECTIF 65 POINTS / 100	MOBILITÉ ACTIVE **NON DOULOUREUSE/ 40 points** **Flexion** 0-30 / 31-60 / 61-90 / 91-120 / 121-150 / 150-180 0 pt 2 pts 4 pts 6 pts 8 pts 10 pts **Abduction :** 0-30 / 31-60 / 61-90 / 91-120 / 121-150 / 150-180 0 pt 2 pts 4 pts 6 pts 8 pts 10 pts **Rotation externe:** Main derrière la tête, coude en avant : 2 pts Main derrière la tête, coude en arrière : 2 pts Main sur la tête, coude en avant : 2 pts Main sur la tête, coude en arrière : 2 pts Elévation complète : 2 pts **Rotation interne** – Dos de la main sur : Cuisse latérale 0 pt · Fesse 2 pts · Sacro-iliaque 4 pts · L3 6 pts · TH 12 8 pts · TH 7 10 pts SUB-TOTAL / 40 points		
	FORCE MUSCULAIRE / **25 points** Mesurée avec un dynamomètre, durant 5 sec, le bras à 90° d'élévation dans le plan de l'omoplate. Noter le meilleur résultat de la force max. pour 3 répétitions. Le résultat est donnée en newton, donc diviser par 9.81 pour résultats en kg. Pts =nombre de kg. x 2. SUB-TOTAL / 25 points	Newtons :	Newtons :
INDICE FONCTIONNEL DE CONSTANT TOTAL / 100 points			

Constant, C. R., et al. 2008. A review of the Constant score: Modifications and guidelines for its use. *Journal Of Shoulder And Elbow Surgery / American Shoulder And Elbow Surgeons* 17 (2), pp. 355-361.

DATE D ADMINISTRATION: ___/___/ 201

Dossier du patient

Cette section vous demande de préciser quelques informations générales vous concernant:

D1. Indiquez svp votre code postal: __ __ __ __

D2. Indiquez svp votre date de naissance (JJ-MM-AAAA)? __ __ / __ __ / 19__ __

D3. Indiquez svp si vous êtes de sexe *(Cochez une case svp)*

- ❑ Féminin [F]
- ❑ Masculin [M]

D4. Quel est le plus haut niveau d'éducation que vous ayez reçu? *(Cochez une seule case svp)*

- ❑ Ecole primaire/ cours élémentaire [P]
- ❑ Ecole secondaire/ collège / apprentissage [S]
- ❑ Lycée/ université ou équivalent [U]
- ❑ Autre: _____ [O]
- ❑ Ne sait plus [X]

D5. Avez-vous déjà rempli une demande d'invalidité (AI) concernant l'épaule opérée?

- ❑ Oui, j'ai rempli une demande AI, j'ai reçu une compensation dans le passé mais plus actuellement [1]
- ❑ Oui, j'ai rempli une demande AI, je reçois une rente actuellement [2]
- ❑ Oui, j'ai rempli une demande AI, j'attends la décision [3]
- ❑ Non [4]

D6. A quel pourcentage travaillez-vous?

- ❑ Plein-temps [F]
- ❑ Mi-temps [P]
- ❑ N'a pas d'activité professionnelle rémunérée [N]

D7. Indiquez svp votre profession :

_____

D10. Combien d'heures, en moyenne, travaillez-vous chaque semaine: ____heures [WH]

D11. Vous considérez-vous comme

❑ Droitier [ED] ❑ Gaucher [EG] ❑ Ambidextre [EA]

Si vous avez répondu *ambidextre* (=qui utilise indifféremment la main droite ou gauche), cochez quelle main vous utilisez pour :

- écrire ❑ Droite ❑ Gauche
- lancer ❑ Droite ❑ Gauche

Cette section permet de nous renseigner sur votre état général (EQ-5D):

Veuillez indiquer, pour chacune des rubriques suivantes, l'affirmation qui décrit le mieux votre état de santé aujourd'hui, en entourant le numéro approprié.

EQ1. **Mobilité**

1. Je n'ai aucun problème pour me déplacer à pied
2. J'ai des problèmes pour me déplacer à pied
3. Je suis obligé(e) de rester alité(e)

EQ2. **Autonomie de la personne**

1. Je n'ai aucun problème pour prendre soin de moi
2. J'ai des problèmes pour me laver ou m'habiller tout(e) seul(e)
3. Je suis incapable de me laver ou de m'habiller tout(e) seul(e)

EQ3. **Activités courantes** *(ex. travail, études, travaux domestiques, activités familiales ou loisirs)*

1. Je n'ai aucun problème pour accomplir mes activités courantes
2. J'ai des problèmes pour accomplir mes activités courantes
3. Je suis incapable d'accomplir mes activités courantes

EQ4. **Douleurs/gêne**

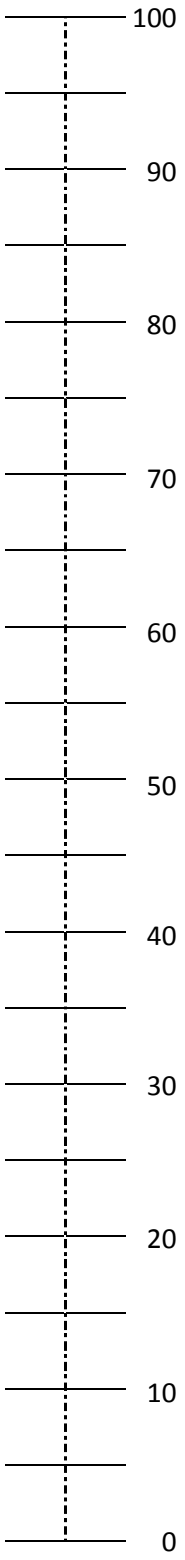
1. Je n'ai ni douleurs, ni gêne
2. J'ai des douleurs ou une gêne modérée(s)
3. J'ai des douleurs ou une gêne extrême(s)

EQ5. **Anxiété/dépression**

1. Je ne suis ni anxieux(se), ni déprimé(e)
2. Je suis modérément anxieux(se) ou déprimé(e)
3. Je suis extrêmement anxieux(se) ou déprimé(e)

DATE D ADMINISTRATION: ___/___/ 201

Meilleur état de santé imaginable



Pire état de santé imaginable

Afin d'évaluer au mieux votre état de santé, nous avons reporté sur cette feuille une échelle en forme de thermomètre où la valeur de 100 correspond à un état de santé le meilleur imaginable et la valeur de 0 correspondant à un état de santé le moins bon imaginable.

Nous vous demandons de tracer une ligne à partir de la lettre **A** et se dirigeant vers la valeur de l'échelle correspondant au mieux à votre état de santé actuel.

Votre état de santé actuel: **A**

DATE D ADMINISTRATION: ___/___/ 201

Valeur subjective de l'épaule

*Indiquez sur l'échelle ci-dessous à combien de % vous coteriez votre épaule **atteinte**, si une épaule complètement normale représente 100%. Cochez une seule case*

☐ ☐

0% 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100%

*Indiquez sur l'échelle ci-dessous à combien de % vous coteriez votre épaule **la plus saine**, si une épaule complètement normale représente 100%.*

☐ ☐

0% 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100%

DATE D ADMINISTRATION: ___/___/ 201

Cette section s'intéresse à ce que vous ressentez et à vos possibilités d'accomplir certaines activités. (QuickDASH)

Veuillez répondre à **toutes** les questions en considérant vos possibilités **au cours des 7 derniers jours**. Si vous n'avez pas eu l'occasion de pratiquer certaines de ces activités au cours des 7 derniers jours, veuillez entourer la réponse qui vous semble la plus exacte si vous aviez dû faire cette tâche. Le côté n'a pas d'importance. Veuillez répondre en fonction du résultat final, sans tenir compte de la façon dont vous y arrivez.

Veuillez évaluer votre capacité à réaliser les activités suivantes **au cours des 7 derniers jours**. Entourez une seule réponse par ligne.

	Aucune Difficulté	Difficulté Légère	Difficulté Moyenne	Difficulté importante	Impossible
1. Dévisser un couvercle serré ou neuf	1	2	3	4	5
2. Effectuer des tâches ménagères lourdes (nettoyage des sols ou des murs)	1	2	3	4	5
3. Porter des sacs de provisions ou une mallette	1	2	3	4	5
4. Se laver le dos	1	2	3	4	5
5. Couper la nourriture avec un couteau	1	2	3	4	5
6. Activités de loisir nécessitant une certaine force ou avec des chocs au niveau de l'épaule du bras ou de la main. (bricolage, tennis, golf, etc..)	1	2	3	4	5

DATE D ADMINISTRATION: ___/___/ 201

	Pas du tout	Légèrement	Moyennement	Beaucoup	Extrêmement
7. Pendant les 7 derniers jours, à quel point votre épaule, votre bras ou votre main vous a-t-elle gêné dans vos relations avec votre famille, vos amis ou vos voisins ? (entourez une seule réponse)	**1**	**2**	**3**	**4**	**5**

	Pas du tout limité	Légèrement limité	Moyennement limité	Très limité	Incapable
8. Avez-vous été limité dans votre travail ou une de vos activités quotidiennes habituelles en raison de problèmes à votre épaule, votre bras ou votre main?	**1**	**2**	**3**	**4**	**5**

Veuillez évaluer la sévérité des symptômes suivants **durant les 7 derniers jours**. (Entourez une réponse sur chacune des lignes)

	Aucune	Légère	Moyenne	Importante	Extrême
9. Douleur de l'épaule, du bras ou de la main	**1**	**2**	**3**	**4**	**5**
10. Picotements ou fourmillements douloureux de l'épaule, du bras ou de la main	**1**	**2**	**3**	**4**	**5**

	Pas du tout perturbé	Légèrement perturbé	Moyennement perturbé	Très perturbé	Tellement perturbé que je ne peux pas dormir
11. Pendant les 7 derniers jours, votre sommeil a-t-il été perturbé par une douleur de votre épaule, de votre bras ou de votre main ? (entourez une seule réponse)	**1**	**2**	**3**	**4**	**5**

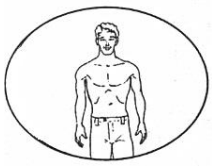
DATE D ADMINISTRATION: ___/___/ 201

La section suivante nous renseignera sur l'état fonctionnel de l'épaule qui vous pose problème (Simple Shoulder Test) :

Veuillez répondre aux rubriques suivantes en marquant d'une croix la bonne réponse.

REMPLIE PAR LE PATIENT

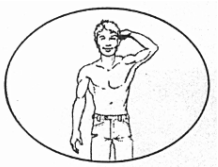
1. Votre épaule est-elle indolore lorsque votre bras est au repos sur le côté ? Oui Non



2. Votre épaule vous permet-elle de dormir confortablement ? Oui Non

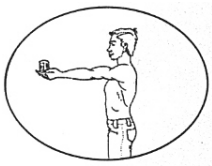
3. Pouvez-vous mettre la main dans le dos pour enfiler votre chemise dans votre pantalon ou votre jupe ? Oui Non

4. Pouvez-vous mettre votre main derrière la tête en mettant complètement le coude sur le côté ? Oui Non



5. Pouvez-vous mettre une pièce de monnaie à hauteur de votre épaule sans plier le coude ? Oui Non

6. Pouvez-vous soulever 500 g (1 boîte de conserves) à hauteur de votre épaule sans plier le coude ? Oui Non



DATE D ADMINISTRATION: ___/___/ 201

Suite Simple Shoulder Test

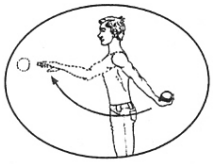
Veuillez répondre aux rubriques suivantes en marquant d'une croix la bonne réponse.

REMPLIE PAR LE PATIENT

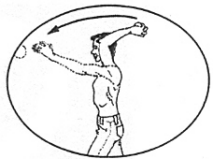
7. Pouvez-vous soulever 4 kilos (1 baril de lessive) jusqu'au niveau de votre tête sans plier le coude ? Oui Non

8. Pouvez-vous porter, du côté atteint, une valise ou un équivalent de 10 kilos ? Oui Non

9. Pensez-vous être capable de lancer une balle de caoutchouc à la façon d'une boule de pétanque à une distance de 10 mètres ? Oui Non

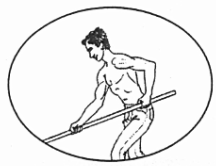


10. Pensez-vous être capable de lancer une balle de caoutchouc à la façon d'une fléchette à une distance de 20 mètres ? Oui Non



11. Pouvez-vous laver l'arrière de l'épaule opposée avec le bras atteint ? Oui Non

12. Votre épaule vous permet-elle de travailler normalement toute la journée dans votre métier ou à la maison ? Oui Non



DATE D ADMINISTRATION: ___/___/ 201

A remplir seulement si votre diagnostic est une instabilité d'épaule (= tendance à la luxation) : questionnaire WOSI

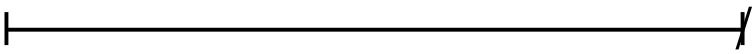
Instructions au patient

Dans les sections A, B, C, et D, il vous sera demandé de noter votre réponse en dessinant un trait en travers de la ligne horizontale,

Si vous mettez un trait à l'extrémité gauche de la ligne, tel que représenté ci-dessous,

vous signifiez alors que vous n'avez aucune douleur

Si vous mettez un trait à l'extrémité droite de la ligne, tel que représenté ci-dessous,



vous signifiez alors que votre douleur est extrême

Plus vous mettez le trait à droite, plus le symptôme que vous ressentez est fort

Plus vous mettez le trait à gauche, moins le symptôme que vous ressentez est fort

Veuillez ne pas inscrire de trait en dehors des lignes horizontales s'il vous plaît

Vous êtes appelés à noter dans ce questionnaire l'intensité des symptômes que vous avez ressentis la semaine passée à l'épaule qui pose problème. Si vous n'êtes pas sûr de savoir de quelle épaule il s'agit ou si vous avez d'autres questions, n'hésitez pas à les poser librement avant de compléter ce questionnaire.

Si une question ne s'applique pas à votre situation ou que vous n'avez pas ressenti le symptôme durant la semaine passée, essayer de vous imaginer dans cette situation afin de répondre au mieux.

DATE D ADMINISTRATION: ___/___/ 201

Section A : Symptômes physiques

Les questions suivantes portent sur les symptômes physiques que vous éprouvez en raison de votre problème d'épaule. Pour chaque question, veuillez indiquer l'intensité du symptôme éprouvé au cours de **la semaine dernière** *(Inscrivez un trait « I » sur l'échelle horizontale).*

1. Quelle intensité de douleur ressentez-vous à l'épaule lors d'activités nécessitant des mouvements au-dessus de la tête?

aucune douleur ——— douleur **extrême**

2. Quelle intensité de douleur continue ou pulsatile éprouvez-vous à l'épaule?

aucune douleur continue ou pulsatile ——— douleur continue ou pulsatile **extrême**

3. Eprouvez-vous une faiblesse ou un manque de force à l'épaule?

aucune faiblesse ——— faiblesse **extrême**

4. Ressentez vous une fatigue ou un manque d'endurance dans votre épaule?

aucune fatigue ——— fatigue **extrême**

5. Ressentez-vous des craquements ou claquements dans votre épaule?

aucun craquement ——— craquements **extrêmes**

DATE D ADMINISTRATION: ___/___/ 201

6. Ressentez vous une raideur de votre épaule?

aucune raideur ——— raideur **extrême**

7. Ressentez vous une gêne au niveau des muscles de la nuque en raison de votre épaule?

aucun inconfort ——— inconfort **extrême**

8. A quel point ressentez vous votre épaule comme instable?

aucune instabilité ——— instabilité **extrême**

9. À quel point compensez-vous la perte fonctionnelle de votre épaule à l'aide d'autres muscles?

aucunement ——— **extrêmement**

10. Quelle est la perte de mobilité au niveau de votre épaule ?

aucune perte ——— perte **extrême**

DATE D ADMINISTRATION: ___/___/ 201

Section B : Sports, loisirs et travail

Les questions suivantes portent sur la manière dont votre problème d'épaule a perturbé le travail, le sport et les activités de loisir durant la semaine passée. Pour chaque question, tracez un trait « l » sur l'échelle horizontale à l'endroit qui correspond à l'intensité de votre symptôme.

11. À quel point votre épaule limite-t-elle votre capacité de participer à des activités sportives ou récréatives?

aucunement limité — limitation **extrême**

12. À quel point votre épaule affecte-t-elle le niveau de performance auquel vous pratiquez votre sport ou effectuez votre travail? (si votre épaule perturbe le sport et le travail, prenez en considération le domaine le plus perturbé)

aucunement affecté — affecté de façon **extrême**

13. À quel point ressentez-vous le besoin de protéger votre bras lorsque vous pratiquez une activité?

aucunement — **extrêmement**

14. À quel point éprouvez-vous de la difficulté lorsque vous soulevez un objet lourd au-dessous de la hauteur de l'épaule?

aucune difficulté — difficulté **extrême**

DATE D ADMINISTRATION: ___/___/ 201

Section C : Mode de vie

Les questions suivantes portent sur la manière dont votre problème d'épaule a perturbé ou changé votre mode de vie. A nouveau, veuillez tracer pour chaque question un trait « I » sur l'échelle horizontale à l'endroit qui correspond à l'intensité de votre symptôme.

15. À quel point craignez-vous de tomber sur votre épaule?

aucunement peur ⊢——————————————————⊣ **extrêmement** peur

16. À quel point éprouvez-vous de la difficulté à maintenir votre niveau de condition physique souhaité?

aucune difficulté ⊢——————————————————⊣ difficulté **extrême**

17. À quel point avez-vous de la difficulté à jouer physiquement (ex : jouer à la lutte, taquiner…) avec votre famille ou vos amis?

aucune difficulté ⊢——————————————————⊣ difficulté **extrême**

18. À quel point avez-vous de la difficulté à dormir à cause de votre épaule?

aucune difficulté ⊢——————————————————⊣ difficulté **extrême**

DATE D ADMINISTRATION: ___/___/ 201

Section D : Émotions

Les questions suivantes demandent comment vous vous êtes senti au cours ***de la semaine dernière*** *quand à votre problème d'épaule. Veuillez indiquer votre réponse par un trait « I » sur l'échelle horizontale à l'endroit qui correspond à l'intensité de votre symptôme.*

19. À quel point êtes-vous focalisé sur votre épaule?

aucunement focalisé ——— **extrêmement** focalisé

20. À quel point craignez-vous que l'état de votre épaule ne s'aggrave?

aucunement préoccupé ——— préoccupation **extrême**

21. À quel point éprouvez-vous de la frustration à cause de votre épaule?

aucune frustration ——— frustration **extrême**

MERCI D'AVOIR COMPLÉTÉ LE QUESTIONNAIRE

Appendix VIII

Phase 2-related published article

PLOS ONE

RESEARCH ARTICLE

Heightened clinical utility of smartphone versus body-worn inertial system for shoulder function B-B score

Claude Pichonnaz[1,2]*, Kamiar Aminian[3], Céline Ancey[1], Hervé Jaccard[1,2], Estelle Lécureux[4], Cyntia Duc[3], Alain Farron[2], Brigitte M. Jolles[2], Nigel Gleeson[5]

1 Physiotherapy Department, Haute Ecole de Santé Vaud (HESAV)//HES-SO, University of Applied Sciences Western Switzerland, Lausanne, Switzerland, **2** Service of Orthopaedics and Traumatology, Department of Musculoskeletal Medicine, University Hospital of Lausanne, Lausanne, Switzerland., CHUV-UNIL, Lausanne, Switzerland, **3** Laboratory of Movement Analysis and Measurement, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, **4** Direction médicale, CHUV-UNIL, Lausanne, Switzerland, **5** School of Health Sciences, Queen Margaret University, Edinburgh, Scotland

* Claude.Pichonnaz@hesav.ch

Citation: Pichonnaz C, Aminian K, Ancey C, Jaccard H, Lécureux E, Duc C, et al. (2017) Heightened clinical utility of smartphone versus body-worn inertial system for shoulder function B-B score. PLoS ONE 12(3): e0174365. https://doi.org/10.1371/journal.pone.0174365

Editor: Antoine Nordez, Universite de Nantes, FRANCE

Received: March 25, 2016

Accepted: March 8, 2017

Published: March 20, 2017

Data Availability Statement: All .xlsx files are available from the Figshare database (https://figshare.com/s/295439a2dba9bf9635be).

Funding: Funded by Swiss National Science Foundation Grant number 135061 http://p3.snf.ch/Project-135061. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: Kamiar Aminian is cofounder of Gait Up company provider of Physilog used as a

Abstract

Background

The B-B Score is a straightforward kinematic shoulder function score including only two movements (hand to the Back + lift hand as to change a Bulb) that demonstrated sound measurement properties for patients for various shoulder pathologies. However, the B-B Score results using a smartphone or a reference system have not yet been compared. Provided that the measurement properties are comparable, the use of a smartphone would offer substantial practical advantages. This study investigated the concurrent validity of a smartphone and a reference inertial system for the measurement of the kinematic shoulder function B-B Score.

Methods

Sixty-five patients with shoulder conditions (with rotator cuff conditions, adhesive capsulitis and proximal humerus fracture) and 20 healthy participants were evaluated using a smartphone and a reference inertial system. Measurements were performed twice, alternating between two evaluators. The B-B Score differences between groups, differences between devices, relationship between devices, intra- and inter-evaluator reproducibility were analysed.

Results

The smartphone mean scores (SD) were 94.1 (11.1) for controls and 54.1 (18.3) for patients ($P < 0.01$). The difference between devices was non-significant for the control ($P = 0.16$) and the patient group ($P = 0.81$). The analysis of the relationship between devices showed 0.97 ICC, −0.6 bias and −13.2 to 12.0 limits of agreement (LOA). The smartphone intra-evaluator ICC was 0.92, the bias 1.5 and the LOA −17.4 to 20.3. The smartphone inter-evaluator ICC was 0.92, the bias 1.5 and the LOA −16.9 to 20.0.

reference system in the study. The other authors have declared that no competing interests exist.

Conclusions

The B-B Score results measured with a smartphone were comparable to those of an inertial system. While single measurements diverged in some cases, the intra- and inter-evaluator reproducibility was excellent and was equivalent between devices. The B-B score measured with a smartphone is straightforward and as efficient as a reference inertial system measurement.

1. Introduction

1.1. Current methods for shoulder function evaluation in clinical settings

The shoulder is the second most frequently affected body site [1]. The quality of tools for the evaluation of shoulder function is of primary interest to adequately address the problems of this large population and therefore limit the impact of shoulder pathologies on patients and society. Shoulder function is usually evaluated using questionnaires. Dozens of evaluation tools exist but most have not undergone a full validation process [2, 3]. Thus the measurement of the shoulder functional outcome remains a controversial issue.

Several reviews of literature have concluded that no single questionnaire of shoulder function offered superiority regarding measurement properties [3–5], while one concluded that the DASH (Disabilities of the Arm, Shoulder and Hand) score compared favourably to other questionnaires [6]. As a consequence, a large variety of outcome measurements tools have been used, hindering the development of scientific evidence about the treatment of shoulder conditions [2].

Clinical questionnaires have the advantages of handiness and low cost. Conversely, they present intrinsic limitations related to language and cultural issues, respondents' interpretations and content validity [7, 8]. The validation of questionnaires's translations into various languages is a time-consuming and cumbersome process. Moreover, the delineation between objective and subjective evaluation is not always clearly defined in questionnaire-based assesssment, with both approaches producing different results [9, 10].

1.2. Computerized shoulder function evaluation

Laboratory-based movement analysis overcomes these limitations and displays high accuracy and precision. It has thus been largely used in research studies aiming at the characterization and evaluation of shoulder motion. Most motion analysis studies have addressed the development of innovative measurement' methods mainly and have investigated differences between healthy and pathological participants' groups. However, none of them had proposed a shoulder function score that could be possibly used to monitor patient clinical evolution, to the best of our knowledge.

Although 3D laboratory motion analysis systems have assumed a growing importance in research, it's their application in clinical settings that has remained likely to be limited by complexity and cost. So, embedded systems, like inertial measurement units (IMU) have also been developed for shoulder evaluation, as their portability and practicality facilitates the procedures for measurement.

Measurements using embedded systems may provide a well-balanced compromise between practicality and reliability. They may thus constitute a valuable alternative to questionnaires or laboratory-based evaluation. The embedded systems' results are highly correlated to laboratory

measurements and display adequate accuracy for clinical evaluation. Also, their use is not restricted to laboratory settings and the measurement completion is easier [11]. Body-worn sensors have been applied with promising results, to measure arm and shoulder movement in various conditions [12–20].

Despite the simplification of the measurement procedures provided by body-worn sensors their use for shoulder function evaluation has remained limited in clinical settings. Several barriers still hinder the wide-spread use of such devices among health professionals. The requirements for the routine application in clinical practice are very demanding as, in addition to measurement properties, time, practicability, user-friendliness and cost are of concern.

Using a smartphone for evaluation purposes might contribute to meeting these requirements and facilitating the regular use of computerized movement analysis in current practice. Like embedded measurement systems, most smartphones are now fitted with built-in accelerometers and gyroscopes. Using a dedicated application, they can thus be used for movement analysis.

1.3. Present smartphone applications for shoulder evaluation

Numerous smartphone applications have been developed for patient evaluation, patient education or to assist health care professionals in their practice. The applications addressing the assessment of shoulder range of motion (ROM) generally demonstrated adequate measurement properties [21–23]. However, ROM is only one component of shoulder function and no smartphone-based assessment score for shoulder function has been validated to our knowledge. The validation of smartphone-based outcomes would be of interest because of the high prevalence of shoulder conditions and of the existing controversy about shoulder function questionnaires.

Smartphone-based evaluation in clinical conditions is valuable only provided that the measurement properties have previously been validated. This is mandatory as important decisions are taken based on clinical outcome. The smartphone results might possibly differ from inertial-based systems as the sensors' features have not been specifically designed for scientific measurement. An extensive validation process is thus needed before clinical implementation.

1.4. Inception of a smartphone application for shoulder function

Coley developed a shoulder function scoring system using inertial sensors. He proposed a relatively simple shoulder function score based on three dimensional measurements of a power-related metric using accelerometers and gyroscopes (P score) [11]. The procedure relied on a sequence of seven functional movements based on the Simple Shoulder Test functional score [24]. This approach demonstrated clinical relevance following rotator cuff and arthroplasty surgery. It clearly discriminated healthy from pathological subjects, was correlated to clinical scores and displayed good responsiveness [11]. However, the full test procedure required around 20 minutes, which precluded routine application in clinical settings.

Körver et al. [25, 26] proposed a kinematic score based on angular rate (AR Score). This score required less than 5 minutes to perform as it included only "arm to the back" and "arm behind the head" movements. It demonstrated high intra- and inter-evaluator reproducibility, with intraclass coefficient of correlation (ICC) of 0.95 and 0.91, respectively. The diagnostic sensitivity was 98% and the specificity 81%. However, the criterion-based validity for shoulder function evaluation was limited, as correlations with the DASH and SST (simple shoulder test) clinical scores were weak [24, 27].

The latter weakness was not found for the B-B Score, a simplified version of P Score including two movements only (hand to the Back & hand upwards as if to change a Bulb) [28]. This

score was developed based on principal component analysis and multiple regression of the P Score original data. The B-B Score results showed no significant difference with the P score during the first year after shoulder surgery and both scores were highly related ($R^2 > .97$). The diagnostic sensitivity was 97% and the specificity 94% for patients following rotator cuff surgery or shoulder arthroplasty. The correlations with current clinical questionnaires ranged from 0.51 to 0.77, indicating that the B-B Score had good criterion-based validity for shoulder function evaluation. Thus, the simplified model is comparable to the P Score but presents practical advantages that facilitate the evaluation of shoulder function in clinical practice.

Pichonnaz et al. [29] investigated the measurement properties of a smartphone-based version of the B-B Score in various shoulder pathologies. Diagnostic power, responsiveness and concurrent validity with shoulder function questionnaires were insufficient for shoulder instability, but were appropriate for patients conservatively treated for rotator cuff conditions or capsulitis, and patients surgically or conservatively treated for proximal humerus fracture, when compared to accepted clinimetric standards.

Despite these promising results, it remains presently unknown if the measurement obtained using a smartphone are comparable those obtained using a reference human movement analysis system and display equivalent reproducibility. If so, the use of a smartphone for the B-B Score measurement might offer a cost-effective and straightforward clinical outcome measurement.

1.5. Study aim and hypotheses

The aims of this study were to investigate the validity and reproducibility of a smartphone-assessed kinematic shoulder function B-B Score, and to compare the performance of the smartphone to a reference inertial system.

Thus, the study hypothesis is that the B-B Score meets the requirements of a valid shoulder function score. This implies that the differences between the control and the pathological group but not the difference between devices should be significant, the ICCs ≥ 0.80 for inter-device, intra-evaluator and inter-evaluator reproducibility, the limits of agreement (LOA) between devices $\leq 10\%$ and the bias $\leq 5\%$ [30, 31]. The B-B Score results should also be coherent with those of shoulder function questionnaires.

2. Materials and methods

2.1. Study sample

A prospective cohort study was conducted between August 2011 and May 2014 at the Department of Traumatology and Orthopaedic Surgery of the University Hospital of Lausanne. Ethical approval was granted by the Human Research Ethics Committee of the Canton of Vaud (CER-VD), protocol number 205/10. Patients gave their signed informed consent for participation in the study. The study was registered under ClinicalTrials.gov Identifier: NCT01431417. Three healthy participants where inadvertently measured within the two weeks preceding the registration date. The measurement protocol was strictly identical for all participants and was in line with study declaration.

The included patients were adults > 18 year old. They presented with one of the following shoulder conditions, as recorded during their first medical consultation at the specialized shoulder consultation unit of the hospital: rotator cuff condition, adhesive capsulitis, proximal humerus fracture i.e. the pathologies for which the B-B score measurement properties were known as appropriate [29]. With the exception of patients with fracture, patients who gave their consent underwent the measurement session within two weeks following medical

consultation. Measurements were performed 6 weeks post stabilisation for patients with humerus fracture, provided that the radiological control showed normal consolidation.

For the rotator cuff condition or capsulitis, patients were selected who required only conservative treatment. As the B-B Score had previously been validated after rotator cuff and arthroplasty surgery [28], it was of interest to explore its validity in different populations. Surgical and conservative fracture treatment were included in the same group as the evolution and functional prognosis is similar in both populations [32].

A group of participants younger than 35 years-old without history of shoulder condition/pain, was also included to evaluate the performance in a healthy population and the stability of the score. These participants were selected purposefully to be younger than the patients to avoid bias related to the high prevalence of asymptomatic rotator cuff tear above 40 years old [33].

The sample size calculation was based on the data of a pilot study that included 7 controls and 16 patients. The calculation was made so that, with a significance level at $P < 0.05$, the power of 0.80 was reached when the minimal standards for acceptable properties of the score were met. Fourty-six patients were required considering a lowest acceptable ICC of 0.80, corresponding to a substantial correlation, and an expected ICC of 0.90 for two measurements [31, 34]. Nine patients were required to get the expected power for the difference between the patients and the control group [35, 36]. A considerably larger sample was enrolled to get precise estimations of results and to allow subsequent subgroup analysis in further investigations.

Exclusion criteria were bilateral shoulder conditions, any concomitant pain or condition involving the upper limb or cervical spine, medical contraindication to execute movements required for score completion, tumour, neurological condition interfering with the test and an insufficient local language level to give truly informed consent or to understand questionnaires.

2.2. B-B Score calculation

The B-B Score was calculated according to the method described in Pichonnaz et al. and Coley at al. [11, 28]. A power-related parameter was extracted from the recorded signals: the range of acceleration was multiplied by the range of angular velocity, with a measurement unit of $[(deg/s) \times (m/s^2)]$, for each movement. This parameter was calculated for each axis and for each movement of the B-B Score ("hand to the Back" movement and "lift hand as to change a Bulb" movement) and added, separately for each side and for each movement. The ratio of the performance of the affected side relative to the healthy side (or the dominant side relative to the non-dominant side for healthy participants), expressed in percentage, was then calculated for each of the two movements. The values of the movements were then weighted using the equation: B-B Score = 16.71 + 0.32 x hand to the Back. + 0.45 x lift hand.

One hundred percent represents a perfect balance in capability between sides and the score decreases in accordance with the severity of functional loss. For example, while a typical healthy person performs near to 100%, the average patient might reach 46% before surgery, 67% at 3 months and 71% at 6 months after surgery.

2.3 Experimental system: Smartphone

A smartphone (iPod®, Apple, Cupertino, USA) was chosen as the support device for the development of the application. It was fitted with 3D built-in sensors (Accelerometers: ± 2 g precision: ± 0.02 g; Gyroscopes: ± 500 deg./s precision: ± 0.2 deg./s; Sampling frequency: 100 Hz) [37]. An application, called iShould (instrumented shoulder test) was programmed in Objective-C [38, 39]. This application enabled the acquisition of the acceleration and angular

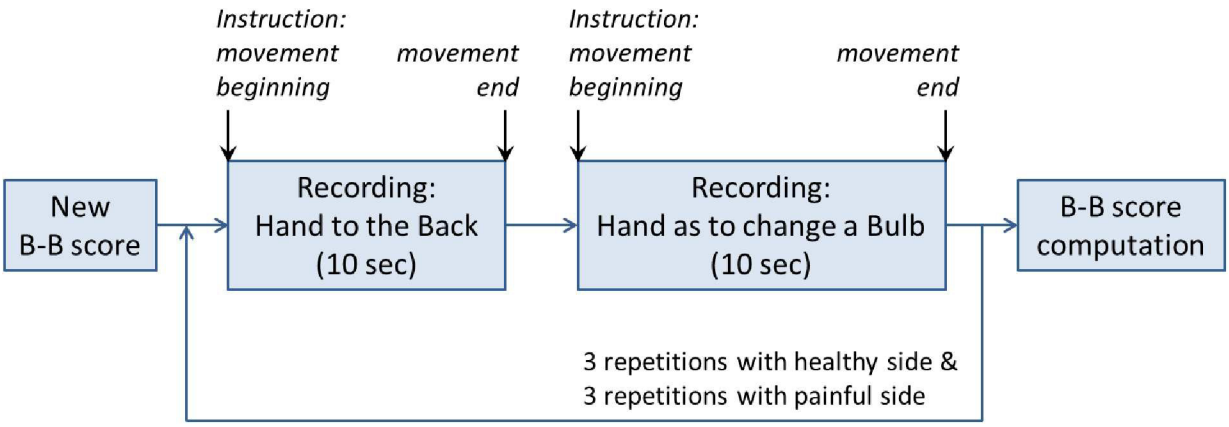


Fig 1. Schema of the application steps for the recording of a B-B score. From: Pichonnaz C, Duc C, Gleeson N, Ancey C, Jaccard H, Lecureux E, et al. Measurement Properties of the Smartphone-Based B-B Score in Current Shoulder Pathologies. Sensors (Basel). 2015;15(10):26801-17.

https://doi.org/10.1371/journal.pone.0174365.g001

velocity signals during the movements of the B-B Score and the computation of the B-B Score value, as described in the Fig 1. Once the application was launched, the smartphone provided instructions to the user, through the smartphone loudspeaker, when to perform a score movement. For each score movement, the application recorded the acceleration and angular velocity signals for a predefined period of 10 sec. The movements were first performed with the healthy side and then repeated with the painful side. At the end of the test, the B-B Score was directly calculated, displayed on the smartphone screen and then stored on the smartphone. The application enabled exporting of all saved data to a computer for its direct comparison with the data from the inertial sensors of the reference system.

2.4 Reference system

The reference system for body-worn movement analysis was composed of 2 inertial sensors and a datalogger system (Physilog®, Gait Up, Lausanne Switzerland).

Each inertial sensor included three dimensional accelerometers and gyroscopes (Accelerometers: Analog device, ADXL 210, ±5 g, precision: ± 0.2% of Full Scale; Gyroscopes: Analog device, ADXRS 250, ±400 deg/s, precision: ± 0.1% of Full Scale). The device resolution was 16 bits and the sampling frequency was 200 Hz.

An inertial measurement system was used as a reference in this study because the B-B Score has been previously developed based on this approach, and because inertial sensors provide direct measurements of angular velocities and accelerations used in the score calculation. Initial study try-outs showed that the influence of measurement errors (offset, sensitivity or drift) was negligible in the study context.

2.5. Measurement procedure

The inertial sensors of the reference system were placed on each humerus, 3 cm above the midpoint of the line connecting the lateral epicondyle (EL) and medial epicondyle (EM). The sensor's axes were aligned to the anatomical frame of the humerus following the ISB recommendations [40, 41]: Yh on the line connecting the gleno-humeral (GH) joint and the midpoint of EL and EM, pointing to GH; Xh on the line perpendicular to the plane formed by EL, EM and GH, pointing forward; Zh on the line perpendicular to Xh and Yh, pointing to the right (Fig 2). The smartphone was also attached to the back of the arm with an armband. The

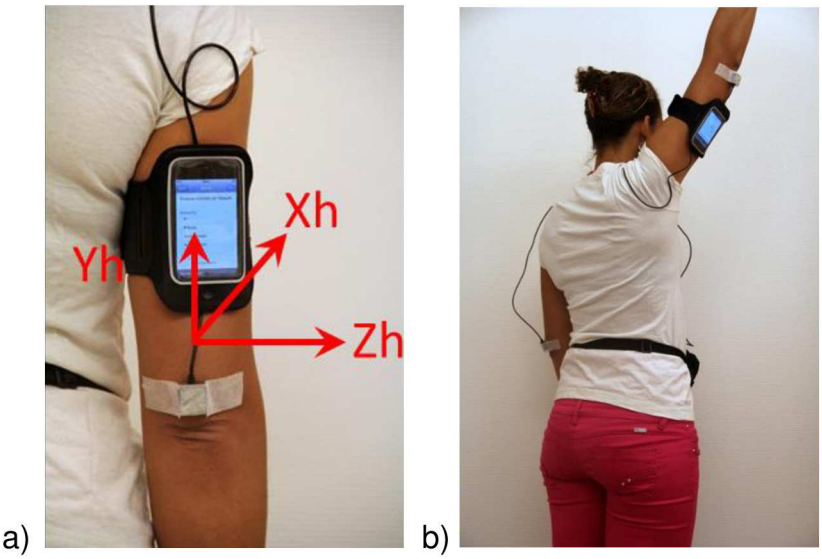


Fig 2. Inertial sensors and smartphone placement and axes. (a) The inertial sensor module (Physilog® reference system) attached to the arm with medical tape and connected by cable to the datalogger carried on wait. The smartphone is attached to the arm with the armband. (b) Test completion of "hand to the ceiling".

https://doi.org/10.1371/journal.pone.0174365.g002

lower edge of the smartphone was set 3 cm above the upper edge of the inertial sensors' module [29]. Similar to previous work angular velocities and accelerations in the sensor frame have been used to calculate the B-B Score [11, 28].

After setting-up of the systems, the participants watched a video-recorded demonstration of the execution of the B-B Score. They were instructed to do the movements in the pain free ROM, at their self-selected speed and in their natural way. The starting position was the arm alongside the body, in a relaxed position. Movements were executed in a standing position following the smartphone-recorded instructions. The patients undertook first 3 repetitions of the two B-B Score movements on the healthy side (put hand to the back + hand to the ceiling as to change a bulb) and then repeated the task on the pathological side. The controls executed the same procedure beginning on the dominant side.

The measurement procedure was repeated twice alternating between two evaluators. All evaluators were experienced physiotherapists engaged in the project, who had previously been trained to the score completion. The first evaluator was randomly assigned. All measurement systems were detached for inter-evaluator administration of assessments to account for the variability induced by possible inconsistent sensors' placement in clinics. The score was calculated based on the mean of the 3 replications because the pilot study showed that the variability was not significantly different with a higher number of repetitions.

Clinical questionnaires were also completed. Three currently used shoulder function questionnaires [Quick Disabilities of the Arm and Shoulder score (QuickDASH), Simple shoulder test (SST), Constant score and Constant relative score (based on an age- and sex-matched normal populations)], the EuroQol generic quality of life questionnaire [EQ-5D] and the pain visual analog scale (VAS) [24, 42–44]. The Constant Score was undertaken according to the modified guidelines of Constant [45]. The shoulder function questionnaires were selected because they represent current standards [3, 4, 46, 47]. They allowed the evaluation of the concurrent validity for the B-B Score but not of its validity against a 'gold standard', due to the controversy surrounding shoulder function evaluation.

2.6. Analysis

Descriptive statistics including mean, standard deviation (SD) and boxplots were performed for patients' characteristics and outcomes of both groups. The difference between the B-B Scores measured by each device was evaluated using the Wilcoxon rank-sum test. The relationship between the B-B Scores of each device, and the intra- and inter-evaluator reproducibility were evaluated using the ICC, measurement error (ME: standard error of the mean difference), standard error of measurement [SEM: $(\text{pooled SD} \times \sqrt{1 - \text{ICC agreement}})$] and Bland and Altman LOA analysis. Intra-evaluator reproducibility was calculated comparing the 1st with the 2nd score obtained by the same evaluator, for the two evaluators. Inter-evaluator reproducibility was calculated comparing the score obtained by one evaluator with the score by the other evaluator, for the 1st and 2nd evaluator's measurement. The Shapiro–Wilk test and Komolgorov-Smirnov tests were used for the normal distribution analysis. The discriminative power was evaluated by the significance level for the differences between groups (Mann-Whitney) and between stages (Wilcoxon).

3. Results

3.1. Study sample

Twenty healthy participants and 65 patients (20 with rotator cuff condition, 23 with fractures, 22 with capsulitis) were included.

The population characteristics and the significance of the differences between groups are described in Table 1.

3.2. Score outcome

The outcomes of the control group and the patient group, for the smartphone and the reference system (Physilog®), respectively, are presented in Table 2 and in Fig 3.

The difference between the control and the patient group was significant for the reference system and the smartphone ($P < 0.01$).

The difference between the reference system and the smartphone was non-significant for the control ($P = 0.16$) and for the patient group ($P = 0.81$).

3.3. Measurement reproducibility

The Shapiro-Wilk and Komolgorov-Smirnov tests confirmed the normal distribution of data ($P > 0.05$) in the patient and in the control group, regardless of device. The numerical and graphical presentations of reproducibility of measurement for inter-devices and intra- and inter-evaluator comparison are presented in Table 3 and Fig 4.

Table 1. Participants' characteristics.

	Patient (n = 65)	Control (n = 20)
Age mean (SD), years	58.5 (14.2)**	28.2 (6.2)
Sex (% women)	63	50
Weight mean (SD), kg	75.2 (15.8)	74.7 (17.4)
Body mass index mean (SD), kg/m²	26.6 (5.8)	24.2 (3.9)
Size mean (SD), m.	1.68 (0.10)	1.75 (0.10)
Hand dominance (% right-handed)	92	90
Affected side (% dominant side)	43	-

** Significant difference between groups with p-value < 0.01.

https://doi.org/10.1371/journal.pone.0174365.t001

Table 2. Mean and standard deviation of B-B Score using the smartphone and the reference system. Unit of scores are % representing the performance of the pathological side compared to the healthy side.

Mean (SD), % Min;max	Reference system	Smartphone
Control	97.0 (13.8)	94.1 (11.1)
	79.5 ; 125.2	71.9 ; 115.7
Patient	54.0 (19.0)	54.1 (18.3)
	21.5; 114.5	21.7; 108.2

Legend: SD: standard deviation; Min: minimum measured value; Max: maximum measured value.

https://doi.org/10.1371/journal.pone.0174365.t002

3.4. Clinical questionnaires

The results of shoulder function, pain and quality of life questionnaires are presented in Table 4.

4. Discussion

This study focused on the development and validation of the shoulder function B-B Score measured by means of a smartphone. Using shoulder function scores derived from a dedicated

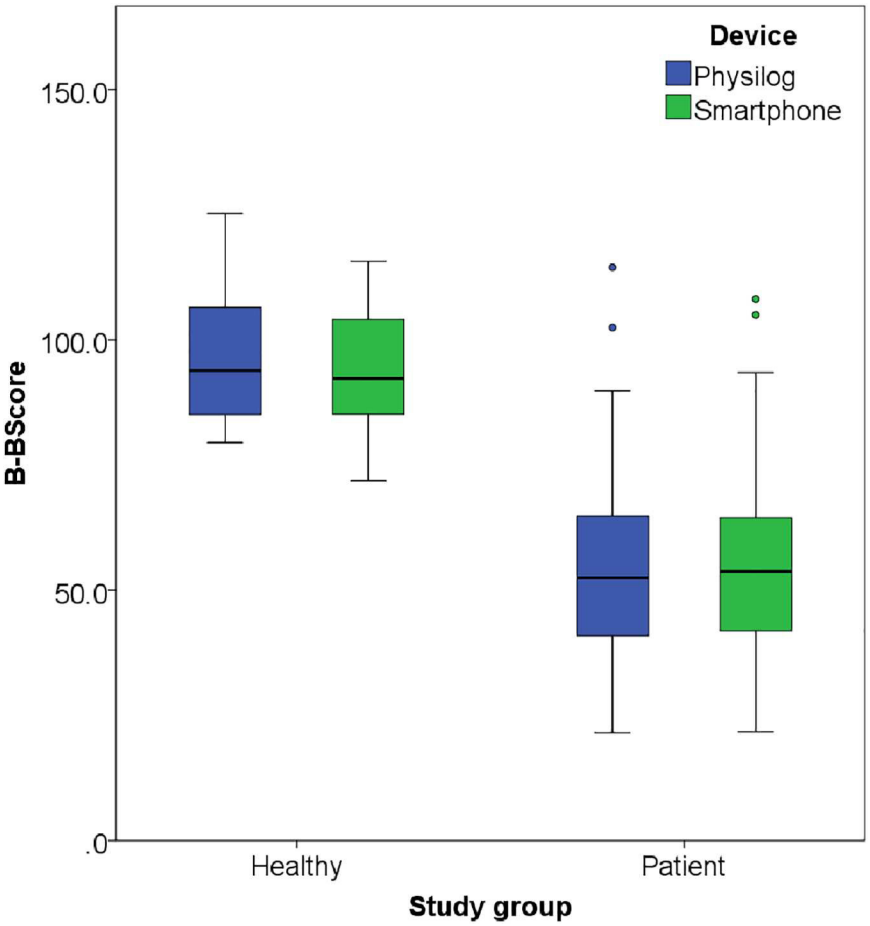


Fig 3. B-B Score outcome in both groups using the reference system (Physilog®) and the smartphone.

https://doi.org/10.1371/journal.pone.0174365.g003

Table 3. Inter-devices and intra- and inter-evaluator reproducibility of the measurements.

	ICC (95% CI)	LOA (%)	Bias (95% CI)	ME (%)	SEM (%)
Inter-devices	0.97 (0.94–0.98)	-13.2 to 12.0	- 0.6 (-0.9 to 1.1)	0.7	4.0
Intra-evaluator					
Smartphone	0.92 (0.89–0.94)	-17.4 to 20.3	1.5 (0.0 to 2.9)	0.7	6.6
Reference System	0.92 (0.89–0.94)	-19.3 to 19.6	0.1 (- 1.4 to 1.6)	0.8	6.6
Inter-evaluator					
Smartphone	0.92 (0.90–0.94)	- 16.9 to 20.0	1.5 (0.1 to 3.0)	0.7	6.6
Reference System	0.93 (0.91–0.95)	- 18.1 to 20.0	1.0 (-0.5 to 2.4)	0.7	6.4

ICC: intraclass coefficient of correlation; 95%CI: 95% confidence interval; LOA: limits of agreement; ME: measurement error; SEM: standard error of measurement

https://doi.org/10.1371/journal.pone.0174365.t003

smartphone application, the study aimed at the technical and clinical validation of them within various shoulder pathologies. Provided that the score is valid, it can offer a valuable alternative to concurrent assessment methods as it is accessible and quickly performed.

4.1. Devices comparison

The reference system (Physilog®) and the smartphone produced comparable B-B Score outcomes regarding group measurements. Although the specificities of the measurement systems were different, e.g. sensors noise, sensor ranges and sampling frequency, the smartphone performance appeared to be sufficient for the scores' proper measurement. The mean differences between the devices were non-significant and of limited magnitude (0.0% for the patient group and 2.9% for the control group). These differences are minor in proportion to the 42.9% and 40% difference between the patient and the control group, for the reference system and the smartphone, respectively.

An excellent relationship was found between measurements from the devices (ICC 0.97). Moreover, the Bland and Altman analysis demonstrated that the systematic error of the smartphone was minor. The ME and SEM were acceptable when considered in relation to the minimum-maximum range of the scores in the study sample. Conversely, the LOA exceeded the 10% criterion that had defined the threshold. Thus, the Physilog and the iPod are interchangeable for group measurement, but the magnitude of the LOA might preclude the devices' routine exchange.

4.2. Groups' comparison

There were no deviations away from the planned sampling for this study. No significant difference was observed between the groups, except for age. The control group was purposefully younger than the patient group as it was of primary importance that the reference population had healthy shoulders. The patient characteristics were representative of the population commonly treated for shoulder pain [1, 48].

The B-B Score difference between the control and the patient groups was highly significant regardless of the device. Hence, the B-B Score clearly discriminated the patient group from the healthy group.

4.3. Score reproducibility

The intra- and inter-evaluator reproducibility was excellent (0.92 to 0.93) and comparable between devices. As shown by the non-significant difference between B-B Scores computed from reference and smartphone devices and by the small bias (<1.5%) derived from the Bland

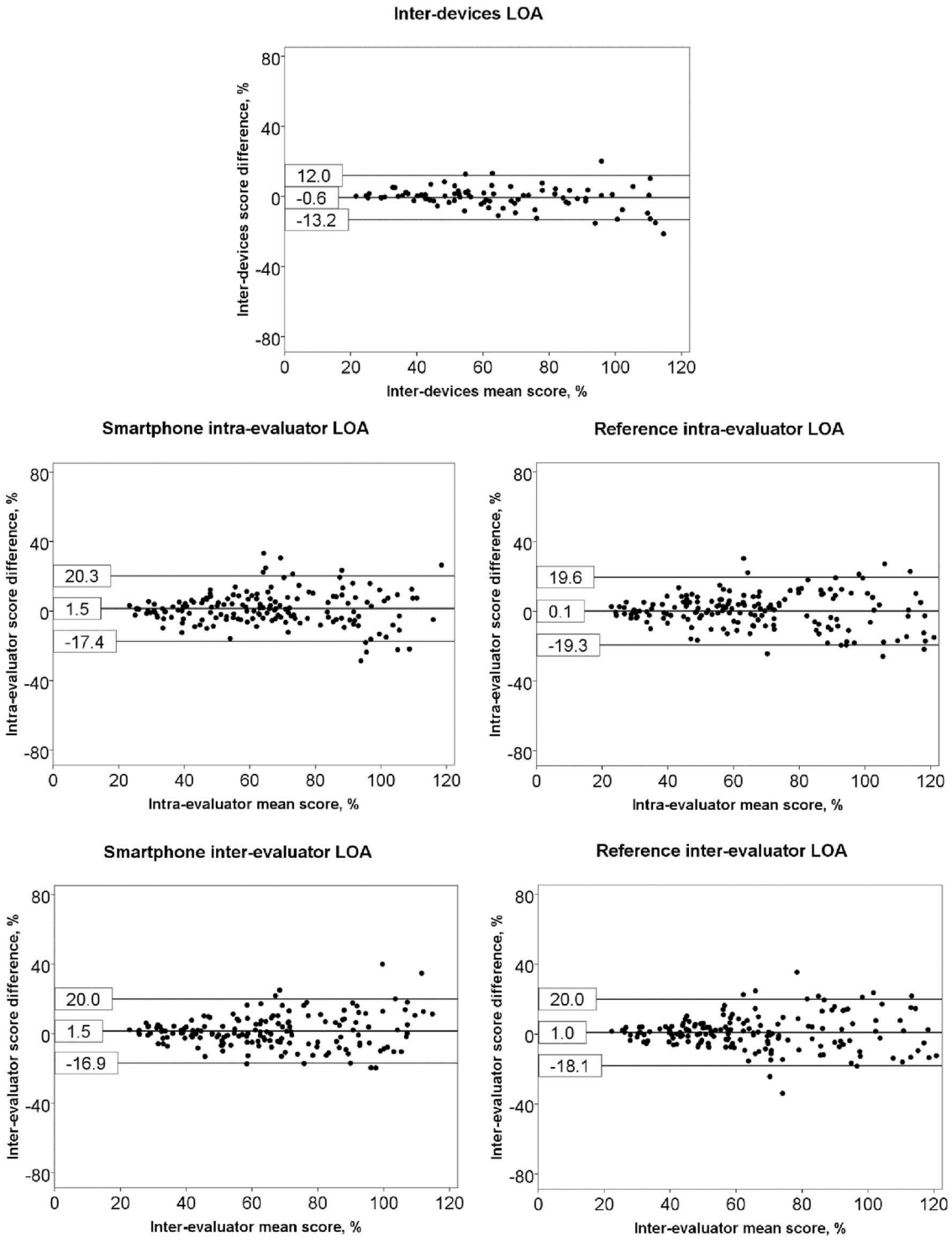


Fig 4. Bland and Altman graphs for inter-devices, intra- and inter-evaluator limits of agreement. Legend: LOA: limits of agreement.

https://doi.org/10.1371/journal.pone.0174365.g004

and Altman analyses, the B-B Score' replication and the evaluator biases were relatively minor, indicating that the systematic errors were negligible.

Table 4. Clinical questionnaires results.

Questionnaires mean (SD) *	Patient	Control
Min;max	(n = 65)	(n = 20)
Constant Score (SD), points	42.8 (17.9)	93.7 (6.6)
	10 ; 85	80 ; 100
Relative Constant Score (SD), %	55.5 (23.9)	97.6 (7.5)
	12 ; 110	82; 108
SST (SD), points	4.6 (3.1)	11.9 (0.2)
	0; 12	11; 12
QuickDASH (SD), %	42.8	1.1 (2.5)
	0.0; 86.4	0.0; 6.8
VAS pain (SD), mm	40.5 (24.2)	0.9 (2.7)
	0; 81	0.0; 10
EQ-5D (SD), index	0.70 (0.19)	1.00 (0.00)
	- 0.18; 1.00	1.00; 1.00
EQ-5D VAS (SD), points	74.3 (18.0)	98.4 (44.9)
	10.0; 100.0	85.0; 100.0

* Best possible scores: Constant 100 points, Relative Constant theoretically no limit (scores in % based on an age-and sex-matched normal population for Constant score), SST 12 points; QuickDASH 0, VAS pain 0, EQ5D 1.00 (index score of a value set derived from the general population sample), EQ5D VAS 100.

https://doi.org/10.1371/journal.pone.0174365.t004

Conversely, for both devices, the LOA for the repeated measurement of a B-B Score had exceeded an arbitrary 10% threshold defining its clinical utility. Thus, the results are comparable between replications and between evaluators for group measurement, but divergences are possible for single measurements when using this study's protocol, i.e. when taking the mean of three repetitions. Measurements relating to the assessment of a single patient is still feasible but would be expected to require acquiring the mean of more than three replications in order to counteract inflated error and establish the requisite precision of measurement [49], as the variability and error in a measurement mean score decreases with the square root of the repetitions number (assuming a normal distribution of error). The simplicity of the procedure for assessing the B-B Score facilitates measurement repetition and largely overcomes this limitation.

4.4. Comparison with clinical scores

The kinematic measurements were also compared to currently-used clinical scores for benchmarking. The clinical scores included shoulder function (Constant, Relative Constant, SST and QuickDASH), pain (VAS) and quality of life (EQ-5D).

In healthy subjects, both clinical questionnaires and the kinematic B-B score were near to the maximum performance for all scores, showing that the reference population had almost perfect shoulder function. For patients, the observed importance of shoulder function loss was also comparable between questionnaires and the B-B score, all scores indicating a substantial function loss in the measured sample. It appeared thus in this study that the B-B score produces coherent results to the shoulder function questionnaires in terms of measured loss of function, regardless of the device used.

These results were in line with published results on the relationship between kinematic scores and clinical questionnaires, which showed moderate to high correlations of the B-B score with the Constant and SST scores and moderate correlations with the QuickDASH for various shoulder pathologies [29].

4.5. Body-worn sensors shoulder function evaluation in the literature

Most previous studies that had investigated the measurement properties of body-worn sensors for shoulder function scores used dedicated inertial-based system [11, 25, 26, 28, 50–55]. All these studies concluded that the inertial-based systems produced a valid evaluation of shoulder function. Similar conclusions have since been drawn by a study using smartphone technologies [29]. However, no comparison with a reference system was reported. To our knowledge the present study has been the first to investigate the concordance and the relationship of a smartphone-based and a reference inertial-based system for shoulder function evaluation. The results are valuable for research and clinics as they demonstrate that the validity of the B-B Score measurement is not altered when using a simple and accessible device.

4.6 Study limitations and further developments

The results apply for a situation in which the measurement has been performed under supervision and at the patient's self-selected speed of movement. Further investigations are needed to determine the validity of the score in other conditions. For example, the relationship between devices might be different if the patients perform movements associated with the B-B Score at their maximum speed due to the difference in sensors' characteristics. Measurement' reliability might also be different if the patient performs the test without supervision.

The results were not detailed for each pathological subgroup in this study. This is a minor limitation with regard to the study's objectives, as the relationship between devices is not likely to be significantly influenced by the pathology. Conversely, the use of a larger group had the advantage of providing more precise estimations of the reproducibility.

Despite the widespread use and the convenience of smartphones, there are also limitations in their use for scientific measurement. The precise features of the device are not fully disclosed by manufacturers due to commercial sensitivities. The users should remain conscious that the characteristics may differ according to smartphone version and brand. An accessible middle-segment smartphone model had been chosen specifically to offer insight into its performance' characteristics. The B-B Score would probably remain robust when faced with minor variations in smartphone technology, as it would have compared the performance of the affected shoulder with that of the healthy one [28], with the score unaffected by systematic errors in measurement affecting both sides.

Based on this study and the body of literature on the subject, it appears that smartphones most likely present measurement properties that are compatible with research requirements for measurements comparing both sides and for range of motion measurements [21–23]. Nevertheless, the validity of using smartphones for more complex measurements, e.g. those associated with 3D kinematic analysis of sport activities, remains unknown to date. Also, the aforementioned variations in smartphones' features imply that further research is needed to investigate and quantify the influence of these variations on the outcome before clinical implementation.

The duration required to conduct the whole procedure using the smartphone was around two minutes. All things being equal, the advantage of the measurement approach used in this study mainly resides in its clinical practicality and low cost. Further development of the smartphone approach is possible to accrue maximum benefit from it clinically. Thus, an android version of the application has recently been made available to the public [56]. Future development may also consider facilitating the communication of clinically-relevant results between stakeholders, producing progression curves of functional improvements and comparing the patient's evolution of performance during care-pathways to benchmark results on a routine basis.

5. Conclusion

This study aimed at the technical and clinical validation of a B-B Score smartphone application for shoulder function evaluation. The results showed that the B-B Score acquired by means of a smartphone was valid and reproducible for the measurement of shoulder function of groups of patients including those presenting with rotator cuff conditions, proximal humerus fractures or adhesive capsulitis. It displayed excellent intra- and inter-evaluator reproducibility and discriminative power. Conversely, single measurements may offer reduced precision in some circumstances. The assessments acquired using either a smartphone or a reference inertial system displayed comparable measurement properties across a wide-range of clinimetrics.

Thus, the B-B Score measured with a smartphone allows valid, user-friendly and low-cost evaluation of shoulder function for research and clinical work. This could facilitate the use of objective measurement methods in routine practice and thus improve the quality of patient follow up. Further research is needed to investigate the influence of the specific characteristics of various smartphone models on results. Further technological developments are also required to achieve maximum benefit from the smartphone approach.

Acknowledgments

This study was funded by the Swiss National Science Foundation—DORE 135061.

The authors would like to thank Jean-Philippe Bassin for his contribution to study design and data collection, Noémie Sauvage Pasche for her contribution to study organization and data collection, Barbara Balmelli, Anne Rothenbacher and Guillaume Christe for their contribution to data collection, Valérie Zoll and Jean Lambert for their contribution to study organization.

Author Contributions

Conceptualization: CP CD KA AF BMJ HJ NG.

Data curation: CP HJ CA.

Formal analysis: EL CP NG.

Funding acquisition: CP KA CD BMJ AF EL.

Investigation: CP CA HJ.

Methodology: CP CD KA AF EL BMJ HJ NG.

Project administration: CP HJ CA.

Resources: KA AF BMJ CD.

Software: CD KA.

Supervision: NG KA AF BMJ.

Validation: CP CD HJ CA.

Visualization: CP NG CD.

Writing – original draft: CP.

Writing – review & editing: CP KA CA HJ EL CD AF BMJ NG.

References

1. Picavet HS, Schouten JS. Musculoskeletal pain in the Netherlands: prevalences, consequences and risk groups, the DMC(3)-study. Pain. 2003; 102(1-2):167–78. PMID: 12620608
2. Harvie P, Pollard TCB, Chennagiri RJ, Carr AJ. The use of outcome scores in surgery of the shoulder. Journal of Bone and Joint Surgery-British Volume. 2005; 87b(2):151–4.
3. Oh JH, Jo KH, Kim WS, Gong HS, Han SG, Kim YH. Comparative evaluation of the measurement properties of various shoulder outcome instruments. Am J Sports Med. 2009; 37(6):1161–8. https://doi.org/10.1177/0363546508330135 PMID: 19403837
4. Roy JS, MacDermid JC, Woodhouse LJ. Measuring shoulder function: a systematic review of four questionnaires. Arthritis Rheum. 2009; 61(5):623–32. https://doi.org/10.1002/art.24396 PMID: 19405008
5. Angst F, Schwyzer HK, Aeschlimann A, Simmen BR, Goldhahn J. Measures of adult shoulder function: Disabilities of the Arm, Shoulder, and Hand Questionnaire (DASH) and its short version (QuickDASH), Shoulder Pain and Disability Index (SPADI), American Shoulder and Elbow Surgeons (ASES) Society standardized shoulder assessment form, Constant (Murley) Score (CS), Simple Shoulder Test (SST), Oxford Shoulder Score (OSS), Shoulder Disability Questionnaire (SDQ), and Western Ontario Shoulder Instability Index (WOSI). Arthritis Care Res (Hoboken). 2011; 63 Suppl 11:S174–88.
6. Bot SD, Terwee CB, van der Windt DA, Bouter LM, Dekker J, de Vet HC. Clinimetric evaluation of shoulder disability questionnaires: a systematic review of the literature. Ann Rheum Dis. 2004; 63(4):335–41. https://doi.org/10.1136/ard.2003.007724 PMID: 15020324
7. Ragab AA. Validity of self-assessment outcome questionnaires: patient-physician discrepancy in outcome interpretation. Biomed Sci Instrum. 2003; 39:579–84. PMID: 12724955
8. Olley LM, Carr AJ. The use of a patient-based questionnaire (the Oxford Shoulder Score) to assess outcome after rotator cuff repair. Ann R Coll Surg Engl. 2008; 90(4):326–31. https://doi.org/10.1308/003588408X285964 PMID: 18492399
9. Krueger D, Kraus N, Pauly S, Chen J, Scheibel M. Subjective and objective outcome after revision arthroscopic stabilization for recurrent anterior instability versus initial shoulder stabilization. Am J Sports Med. 2011; 39(1):71–7. https://doi.org/10.1177/0363546510379336 PMID: 20855555
10. Moustgaard H, Bello S, Miller FG, Hrobjartsson A. Subjective and objective outcomes in randomized clinical trials: definitions differed in methods publications and were often absent from trial reports. J Clin Epidemiol. 2014; 67(12):1327–34. https://doi.org/10.1016/j.jclinepi.2014.06.020 PMID: 25263546
11. Coley B, Jolles BM, Farron A, Bourgeois A, Nussbaumer F, Pichonnaz C, et al. Outcome evaluation in shoulder surgery using 3D kinematics sensors. Gait Posture. 2007; 25(4):523–32. https://doi.org/10.1016/j.gaitpost.2006.06.016 PMID: 16934979
12. Liu ZH J; Shi J; Tao R; Zhou W; Zhang L. Characterizing and estimating rice brown spot disease severity using stepwise regression, principal component regression and partial least-square regression. Journal of Zhejiang University-Science B. 2007; 8(10):738–44. https://doi.org/10.1631/jzus.2007.B0738 PMID: 17910117
13. Luinge HJ, Veltink PH, Baten CT. Ambulatory measurement of arm orientation. J Biomech. 2007; 40 (1):78–85. https://doi.org/10.1016/j.jbiomech.2005.11.011 PMID: 16455089
14. Wong WY, Wong MS, Lo KH. Clinical applications of sensors for human posture and movement analysis: a review. Prosthet Orthot Int. 2007; 31(1):62–75. https://doi.org/10.1080/03093640600983949 PMID: 17365886
15. Coley B, Jolles BM, Farron A, Pichonnaz C, Bassin JP, Aminian K. Estimating dominant upper-limb segments during daily activity. Gait Posture. 2008; 27(3):368–75. https://doi.org/10.1016/j.gaitpost.2007.05.005 PMID: 17582769
16. Ludewig PM, Reynolds JF. The association of scapular kinematics and glenohumeral joint pathologies. J Orthop Sports Phys Ther. 2009; 39(2):90–104. https://doi.org/10.2519/jospt.2009.2808 PMID: 19194022
17. Ludewig PM, Cook TM. Alterations in shoulder kinematics and associated muscle activity in people with symptoms of shoulder impingement. Phys Ther. 2000; 80(3):276–91. PMID: 10696154
18. Rundquist PJ, Anderson DD, Guanche CA, Ludewig PM. Shoulder kinematics in subjects with frozen shoulder. Arch Phys Med Rehabil. 2003; 84(10):1473–9. PMID: 14586914
19. Rundquist PJ, Ludewig PM. Correlation of 3-dimensional shoulder kinematics to function in subjects with idiopathic loss of shoulder range of motion. Phys Ther. 2005; 85(7):636–47. PMID: 15982170
20. Rundquist PJ, Ludewig PM. Patterns of motion loss in subjects with idiopathic loss of shoulder range of motion. Clin Biomech (Bristol, Avon). 2004; 19(8):810–8.

21. Shin SH, Ro du H, Lee OS, Oh JH, Kim SH. Within-day reliability of shoulder range of motion measurement with a smartphone. Man Ther. 2012; 17(4):298–304. https://doi.org/10.1016/j.math.2012.02.010 PMID: 22421186

22. Werner BC, Holzgrefe RE, Griffin JW, Lyons ML, Cosgrove CT, Hart JM, et al. Validation of an innovative method of shoulder range-of-motion measurement using a smartphone clinometer application. J Shoulder Elbow Surg. 2014; 23(11):e275–82. https://doi.org/10.1016/j.jse.2014.02.030 PMID: 24925699

23. Mitchell K, Gutierrez SB, Sutton S, Morton S, Morgenthaler A. Reliability and validity of goniometric iPhone applications for the assessment of active shoulder external rotation. Physiother Theory Pract. 2014; 30(7):521–5. https://doi.org/10.3109/09593985.2014.900593 PMID: 24654927

24. Lippitt SBH, D. T.; Matsen F. A. A practical tool for evaluating function: the Simple Shoulder Test. In: Matsen PA F, FH.; Hawkins RJ, editor. The shoulder: a balance of mobility and stability. Rosemont: American Academy of Orthopaedic Surgery; 1993. p. 501–18.

25. Korver RJ, Heyligers IC, Samijo SK, Grimm B. Inertia based functional scoring of the shoulder in clinical practice. Physiol Meas. 2014; 35(2):167–76. https://doi.org/10.1088/0967-3334/35/2/167 PMID: 24398361

26. Korver RJ, Senden R, Heyligers IC, Grimm B. Objective outcome evaluation using inertial sensors in subacromial impingement syndrome: a five-year follow-up study. Physiol Meas. 2014; 35(4):677–86. https://doi.org/10.1088/0967-3334/35/4/677 PMID: 24622109

27. Jester A, Harth A, Wind G, Germann G, Sauerbier M. Disabilities of the arm, shoulder and hand (DASH) questionnaire: Determining functional activity profiles in patients with upper extremity disorders. Journal of hand surgery. 2005; 30(1):23–8.

28. Pichonnaz C, Lecureux E, Bassin JP, Duc C, Farron A, Aminian K, et al. Enhancing clinically-relevant shoulder function assessment using only essential movements. Physiol Meas. 2015; 36(3):547–60. https://doi.org/10.1088/0967-3334/36/3/547 PMID: 25690269

29. Pichonnaz C, Duc C, Gleeson N, Ancey C, Jaccard H, Lecureux E, et al. Measurement properties of the smartphone-based B-B Score in current shoulder pathologies. Sensors (Basel). 2015; 15(10):26801–17.

30. Portney LGW, Mary P. Foundations of Clinical Research: Applications to Practice. Upper Saddle River N.J., USA: Prentice Hall Health; 2009.

31. Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. Stat Med. 1998; 17(1):101–10. PMID: 9463853

32. Handoll HH, Ollivere BJ, Rollins KE. Interventions for treating proximal humeral fractures in adults. Cochrane Database Syst Rev. 2012; 12:CD000434. https://doi.org/10.1002/14651858.CD000434.pub3 PMID: 23235575

33. Sher JS, Uribe JW, Posada A, Murphy BJ, Zlatkin MB. Abnormal findings on magnetic resonance images of asymptomatic shoulders. J Bone Joint Surg Am. 1995; 77(1):10–5. PMID: 7822341

34. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33(1):159–74. PMID: 843571

35. Soper DS. Statistics Calculators 2004 [cited 2015 Archived on 12 May 2015]. Available from: http://www.webcitation.org/6ZEMd2NIS.

36. Lenth RV. Java applets for power and sample size 2010 [cited 2010 Archived on 12 May 2015]. Available from: http://www.webcitation.org/6ZEMrvmpu.

37. Mark DN, Jack; LaMarche Jeff. Beginning iOS 5 Development: Exploring the iOS SDK: Apress; 2011.

38. Oïhénart L, Duc C, Aminian K. iShould: Functional evaluation of the shoulder using a Smartphone. Gait & Posture. 2012; 36(0):S61–S2.

39. Laboratory of Movement Analysis and Measurement—Swiss Institute of Technology of Lausanne. Smartphone App iShould 2015 [cited 2016 19 February 2016]. Available from: http://lmam.epfl.ch/smartphone/ishould (archived at http://www.webcitation.org/6cUIEWqoL on 23 October 2015).

40. Wu G, van der Helm FCT, Veeger HEJ, Makhsous M, Van Roy P, Anglin C, et al. ISB recommendation on definitions of joint coordinate systems of various joints for the reporting of human joint motion—Part II: shoulder, elbow, wrist and hand. Journal of Biomechanics. 2005; 38(5):981–92. PMID: 15844264

41. Coley B, Jolles BM, Farron A, Aminian K. Detection of the movement of the humerus during daily activity. Med Biol Eng Comput. 2009; 47(5):467–74. https://doi.org/10.1007/s11517-009-0464-x PMID: 19277750

42. American Academy of Orthopaedic Surgeons. The DASH Outcome Measure 2009 [updated 2009/10/19/; cited 2009 Archived on 12 May 2015]. Available from: http://www.webcitation.org/6ZEN143eU.

43. EuroQol G. EQ-5D a standardised instrument for use as a measure of health outcome 2009 [updated 2009/10/20/; cited 2009 Archived on 12 May 2015]. Available from: http://www.webcitation.org/6ZEN3QDth.

44. Richards RR, An KN, Bigliani LU, Friedman RJ, Gartsman GM, Gristina AG, et al. A standardized method for the assessment of shoulder function. J Shoulder Elbow Surg. 1994; 3(6):347–52. https://doi.org/10.1016/S1058-2746(09)80019-0 PMID: 22958838

45. Constant CR, Gerber C, Emery RJ, Sojbjerg JO, Gohlke F, Boileau P. A review of the Constant score: modifications and guidelines for its use. J Shoulder Elbow Surg. 2008; 17(2):355–61. https://doi.org/10.1016/j.jse.2007.06.022 PMID: 18218327

46. Kirkley A, Griffin S, Dainty K. Scoring systems for the functional assessment of the shoulder. Arthroscopy: the journal of arthroscopic & related surgery: official publication of the Arthroscopy Association of North America and the International Arthroscopy Association. 2003; 19(10):1109–20.

47. Beaton DE, Richards RR. Measuring function of the shoulder. A cross-sectional comparison of five questionnaires. J Bone Joint Surg Am. 1996; 78(6):882–90. PMID: 8666606

48. van der Windt DA, Koes BW, de Jong BA, Bouter LM. Shoulder disorders in general practice: incidence, patient characteristics, and management. Ann Rheum Dis. 1995; 54(12):959–64. PMID: 8546527

49. Mercer TH, Gleeson NP. The efficacy of measurement and evaluation in evidence-based clinical practice. Physical Therapy in Sport. 2002; 3(1):27–36.

50. Jolles BM, Duc C, Coley B, Aminian K, Pichonnaz C, Bassin JP, et al. Objective evaluation of shoulder function using body-fixed sensors: a new way to detect early treatment failures? J Shoulder Elbow Surg. 2011; 20(7):1074–81. https://doi.org/10.1016/j.jse.2011.05.026 PMID: 21925353

51. Duc C, Farron A, Pichonnaz C, Jolles BM, Bassin JP, Aminian K. Distribution of arm velocity and frequency of arm usage during daily activity: objective outcome evaluation after shoulder surgery. Gait Posture. 2013; 38(2):247–52. https://doi.org/10.1016/j.gaitpost.2012.11.021 PMID: 23266045

52. Duc C, Pichonnaz C, Bassin JP, Farron A, Jolles BM, Aminian K. Evaluation of muscular activity duration in shoulders with rotator cuff tears using inertial sensors and electromyography. Physiol Meas. 2014; 35(12):2389–400. https://doi.org/10.1088/0967-3334/35/12/2389 PMID: 25390457

53. Luinge HJ, Veltink PH. Measuring orientation of human body segments using miniature gyroscopes and accelerometers. Medical & Biological Engineering & Computing. 2005; 43(2):273–82.

54. Cutti AG, Giovanardi A, Rocchi L, Davalli A, Sacchetti R. Ambulatory measurement of shoulder and elbow kinematics through inertial and magnetic sensors. Medical & Biological Engineering & Computing. 2008; 46(2):169–78.

55. de Vries WH, Veeger HE, Baten CT, van der Helm FC. Can shoulder joint reaction forces be estimated by neural networks? J Biomech. 2016; 49(1):73–9. https://doi.org/10.1016/j.jbiomech.2015.11.019 PMID: 26654109

56. Gait Up. Hands Up shoulder testing App is now available 2016 [updated 6 July 2016; cited 2016 28 October]. Available from: http://www.gaitup.com/hands-up-for-android-available-15716/.

Appendix IX

ClinicalTrials registry receipt of registration

Phase 2 and 3 studies

ClinicalTrials.gov Protocol Registration and Results System (PRS) Receipt
Release Date: June 2, 2015

ClinicalTrials.gov ID: NCT01431417

Study Identification

Unique Protocol ID:	FNS-DORE 13DPD6_135061
Brief Title:	Validation of a Kinematic Functional Shoulder Score Including Only Essential Movements
Official Title:	Validation of a Kinematic Functional Shoulder Score Including Only Essential Movements
Secondary IDs:	

Study Status

Record Verification:	June 2015
Overall Status:	Completed
Study Start:	August 2011 []
Primary Completion:	December 2013 [Actual]
Study Completion:	May 2014 [Actual]

Sponsor/Collaborators

Sponsor:	Haute Ecole Cantonale Vaudoise de Santé
Responsible Party:	Sponsor
Collaborators:	Swiss National Science Foundation

Oversight

U.S. FDA-regulated Drug:	
U.S. FDA-regulated Device:	
U.S. FDA IND/IDE:	No
Human Subjects Review:	Board Status: Approved Approval Number: 205/10 Board Name: Commission d'éthique de la recherche clinique FBM-UNIL, Lausanne Board Affiliation: Département de la Santé et de l'action sociale du Canton de Vaud Phone: ++ 41 21 314 55 98 Email: secretariatcervd@unil.ch Address: Commission cantonale (VD) d'éthique

de la recherche sur l'être humain
Secrétariat central
Rue César-Roux 19
CH-1005 Lausanne

Data Monitoring: No

FDA Regulated Intervention: No

Study Description

Brief Summary: A lot of shoulder function evaluation scores exist but none has been universally accepted as a gold standard.

Recent studies have demonstrated the potential of computerized movement analysis with embedded sensors for objective evaluation of shoulder functional outcome following surgery.

A very simple testing procedure is possible as just a few repetitions of two simple shoulder movements are sufficient. This could potentially facilitate implementation of shoulder function movement analysis in current clinical practice.

However, at the present stage of development, the method needs to be extensively validated. This means that the research will intend to determine precisely for which current shoulder pathology it can be applied, what the outcome of healthy people is, what the reliability of the score is and how it can monitor patient evolution.

Detailed Description: Measurement of shoulder function is a controversial issue. There is a great variety of measurement tools but none of them has been universally accepted. There is therefore a need to develop extensively validated and convenient measurement tools.

Embedded computerized movement analysis can potentially meet these requirements for measurement of shoulder function. Ambulatory measurement devices allow application in various clinical conditions, display adequate precision and accuracy, and are considerably more straightforward than laboratory-based systems.

Using a Physilog ® II embedded system, Coley (2007) developed a relatively simple score of shoulder function (P Score). The method is based on arm power measurement by three-dimensional accelerometers and gyroscopes during seven consecutive shoulder movements. It demonstrated reliability, responsiveness and criterion-based validity. However, additional knowledge and technological progress could now contribute to further simplification of the testing procedure.

Indeed, a secondary analysis of Coley's study data based on principal component analysis and multiple regressions highlighted that a procedure including only two selected movements produces comparable results to P Score. Moreover, the development of wireless systems considerably simplifies set up. Consequently, simpler but equivalent measurement procedure can now be considered.

A pilot study (ClinicalTrials.gov identifier: NCT01281085) has been conducted to prepare this study. it contributed to determine the number of replications of movements needed and to refine the testing procedure.

The aim of this study is to proceed to an extensive validation study of the simplified testing procedure. Kinematic measurements will be carried out with four groups of patients presenting with frequent shoulder conditions (rotator cuff condition, shoulder instability, diaphyseal or subcapital humerus fracture, frozen

shoulder) and a group of healthy people. Measurement procedure includes two consecutive measurements, alternatively conducted by two evaluators and measured simultaneously by two different movement analysis systems. Currently used functional questionnaires will be completed at both stages for comparison. Measurement will be performed at baseline and 6 months later.

Statistical analysis will address reproducibility, responsiveness, minimal clinically important difference and correlation with current clinical scores.

Conditions

Conditions:	Rotator Cuff, Syndrome Frozen Shoulder Humerus, Fracture Other Instability of Joint, Shoulder Region
Keywords:	Shoulder Outcome treatment Validation Studies as topic Biomechanics Shoulder, Joint Instability

Study Design

Study Type:	Observational
Observational Study Model:	Cohort
Time Perspective:	Prospective
Biospecimen Retention:	None Retained
Biospecimen Description:	
Enrollment:	108 [Actual]
Number of Groups/Cohorts:	5

Groups and Interventions

Groups/Cohorts	Interventions
Healthy volunteers Healthy volunteers, less than 35 years old and presenting with no shoulder condition	
Patients with rotator cuff condition Patients with rotator cuff condition, conservative treatment indicated	
Patients with shoulder instability Patients with shoulder instability, conservative treatment indicated	
Patients with proximal humerus fracture Patients with diaphyseal humerus fracture or subcapital humerus fracture treated surgically or conservatively, at 6 weeks post stabilization. (Surgical and conservative treatment will be considered as the same population from the functional point of view as functional outcome is similar) (Handoll et al. 2003).	
Patients with frozen shoulder Patients with frozen shoulder, conservative treatment indicated	

Outcome Measures

Primary Outcome Measure:

1. Kinematic functional score
   The kinematic functional score will be determined as the percentage of power of the pathological shoulder compared to the healthy shoulder (e.g. 70% means that the power developed during the movement of the pathological shoulder reaches 70% of the power developed on the healthy side)

   [Time Frame: Baseline]

2. Changes in kinematic functional shoulder scores
   Aforementioned score will be measured again 6 months after baseline to evaluate its responsiveness to patients' evolution

   [Time Frame: Change from Baseline in kinematic functional shoulder scores at 6 months]

Secondary Outcome Measure:

3. Functional scores as determined by several currently used shoulder scores
   Questionnaires include Constant score, Quick DASH, subjective shoulder value, Simple shoulder test, WOSI (Western Ontario Shoulder Instability Index; when relevant i.e. for shoulder instability), stiffness and pain EVA

   [Time Frame: Baseline]

4. Changes in functional shoulder scores
   All aforementioned scores will be measured again 6 months after baseline to compare their respective responsiveness to patients' evolution

   [Time Frame: Change from Baseline in functional shoulder scores at 6 months]

Eligibility

Study Population:	Patients consulting at the specialized shoulder consultation of the University Hospital of Lausanne
Sampling Method:	Non-Probability Sample
Minimum Age:	18 Years
Maximum Age:	
Sex:	All
Gender Based:	
Accepts Healthy Volunteers:	Yes
Criteria:	Inclusion Criteria:

- Rotator cuff condition, conservative treatment indicated
- Shoulder instability, conservative treatment indicated
- Diaphyseal humerus fracture or subcapital humerus fracture treated surgically or conservatively, at 6 weeks post stabilization. (Surgical and conservative treatment will be considered as the same population from the functional point of view as functional outcome is similar) (Handoll et al. 2003).
- Frozen shoulder, conservative treatment indicated

Exclusion Criteria:

- Bilateral shoulder condition or other shoulder condition than the ones mentioned in inclusion criteria
- Any concomitant pain or condition involving upper limb
- Cervical spine condition involving upper limb pain or mobility restriction

- Insufficient French language level to understand patient information form, consent form or questionnaires
- Insufficient ability to give truly informed consent or to understand questionnaires. It will be proceeded to a Mini Mental State score in case of uncertainty, with exclusion criteria at 24 points/30 (ANAES 2000).
- Medical contraindication to execute movements required for score completion
- Tumor
- Neurological condition interfering with test

Contacts/Locations

Central Contact Person: Claude A. Pichonnaz, PT MSc
Telephone: ++ 41 21 316 81 26
Email: claude.pichonnaz@hesav.ch

Central Contact Backup: Jean-Philippe Bassin, PT OMT MSc
Telephone: ++ 41 21 316 81 33
Email: Jean-Philippe.BASSIN@hesav.ch

Study Officials: Claude A. Pichonnaz, PT MSc
Study Director
HESAV and University Hospital of Lausanne

Farron Alain, MER PD
Study Chair
University Hospital of Lausanne

Locations: Switzerland
Département de l'Appareil Locomoteur - CHUV
Lausanne, Switzerland, 1005
Contact: Claude A Pichonnaz, PT MSc 0041 21 3168126 claude.pichonnaz@hesav.ch
Contact: Bassin Jean-Philippe, PT OMT 0041 21 3168133 Jean-Philippe.BASSIN@hesav.ch
Sub-Investigator: Jean-Philippe Bassin, PT OMT MSc
Sub-Investigator: Hervé Jaccard, PT
Sub-Investigator: Barbara Balmelli, PT
Sub-Investigator: Bovey Anne, PT
Sub-Investigator: Céline Ancey, PT

IPDSharing

Plan to Share IPD:

References

Citations: Coley B, Jolles BM, Farron A, Bourgeois A, Nussbaumer F, Pichonnaz C, Aminian K. Outcome evaluation in shoulder surgery using 3D kinematics sensors. Gait Posture. 2007 Apr;25(4):523-32. Epub 2006 Aug 28. PubMed 16934979

Jolles, BM, Duc C, Coley B, Aminian K, Pichonnaz C., Bassin J-P, Farron A. Objective evaluation of shoulder function using body-fixed sensors: a new way to detect early treatment failures? Journal of Shoulder and Elbow Surgery, 20(7), 1074-1081, 2011.

Links:	URL: http://p3.snf.ch/project-135061 Description Lay summary in French language on the sponsor's website (Swiss National Science Foundation)
Available IPD/Information:	

Appendix X

Acceptance letter of the Ethical Commission of the Faculty of biology and medicine of the University of Lausanne for access to medical information

canton de vaud
LIBERTÉ ET PATRIE

Commission cantonale
d'éthique de la recherche
sur l'être humain
Ch. des Falaises 1
1005 Lausanne

Prof. R. Darioli
Président

Secrétariat central
Tél. 021 314 5598/5601/8622
Fax 021 314 76 01
E-mail: secretariatcervd@unil.ch

Sous-Commission I
Président a.i. Prof. R. Darioli
Tél. 021 314 5629

Sous-Commission II
Président Prof. R. Darioli
Tél. 021 314 5629

Sous-Commission III (Psychiatrie)
Président Prof. F. Stiefel
Tél. 021 314 0234

M. Claude Pichonnaz
Enseigant
DAL6 - Physiothérapie DAL
NES 04-322
1011 Lausanne

Lausanne, le 5 juillet 2012
RD/ag

Avis de la Commission cantonale (VD) d'éthique de la recherche sur l'être humain

Monsieur,

Après réception du Formulaire de Demande d'accès aux informations médicales dans le cadre de la recherche, la CE vous fait part de son avis :

Protocole 270/12 : Validation d'un test cinématique simplifié de l'épaule

Investigateur principal :

M. Claude Pichonnaz
Enseigant
DAL6 - Physiothérapie DAL
NES 04-322
1011 Lausanne

Cet avis est fondé sur l'examen des documents reçus le 25 juin 2012 :

1. Formulaire de Demande d'accès aux informations médicales dans le cadre de la recherche

Type de procédure:

☐ procédure ordinaire ☐ ré-évaluation ☐ procédure ordinaire CED

☐ procédure simplifiée ☒ Avis présidentiel ☐ Avis présidentiel CEL

La Commission arrête l'avis suivant:

☒ **positif**[1] **sous réserve de l'approbation conjointe du Directeur médical du CHUV qui devrait vous parvenir sous peu.**

☐ **avis conditionnel**[2] (conditions à remplir avant approbation)

- ☐ Les documents révisés seront réévalués en procédure ordinaire (nombre de copies: 13)
- ☐ Révision des documents et information écrite à la Commission d'éthique (nombre de copies: 1)

☐ **négatif**[3] (motivé)

☐ **avis justifié de ne pas entrer en matière**[4]

.....

signifie

[1] L'étude peut être soumise aux autorités fédérales compétentes (Swissmedic / OFSP / OFEFP) pour notification. L'étude peut être entreprise (s'il s'agit d'une étude non régie par la Loi sur les produits thérapeutiques, la Loi sur la transplantation, la Loi relative à la recherche sur les cellules souches ou l'Ordonnance sur la radioprotection).

[2] Les documents concernés doivent être révisés avant soumission à la Commission d'éthique. L'étude ne peut ni débuter ni être notifiée avant d'avoir obtenu l'avis positif de la Commission d'éthique.

[3] Dans sa forme actuelle, l'étude ne peut pas être mise en route.

[4] La CE n'est légalement pas compétente pour évaluer cette étude. Soit une autre CE est habilitée à l'évaluer, soit l'étude ne nécessite pas d'approbation par une CE.

Emoluments perçus pour chaque dossier soumis à la Commission pour évaluation, selon barème ci-joint: **CHF 50.- (code 4.5).** Vous recevrez une facture ultérieurement.

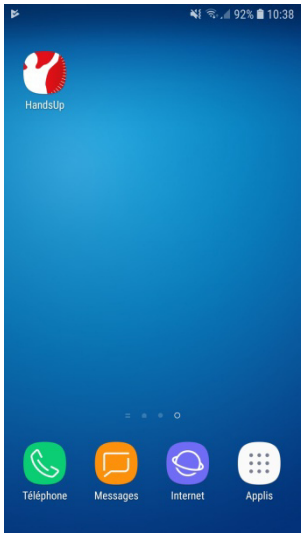
Remarques :

- La CE atteste qu'elle accomplit son travail conformément aux recommandations ICH-GCP.
- Veuillez SVP surligner les modifications apportées au document.
- Droit de recours dans le cadre de la Commission d'éthique.
- L'avis s'applique également aux autres investigateurs(trices) mentionné(e)s dans la demande d'évaluation qui travaillant dans des sites de recherche relevant du champ de compétence de la CE (doivent figurer sur une liste séparée).

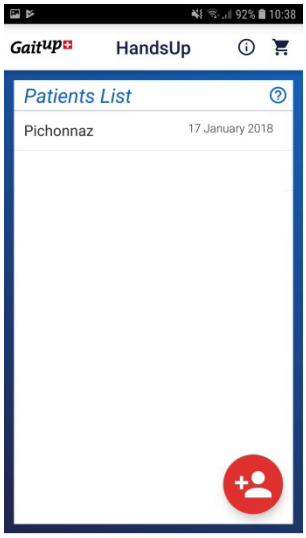
Prof. Roger Darioli
Président

Appendix XI

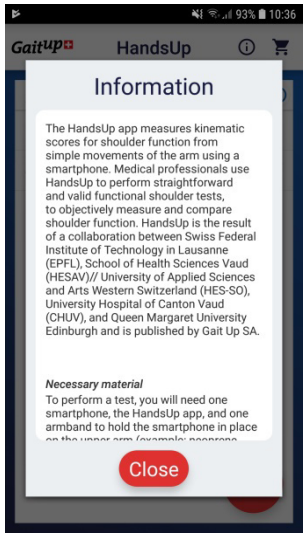
Presentation of the B-B Score application features



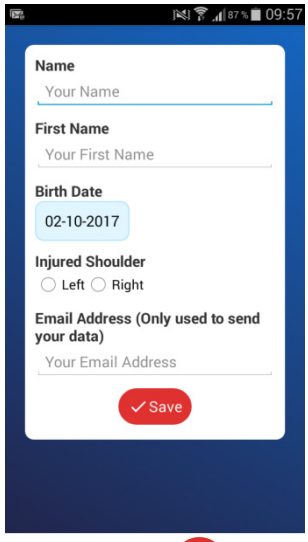
Appplication icon



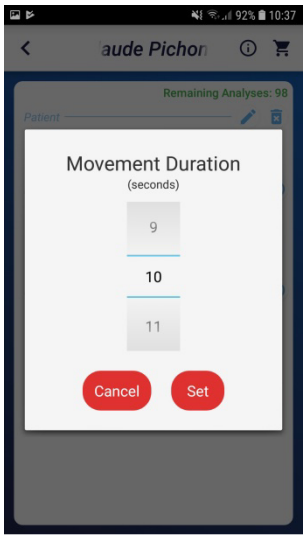
Home page, where patients' files can be created



By pressing ⓘ detailled information on the score can be read: Generalities on conception and involved partners, necessary material, scientific backgroung, score presentation, disclosure and published references

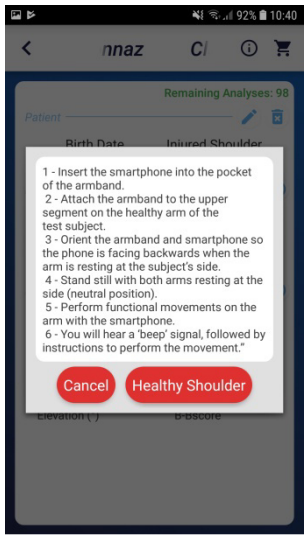


By pressing  patient's information can be entered. For the sake of confidentiality, the data can be transferred only using an email. No data are stored elsewhere that on the smartphone or communicated to anybody. As required by the legal obligation to maintain professional secrecy, the results must not be transferred to anyone without the patient's previous agreement.

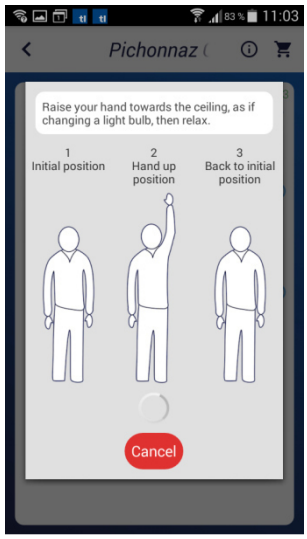


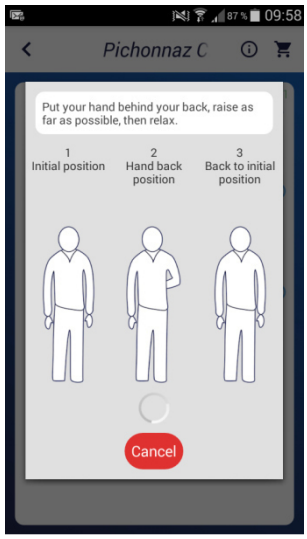
The movement duration can be adjusted to the patient's capacities.

The smarphone is then attached to the back of the arm by means of an armband (please see Figure 3.2 within Chapter three, p. 128)

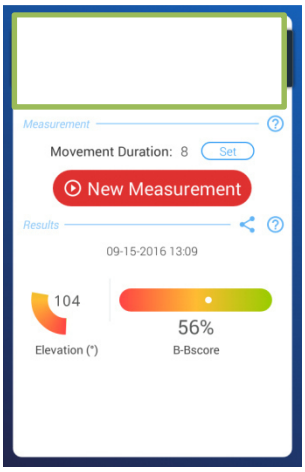


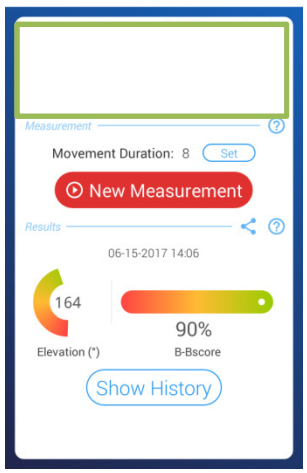
A description of the measurement procedure is shown before the beginning of a new measurement. The measurement begins when the "healthy shoulder" button is touched.

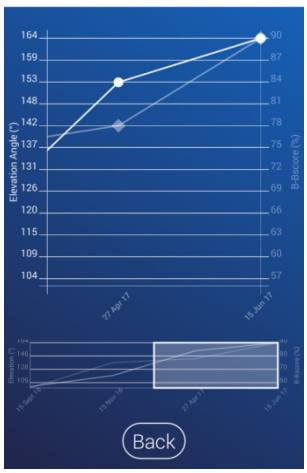




Three replications of the series of movements "hand to the back" and "hand to the ceiling" are completed. On 'launching' and initiating the software application, the smartphone provides instructions to the user, through the smartphone's loudspeaker, when to perform a B-B Score-related movement. For each movement, the application records the acceleration and angular velocity signals. The movements are first performed using the healthy side of the body and then repeated with the painful side.







For illustation, the outcomes of a patient treated in physiothepray for a capsulitis (patient's information on the top of the screen have been blinded). The B-B Score change over time can be monitored (1st: 56%; 2nd 75%; 3rd 90%). The maximum range of motion of the shoulder in elevation on the affected side is also recorded (1st: 104%; 2nd 119%; 3rd 164%). Progression curves can be inspected (faded line and numbers on the right scale are related to the B-B Score; white line and numbers are related to shoulder elevation).
The faded cursor at the bottom of the screen allows to navigate the time scale in order to see a portions of interest of the progression curve.

Appendix XII

Phase 3 related published article

Sensors **2015**, *15*, 26801-26817; doi:10.3390/s151026801

OPEN ACCESS

sensors

ISSN 1424-8220

www.mdpi.com/journal/sensors

Article

Measurement Properties of the Smartphone-Based B-B Score in Current Shoulder Pathologies

Claude Pichonnaz [1,2,*], Cyntia Duc [3], Nigel Gleeson [4], Céline Ancey [1], Hervé Jaccard [1,2], Estelle Lécureux [5], Alain Farron [2], Brigitte M. Jolles [2] and Kamiar Aminian [3]

1 Haute Ecole de Santé Vaud (HESAV)//HES-SO, University of Applied Sciences Western Switzerland, Physiotherapy Department, Avenue de Beaumont 21, Lausanne 1011, Switzerland; E-Mail: Celine.Ancey@hesav.ch

2 CHUV-UNIL, Orthopedics and Traumatology Department, Avenue Pierre-Decker 4, Lausanne 1011, Switzerland; E-Mails: Hervé.Jaccard@chuv.ch (H.J.); Alain.Farron@chuv.ch (A.F.); Brigitte.Jolles-Haeberli@chuv.ch (B.M.J.)

3 Laboratory of Movement Analysis and Measurement, Ecole Polytechnique Fédérale de Lausanne (EPFL), ELH 135 (Bâtiment ELH), Station 11, Lausanne 1015, Switzerland; E-Mails: cyntia.duc@a3.epfl.ch (C.D.); kamiar.aminian@epfl.ch (K.A.)

4 School of Health Sciences, Queen Margaret University, Edinburgh EH21 6UU, UK; E-Mail: ngleeson@qmu.ac.uk

5 CHUV-UNIL, direction médicale, Rue du Bugnon 46, Lausanne 1011, Switzerland; E-Mail: Estelle.Lecureux@chuv.ch

* Author to whom correspondence should be addressed; E-Mail: Claude.Pichonnaz@hesav.ch; Tel.: +4-121-316-8126; Fax: +4-121-316-8102.

Academic Editor: Ki H. Chon

Received: 15 June 2015 / Accepted: 12 October 2015 / Published: 22 October 2015

Abstract: This study is aimed at the determination of the measurement properties of the shoulder function B-B Score measured with a smartphone. This score measures the symmetry between sides of a power-related metric for two selected movements, with 100% representing perfect symmetry. Twenty healthy participants, 20 patients with rotator cuff conditions, 23 with fractures, 22 with capsulitis, and 23 with shoulder instabilities were measured twice across a six-month interval using the B-B Score and shoulder function questionnaires. The discriminative power, responsiveness, diagnostic power, concurrent validity, minimal detectable change (MDC), minimal clinically important improvement (MCII), and patient acceptable symptom state (PASS) were evaluated. Significant

differences with the control group and significant baseline—six-month differences were found for the rotator cuff condition, fracture, and capsulitis patient groups. The B-B Score was responsive and demonstrated excellent diagnostic power, except for shoulder instability. The correlations with clinical scores were generally moderate to high, but lower for instability. The MDC was 18.1%, the MCII was 25.2%, and the PASS was 77.6. No floor effect was observed. The B-B Score demonstrated excellent measurement properties in populations with rotator cuff conditions, proximal humerus fractures, and capsulitis, and can thus be used as a routine test to evaluate those patients.

Keywords: shoulder; shoulder function; measurement properties; outcome assessment; validation studies; smartphone sensors; body-worn sensors; kinematics

1. Introduction

1.1. Measurement Properties in Shoulder Function Evaluation

The prevalence of shoulder pain is estimated at 26.9% [1]. This places the shoulder as the second most frequently affected body site behind the lower back. Despite the high occurrence of shoulder conditions, there is an on-going controversy about the best methods to evaluate the impact of pathologies on shoulder function. Numerous clinical questionnaires exist but the methodological and reporting quality of the validation studies is generally low [2]. As a consequence, none has been recognized as a universal standard [3–5]. Computerized movement analysis might be an alternative due to its precision and reliability. However, the use of computerized systems is restricted to research for reasons of cost, training, practicality, and accessibility. The use of smartphones allows these limitations to be largely overcome, as they are fitted with built-in movement sensors, working in three dimensions but are affordable and have become items of everyday life. However, the use of smartphones for scientific purposes requires prior scientific validation.

Clinicians and clinical researchers need thoroughly validated measurement methods to correctly evaluate the patient's performance and the efficiency of therapeutic interventions. It is essential that the measurement properties of evaluation tools are extensively established to allow a correct interpretation of the outcome. In addition to the validity and the reliability, the investigation of the responsiveness and the definition of the clinically-important values are fundamental to correctly interpret the progress over time. This work requires a methodical process as the measurement properties are context-dependent. Thus, the investigations have to be performed in a large variety of situations to provide specific values for the clinicians to be able to tackle the wide range of conditions encountered in their practice [6].

Computer-based kinematic evaluation showed promising results for objective function evaluation but has remained too cumbersome for routine clinical application. Based on nine functional tests inspired from the Simple Shoulder Test (SST) [7], Coley *et al.* developed different scores (P, RAV, and M scores) using arm acceleration and angular velocity [8]. The kinematics have been recorded with arm-attached inertial sensors, with the aim to produce a valid and clinically-applicable kinematic score that can be straightforwardly performed in clinical settings. Recently the functional tests were

simplified to provide a shoulder function score, named the B-B Score by including only two basic arm movements (hand to the Back + hand upwards as to change a Bulb) [9].

Considering the simplicity of the B-B Score and the inertial sensing facility provided by smartphones, the measurement of this score using a smartphone might make computerized shoulder evaluation much more accessible for clinicians and researchers. We have investigated the validity and the reliability of the shoulder function B-B Score measured with a smartphone in a preliminary phase of the present study. It was demonstrated that a smartphone produced comparable group measurements to an inertial sensor-based body-worn system [10]. However, the ability of the score to evaluate the patient's progression and to differentiate the results according to pathologies have not been investigated yet. The responsiveness, minimal detectable change (MDC), minimal clinically important improvement (MCII), and patient acceptable symptom state (PASS) need to be evaluated to allow a well-substantiated interpretation of the results during the patient follow-up [11–13]. The MDC is the lowest value that can be considered as above the bounds of measurement error for an instrument [12]. The MCII is the smallest change in measurement that signifies an important improvement for the patient, and the PASS is the symptom state that the patients consider acceptable [11].

1.2. Influence of Shoulder Pathologies on Physiological Movement

The measurement properties for the B-B Score need to be determined first for conservatively-treated shoulder conditions, as they are much more frequent than surgically-treated conditions. Overall, only one in every sixteen patients presenting with shoulder pain requires surgery [14]. Moreover, some results were already available for the postsurgical context as the B-B Score was developed in a population who had undergone rotator cuff and arthroplasty surgery [9]. It has been established that the B-B Score produces comparable results to the kinematic P Score, which is valid and responsive following shoulder surgery [8,15,16].

Patients with rotator cuff conditions, proximal humerus fractures, adhesive capsulitis, and shoulder instabilities are frequently encountered in shoulder consultations [17–22]. It is, thus, essential to investigate the measurement properties of the B-B Score for these conditions. The validity and measurement properties of kinematic analysis may differ according to the type of pathology which affects the movement in a specific way. Thus, the B-B Score has to be validated separately for each pathology.

Conditions associated with the shoulder's rotator cuff musculature are the most common source of shoulder pain (65%). They are caused by rotator cuff tendinopathy, rotator cuff tears, subacromial impingement or subacromial bursitis [23]. Rotator cuff tendinitis affects 29% of patients presenting with shoulder pain in general practice [19]. Rotator cuff tear prevalence is also very high and is strongly related to age. Tears are present in 2.5% of the general population in their 30 s, 25% in their 60 s, and 50% in their 80 s [18]. A painful arc during arm elevation is typical of rotator cuff conditions [24]. However, clinical presentation of rotator cuff conditions varies considerably. Range of motion (ROM) limitations may or may not be observed, and tears may remain asymptomatic despite the anatomical lesions [25].

Adhesive capsulitis, also named frozen shoulder, represents the second most prevalent cause of shoulder pain (22%) [18]. It is an idiopathic disease of the joint capsule causing mainly pain and

stiffness [23]. The adhesive capsulitis is usually considered a 12- to 18-month self-limiting process, but mild symptoms may persist longer [26].

Proximal humeral fractures are also common, as they account for 6% of all adult fractures [20]. The incidence of this type of fracture in Western countries is growing due to the increasing age of the population. The movement is altered during the rehabilitation phase by pain, stiffness, and loss of strength. The recovery at one year is generally good for the conservative and the surgical approach [27].

Finally, the shoulder instability is also a frequent cause of medical consultation in younger populations. It is characterized by the inability to maintain the humeral head in the glenoid fossa of the scapula, so that the humerus slides partially or completely out of its socket. The shoulder instability's one-year incidence is 0.56‰ individuals per year in the general population, but reaches 2.8% in a physically active young population [21,22]. Instability is problematic because it frequently leads to recurrent shoulder dislocation, apprehension, and loss of quality of life [28,29]. The movement is altered in the less stable positions of the glenohumeral joint. Typically, the patient experiences apprehension at the end of ROM while undertaking combined movements but can perform activities without problem in stable glenohumeral joint positions.

1.3. Study Aim and Hypotheses

This study is aimed at the determination of the measurement properties of the smartphone B-B Score for the assessment of the progression of current shoulder pathologies (rotator cuff condition, capsulitis, proximal humerus fracture, and shoulder instability).

Based on two assessments acquired over a six-month period, it was hypothesized that:

- the score would remain stable in the control group while it would progress significantly ($p < 0.05$) over time in each pathological group,
- the responsiveness would be comparable to that of validated clinical questionnaires,
- the area under the receiver operating characteristic (ROC) curve indicative of diagnostic power, would be at least adequate (≥0.70),
- the correlations with clinical questionnaires would be at least moderate ($r > 0.50$) [6,30].

No hypothesis was made about the MDC, MCII, and PASS values as these investigations primarily aimed at the determination of these values for the needs of clinical evaluation.

2. Experimental Section

2.1. Participants

A prospective cohort study was conducted between August 2011 and May 2014 at the Department of Traumatology and Orthopaedic Surgery of the University Hospital of Lausanne. Ethical approval was granted by the Human Research Ethics Committee of the Canton of Vaud (CER-VD). Patients gave their signed informed consent for the participation in the study.

Patients were adults (>18 years old). They presented with one of the following shoulder conditions, as stated during their first medical consultation at the specialized shoulder consultation unit of the hospital: a rotator cuff condition, shoulder instability, adhesive capsulitis, proximal humerus fracture.

With the exception of patients with fractures, patients who gave their consent underwent a baseline measurement session within two weeks following the medical consultation, and a second session six months later. For patients with humerus fractures, measurements were performed six weeks post-stabilisation and six months later, provided that the radiological control showed normal healing.

Only patients who required conservative treatment were selected in the rotator cuff condition, capsulitis or instability groups. Patients undergoing surgical and conservative fracture treatments were included as the progress and functional prognosis is similar in both populations [27].

A group of participants younger than 35 years old without a history of shoulder condition/pain, was also included to evaluate the performance in a healthy population and the stability of the score. These participants were purposefully younger than the patients to avoid bias related to the high prevalence of asymptomatic rotator cuff tear above 40 years old [25].

The sample size calculation was based on the data of a pilot study that included seven controls and 16 patients. The calculation was made so that, with a significance level at $p < 0.05$, the power of 0.80 was reached when the minimal standards for acceptable properties of the score were met. Eighteen patients per group were needed for a significant correlation when $r > 0.50$, 11 patients for an area under a ROC curve of 0.80 with a standard error of 0.1, and nine patients for a significant difference between the patients and the control group [31,32]. According to these estimations, 20 participants were enrolled in each group of pathology and in the control group.

Exclusion criteria were a bilateral shoulder condition, any concomitant pain or condition involving the upper limb or cervical spine, medical contraindication to execute movements required for score completion, tumour, neurological conditions interfering with the test, and an insufficient local language level to give truly informed consent or to understand questionnaires.

2.2. Measurement Protocol Heading

Patients were measured using a smartphone (iPod®, Apple, Cupertino, CA, USA) attached to the back of the arm with an armband (Figure 1). The lower edge of the smartphone was set 3 cm above the upper edge of olecranon. The iPod was fitted with 3D built-in sensors (accelerometers: ±2 g precision: ±0.02 g; gyroscopes: ±500°/s precision: ±0.2°/s; sampling frequency: 100 Hz) [33].

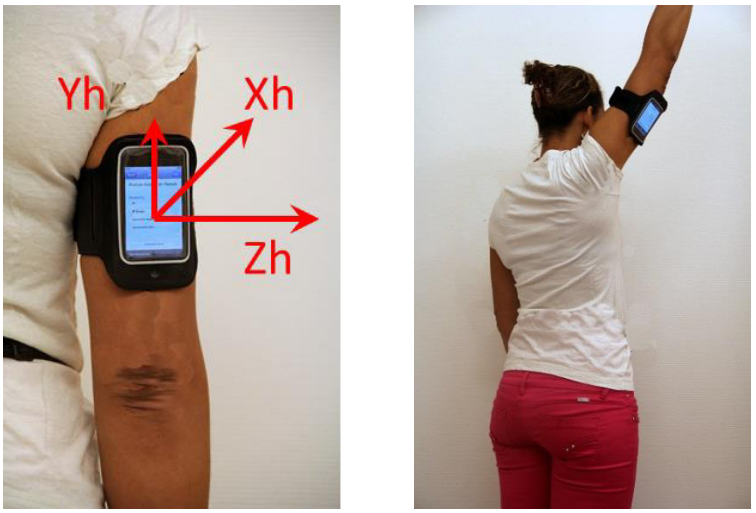


Figure 1. iPod® attached to the arm during the test completion.

After the setting-up of the system, the participants watched a video-recorded demonstration of the execution of the B-B Score. They were instructed to do the movements in the pain free ROM at their self-selected speed. Movements were executed in a standing position following smartphone-recorded instructions. The patients first undertook three repetitions of the two B-B Score movements on the healthy side (put hand to the back + hand to the ceiling as to change a bulb) and then repeated the task on the pathological side. The controls executed the same procedure beginning on the dominant side.

The B-B Score was computed as the ratio of a power-related unit [(deg/s) × (m/s^2)] of the affected side relative to the healthy side, expressed as a percentage [8]. It was calculated along the method described in Pichonnaz [9].

An application, called iShould (instrumented shoulder test) was programmed in Objective-C [34,35]. This application enabled the acquisition of the acceleration and angular velocity signals during the movements of the shoulder, and the computation of the B-B Score value, as described in the Figure 2. Once the application had been initiated at the start of the assessment, the smartphone provided instructions to the user, through the smartphone loudspeaker, as to when the user should perform a movement associated with the B-B Score. For each score' movement, the application recorded the acceleration and angular velocity signals for a predefined period of 10 s. The movements were first performed with the healthy side and then repeated with the painful side. At the end of the test, the B-B Score was directly calculated, displayed on the smartphone screen, and then stored on the smartphone. The application enabled exporting of all saved data to a computer for its direct comparison with the data from the inertial sensors of the reference system.

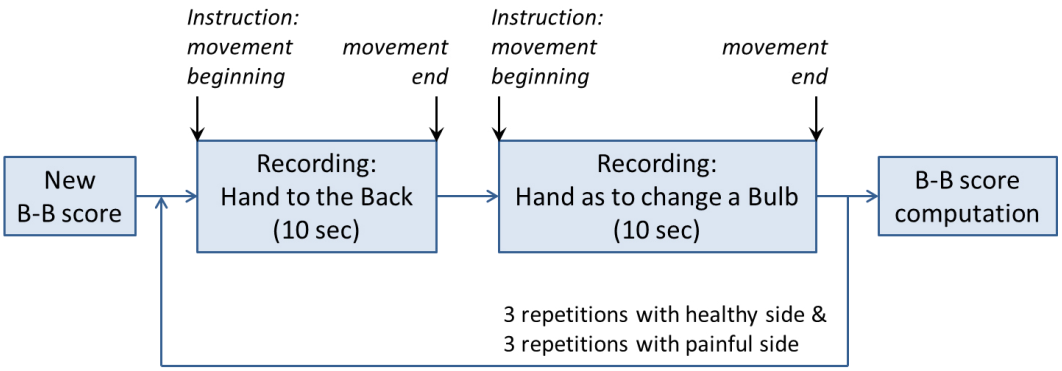


Figure 2. Schema of the application steps for the recording of a B-B Score.

One hundred percent represented a perfect balance between sides and the score decreases according to the severity of the functional loss. The score was calculated based on the mean over the three replications.

Clinical questionnaires were also completed. Four currently-used shoulder function questionnaires (Quick Disabilities of the Arm and Shoulder score (QuickDASH), Simple shoulder test (SST), Constant score and Constant relative score (based on an age-and sex-matched normal populations)), a specific shoulder instability questionnaire (Western Ontario Shoulder Instability Index (WOSI)), the EuroQol quality of life questionnaire (EQ-5D), and the pain visual analog scale (VAS) were completed [7,36–40]. The Constant Score was completed according to the modified guidelines [41]. The shoulder function questionnaires were selected because they represent current standards [5,42–44]. They

allowed the evaluation of the concurrent validity for the B-B Score but not of its validity against a "gold standard", due to the controversy surrounding shoulder function evaluation.

2.3. Analysis

Descriptive statistics were calculated for the patients' characteristics and the outcomes at baseline and at six months. The differences between groups were analyzed using the Mann-Whitney or the chi-square tests as applicable, and the differences between stages were tested for each pathological group using the Wilcoxon signed rank test. The responsiveness for the baseline—six months evolution was calculated using Cohen's *d* effect size with a 95% confidence interval. The diagnostic power for shoulder pathology detection was calculated using the ROC curve analysis. The area under the curve (AUC), sensitivity, specificity, and optimal detection threshold (highest sensitivity-specificity ratio) were calculated. The Spearman correlations were used to assess the strength of relationship between the B-B Score and the questionnaires for each of the pathologies. It was considered that a floor effect existed if >15% of patients scored less than 0 + MDC at baseline [13,45]. No ceiling effect was calculated as the score has theoretically no upper limit.

The MCII and PASS were determined for the patient group using the anchor-based method as described in Tubach *et al.* [11]. The MDC was calculated as described in Beaton *et al.* [12].

3. Results

One hundred and eight participants were tested at baseline (20 healthy participants, 20 patients with rotator cuff condition, 23 with fractures, 22 with capsulitis, and 23 with shoulder instability). All controls were measured at six months. Four patients could not be contacted at six months and four refused to participate for reasons without relationship with the study.

Drop-out rate was low (7%) and the number of patients lost at follow up were compensated to reach the planned sample size.

The population characteristics and the significance of the differences between groups are described in Table 1.

Table 1. Participants' characteristics by group.

	Rotator Cuff (*n* = 20)	**Fracture (*n* = 23)**	**Capsulitis (*n* = 22)**	**Instability (*n* = 23)**	**Control (*n* = 20)**
Age mean (SD), Years	63.5 * (10.6)	60.1 * (15.6)	52.5 * (13.8)	32.1 (14.1)	28.2 (6.2)
Sex, % Women	50	78	60	43	50
Weight Mean (SD), kg	78.3 (18.2)	69.6 (15.1)	78.3 (15.1)	70.8 (12.9)	74.7 (17.4)
Body Mass Index Mean (SD), kg/m^2	25.8 (5.4)	25.8 (5.4)	25.8 (5.4)	25.8 (5.4)	24.2 (3.9)
Size Mean (SD), m.	164.0 * (7.4)	167.7 (9.7)	172.4 (10.9)	172.6 (9.4)	175.0 (10.3)
Hand Dominance, % Right-Handed	90	87	100	87	90
Affected Side, % Dominant Side	70	25	45	52	-

* Significant difference with control group.

The outcomes of the B-B Score for the control group, and for the patient group by pathologies are presented in Table 2 and Figure 3. The differences between the control group and the rotator cuff condition, fracture, and capsulitis patient groups were significant ($p < 0.01$). The difference between the shoulder instability group and the control group, was non-significant ($p = 0.06$).

Table 2. Mean and standard deviation of the B-B Score. Unit of scores are % representing the performance of the pathological side compared to the healthy side.

Pathology		Control	Rotator Cuff	Humerus Fracture	Capsulitis	Shoulder Instability
Baseline	Mean (SD)	94.1 (11.1) *	63.1 (19.7) *	46.3 (17.5) *	54.4 (14.6) *	84.5 (22.6)
	Sample size	20	20	23	22	23
6 months	Mean (SD)	96.0 (8.3) *	77.6 (21.1) *,†	78.9 (15.1) *,†	75.3 (20.5) *,†	91.2 (15.6)
	Sample size	20	19	20	21	20

* Significant difference with the control group ($p < 0.01$); † Significant difference with baseline ($p < 0.01$).

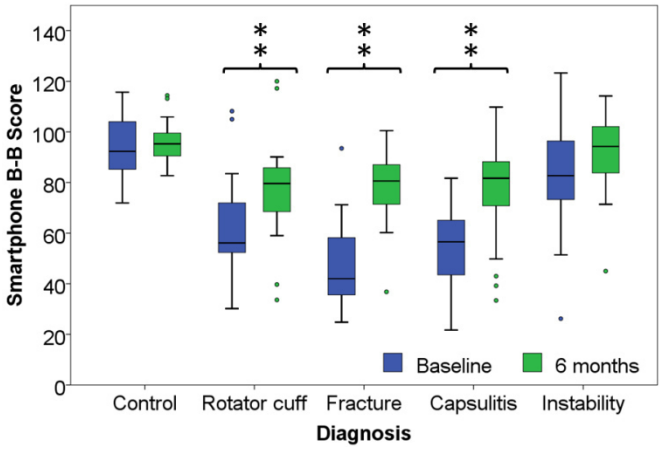


Figure 3. Outcome of the B-B Score for the control group and the pathology groups. **: significant difference with the control group $p < 0.01$.

The effect size and 95% confidence intervals are presented in Table 3 for the B-B Score, Constant and Constant relative score, the SST, QuickDASH, and WOSI. The area under the curve (AUC) with 95% CI and the cut-off for optimal sensitivity-specificity ratio are detailed in Table 4. The correlations between the shoulder function questionnaires are presented for each pathologies in Table 5.

Table 3. Effect Size (95% CI) for each score and each pathology.

Outcome Measure	Rotator Cuff	Fracture	Capsulitis	Instability
	Effect Size (95% CI)			
B-B Score	0.69 (0.02–1.33)	1.94 (1.14–2.67)	1.16 (0.49–1.79)	0.10 (−0.52–0.72)
Constant	0.54 (−0.12–1.18)	2.09 (1.26–2.83)	1.05 (0.38–1.67)	0.21 (−0.42–0.82)
Relative Constant	0.50 (−0.15–1.14)	2.10 (1.27–2.84)	1.04 (0.38–1.67)	0.27 (−0.36–0.89)
SST	0.52 (−0.13–1.16)	1.65 (0.89–2.35)	0.86 (0.22–1.48)	0.10 (−0.53–0.71)
QuickDASH	0.35 (−0.30–0.98)	1.25 (0.53–1.91)	0.55 (−0.08–1.16)	0.01 (−0.61–0.63)
WOSI	-	-	-	0.47 (0.17–1.09)
EQ-5D	0.23 (−0.42–0.86)	0.76 (0.09–1.40)	0.34 (−0.27–0.94)	0.37 (−0.26–0.99)
EQ-5D VAS	0.07 (−0.57–0.70)	0.37(−0.26–0.99)	0.06 (−0.55–0.66)	0.11 (−0.51–0.73)

Table 4. ROC curve analysis results for classifying pathologies.

	AUC (95% CI)	B-B Score Threshold (%)	Sensitivity (%)	Specificity (%)
Rotator Cuff	0.90 (0.78–1.00)	83.6	90	90
Humerus Fracture	0.98 (0.94–1.00)	71.6	100	96
Capsulitis	0.99 (0.98–1.00)	82.1	95	100
Shoulder Instability	0.67 (0.50–0.84)	81.6	95	48

Table 5. Spearman correlation coefficients between B-B Score and clinical questionnaires.

	Rotator Cuff	Humerus Fracture	Capsulitis	Shoulder Instability
Constant	0.82 **	0.70 **	0.68 **	0.46 *
Relative Constant	0.84 **	0.69 **	0.69 **	0.43 *
SST	0.63 **	0.66 **	0.76 **	0.52 *
QuickDASH	−0.55 *	−0.40	−0.64 **	−0.57 **
WOSI	-	-	-	0.58
VAS pain	−0.50 *	−0.07	−0.39	−0.19
EQ5D	0.33	0.18	0.63 **	0.46 *
EQ5D-VAS	0.16	−0.30	0.44 *	0.47 *

SST: simple shoulder test; QuickDASH: Quick Disabilities of the Arm, Shoulder and Hand score; WOSI: Western Ontario Shoulder Instability Index; SSV: Subjective Shoulder Value; VAS: Visual Analog Scale. * significant correlation ($p < 0.05$); ** significant correlation ($p < 0.01$).

The MDC was 18.1%. The MCII of the B-B Score was 25.2% and the PASS was 77.6. No floor effect was observed as all patients performed above the MDC.

4. Discussion

This study aimed at the determination of the measurement properties of the smartphone B-B Score in current shoulder pathologies (rotator cuff conditions, capsulitis, proximal humerus fractures, and shoulder instabilities).

4.1. Results Interpretation

Participants younger than 40 years old were purposefully enrolled in the control group to prevent the inclusion of people with undetected rotator cuff conditions [25]. Thus, the significant difference in patient size between the rotator cuff group and the control group reflects the age-related decrease in size [46]. It is not likely to have an impact on this study's results as age is not known to have an influence on symmetry in arm movement, as measured by the B-B Score. The high proportion of women in the fracture group is representative of gender prevalence in the wider population [20]. The low proportion of patients affected on the dominant side in the same group is of minor importance, as the outcome is not influenced by the fracture side [47]. Further, the influence of dominance on the B-B Score is minimal, as observed in the control group and in a previous study [9].

The B-B Score differences between the control and the patient groups were highly significant with the exception of the shoulder instability group. The functional loss was, in order of importance, more

marked for patient with a fracture, a capsulitis, and a rotator cuff condition than for instability. Hence, the B-B Score clearly discriminated the three first groups from the healthy group but displayed a lower discriminative power for shoulder instability.

Shoulder instability is characterised by apprehension in the arm positions that exposes the patient to a glenohumeral dislocation risk [29]. It might be that the B-B Score is not challenging enough for these patients, as it is executed in the pain-free ROM and involved a self-chosen speed. Thus, the movement of the involved shoulder is not affected by the instability in the normal testing conditions of the B-B Score. Consequently, the functional loss may remain undetected. A more challenging version of the score inducing apprehension is hardly conceivable, as it might put the patient in a situation of actual dislocation likelihood. These results highlight that shoulder instability affects movement in a different way than other shoulder pathologies and should, thus, be evaluated using a specific tool, like the WOSI, for example.

The non-significant baseline to six-month progression in the control group indicated that the B-B Score is stable over time during which the participant's performance can reasonably be expected to have remained unchanged. The significant differences over time observed in the rotator cuff condition, humerus fracture, and capsulitis groups indicate that the B-B Score discriminates amongst clinical stages for these pathologies. Conversely, no significant difference was found in the shoulder instability group.

It should be noted that the treatments were not standardized in this study as the aim was to evaluate the score properties but not the treatment's efficacy. Thus, the observed results reflect the combination of the natural evolution and of the individualized treatment received by the patients.

The effect size measured in this study should be considered as indicative, as the confidence intervals were large. The effect sizes were larger, in order of importance, for the rotator cuff, humerus fracture, and capsulitis conditions, than for the shoulder instability condition. These results are essentially related to the respective baseline to six-month progression in each one of these pathologies. As a consequence, the absolute value of the effect size is relative to the context of measurement and, hence, the reference to cut-off values can be misleading [48].

Conversely, the comparison of the effect sizes of concurrent measurement methods for a given condition is informative towards the respective responsiveness of a score. The B-B Score was the most responsive score for the rotator cuff and capsulitis groups. The Constant and Constant relative score displayed the better responsiveness for humerus fracture, followed by the B-B Score. The B-B Score nevertheless constitutes a reasonable alternative to the Constant score for fracture evaluation, when the patient is unable to perform the strength measurement (as is the case in 51.9% of patients referred for shoulder surgery), and when the administrative burden is of concern [4]. All shoulder function evaluation methods showed better responsiveness than the EQ-5D generic quality of life questionnaire. No floor effect was observed for the B-B Score as all patients performed above the MDC value.

Similarly, to the Constant, DASH, and SST, the B-B Score demonstrated a poor responsiveness for shoulder instability. The WOSI displayed the best responsiveness for the evaluation of the shoulder instability. The limited responsiveness of the Constant, DASH, and SST for this patient population has previously been reported in the literature [40,49,50].

The AUC were excellent ($\geq$0.90) for all pathologies except shoulder instability. The diagnostic power of the B-B Score was higher for fractures and capsulitis (0.98 to 0.99) than for rotator cuff

conditions (0.90). The sensitivity and specificity at the optimal threshold were excellent for these three pathologies. Conversely, the diagnostic power was insufficient in the instability group as the AUC was lower than the 0.70 threshold [51]. Thus, the B-B Score is highly efficient for detecting loss of shoulder function in rotator cuff, fracture, and capsulitis. However, although the score is able to detect that pathology impairs shoulder function, it is not able to differentiate amongst pathologies. Further research should investigate to what extent alterations in specific movement patterns might allow discrimination amongst pathologies.

The correlations for the B-B Score with the Constant, Constant relative, and SST were moderate to high (0.63 to 0.82) for rotator cuff conditions, factures, and capsulitis [30]. In contrast, the relationship with the QuickDASH was generally lower (0.36–0.64) and non-significant in some cases. The merely objective nature of the B-B Score and the merely subjective nature of the QuickDASH may explain the lower relation with this questionnaire. The lower correlations with the VAS pain scale indicated that the B-B Score is essentially a measure of shoulder function.

Moderate to low correlations were found between the B-B Score and shoulder function questionnaires when considering instability. These results indicated that the relation to function was limited for this pathology. Conversely, the B-B Score adequately captured shoulder function for rotator cuff, fracture, and capsulitis. The absence of a floor effect indicated that the responsiveness was not altered for patients performing at a low functional level.

Some clinically useful values (MDC, MCII, and PASS) were also calculated in this study. No differentiation between pathologies was made due to the limited sample size. The MDC reflects the magnitude of change that is needed to consider that the change is greater than the measurement error for an instrument [12]. The MDC of the B-B Score using a smartphone indicated that the score difference needs to be greater than 18.1% to ensure that it is a real variation of a patient's state. The MCII characterizes which level of score improvement reflects a meaningful progress for the patient [52]. Based on the MCII value, the B-B Score improvement between two stages needs to be greater than 25.2% for the patient to consider the improvement as meaningful. The PASS is the value beyond which patients consider themselves well [53]. Patients performing above the 77.6% will usually consider that the function loss is acceptable.

4.2. Limitations and Further Developments

Limitations are related to the limited sample size of each patient group. Though the group size was sufficient to compare the measurement properties of the B-B Score with those of concurrent scores, larger sample sizes would be needed to get more precise estimations. Additionally, the MDC, MCII, and PASS could not be calculated separately for each pathology group.

Though the B-B Score was compared to frequently-used shoulder function questionnaires, none of them is considered as a gold standard for shoulder function evaluation. Thus, the results of this study could solely investigate the concurrent validity but not the validity of the new score by comparison to a gold standard. The use of other questionnaires would have provided a different benchmark for the comparisons. It can nevertheless be considered that the questionnaires used in this study are fair comparators as no concurrent questionnaire has demonstrated its superiority over them [2].

The results found in this study demonstrated that the B-B Score has limitations for the evaluation of patients with shoulder instability. The score discriminated neither the instability from the control group, nor the stages within the instability group. Additionally, the responsiveness was lower than that of the WOSI and the diagnostic power was poor [54]. Based on these results, the B-B Score should not be used for the evaluation of shoulder function in a shoulder instability population. Conversely, all minimum requirements were met for rotator cuff conditions, proximal humerus fractures, and adhesive capsulitis.

Based on this study, it can be considered that clinically-important measurement properties of the smartphone-based B-B Score have been defined. The determination of the clinically useful values for the shoulder pathologies considered in this study provides a background for adequate interpretation of the results in research and clinics. However, a benchmark with a reference measurement system has not been provided in this study. Future studies should compare the results, reproducibility, and diagnostic power of a smartphone and a scientific measurement device. More research is also needed in patient populations that were not investigated in this study. For example, robust validation of the B-B Score is needed within populations experiencing glenohumeral osteoarthritis, shoulder arthroplasty, and rotator cuff surgery that have been the focus of validation studies in the past [9].

A middle segment smartphone model was chosen to have an insight into the performance of an accessible model. As a wide range of smartphones have similar or better quality sensors, the results from these models should, theoretically, be comparable to those found in this study. The B-B Score is probably robust to device variations, as it compares the performance of the affected shoulder with that of the healthy one. Thus, systematic errors in measurement affecting both sides will not affect the score. However, the influence of the characteristics of each smartphone on the outcome has to be investigated and quantified before clinical implementation.

The scientific value of a novel and objective test of shoulder function, the smartphone B-B Score technique, has been endorsed by the findings of this study, but no cost analysis was conducted at this stage of development. Further studies reproducing routine working conditions should evaluate this aspect. Given the reasonable material costs and the simplicity of the procedure, there would be a reasonable expectation for a favorable outcome following scrutiny by a formal cost-analysis.

Information and communication technologies developments were not considered in this study but may be possible at a later stage. The use of a smartphone makes the measurement much more accessible for clinicians or event patients. Thus, larger scale data collection could be performed by more evaluators at a lower cost. The smartphone B-B Score measurement might, for example, be used in telemedicine due to its simplicity and accessibility. It could also facilitate the centralization of data collected in a large number of settings at an acceptable cost, thus facilitating data collection for multicentric studies and registries.

5. Conclusions

The smartphone B-B Score demonstrated excellent measurement properties in populations with a rotator cuff condition, proximal humerus fracture, and capsulitis. The diagnostic and discriminative power were excellent for these populations. The correlations with the clinical questionnaires indicated that the B-B Score is valid for shoulder function evaluation. The responsiveness compared favourably

with clinical questionnaires and no floor effect was detected. The determination of the MDC, MCII, and PASS provided a robust basis for the clinical interpretation of the outcome.

This opens interesting perspectives for routine objective shoulder function measurement in clinics, as this validated score can quickly be performed with an inexpensive device. The affordable measurement of large cohorts of participants may also be facilitated. However, the performance of the smartphones should first be compared to that of scientific measurement devices. Further investigation is also needed to devise a kinematics smartphone-based score for the evaluation of shoulder instability where the B-B Score did not meet the minimal requirements. Moreover, the measurement properties of the B-B Score should be further investigated in patient populations presenting other shoulder conditions. Studies could also explore the possibility to use the smartphone B-B Score for remote follow-ups and for early detection of suboptimal recovery.

Acknowledgments

This study was funded by the Swiss National Science Foundation—DORE 135061.

The authors would like to thank Jean-Philippe Bassin for his contribution to study design and data collection, Noémie Sauvage Pasche for her contribution to study organization and data collection, Barbara Balmelli, Anne Rothenbacher and Guillaume Christe for their contribution to data collection, Valérie Zoll and Jean Lambert for their contribution to study organization.

Author Contributions

C.P., C.D., N.G., E.L., A.F., B.M.J. and K.A. conceived and designed the experiments. C.P., C.A. and H.J. performed the experiments. C.P. and E.L. analyzed the data; C.D., N.G., C.A., H.J., A.F., B.M.J. and K.A. contributed to results interpretation. C.D. and K.A. contributed to analysis tools development. C.P. wrote the paper, all authors contributed to paper writing and revision.

Conflicts of Interest

The Authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Picavet, H.S.; Schouten, J.S. Musculoskeletal pain in the Netherlands: Prevalences, consequences and risk groups, the DMC(3)-study. *Pain* **2003**, *102*, 167–178.
2. Huang, H.; Grant, J.A.; Miller, B.S.; Mirza, F.M.; Gagnier, J.J. A systematic review of the psychometric properties of patient-reported outcome instruments for use in patients with rotator cuff disease. *Am. J. Sports Med.* **2015**, *43*, 2572–2582.
3. Fayad, F.; Mace, Y.; Lefevre-Colau, M.M. Les échelles d'incapacité fonctionnelle de l'épaule: Revue systématique. *Ann. Réadaptation Méd. Phys.* **2005**, *48*, 298–306.

4. Christie, A.; Hagen, K.B.; Mowinckel, P.; Dagfinrud, H. Methodological properties of six shoulder disability measures in patients with rheumatic diseases referred for shoulder surgery. *J. Shoulder Elbow Surg.* **2009**, *18*, 89–95.
5. Oh, J.H.; Jo, K.H.; Kim, W.S.; Gong, H.S.; Han, S.G.; Kim, Y.H. Comparative evaluation of the measurement properties of various shoulder outcome instruments. *Am. J. Sports Med.* **2009**, *37*, 1161–1168.
6. Portney, L.G.; Watkins, M.P. *Foundations of Clinical Research: Applications to Practice*; Prentice Hall Health: Upper Saddle River, NJ, USA, 2009.
7. Lippitt, S.B.; Harryman, D.T.; Matsen, F.A. A practical tool for evaluating function: The simple shoulder test. In *The Shoulder: A Balance of Mobility and Stability*; Matsen, American Academy of Orthopaedic Surgery: Rosemont, IL, USA, 1993; pp. 501–518.
8. Coley, B.; Jolles, B.M.; Farron, A.; Bourgeois, A.; Nussbaumer, F.; Pichonnaz, C.; Aminian, K. Outcome evaluation in shoulder surgery using 3D kinematics sensors. *Gait Posture* **2007**, *25*, 523–532.
9. Pichonnaz, C.; Lecureux, E.; Bassin, J.P.; Duc, C.; Farron, A.; Aminian, K.; Jolles, B.M.; Gleeson, N. Enhancing clinically-relevant shoulder function assessment using only essential movements. *Physiol. Meas.* **2015**, *36*, 547–560.
10. Pichonnaz, C.; Duc, C.; Jaccard, H.; Ancey, C.; Lécureux, E.; Aminian, K.; Farron, A.; Jolles, B.M.; Gleeson, N. Comparison of a dedicated body-worn inertial system and a smartphone for shoulder function and arm elevation evaluation. *Physiotherapy* **2015**, *101*, 1205–1206.
11. Tubach, F.; Ravaud, P.; Beaton, D.; Boers, M.; Bombardier, C.; Felson, D.T.; van der Heijde, D.; Wells, G.; Dougados, M. Minimal clinically important improvement and patient acceptable symptom state for subjective outcome measures in rheumatic disorders. *J. Rheumatol.* **2007**, *34*, 1188–1193.
12. Beaton, D.E.; Bombardier, C.; Katz, J.N.; Wright, J.G.; Wells, G.; Boers, M.; Strand, V.; Shea, B. Looking for important change/differences in studies of responsiveness. *J. Rheumatol.* **2001**, *28*, 400–405.
13. Terwee, C.B.; Bot, S.D.M.; de Boer, M.R.; van der Windt, D.A.; Knol, D.L.; Dekker, J.; Bouter, L.M.; de Vet, H.C. Quality criteria were proposed for measurement properties of health status questionnaires. *J. Clin. Epidemiol.* **2007**, *60*, 34–42.
14. Colvin, A.C.; Egorova, N.; Harrison, A.K.; Moskowitz, A.; Flatow, E.L. National trends in rotator cuff repair. *J. Bone Joint Surg. Am.* **2012**, *94*, 227–233.
15. Coley, B. *Shoulder Function and Outcome Evaluation after Surgery Using 3D Inertial Sensors*; Doctorate ès Sciences, Swiss Institute of Technology: Lausanne, Switzerland, 2007.
16. Jolles, B.M.; Duc, C.; Coley, B.; Aminian, K.; Pichonnaz, C.; Bassin, J.P.; Farron, A. Objective evaluation of shoulder function using body-fixed sensors: A new way to detect early treatment failures? *J. Shoulder Elbow Surg.* **2011**, *20*, 1074–1081.
17. Van der Windt, D.A.; Koes, B.W.; Boeke, A.J.; Deville, W.; De Jong, B.A.; Bouter, L.M. Shoulder disorders in general practice: Prognostic indicators of outcome. *Br. J. Gen. Pract.* **1996**, *46*, 519–523.

18. Yamamoto, A.; Takagishi, K.; Osawa, T.; Yanagawa, T.; Nakajima, D.; Shitara, H.; Kobayashi, T. Prevalence and risk factors of a rotator cuff tear in the general population. *J. Shoulder Elbow Surg.* **2010**, *19*, 116–120.
19. Van der Windt, D.A.; Koes, B.W.; de Jong, B.A.; Bouter, L.M. Shoulder disorders in general practice: Incidence, patient characteristics, and management. *Ann. Rheum. Dis.* **1995**, *54*, 959–964.
20. Court-Brown, C.M.; Caesar, B. Epidemiology of adult fractures: A review. *Injury* **2006**, *37*, 691–697.
21. Liavaag, S.; Svenningsen, S.; Reikeras, O.; Enger, M.; Fjalestad, T.; Pripp, A.H.; Brox, J.I. The epidemiology of shoulder dislocations in oslo. *Scand. J. Med. Sci. Sports* **2011**, *21*, 334–340.
22. Owens, B.D.; Duffey, M.L.; Nelson, B.J.; DeBerardino, T.M.; Taylor, D.C.; Mountcastle, S.B. The incidence and characteristics of shoulder instability at the United States Military Academy. *Am. J. Sports Med.* **2007**, *35*, 1168–1173.
23. Mitchell, C.; Adebajo, A.; Hay, E.; Carr, A. Shoulder pain: Diagnosis and management in primary care. *BMJ* **2005**, *331*, 1124–1128.
24. O'Kane, J.W.; Toresdahl, B.G. The evidenced-based shoulder evaluation. *Curr. Sports Med. Rep.* **2014**, *13*, 307–313.
25. Moosmayer, S.; Smith, H.J.; Tariq, R.; Larmo, A. Prevalence and characteristics of asymptomatic tears of the rotator cuff: An ultrasonographic and clinical study. *J. Bone Joint Surg. Br.* **2009**, *91*, 196–200.
26. Kelley, M.J.; Shaffer, M.A.; Kuhn, J.E.; Michener, L.A.; Seitz, A.L.; Uhl, T.L.; Godges, J.J.; McClure, P.W. Shoulder pain and mobility deficits: Adhesive capsulitis. *J. Orthop. Sports Phys. Ther.* **2013**, *43*, A1–A31.
27. Handoll, H.H.; Ollivere, B.J.; Rollins, K.E. Interventions for treating proximal humeral fractures in adults. *Cochrane Database Syst. Rev.* **2012**, *12*, doi:10.1002/14651858.CD000434.pub3.
28. Handoll, H.H.; Almaiyah, M.A.; Rangan, A. Surgical *versus* non-surgical treatment for acute anterior shoulder dislocation. *Cochrane Database Syst. Rev.* **2004**, doi:10.1002/14651858.
29. Rouleau, D.M.; Faber, K.; MacDermid, J.C. Systematic review of patient-administered shoulder functional scores on instability. *J. Shoulder Elbow Surg.* **2010**, *19*, 1121–1128.
30. Munro, B.H. *Statistical Methods for Health Care Research*; Lippincott Williams & Wilkins: Philadelphia, PA, USA, 2005.
31. Soper, D.S. Statistics Calculators. Available online: http://www.webcitation.org/6ZEMd2NIS (accessed on 12 May 2015).
32. Lenth, R.V. Java Applets for Power and Sample Size. Available online: http://www.webcitation.org/6ZEMrvmpu (accessed on 12 May 2015).
33. Mark, D.; Nutting, J.; LaMarche, J. *Beginning iOS 5 Development: Exploring the iOS SDK*; Apress: New York, NY, USA, 2011.
34. Oïhénart, L.; Duc, C.; Aminian, K. iShould: Functional evaluation of the shoulder using a Smartphone. *Gait Posture* **2012**, *36*, 61–62.
35. Smartphone App iShould. Available online: http://lmam.epfl.ch/smartphone/ishould (accessed on 23 October 2015)

36. Constant, C.R.; Murley, A.H. A clinical method of functional assessment of the shoulder. *Clin. Orthop. Relat. Res.* **1987**, *214*, 160–164.
37. Richards, R.R.; An, K.N.; Bigliani, L.U.; Friedman, R.J.; Gartsman, G.M.; Gristina, A.G.; Iannotti, J.P.; Mow, V.C.; Sidles, J.A.; Zuckerman, J.D. A standardized method for the assessment of shoulder function. *J. Shoulder Elbow Surg.* **1994**, *3*, 347–352.
38. American Academy of Orthopaedic Surgeons. The DASH Outcome Measure. Available online: http://www.webcitation.org/6ZEN143eU (accessed on 12 May 2015).
39. EuroQol Group. EQ-5D a Standardised Instrument for Use as a Measure of Health Outcome. Available online: http://www.webcitation.org/6ZEN3QDth (accessed on 12 May 2015).
40. Kirkley, A.; Griffin, S.; McLintock, H.; Ng, L. The development and evaluation of a disease-specific quality of life measurement tool for shoulder instability. *Am. J. Sports Med.* **1998**, *26*, 764–772.
41. Constant, C.R.; Gerber, C.; Emery, R.J.H.; Søjbjerg, J.O.; Gohlke, F.; Boileau, P. A review of the Constant score: Modifications and guidelines for its use. *J. Shoulder Elbow Surg.* **2008**, *17*, 355–361.
42. St-Pierre, C.; Desmeules, F.; Dionne, C.E.; Fremont, P.; MacDermid, J.C.; Roy, J.S. Psychometric properties of self-reported questionnaires for the evaluation of symptoms and functional limitations in individuals with rotator cuff disorders: A systematic review. *Disabil. Rehabil.* **2015**, doi:10.3109/09638288.2015.1027004.
43. Gartsman, G.M.; Morris, B.J.; Unger, R.Z.; Laughlin, M.S.; Elkousy, H.A.; Edwards, T.B. Characteristics of clinical shoulder research over the last decade: A review of shoulder articles in the Journal of Bone & Joint Surgery from 2004 to 2014. *J. Bone Joint Surg. Am.* **2015**, *97*, doi:10.2106/jbjs.n.00831.
44. Angst, F.; Schwyzer, H.K.; Aeschlimann, A.; Simmen, B.R.; Goldhahn, J. Measures of adult shoulder function: Disabilities of the Arm, Shoulder, and Hand Questionnaire (DASH) and its short version (QuickDASH), Shoulder Pain and Disability Index (SPADI), American Shoulder and Elbow Surgeons (ASES) Society standardized shoulder assessment form, Constant (Murley) Score (CS), Simple Shoulder Test (SST), Oxford Shoulder Score (OSS), Shoulder Disability Questionnaire (SDQ), and Western Ontario Shoulder Instability Index (WOSI). *Arthritis Care Res.* **2011**, *63*, 174–188.
45. McHorney, C.A.; Tarlov, A.R. Individual-patient monitoring in clinical practice: Are available health status surveys adequate? *Qual. Life Res.* **1995**, *4*, 293–307.
46. Cline, M.G.; Meredith, K.E.; Boyer, J.T.; Burrows, B. Decline of height with age in adults in a general population sample: Estimating maximum height and distinguishing birth cohort effects from actual loss of stature with aging. *Hum. Biol.* **1989**, *61*, 415–425.
47. Torrens, C.; Sanchez, J.F.; Isart, A.; Santana, F. Does fracture of the dominant shoulder have any effect on functional and quality of life outcome compared with the nondominant shoulder? *J. Shoulder Elbow Surg.* **2015**, *25*, 677–681.
48. Baguley, T. Standardized or simple effect size: What should be reported? *Br. J. Psychol.* **2009**, *100*, 603–617.
49. Godfrey, J.; Hamman, R.; Lowenstein, S.; Briggs, K.; Kocher, M. Reliability, validity, and responsiveness of the simple shoulder test: Psychometric properties by age and injury type. *J. Shoulder Elbow Surg.* **2007**, *16*, 260–267.

50. Dawson, J.; Fitzpatrick, R.; Carr, A. The assessment of shoulder instability. *J. Bone Joint Surg. Br.* **1999**, *81*, 420–426.
51. Jimerson, S.R.; Burns, M.K.; VanDerHeyden, A. *Handbook of Response to Intervention: The Science and Practice of Assessment and Intervention*; Springer Science & Business Media: Berlin, Germany, 2007.
52. Tubach, F.; Ravaud, P.; Baron, G.; Falissard, B.; Logeart, I.; Bellamy, N.; Bombardier, C.; Felson, D.; Hochberg, M.; van der Heijde, D.; *et al*. Evaluation of clinically relevant changes in patient reported outcomes in knee and hip osteoarthritis: The minimal clinically important improvement. *Ann. Rheum. Dis.* **2005**, *64*, 29–33.
53. Tubach, F.; Ravaud, P.; Baron, G.; Falissard, B.; Logeart, I.; Bellamy, N.; Bombardier, C.; Felson, D.; Hochberg, M.; van der Heijde, D.; *et al*. Evaluation of clinically relevant states in patient reported outcomes in knee and hip osteoarthritis: The patient acceptable symptom state. *Ann. Rheum. Dis.* **2005**, *64*, 34–37.
54. McDowell, I. *Measuring Health: A Guide to Rating Scales and Questionnaires*; Oxford University Press: Oxford, UK, 2006.

Appendix XIII

PROSPERO registry receipt of registration

Systematic review

Please complete all mandatory fields below (marked with an asterisk *) and as many of the non-mandatory fields as you can then click *Submit* to submit your registration. You don't need to complete everything in one go, this record will appear in your *My PROSPERO* section of the web site and you can continue to edit it until you are ready to submit. Click *Show help* below or click on the icon
to see guidance on completing each section.
This record cannot be edited because it has been rejected

1. * Review title.

Give the working title of the review, for example the one used for obtaining funding. Ideally the title should state succinctly the interventions or exposures being reviewed and the associated health or social problems. Where appropriate, the title should use the PI(E)COS structure to contain information on the Participants, Intervention (or Exposure) and Comparison groups, the Outcomes to be measured and Study designs to be included.
Comparison of patient-reported outcome measures (PROMs) and kinematic scores' measurement properties for shoulder function evaluation: a systematic review

2. Original language title.

For reviews in languages other than English, this field should be used to enter the title in the language of the review. This will be displayed together with the English language title.
Comparison des propriétés de mesure des scores subjectifs et des scores cinématiques de la fonction de l'épaule: une revue systématique

3. * Anticipated or actual start date.

Give the date when the systematic review commenced, or is expected to commence.
18/07/2018

4. * Anticipated completion date.

Give the date by which the review is expected to be completed.
31/12/2018

5. * Stage of review at time of this submission.

Indicate the stage of progress of the review by ticking the relevant Started and Completed boxes. Additional information may be added in the free text box provided.
Please note: Reviews that have progressed beyond the point of completing data extraction at the time of initial registration are not eligible for inclusion in PROSPERO. Should evidence of incorrect status and/or completion date being supplied at the time of submission come to light, the content of the PROSPERO record will be removed leaving only the title and named contact details and a statement that inaccuracies in the stage of the review date had been identified.
This field should be updated when any amendments are made to a published record and on completion and publication of the review.

The review has not yet started: No

Review stage	Started	Completed
Preliminary searches	No	Yes
Piloting of the study selection process	No	Yes
Formal screening of search results against eligibility criteria	No	No
Data extraction	No	No
Risk of bias (quality) assessment	No	No
Data analysis	No	No

Provide any other relevant information about the stage of the review here (e.g. Funded proposal, protocol not yet finalised).

6. * Named contact.

The named contact acts as the guarantor for the accuracy of the information presented in the register record.
Claude Pichonnaz

Email salutation (e.g. "Dr Smith" or "Joanne") for correspondence:

Professor Pichonnaz

7. * Named contact email.

Give the electronic mail address of the named contact.
claude.pichonnaz@hesav.ch

8. Named contact address

Give the full postal address for the named contact.
HESAV\Av. de Beaumont 21\1011 Lausanne\Switzerland

9. Named contact phone number.

Give the telephone number for the named contact, including international dialling code.
++41 21 318 81 26

10. * Organisational affiliation of the review.

Full title of the organisational affiliations for this review and website address if available. This field may be completed as 'None' if the review is not affiliated to any organisation.
Haute Ecole de Santé Vaud (HESAV)

Organisation web address:

http://www.hesav.ch/

11. Review team members and their organisational affiliations.

Give the title, first name, last name and the organisational affiliations of each member of the review team. Affiliation refers to groups or organisations to which review team members belong.

Professor Claude Pichonnaz. Haute Ecole de SantÃ© Vaud (HESAV)
Professor Nigel Gleeson. Queen Margaret University
Mr Pierre Balthazard. HESAV

12. * Funding sources/sponsors.

Give details of the individuals, organizations, groups or other legal entities who take responsibility for initiating, managing, sponsoring and/or financing the review. Include any unique identification numbers assigned to the review by the individuals or bodies listed.
None

13. * Conflicts of interest.

List any conditions that could lead to actual or perceived undue influence on judgements concerning the main topic investigated in the review.

None

14. Collaborators.

Give the name and affiliation of any individuals or organisations who are working on the review but who are not listed as review team members.

15. * Review question.

State the question(s) to be addressed by the review, clearly and precisely. Review questions may be specific or broad. It may be appropriate to break very broad questions down into a series of related more specific questions. Questions may be framed or refined using PI(E)COS where relevant.
Are the measurement properties of currently used shoulder function PROMs and kinematic shoulder function scores comparable for current shoulder pathologies evaluation?

16. * Searches.

Give details of the sources to be searched, search dates (from and to), and any restrictions (e.g. language or publication period). The full search strategy is not required, but may be supplied as a link or attachment.
Databases: PubMed/MEDLINE, Web of Science, Embase, CINAHL, Pedro
Languages: English and French
No time limitation

17. URL to search strategy.

Give a link to the search strategy or an example of a search strategy for a specific database if available (including the keywords that will be used in the search strategies).

Alternatively, upload your search strategy to CRD in pdf format. Please note that by doing so you are consenting to the file being made publicly accessible.

Do not make this file publicly available until the review is complete

18. * Condition or domain being studied.

Give a short description of the disease, condition or healthcare domain being studied. This could include health and wellbeing outcomes.
Shoulder function in current conditions (rotator cuff tear, proximal humerus fracture, capsulitis, osteoarthritis, glenohumeral instability)

19. * Participants/population.

Give summary criteria for the participants or populations being studied by the review. The preferred format includes details of both inclusion and exclusion criteria.
Included: studies on the measurement properties of PROMs or kinematic scores for patients with cuff tear, proximal humerus fracture, capsulitis, osteoarthritis, glenohumeral instability, conservatively or surgicaly treated

Excluded: any other study

20. * Intervention(s), exposure(s).

Give full and clear descriptions or definitions of the nature of the interventions or the exposures to be reviewed.
No intervention, only shoulder function outcome measurement

21. * Comparator(s)/control.

Where relevant, give details of the alternatives against which the main subject/topic of the review will be compared (e.g. another intervention or a non-exposed control group). The preferred format includes details of both inclusion and exclusion criteria.
Measurement properties of currently used shoulder function PROMs vs. measurement properties of shoulder function kinematic scores for similar conditions and treatment

22. * Types of study to be included.

Give details of the types of study (study designs) eligible for inclusion in the review. If there are no restrictions on the types of study design eligible for inclusion, or certain study types are excluded, this should be stated. The preferred format includes details of both inclusion and exclusion criteria.
Included: any published study that aimed at the determination of one or several measurement properties of currently used shoulder function PROMs or kinematic scores. Currently used PROMs have been selected based on an exploratory bibliographic search : Constant score and relative Constant score, DASH/QuickDASH, SST, ASES, WOSI
Excluded: studies on other PROMs, validation studies that do not adress any of the measurement properties mentioned in primary outcomes below, studies on kinematic parameters not related to shoulder function, studies mentionning only the ability to discriminate the patient from the control group but no other measurement property

23. Context.

Give summary details of the setting and other relevant characteristics which help define the inclusion or exclusion criteria.
There is an ongoing controversy on the validity and measurement properties of shoulder function PROMs. In parallel, a lot of research on kinematic evaluation of shoulder function has been conducted. Thus, kinematic scores might be a possible alternative to PROMs, but their respective properties have never been compared.

24. * Primary outcome(s).

Give the pre-specified primary (most important) outcomes of the review, including details of how the outcome is defined and measured and when these measurement are made, if these are part of the review inclusion criteria.
The following measurement properties of current PROMs and kinematic shoulder function scores specificaly for rotator cuff tear, proximal humerus fracture, capsulitis, osteoarthritis and glenohumeral instability, and specificaly for surgical or conservative treatment
- Reliability : test-retest, intra- and inter-evaluator reproducibility
- Responsiveness: effect size, standardised response mean, floor and ceiling effect, area under the ROC curve
- Critical values: standard error of measurement, minimal detectable change, minimal clinically important change/improvement, patient acceptable symptoms state, limits of agreement
- Normal performance

Timing and effect measures

25. * Secondary outcome(s).

List the pre-specified secondary (additional) outcomes of the review, with a similar level of detail to that

required for primary outcomes. Where there are no secondary outcomes please state 'None' or 'Not applicable' as appropriate to the review
None

Timing and effect measures

26. Data extraction (selection and coding).

Give the procedure for selecting studies for the review and extracting data, including the number of researchers involved and how discrepancies will be resolved. List the data to be extracted.
Titles and/or abstracts of studies retrieved using the search strategy and those from additional sources will be screened independently by two review authors (CP and PB) to identify studies that potentially meet the inclusion criteria outlined above. The full text of these potentially eligible studies will be retrieved and independently assessed for eligibility by two review team members (CP and PB). Any disagreement between them over the eligibility of particular studies will be resolved through discussion with a third reviewer (NG).
No fromal assessment of study quality and no quantitative evidence synthesis will be conducted, due to the lack of a study assessement method suitable to the review specific context and the lack of well-established methods for measurement properties meta-analyses. A qualitative study assessement will be conducted to explain results discrepancies between studies. Ranges of extracted measurement properties will be provided when several studies investigated a measuremnt property.
A standardised previously conceived spreadsheet will be used to extract data from the included studies
Two review authors (CP and PB) will extract data independently, discrepancies will be identified and resolved through discussion (with a third author (NG) where necessary).

27. * Risk of bias (quality) assessment.

State whether and how risk of bias will be assessed (including the number of researchers involved and how discrepancies will be resolved), how the quality of individual studies will be assessed, and whether and how this will influence the planned synthesis.
Risk of bias assessment is not applicable using a checklist that would be suitable to all investigated measurement properties. Methods of retrieved articles will be inspected by two authors (CP and PB) and potential sources of bias will be noted. Disagreements between the review authors over the risk of bias in particular studies will be resolved by discussion, with involvement of a third review author where necessary.

28. * Strategy for data synthesis.

Give the planned general approach to synthesis, e.g. whether aggregate or individual participant data will be used and whether a quantitative or narrative (descriptive) synthesis is planned. It is acceptable to state that a quantitative synthesis will be used if the included studies are sufficiently homogenous.
Ranges of extracted measurement properties will be reported, and the reasons for results diverging between studies will be discussed. The results will not be aggregated because inhomogeneity in methods is expected and well-established methods do not exist for each investigated measurement property.

29. * Analysis of subgroups or subsets.

Give details of any plans for the separate presentation, exploration or analysis of different types of participants (e.g. by age, disease status, ethnicity, socioeconomic status, presence or absence or co-morbidities); different types of intervention (e.g. drug dose, presence or absence of particular components of intervention); different settings (e.g. country, acute or primary care sector, professional or family care); or different types of study (e.g. randomised or non-randomised).
None

30. * Type and method of review.

Select the type of review and the review method from the lists below. Select the health area(s) of interest for your review.

Type of review

Cost effectiveness
No

Diagnostic
No

Epidemiologic
No

Individual patient data (IPD) meta-analysis
No

Intervention
No

Meta-analysis
No

Methodology
No

Network meta-analysis
No

Pre-clinical
No

Prevention
No

Prognostic
No

Prospective meta-analysis (PMA)
No

Qualitative synthesis
No

Review of reviews
No

Service delivery
No

Systematic review
No

Other
Yes

Properties of measurement tools

Health area of the review

Alcohol/substance misuse/abuse
No

Blood and immune system
No

Cancer
No

Cardiovascular
No

Care of the elderly
No

Child health
No

Complementary therapies

No

Crime and justice
No

Dental
No

Digestive system
No

Ear, nose and throat
No

Education
No

Endocrine and metabolic disorders
No

Eye disorders
No

General interest
No

Genetics
No

Health inequalities/health equity
No

Infections and infestations
No

International development
No

Mental health and behavioural conditions
No

Musculoskeletal
Yes

Neurological
No

Nursing
No

Obstetrics and gynaecology
No

Oral health
No

Palliative care
No

Perioperative care
No

Physiotherapy
Yes

Pregnancy and childbirth
No

Public health (including social determinants of health)
No

Rehabilitation
Yes

Respiratory disorders
No

Service delivery
No

Skin disorders
No

Social care
No

Surgery
No

Tropical Medicine
No

Urological
No

Wounds, injuries and accidents
No

Violence and abuse
No

31. Language.

Select each language individually to add it to the list below, use the bin icon to remove any added in error.
English
French

There is an English language summary.

32. Country.

Select the country in which the review is being carried out from the drop down list. For multi-national collaborations select all the countries involved.
Scotland
Switzerland

33. Other registration details.

Give the name of any organisation where the systematic review title or protocol is registered (such as with The Campbell Collaboration, or The Joanna Briggs Institute) together with any unique identification number assigned. (N.B. Registration details for Cochrane protocols will be automatically entered). If extracted data will be stored and made available through a repository such as the Systematic Review Data Repository (SRDR), details and a link should be included here. If none, leave blank.
None

34. Reference and/or URL for published protocol.

Give the citation and link for the published protocol, if there is one
None

Give the link to the published protocol.

Alternatively, upload your published protocol to CRD in pdf format. Please note that by doing so you are consenting to the file being made publicly accessible.

No I do not make this file publicly available until the review is complete

Please note that the information required in the PROSPERO registration form must be completed in full even if access to a protocol is given.

35. Dissemination plans.

Give brief details of plans for communicating essential messages from the review to the appropriate audiences.
Publication in peer-reviewed journal and congress presentation

Do you intend to publish the review on completion?

Yes

36. Keywords.

Give words or phrases that best describe the review. Separate keywords with a semicolon or new line. Keywords will help users find the review in the Register (the words do not appear in the public record but are included in searches). Be as specific and precise as possible. Avoid acronyms and abbreviations unless these are in wide use.
Shoulder; measurement tool; measurement properties; function; patient-reported outcome measure; kinematics

37. Details of any existing review of the same topic by the same authors.

Give details of earlier versions of the systematic review if an update of an existing review is being registered, including full bibliographic reference if possible.
None

38. * Current review status.

Review status should be updated when the review is completed and when it is published.
Please provide anticipated publication date

Review_Ongoing

39. Any additional information.

Provide any other information the review team feel is relevant to the registration of the review.

40. Details of final report/publication(s).

This field should be left empty until details of the completed review are available.

Give the link to the published review.

Appendix XIV

Equation used on Pubmed for scores selection

History of Pubmed search for scores selection

Recent queries

Search	Add to builder	Query	Items found	Time
#17	Add	Search **(#3 AND #10)** Filters: **Publication date from 2012/09/15 to 2017/09/15**	79	06:13:40
#16	Add	Search **(#3 AND #9)** Filters: **Publication date from 2012/09/15 to 2017/09/15**	95	06:01:28
#15	Add	Search **(#3 AND #8)** Filters: **Publication date from 2012/09/15 to 2017/09/15**	201	06:01:23
#14	Add	Search **(#3 AND #7)** Filters: **Publication date from 2012/09/15 to 2017/09/15**	376	06:01:17
#13	Add	Search **(#3 AND #6)** Filters: **Publication date from 2012/09/15 to 2017/09/15**	605	06:01:10
#12	Add	Search **(#3 AND #5)** Filters: **Publication date from 2012/09/15 to 2017/09/15**	1447	06:01:05
#11	Add	Search **(#3 AND #4)** Filters: **Publication date from 2012/09/15 to 2017/09/15**	1080	06:00:58
#10	Add	Search **WORC OR "Western Ontario Rotator Cuff Index"** Filters: **Publication date from 2012/09/15 to 2017/09/15**	142	06:00:22
#9	Add	Search **WOSI OR "Western Ontario Shoulder Instability Index"** Filters: **Publication date from 2012/09/15 to 2017/09/15**	95	06:00:14
#8	Add	Search **SPADI OR "Shoulder Pain and Disability Index"** Filters: **Publication date from 2012/09/15 to 2017/09/15**	216	06:00:07
#7	Add	Search **SST OR "Simple shoulder test"** Filters: **Publication date from 2012/09/15 to 2017/09/15**	1643	05:59:59
#6	Add	Search **"ASES" OR "American Shoulder and elbow surgeons score"** Filters: **Publication date from 2012/09/15 to 2017/09/15**	710	05:59:52
#5	Add	Search **(DASH OR QuickDASH OR "Disabilities of the arm, shoulder and hand")** Filters: **Publication date from 2012/09/15 to 2017/09/15**	4214	05:59:20
#4	Add	Search **("Constant Score" OR CS score OR "Constant-Murley" OR "CSM score")** Filters: **Publication date from 2012/09/15 to 2017/09/15**	2239	05:59:09
#3	Add	Search **#1 AND #2** Filters: **Publication date from 2012/09/15 to 2017/09/15**	47932	05:57:36
#2	Add	Search **((((((((((((((((((((((((((((((("Fractures, Bone"[Mesh]) OR "Fracture Dislocation"[Mesh]) OR "Joint Dislocations"[Mesh]) OR "Joint Instability"[Mesh]) OR "Wounds and Injuries"[Mesh:noexp]) OR injur*) OR fracture*) OR dislocation*) OR instability*) OR tear) OR repair) OR surger*) OR surgical) OR "shoulder trauma*") OR "Shoulder Impingement Syndrome"[Mesh]) OR impingement*) OR "Bursitis"[Mesh]) OR "Pain"[Mesh]) OR pain*) OR "Tendinopathy"[Mesh:noexp]) OR tendin*) OR tendoni*) OR "Tendon Injuries"[Mesh:noexp]) OR Arthritis[Mesh:noexp] OR "Arthroplasty"[Mesh]) OR arthroplast*) OR prosthes*) OR "Pathology"[Mesh:noexp]) OR "Disease"[Mesh]) OR condition*) OR disorder) OR disorders) OR "adhesive capsulitis") OR "frozen shoulder*") OR disabilit*) OR dysfunction*)** Filters: **Publication date from 2012/09/15 to 2017/09/15**	2637475	05:56:51
#1	Add	Search **((((((((((("Shoulder" [Mesh] OR "Shoulder Joint"[Mesh]) OR "Upper Extremity"[Mesh:noexp]) OR "Arm"[Mesh]) OR shoulder*) OR "upper extremity") OR "upper extremities") OR "upper limb") OR arm) OR glenohumeral) OR "rotator cuff"))** Filters: **Publication date from 2012/09/15 to 2017/09/15**	66133	05:56:38

Note : differences with results reported in the thesis are due to the selection of articles that focus on shoulder disorders. This difference was limited for all scores to the exception of the DASH/QuickDASH, because this score may potentially address a variety of upper limb condition including also hand, wrist, forearm and elbow conditions.

Appendix XV

Equations used on Pubmed, Cinhal, Embase, Web of Knowledge and Pedro for measurement properties of PROMs and kinematic scores

Medline/Pubmed

PROMs

Recent queries

Search	Add to builder	Query	Items found	Time
#5	Add	Search **(#1 AND #2 AND #3 AND #4)** Filters: **Publication date to 2017/05/05**	1821	03:06:12
#4	Add	Search **((((((((((((DASH) OR QuickDASH) OR "Disabilities of the arm, shoulder and hand") OR "Constant Score") OR CS score) OR "Constant-Murley") OR "CSM score") OR SST) OR "Simple shoulder test") OR WOSI) OR "Western Ontario Shoulder Instability Index") OR "ASES score") OR "American Shoulder and elbow surgeons score"))** Filters: **Publication date to 2017/05/05**	16728	03:05:37
#3	Add	Search **AND ((((((((((((((((((((((((((((("Sensitivity and Specificity"[Mesh]) OR "Reproducibility of Results"[Mesh]) OR "ROC Curve"[Mesh]) OR "Psychometrics"[Mesh]) OR valid*) OR propert*) OR clinimetric*) OR metrolog*) OR psychometric*) OR reliab*) OR reproducib*) OR "test-retest") OR responsiveness) OR "minimal detectable change*") OR MDC) OR "standard error of measurement*") OR SEM) OR "minimal clinically important improvement*") OR MCII) OR "minimal clinically important difference*") OR MCID) OR sensitivit*) OR specificit*) OR "likelihood ratio*") OR "roc curve*") OR detect*) OR correlat*) OR accura*) OR precis*) OR discern*) OR discrimin*) OR "floor effect*") OR "ceiling effect*")))**	6922969	03:03:54
#2	Add	Search **((((((((((((((((((((((((((((("Fractures, Bone"[Mesh]) OR "Fracture Dislocation"[Mesh]) OR "Joint Dislocations"[Mesh]) OR "Joint Instability"[Mesh]) OR "Wounds and Injuries"[Mesh:noexp]) OR injur*) OR fracture*) OR dislocation*) OR instability*) OR tear) OR repair) OR surger*) OR surgical) OR shoulder trauma*) OR "Shoulder Impingement Syndrome"[Mesh]) OR impingement*) OR "Bursitis"[Mesh]) OR "Pain"[Mesh]) OR pain*) OR "Tendinopathy"[Mesh:noexp]) OR tendin*) OR tendoni*) OR "Tendon Injuries"[Mesh:noexp]) OR Arthritis[Mesh:noexp] OR "Arthroplasty"[Mesh]) OR arthroplast*) OR prosthes*) OR "Pathology"[Mesh:noexp]) OR "Disease"[Mesh]) OR condition*) OR disorder) OR disorders) OR "adhesive capsulitis") OR "frozen shoulder*") OR disabilit*) OR dysfunction*)))**	11531600	03:03:43
#1	Add	Search **((((((((((((("Shoulder"[Mesh] OR "Shoulder Joint"[Mesh]) OR "Upper Extremity"[Mesh:noexp]) OR "Arm"[Mesh]) OR shoulder*) OR upper extremity) OR upper extremities) OR upper limb) OR arm) OR glenohumeral) OR rotator cuff))**		

Kinematic scores

Recent queries

Search	Add to builder	Query	Items found	Time
#5	Add	Search **(#1 AND #2 AND #3 AND #4)** Filters: **Publication date to 2017/05/05**	1707	03:17:23
#4	Add	Search **((((((((((((((((((((("Accelerometry"[Mesh:noexp] OR "Smartphone"[Mesh]) OR "Monitoring, Physiologic"[Mesh:noexp]) OR "Torque"[Mesh]) OR inertial sensor*) OR "ultrasound-based") OR "video-based") OR "motion capture") OR gyroscop*) OR acceleromet*) OR "wearable sensors*") OR IMU) OR "inertial measurement") OR infrared camera*) OR "motion analysis") OR "movement analysis") OR smartphone*) OR motion tracker* OR magnetometer*) OR magnetic system*) OR acceleration*) OR angular velocit*) OR fluidity)))** Filters: **Publication date to 2017/05/05**	150378	03:16:59
#3	Add	Search **((((((((((((((((((((((((((((((("Sensitivity and Specificity"[Mesh]) OR "Reproducibility of Results"[Mesh]) OR "ROC Curve"[Mesh]) OR "Psychometrics"[Mesh]) OR valid*) OR propert*) OR clinimetric*) OR metrolog*) OR psychometric*) OR reliab*) OR reproducib*) OR "test-retest") OR responsiveness) OR "minimal detectable change*") OR MDC) OR "standard error of measurement*") OR SEM) OR "minimal clinically important improvement*") OR MCII) OR "minimal clinically important difference*") OR MCID) OR sensitivit*) OR specificit*) OR "likelihood ratio*") OR "roc curve*") OR detect*) OR correlat*) OR accura*) OR precis*) OR discern*) OR discrimin*) OR "floor effect*") OR "ceiling effect*")))**	6922969	03:14:05
#2	Add	Search **((((((((((((((((((((((((((((((("Fractures, Bone"[Mesh]) OR "Fracture Dislocation"[Mesh]) OR "Joint Dislocations"[Mesh]) OR "Joint Instability"[Mesh]) OR "Wounds and Injuries"[Mesh:noexp]) OR injur*) OR fracture*) OR dislocation*) OR instability*) OR tear) OR repair) OR surger*) OR surgical) OR shoulder trauma*) OR "Shoulder Impingement Syndrome"[Mesh]) OR impingement*) OR "Bursitis"[Mesh]) OR "Pain"[Mesh]) OR pain*) OR "Tendinopathy"[Mesh:noexp]) OR tendin*) OR tendoni*) OR "Tendon Injuries"[Mesh:noexp]) OR Arthritis[Mesh:noexp] OR "Arthroplasty"[Mesh]) OR arthroplast*) OR prosthes*) OR "Pathology"[Mesh:noexp]) OR "Disease"[Mesh]) OR condition*) OR disorder) OR disorders) OR "adhesive capsulitis") OR "frozen shoulder*") OR disabilit*) OR dysfunction*)))**	11531600	03:13:52
#1	Add	Search **((((((((((((((("Shoulder"[Mesh] OR "Shoulder Joint"[Mesh]) OR "Upper Extremity"[Mesh:noexp]) OR "Arm"[Mesh]) OR shoulder*) OR upper extremity) OR upper extremities) OR upper limb) OR arm) OR glenohumeral) OR rotator cuff))))**	347540	03:13:39

CINHAL

PROMs

Search ID#	Search Terms	Search Options	Actions
S6	S1 AND S2 AND S3 AND S4 AND S5	**Limiters** - Published Date: - 20170531 **Search modes** - Find all my search terms	**View Results** (431) **Interface** - EBSCOhost Research Databases **Search Screen** - Advanced Search **Database** - CINAHL Complete
S5	“DASH” OR “QuickDASH” OR "Disabilities of the arm, shoulder and hand" OR "Constant Score" OR “CS score” OR "Constant-Murley" OR "CSM score" OR “SST” OR "Simple shoulder test" OR “WOSI” OR "Western Ontario Shoulder Instability Index" OR "ASES score" OR "American Shoulder and elbow surgeons score" OR "UCLA Shoulder" OR "UCLA score"	**Search modes** - Find all my search terms	
S4	(MH "Sensitivity”) OR (MH "Specificity") OR "sensitiv*" OR "specific*" OR "likelihood ratio*" OR (MH "ROC Curve") OR "roc curve*" OR "reproduci*" OR (MH "Psychometrics") OR "psychometric" OR "clinimetric" OR "metrolog*" OR (MH "Validity+") OR (MH "Reliability+") OR "reliab*" OR "test-retest" OR (MH "Measurement Error") OR (MH "Test-Retest Reliability") OR (MH "Intrarater Reliability") OR (MH "Interrater Reliability") OR "responsiv*" OR (MH "Intraclass Correlation Coefficient") OR "intraclass" OR "discrimina*" OR "propert*" OR "clinimetric*" OR "metrolog*" OR “responsiveness” OR "floor effect*" OR "ceiling effect*"OR minimal detectable change* OR "MDC" OR "standard error of measurement*" OR "SEM" OR "minimal clinically important improvement*" OR "MCII" OR "minimal clinically important difference*" OR "MCID" OR "detect*" OR "correlat*" OR "accura*" OR "precis*" OR "discrimin*"(MH "Sensitivity”) OR (MH "Specificity") OR "sensitiv*" OR "specific*" OR "likelihood ratio*" OR (MH "ROC Curve") OR "roc curve*" OR "reproduci*" OR (MH "Psychometrics") OR "psychometric" OR "clinimetric" OR "metrolog*" OR (MH "Validity+") OR (MH "Reliability+") OR "reliab*" OR "test-retest" OR (MH "Measurement Error") OR (MH "Test-Retest Reliability") OR (MH "Intrarater Reliability") OR (MH "Interrater Reliability") OR "responsiv*" OR (MH "Intraclass Correlation Coefficient") OR "intraclass" OR "discrimina*" OR "propert*" OR "clinimetric*" OR "metrolog*" OR	**Search modes** - Find all my search terms	

	“responsiveness” OR "floor effect*" OR "ceiling effect*"OR minimal detectable change* OR "MDC" OR "standard error of measurement*" OR "SEM" OR "minimal clinically important improvement*" OR "MCII" OR "minimal clinically important difference*" OR "MCID" OR "detect*" OR "correlat*" OR "accura*" OR "precis*"		
S3	(MH "Validation Studies") OR (“Instrument Validation”) OR (MH “Measurement Issues and Assessments”) OR "valid*" OR (MH "Clinical Assessment Tools+") OR "assessment tools" OR (MH "Outcome Assessment") OR (MH "Reliability and Validity") OR (MH "Research Measurement") OR (MH "Reproducibility of Results") OR (MH "Sensitivity and Specificity")	**Search modes** - Find all my search terms	
S2	(MH "Shoulder Pain") OR (MH "Pain+") OR (MH "Shoulder Impingement Syndrome") OR "impingement" OR “subacromial” OR (MH "Tendinopathy+") OR "tendinitis" OR "tendon*" OR (MH "Bursitis+") OR (MH "Rotator Cuff Injuries") OR (MH "Tears") OR “rotator cuff tear” OR "tear" OR "repair" OR (MH "Arthritis") OR (MH "Arthroplasty+") OR “arthroplasty” OR “prosthesis” OR (MH "Joint Instability+") OR (MH "Shoulder Dislocation") OR "instability" OR "dislocation" OR (MH "Adhesive Capsulitis+") OR "frozen shoulder" OR (MH "Shoulder Injuries+") OR (MH "Shoulder Fractures+") OR (MH "Fractures+") OR "fracture*" OR (MH "Pathology+") OR (MH "Disease+") OR (MH "Trauma+") OR "condition" OR "disorder*" OR "surgery" OR "surgical"(MH "Shoulder Pain") OR (MH "Pain+") OR (MH "Shoulder Impingement Syndrome") OR "impingement" OR “subacromial” OR (MH "Tendinopathy+") OR "tendinitis" OR "tendon*" OR (MH "Bursitis+") OR (MH "Rotator Cuff Injuries") OR (MH "Tears") OR “rotator cuff tear” OR "tear" OR "repair" OR (MH "Arthritis") OR (MH "Arthroplasty+") OR “arthroplasty” OR “prosthesis” OR (MH "Joint Instability+") OR (MH "Shoulder Dislocation") OR "instability" OR "dislocation" OR (MH "Adhesive Capsulitis+") OR "frozen shoulder" OR (MH "Shoulder Injuries+") OR (MH "Shoulder Fractures+") OR (MH "Fractures+") OR "fracture*" OR (MH "Pathology+") OR (MH "Disease+") OR (MH "Trauma+") OR "condition" OR "disorder*" OR "surgery" OR "surgical"	**Search modes** - Find all my search terms	
S1	(MH "Shoulder") OR "shoulder" OR (MH "Shoulder Joint+") OR (MH "Glenohumeral Joint") OR (MH "Upper Extremity+") OR (MH "Arm") OR ("upper limb") OR (MH "Rotator Cuff+") OR (MH "Rotator Cuff+") OR "rotator cuff" OR (MH "Humerus") OR "Humer*"	**Search modes** - Find all my search terms	

Kinematic scores

Search ID#	Search Terms	Search Options	Actions
S8	S1 AND S2 AND S3 AND S4 AND S7	**Limiters** - Published Date: -20170531 **Search modes** - Find all my search terms	**View Results** (511) **View Details** **Edit** **Interface** - EBSCOhost Research Databases **Search Screen** - Advanced Search **Database** - CINAHL Complete
S7	S5 OR S6	**Search modes** - Find all my search terms	
S6	TX "motion analysis*" OR TX "motion capture" OR TX "motion tracker*" OR TX "movement analysis" OR TX magnetomet* OR TX "magnetic system*" OR TX "angular velocit*" OR TX fluidity OR "infrared camera*"	**Search modes** - Find all my search terms	
S5	TX acceleromet* OR TX accelerat* OR TX smartphone* OR MH monitoring, physiologic OR "monitoring, physiologic" OR TX torque OR TX "inertial sensor*" OR TX "inertial measurement unit" OR imu OR "wearable sensor*" OR TX ultrasound OR TX video	**Search modes** - Find all my search terms	
S4	(MH "Sensitivity”) OR (MH "Specificity") OR "sensitiv*" OR "specific*" OR "likelihood ratio*" OR (MH "ROC Curve") OR "roc curve*" OR "reproduci*" OR (MH "Psychometrics") OR "psychometric" OR "clinimetric" OR "metrolog*" OR (MH "Validity+") OR (MH "Reliability+") OR "reliab*" OR "test-retest" OR (MH "Measurement Error") OR (MH "Test-Retest Reliability") OR (MH "Intrarater Reliability") OR (MH "Interrater Reliability") OR "responsiv*" OR (MH "Intraclass Correlation Coefficient") OR "intraclass" OR "discrimina*" OR "propert*" OR "clinimetric*" OR "metrolog*" OR “responsiveness” OR "floor effect*" OR "ceiling effect*"OR minimal detectable change* OR "MDC" OR "standard error of measurement*" OR "SEM" OR "minimal clinically important improvement*" OR "MCII" OR "minimal clinically important difference*" OR "MCID" OR "detect*" OR "correlat*" OR "accura*" OR "precis*" OR "discrimin*"(MH "Sensitivity”) OR (MH "Specificity") OR "sensitiv*" OR "specific*" OR "likelihood ratio*" OR (MH "ROC Curve") OR "roc curve*" OR "reproduci*" OR (MH "Psychometrics") OR "psychometric" OR "clinimetric" OR "metrolog*" OR (MH "Validity+") OR (MH "Reliability+") OR "reliab*" OR "test-retest" OR (MH "Measurement Error") OR (MH "Test-Retest	**Search modes** - Find all my search terms	

	Reliability") OR (MH "Intrarater Reliability") OR (MH "Interrater Reliability") OR "responsiv*" OR (MH "Intraclass Correlation Coefficient") OR "intraclass" OR "discrimina*" OR "propert*" OR "clinimetric*" OR "metrolog*" OR “responsiveness” OR "floor effect*" OR "ceiling effect*"OR minimal detectable change* OR "MDC" OR "standard error of measurement*" OR "SEM" OR "minimal clinically important improvement*" OR "MCII" OR "minimal clinically important difference*" OR "MCID" OR "detect*" OR "correlat*" OR "accura*" OR "precis*"		
S3	(MH "Validation Studies") OR (“Instrument Validation”) OR (MH “Measurement Issues and Assessments”) OR "valid*" OR (MH "Clinical Assessment Tools+") OR "assessment tools" OR (MH "Outcome Assessment") OR (MH "Reliability and Validity") OR (MH "Research Measurement") OR (MH "Reproducibility of Results") OR (MH "Sensitivity and Specificity")	**Search modes** - Find all my search terms	
S2	(MH "Shoulder Pain") OR (MH "Pain+") OR (MH "Shoulder Impingement Syndrome") OR "impingement" OR “subacromial” OR (MH "Tendinopathy+") OR "tendinitis" OR "tendon*" OR (MH "Bursitis+") OR (MH "Rotator Cuff Injuries") OR (MH "Tears") OR “rotator cuff tear” OR "tear" OR "repair" OR (MH "Arthritis") OR (MH "Arthroplasty+") OR “arthroplasty” OR “prosthesis” OR (MH "Joint Instability+") OR (MH "Shoulder Dislocation") OR "instability" OR "dislocation" OR (MH "Adhesive Capsulitis+") OR "frozen shoulder" OR (MH "Shoulder Injuries+") OR (MH "Shoulder Fractures+") OR (MH "Fractures+") OR "fracture*" OR (MH "Pathology+") OR (MH "Disease+") OR (MH "Trauma+") OR "condition" OR "disorder*" OR "surgery" OR "surgical"(MH "Shoulder Pain") OR (MH "Pain+") OR (MH "Shoulder Impingement Syndrome") OR "impingement" OR “subacromial” OR (MH "Tendinopathy+") OR "tendinitis" OR "tendon*" OR (MH "Bursitis+") OR (MH "Rotator Cuff Injuries") OR (MH "Tears") OR “rotator cuff tear” OR "tear" OR "repair" OR (MH "Arthritis") OR (MH "Arthroplasty+") OR “arthroplasty” OR “prosthesis” OR (MH "Joint Instability+") OR (MH "Shoulder Dislocation") OR "instability" OR "dislocation" OR (MH "Adhesive Capsulitis+") OR "frozen shoulder" OR (MH "Shoulder Injuries+") OR (MH "Shoulder Fractures+") OR (MH "Fractures+") OR "fracture*" OR (MH "Pathology+") OR (MH "Disease+") OR (MH "Trauma+") OR "condition" OR "disorder*" OR "surgery" OR "surgical"	**Search modes** - Find all my search terms	
S1	(MH "Shoulder") OR "shoulder" OR (MH "Shoulder Joint+") OR (MH "Glenohumeral Joint") OR (MH "Upper Extremity+") OR (MH "Arm") OR ("upper limb") OR (MH "Rotator Cuff+") OR (MH "Rotator Cuff+") OR "rotator cuff" OR (MH "Humerus") OR "Humer*"	**Search modes** - Find all my search terms	

Embase

PROMS

No.	Query	Results
#7	#6 AND ('article'/it OR 'review'/it)	**332**
#6	#1 AND #2 AND #3 AND #4 AND #5 AND [1-1-1800]/sd NOT [7-5-2017]/sd	**372**
#5	'disabilities of the arm, shoulder and hand (score)'/exp OR 'disabilities of the arm, shoulder and hand (score)' OR 'quickdash'/exp OR quickdash OR 'constant murley (score)'/exp OR 'constant murley (score)' OR 'constant score'/exp OR 'constant score' OR 'cs score' OR 'csm score' OR 'simple shoulder test'/exp OR 'simple shoulder test' OR sst OR 'western ontario shoulder instability index'/exp OR 'western ontario shoulder instability index' OR 'western ontario shoulder instability score' OR wosi OR 'american shoulder and elbow surgeons score'/exp OR 'american shoulder and elbow surgeons score' OR 'american shoulder and elbow surgeon score'/exp OR 'american shoulder and elbow surgeon score' OR 'ases score'/exp OR 'ases score' OR ases	**10929**
#4	'specificity'/exp OR specificity OR 'sensitivity and sensibility'/exp OR 'sensitivity and sensibility' OR 'reproducibility'/exp OR reproducibility OR 'receiver operating characteristic'/exp OR 'receiver operating characteristic' OR 'psychometry'/exp OR psychometry OR 'psychometric properties'/exp OR 'psychometric properties' OR 'validation'/exp OR validation OR 'validity'/exp OR validity OR 'measurement precision'/exp OR 'measurement precision' OR 'measurement properties' OR 'clinimetrics'/exp OR clinimetrics OR 'metrology'/exp OR metrology OR 'reliability'/exp OR reliability OR 'test retest reliability'/exp OR 'test retest reliability' OR 'test retest variability'/exp OR 'test retest variability' OR 'responsiveness'/exp OR responsiveness OR responsiv* OR 'minimal detectable change'/exp OR 'minimal detectable change' OR mdc OR 'standard error of measurement'/exp OR 'standard error of measurement' OR 'sem'/exp OR sem OR 'minimal clinically important difference'/exp OR 'minimal clinically important difference' OR 'minimal clinically important improvement'/exp OR 'minimal clinically important improvement' OR mcii OR mcid OR 'likehood ratio' OR 'roc curve*' OR 'detection'/exp OR detection OR 'correlation coefficient'/exp OR 'correlation coefficient' OR 'accuracy'/exp OR accuracy OR 'precision'/exp OR precision OR discern* OR 'discrimination'/exp OR discrimination OR 'floor effect' OR 'ceiling effect'/exp OR 'ceiling effect'	**3289470**
#3	'evaluation study'/exp OR 'evaluation study' OR 'evaluation methodology'/exp OR 'evaluation methodology' OR 'symptom assessment'/exp OR 'symptom assessment' OR 'validation study'/exp OR 'validation study' OR 'validation'/exp OR validation OR 'reproducibility'/exp	**3059475**

No.	Query	Results
	OR reproducibility OR 'outcome assessment'/exp OR 'outcome assessment' OR 'measurement'/exp OR measurement	
#2	'fracture'/exp OR fracture OR 'dislocation'/exp OR dislocation OR 'limb disease'/exp OR 'limb disease' OR 'bursitis'/exp OR bursitis OR 'pain'/exp OR pain OR 'joint instability'/exp OR 'joint instability' OR 'tendinitis'/exp OR tendinitis OR 'tendon injury'/exp OR 'tendon injury' OR 'arthritis'/exp OR arthritis OR 'arthroplasty'/exp OR arthroplasty OR 'pathology'/exp OR pathology OR 'diseases'/exp OR diseases OR 'injury'/exp OR injury OR 'tear'/exp OR tear OR 'condition'/exp OR condition OR 'repair'/exp OR repair OR surg* OR 'humeroscapular periarthritis'/exp OR 'humeroscapular periarthritis' OR 'frozen shoulder'/exp OR 'frozen shoulder' OR dysfunction* OR 'prostheses and orthoses'/exp OR 'prostheses and orthoses'	**21433437**
#1	'shoulder' OR 'shoulder'/exp OR shoulder OR 'arm' OR 'arm'/exp OR arm OR 'rotator cuff'/exp OR 'rotator cuff' OR 'upper extremit*' OR 'upper limb*'	**427375**

Kinematic

No.	Query	Results
#7	#6 AND ('article'/it OR 'review'/it) AND [1-1-1800]/sd NOT [6-5-2017]/sd	**662**
#6	#1 AND #2 AND #3 AND #4 AND #5	**925**
#5	'acceleromet*' OR 'smartphone*' OR 'physiologic monitoring' OR 'torque' OR 'inertial sensor*' OR 'inertial measurement unit*' OR 'ultrasound-based' OR 'video' OR 'motion capture' OR 'gyroscope*' OR 'wearable sensor*' OR 'infrared camera*' OR 'motion analysis' OR 'movement analysis' OR 'motion tracker' OR 'magnetometer*' OR 'magnetic system' OR 'acceleration*' OR 'angular velocit*' OR 'fluidity'	**232314**
#4	'specificity'/exp OR specificity OR 'sensitivity and sensibility'/exp OR 'sensitivity and sensibility' OR 'reproducibility'/exp OR reproducibility OR 'receiver operating characteristic'/exp OR 'receiver operating characteristic' OR 'psychometry'/exp OR psychometry OR 'psychometric properties'/exp OR 'psychometric properties' OR 'validation'/exp OR validation OR 'validity'/exp OR validity OR 'measurement precision'/exp OR 'measurement precision' OR 'measurement properties' OR 'clinimetrics'/exp OR clinimetrics OR 'metrology'/exp OR metrology OR 'reliability'/exp OR reliability OR 'test retest reliability'/exp OR 'test retest reliability' OR 'test retest variability'/exp OR 'test retest variability' OR 'responsiveness'/exp OR responsiveness OR responsiv* OR 'minimal detectable change'/exp OR 'minimal detectable change' OR mdc OR 'standard error of measurement'/exp OR 'standard error of measurement' OR 'sem'/exp OR sem OR 'minimal clinically important difference'/exp OR 'minimal clinically important difference' OR 'minimal clinically important improvement'/exp OR 'minimal clinically important improvement' OR mcii OR mcid OR 'likehood ratio' OR 'roc curve*' OR 'detection'/exp OR detection OR 'correlation coefficient'/exp OR 'correlation coefficient' OR 'accuracy'/exp OR accuracy OR 'precision'/exp OR precision OR discern* OR 'discrimination'/exp OR discrimination OR 'floor effect' OR 'ceiling effect'/exp OR 'ceiling effect'	**3316791**
#3	'evaluation study'/exp OR 'evaluation study' OR 'evaluation methodology'/exp OR 'evaluation methodology' OR 'symptom assessment'/exp OR 'symptom assessment' OR 'validation study'/exp OR 'validation study' OR 'validation'/exp OR validation OR 'reproducibility'/exp OR reproducibility OR 'outcome assessment'/exp OR 'outcome assessment' OR 'measurement'/exp OR measurement	**3086036**
#2	'fracture'/exp OR fracture OR 'dislocation'/exp OR dislocation OR 'limb disease'/exp OR 'limb disease' OR 'bursitis'/exp OR bursitis OR 'pain'/exp OR pain OR 'joint instability'/exp OR 'joint instability' OR 'tendinitis'/exp OR	**21551637**

No.	Query	Results
	tendinitis OR 'tendon injury'/exp OR 'tendon injury' OR 'arthritis'/exp OR arthritis OR 'arthroplasty'/exp OR arthroplasty OR 'pathology'/exp OR pathology OR 'diseases'/exp OR diseases OR 'injury'/exp OR injury OR 'tear'/exp OR tear OR 'condition'/exp OR condition OR 'repair'/exp OR repair OR surg* OR 'humeroscapular periarthritis'/exp OR 'humeroscapular periarthritis' OR 'frozen shoulder'/exp OR 'frozen shoulder' OR dysfunction* OR 'prostheses and orthoses'/exp OR 'prostheses and orthoses'	
#1	'shoulder' OR 'shoulder'/exp OR shoulder OR 'arm' OR 'arm'/exp OR arm OR 'rotator cuff'/exp OR 'rotator cuff' OR 'upper extremit*' OR 'upper limb*'	**430398**

Web of Knowledge

PROMs

Set	Results	Save History / Create Alert Open Saved History
# 6	1,822	(#5) *AND* **LANGUAGE:** (English OR French) *Indexes=SCI-EXPANDED, ESCI Timespan=1900-2017*
# 5	1,915	#1 AND #2 AND #3 AND #4 *Indexes=SCI-EXPANDED, ESCI Timespan=1900-2017*
# 4	28,562	TS=(DASH) OR TS=(QuickDASH) OR TS=("Disabilities of the arm, shoulder and hand") OR TS=("Constant Score") OR TS=("CS Score") OR TS=("Constant-Murley") OR TS=("CSM score") OR TS=(SST) OR TS=("simple shoulder test") OR TS=(WOSI) OR TS=("Western Ontario Shoulder Instability Index") OR TS=("ASES score") AND TS=("American Shoulder and elbow surgeons score") AND TS=("UCLA Shoulder") AND TS=("UCLA score") *Indexes=SCI-EXPANDED, ESCI Timespan=1900-2017*
# 3	10,818,476	TS=(sensitiv*) OR TS=(specifici*) OR TS=(reproducib*) OR TS=("ROC curve") OR TS=(psychometric*) OR TS=(valid*) OR TS=(propert*) OR TS=(clinimetric*) OR TS=(metrolog*) OR TS=(reliab*) OR TS=(test-retest) OR TS=(responsiv*) OR TS=("minimal detectable change*") OR TS=(MDC) OR TS=("standard error of measurement*") OR TS=(SEM) OR TS=("minimal clinically important improvement*") OR TS=(MCII) OR TS=("minimal clinically important difference*") OR TS=(MCID) OR TS=("likelihood ratio*") OR TS=(detect*) OR TS=(intraclass) OR TS=(correl*) OR TS=(accura*) OR TS=(precis*) OR TS=(discern*) OR TS=(discrim*) OR TS=("floor effect*") OR TS=("ceiling effect*") *Indexes=SCI-EXPANDED, ESCI Timespan=1900-2017*
# 2	10,090,842	TS=(pain) OR TS=(impingement) OR TS=(bursitis) OR TS=(tendon*) OR TS=(tendin*) OR TS=(fracture*) OR TS=(dislocation*) OR TS=(instab*) OR TS=(arthritis) OR TS=(arthroplast*) OR TS=(injur*) OR TS=(patholog*) OR TS=(disease) OR TS=(trauma*) OR TS=(tear*) OR TS=(condition*) OR TS=(disorder*) OR TS=(repair*) OR TS=(surger*) OR TS=(surgical) OR TS=("adhesive capsulitis") OR TS=("frozen shoulder") OR TS=(dysfunction) OR TS=(prosthes*) *Indexes=SCI-EXPANDED, ESCI Timespan=1900-2017*
# 1	374,539	TS=(shoulder*) OR TS=(upper extremit*) OR TS=(upper limb*) OR TS=(arm*) OR TS=(glenohumeral) OR TS=(humer*) OR TS=(rotator cuff*) *Indexes=SCI-EXPANDED, ESCI Timespan=1900-2017*

Kinematic scores

Set	Results	Save History / Create Alert Open Saved History
# 7	2,076	#4 AND #3 AND #2 AND #1 **Refined by: DOCUMENT TYPES:** (ARTICLE OR REVIEW) AND **LANGUAGES:** (ENGLISH OR FRENCH) *Indexes=SCI-EXPANDED, ESCI Timespan=1900-2017*
# 6	2,104	#4 AND #3 AND #2 AND #1 **Refined by: DOCUMENT TYPES:** (ARTICLE OR REVIEW) *Indexes=SCI-EXPANDED, ESCI Timespan=1900-2017*
# 5	2,109	#4 AND #3 AND #2 AND #1 *Indexes=SCI-EXPANDED, ESCI Timespan=1900-2017*
# 4	287,283	TS=(acceleromet*) OR TS=(smartphone*) OR TS=("physiologic monitoring") OR TS=(torque) OR TS=("inertial sensor*") OR TS=(IMU) OR TS="inertial measurement" OR TS=("wearable sensor*") OR TS=("ultrasound-based") OR TS=("video-based") OR TS=("motion capture") OR TS=(gyroscop*) OR TS=("infrared-camera*") OR TS=("motion analysis") OR TS=("movement analysis") OR TS=("motion tracker*") OR TS=(magnetometer*) OR TS=("magnetic system") OR TS=(acceleration*) OR TS=("angular velocity") OR TS=(fluidity) *Indexes=SCI-EXPANDED, ESCI Timespan=1900-2017*
# 3	10,818,476	TS=(sensitiv*) OR TS=(specifici*) OR TS=(reproducib*) OR TS=("ROC curve") OR TS=(psychometric*) OR TS=(valid*) OR TS=(propert*) OR TS=(clinimetric*) OR TS=(metrolog*) OR TS=(reliab*) OR TS=(test-retest) OR TS=(responsiv*) OR TS=("minimal detectable change*") OR TS=(MDC) OR TS=("standard error of measurement*") OR TS=(SEM) OR TS=("minimal clinically important improvement*") OR TS=(MCII) OR TS=("minimal clinically important difference*") OR TS=(MCID) OR TS=("likelihood ratio*") OR TS=(detect*) OR TS=(intraclass) OR TS=(correl*) OR TS=(accura*) OR TS=(precis*) OR TS=(discern*) OR TS=(discrim*) OR TS=("floor effect*") OR TS=("ceiling effect*") *Indexes=SCI-EXPANDED, ESCI Timespan=1900-2017*
# 2	10,090,842	TS=(pain) OR TS=(impingement) OR TS=(bursitis) OR TS=(tendon*) OR TS=(tendin*) OR TS=(fracture*) OR TS=(dislocation*) OR TS=(instab*) OR TS=(arthritis) OR TS=(arthroplast*) OR TS=(injur*) OR TS=(patholog*) OR TS=(disease) OR TS=(trauma*) OR TS=(tear*) OR TS=(condition*) OR TS=(disorder*) OR TS=(repair*) OR TS=(surger*) OR TS=(surgical) OR TS=("adhesive capsulitis") OR TS=("frozen shoulder") OR TS=(dysfunction) OR TS=(prosthes*) *Indexes=SCI-EXPANDED, ESCI Timespan=1900-2017*
# 1	374,539	TS=(shoulder*) OR TS=(upper extremit*) OR TS=(upper limb*) OR TS=(arm*) OR TS=(glenohumeral) OR TS=(humer*) OR TS=(rotator cuff*) *Indexes=SCI-EXPANDED, ESCI Timespan=1900-2017*

PEDro Physiotherapy Evidence Database

This database does not allow complex combination of search terms. As the “advanced search” feature allow several thematic selections, the searches were limited to “clinical trials” and “upper arm, shoulder or shoulder girdle”. Then simple keywords combination were used to investigate the content of the database.

PROMs

Keywords	Results
shoulder DASH	38
« Disabilities of the Arm, Shoulder and Hand »	36
shoulder QuickDASH	4
shoulder Constant	117
shoulder ASES	7
« American Shoulder And Elbow Surgeons »	16
shoulder SST	12
simple shoulder test	18

None of the retrieved references addressed the measurement properties of shoulder function PROMs

Kinematic

Keywords	Results
shoulder kinematic	27
shoulder movement analysis	25
shoulder motion analysis	57
shoulder sensor*	20
shoulder inertia*	1
shoudler infrared	7
shoulder magnet*	22
shoulder kinematic*	27

None of the retrieved references addressed the measurement properties of shoulder function kinematic scores

Appendix XVI

References of selected articles

PROMs outcome measures

ANGST, F., GOLDHAHN, J., DRERUP, S., AESCHLIMANN, A., SCHWYZER, H. K. & SIMMEN, B. R. 2008. Responsiveness of six outcome assessment instruments in total shoulder arthroplasty. *Arthritis Rheum,* 59**,** 391-8.

ANGST, F., GOLDHAHN, J., DRERUP, S., FLURY, M., SCHWYZER, H. K. & SIMMEN, B. R. 2009. How sharp is the short QuickDASH? A refined content and validity analysis of the short form of the disabilities of the shoulder, arm and hand questionnaire in the strata of symptoms and function and specific joint conditions. *Qual Life Res,* 18**,** 1043-51.

ANGST, F., PAP, G., MANNION, A. F., HERREN, D. B., AESCHLIMANN, A., SCHWYZER, H. K. & SIMMEN, B. R. 2004. Comprehensive assessment of clinical outcome and quality of life after total shoulder arthroplasty: usefulness and validity of subjective outcome measures. *Arthritis Rheum,* 51**,** 819-28.

BASAR, S., GUNAYDIN, G., HAZAR KANIK, Z., SOZLU, U., ALKAN, Z. B., PALA, O. O., CITAKER, S. & KANATLI, U. 2017. Western Ontario Shoulder Instability Index: cross-cultural adaptation and validation of the Turkish version. *Rheumatology International***,** 1-7.

BEATON, D. E., VAN EERD, D., SMITH, P., VAN DER VELDE, G., CULLEN, K., KENNEDY, C. A. & HOGG-JOHNSON, S. 2011. Minimal change is sensitive, less specific to recovery: a diagnostic testing approach to interpretability. *Journal of Clinical Epidemiology,* 64**,** 487-496.

BEATON, D. E., WRIGHT, J. G. & KATZ, J. N. 2005. Development of the QuickDASH: comparison of three item-reduction approaches. *J Bone Joint Surg Am,* 87**,** 1038-46.

BEATON, D. R., ROBIN R. 1998. Assessing the reliability and responsiveness of 5 shoulder questionnaires. *Journal of Shoulder and Elbow Surgery,* 7**,** 565-572.

BECKMANN, J. T., HUNG, M., BOUNSANGA, J., WYLIE, J. D., GRANGER, E. K. & TASHJIAN, R. Z. 2015. Psychometric evaluation of the PROMIS Physical Function Computerized Adaptive Test in comparison to the American Shoulder and Elbow Surgeons score and Simple Shoulder Test in patients with rotator cuff disease. *J Shoulder Elbow Surg,* 24**,** 1961-7.

BLONNA, D., SCELSI, M., MARINI, E., BELLATO, E., TELLINI, A., ROSSI, R., BONASIA, D. E. & CASTOLDI, F. 2012. Can we improve the reliability of the Constant-Murley score? *J Shoulder Elbow Surg,* 21**,** 4-12.

CACCHIO, A., PAOLONI, M., GRIFFIN, S. H., ROSA, F., PROPERZI, G., PADUA, L., PADUA, R., CARNELLI, F., CALVISI, V. & SANTILLI, V. 2012. Cross-cultural adaptation and measurement properties of an Italian version of the Western Ontario Shoulder Instability Index (WOSI). *J Orthop Sports Phys Ther,* 42**,** 559-67.

CELIK, D. 2016. Turkish version of the modified Constant-Murley score and standardized test protocol: reliability and validity. *Acta Orthop Traumatol Turc,* 50**,** 69-75.

CELIK, D., ATALAR, A. C., DEMIRHAN, M. & DIRICAN, A. 2013. Translation, cultural

adaptation, validity and reliability of the Turkish ASES questionnaire. *Knee Surg Sports Traumatol Arthrosc,* 21**,** 2184-9.

CHRISTIANSEN, D. H. F., POUL; FALLA, DEBORAH; HAAHR, JENS PEDER; FRICH, LARS HENRIK; SVENDSEN SUSANNE WULFF 2015. Responsiveness and Minimal Clinically Important Change: A Comparison Between Two Shoulder Outcome Measures. *Journal of Orthopaedic & Sports Physical Therapy,* 45**,** 1-19.

CHRISTIE, A., DAGFINRUD, H., GARRATT, A. M., RINGEN OSNES, H. & HAGEN, K. B. 2011. Identification of shoulder-specific patient acceptable symptom state in patients with rheumatic diseases undergoing shoulder surgery. *J Hand Ther,* 24**,** 53-60; quiz 61.

CHRISTIE, A., HAGEN, K. B., MOWINCKEL, P. & DAGFINRUD, H. 2009. Methodological properties of six shoulder disability measures in patients with rheumatic diseases referred for shoulder surgery. *J Shoulder Elbow Surg,* 18**,** 89-95.

CONBOY, V. B., MORRIS, R. W., KISS, J. & CARR, A. J. 1996. An evaluation of the Constant-Murley shoulder assessment. *J Bone Joint Surg Br,* 78**,** 229-32.

COOK, K. F., RODDEY, T. S., GARTSMAN, G. M. & OLSON, S. L. 2003. Development and psychometric evaluation of the Flexilevel Scale of Shoulder Function. *Med Care,* 41**,** 823-35.

COOK, K. F., RODDEY, T. S., OLSON, S. L., GARTSMAN, G. M., VALENZUELA, F. F. & HANTEN, W. P. 2002. Reliability by surgical status of self-reported outcomes in patients who have shoulder pathologies. *J Orthop Sports Phys Ther,* methods

32**,** 336-46.

CORONA, K., CERCIELLO, S., MORRIS, B. J., VISONA, E., MEROLLA, G. & PORCELLINI, G. 2016. Cross-cultural adaptation and validation of the Italian version of the Western Ontario Osteoarthritis of the Shoulder index (WOOS). *J Orthop Traumatol*.

DAWSON, J., FITZPATRICK, R. & CARR, A. 1999. The assessment of shoulder instability. The development and validation of a questionnaire. *J Bone Joint Surg Br,* 81**,** 420-6.

DE WITTE, P. B., HENSELER, J. F., NAGELS, J., VLIET VLIELAND, T. P. & NELISSEN, R. G. 2012. The Western Ontario rotator cuff index in rotator cuff disease patients: a comprehensive reliability and responsiveness validation study. *Am J Sports Med,* 40**,** 1611-9.

DINIZ LOPES, A., CICONELLI, R. M., CARRERA, E. F., GRIFFIN, S., FALOPPA, F. & BALDY DOS REIS, F. 2009. Comparison of the responsiveness of the Brazilian version of the Western Ontario Rotator Cuff Index (WORC) with DASH, UCLA and SF-36 in patients with rotator cuff disorders. *Clin Exp Rheumatol,* 27**,** 758-64.

EBRAHIMZADEH, M. H., VAHEDI, E., BARADARAN, A., BIRJANDINEJAD, A., SEYYED-HOSEINIAN, S. H., BAGHERI, F. & KACHOOEI, A. R. 2016. Psychometric Properties of the Persian Version of the Simple Shoulder Test (SST) Questionnaire. *Arch Bone Jt Surg,* 4**,** 387-392.

FAYAD, F., LEFEVRE-COLAU, M. M., GAUTHERON, V., MACE, Y., FERMANIAN, J., MAYOUX-BENHAMOU, A., ROREN, A., RANNOU, F., ROBY-BRAMI, A., REVEL, M. & POIRAUDEAU, S. 2009. Reliability, validity and responsiveness of the French version of the questionnaire Quick Disability of the Arm, Shoulder and Hand in shoulder disorders. *Man Ther,* 14**,** 206-12.

FAYAD, F., LEFEVRE-COLAU, M. M., MACE, Y., FERMANIAN, J., MAYOUX-BENHAMOU, A., ROREN, A., RANNOU, F., ROBY-BRAMI, A., GAUTHERON, V., REVEL, M. & POIRAUDEAU, S. 2008. Validation of the French version of the Disability of the Arm, Shoulder and Hand questionnaire (F-DASH). *Joint Bone Spine,* 75**,** 195-200.

FAYAD, F., LEFEVRE-COLAU, M. M., MACE, Y., GAUTHERON, V., FERMANIAN, J., ROREN, A., ROBY-BRAMI, A., REVEL, M. & POIRAUDEAU, S. 2008. Responsiveness of the French version of the Disability of the Arm, Shoulder and Hand questionnaire (F-DASH) in patients with orthopaedic and medical shoulder disorders. *Joint Bone Spine,* 75**,** 579-84.

GAUDELLI, C., BALG, F., GODBOUT, V., PELET, S., DJAHANGIRI, A., GRIFFIN, S. & ROULEAU, D. M. 2014. Validity, reliability and responsiveness of the French language translation of the Western Ontario Shoulder Instability Index (WOSI). *Orthop Traumatol Surg Res,* 100**,** 99-103.

GE, Y., CHEN, S., CHEN, J., HUA, Y. & LI, Y. 2013. The development and evaluation of a new shoulder scoring system based on the view of patients and physicians: the Fudan University shoulder score. *Arthroscopy,* 29**,** 613-22.

GODFREY, J., HAMMAN, R., LOWENSTEIN, S., BRIGGS, K. & KOCHER, M. 2007. Reliability, validity, and responsiveness of the simple shoulder test: psychometric properties by age and injury type. *J Shoulder Elbow Surg,* 16**,** 260-7.

GOLDHAHN, J., ANGST, F., DRERUP, S., PAP, G., SIMMEN, B. R. & MANNION, A. F. 2008. Lessons learned during the cross-cultural adaptation of the American Shoulder and Elbow Surgeons shoulder form into German. *J Shoulder Elbow Surg,* 17**,** 248-54.

HALDORSEN, B., SVEGE, I., ROE, Y. & BERGLAND, A. 2014. Reliability and validity of the Norwegian version of the Disabilities of the Arm, Shoulder and Hand questionnaire in patients with shoulder impingement syndrome. *BMC Musculoskelet Disord,* 15**,** 78.

HATTA, T. S., N.; OMI, R.; SANO, H.; YAMAMOTO, N.; ANDO, A.; SUGAYA, H.; AIZAWA, T.; KURIYAMA, S.; ITOI, E. 2011. Reliability and validity of the Western Ontario Shoulder Instability Index (WOSI) in the Japanese population. *J Orthop Sci,* 16**,** 732-6.

HENSELER, J. F., KOLK, A., VAN DER ZWAAL, P., NAGELS, J., VLIET VLIELAND, T. P. & NELISSEN, R. G. 2015. The minimal detectable change of the Constant score in impingement, full-thickness tears, and massive rotator cuff tears. *J Shoulder Elbow Surg,* 24**,** 376-81.

HOFSTAETTER, J. G., HANSLIK-SCHNABEL, B., HOFSTAETTER, S. G., WURNIG, C. & HUBER, W. 2010. Cross-cultural adaptation and validation of the German version of the Western Ontario Shoulder Instability index. *Arch Orthop Trauma Surg,* 130**,** 787-96.

HOLMGREN, T., OBERG, B., ADOLFSSON, L., BJORNSSON HALLGREN, H. & JOHANSSON, K. 2014. Minimal important changes in the Constant-Murley score in patients with subacromial pain. *J Shoulder Elbow Surg,* 23**,** 1083-90.

HOLTBY, R. & RAZMJOU, H. 2005. Measurement properties of the Western Ontario rotator cuff outcome measure: a preliminary report. *J Shoulder Elbow Surg,* 14**,** 506-10.

HUNSAKER, F. G., CIOFFI, D. A., AMADIO, P. C., WRIGHT, J. G. & CAUGHLIN, B. 2002. The American academy of orthopaedic surgeons outcomes instruments: normative values from the general population. *J Bone Joint Surg Am,* 84-A**,** 208-15.

KATOLIK, L. R., AA; COLE, BJ; VERMA, NN; HAYDEN, JK; BACH, BR 2005. Normalization of the Constant score. *Journal of Shoulder and Elbow Surgery,* 14**,** 279-285.

KIRKLEY, A. G., S.; MCLINTOCK, H.; NG, L. 1998. The development and evaluation of a disease-specific quality of life measurement tool for shoulder instability. The Western Ontario Shoulder Instability Index (WOSI). *Am J Sports Med,* 26**,** 764-72.

KOCHER, M. S. H., MARILEE P.; BRIGGS, KAREN K.; RICHARDSON, TYLER R.; O'HOLLERAN, JAMES; HAWKINS, RICHARD J. 2005. Reliability, Validity, and Responsiveness of the American Shoulder and Elbow Surgeons Subjective Shoulder Scale in Patients with Shoulder Instability, Rotator Cuff Disease, and Glenohumeral Arthritis. *J Bone Joint Surg Am,* 87**,** 2006-2011.

KUKKONEN, J., KAUKO, T., VAHLBERG, T., JOUKAINEN, A. & AARIMAA, V. 2013. Investigating minimal clinically important difference for Constant score in patients undergoing rotator cuff surgery. *J Shoulder Elbow Surg,* 22**,** 1650-5.

LO, I. K., GRIFFIN, S. & KIRKLEY, A. 2001. The development of a disease-specific quality of life measurement tool for osteoarthritis of the shoulder: The Western Ontario Osteoarthritis of the Shoulder (WOOS) index. *Osteoarthritis Cartilage,* 9**,** 771-8.

LUNDQUIST, C. B., DOSSING, K. & CHRISTIANSEN, D. H. 2014. Responsiveness of a Danish version of the Disabilities of the Arm, Shoulder and Hand (DASH) questionnaire. *Dan Med J,* 61**,** A4813.

MACDERMID, J. C., DROSDOWECH, D. & FABER, K. 2006. Responsiveness of self-report scales in patients recovering from rotator cuff surgery. *J Shoulder Elbow Surg,* 15**,** 407-14.

MACDERMID, J. C., KHADILKAR, L., BIRMINGHAM, T. B. & ATHWAL, G. S. 2015. Validity of the QuickDASH in patients with shoulder-related disorders undergoing surgery. *J Orthop Sports Phys Ther,* 45**,** 25-36.

MAHABIER, K. C., DEN HARTOG, D., THEYSKENS, N., VERHOFSTAD, M. H. J., VAN LIESHOUT, E. M. M. & INVESTIGATORS, H. T. 2017. Reliability, validity, responsiveness, and minimal important change of the Disabilities of the Arm, Shoulder and Hand and Constant-Murley scores in patients with a humeral shaft fracture. *J Shoulder Elbow Surg,* 26**,** e1-e12.

MEHTA, S. P., TIRUTTANI, R., KAUR, M. N., MACDERMID, J. & KARIM, R. 2015.

Psychometric Properties of the Hindi Version of the Disabilities of Arm, Shoulder, and Hand: A Pilot Study. *Rehabil Res Pract,* 2015**,** 482378.

MEMBRILLA-MESA, M. D., TEJERO-FERNANDEZ, V., CUESTA-VARGAS, A. I. & ARROYO-MORALES, M. 2015. Validation and reliability of a Spanish version of Simple Shoulder Test (SST-Sp). *Qual Life Res,* 24**,** 411-6.

MICHENER, L. A., SNYDER VALIER, A. R. & MCCLURE, P. W. 2013. Defining substantial clinical benefit for patient-rated outcome tools for shoulder impingement syndrome. *Arch Phys Med Rehabil,* 94**,** 725-30.

MICHENER, L. A. M., P. W.; SENNETT, B. J. 2002. American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form, patient self-report section: reliability, validity, and responsiveness. *J Shoulder Elbow Surg,* 11**,** 587-94.

MINTKEN, P. E., GLYNN, P. & CLELAND, J. A. 2009. Psychometric properties of the shortened disabilities of the Arm, Shoulder, and Hand Questionnaire (QuickDASH) and Numeric Pain Rating Scale in patients with shoulder pain. *J Shoulder Elbow Surg,* 18**,** 920-6.

MOELLER, A. D., THORSEN, R. R., TORABI, T. P., BJOERKMAN, A. S., CHRISTENSEN, E. H., MARIBO, T. & CHRISTIANSEN, D. H. 2014. The Danish version of the modified Constant-Murley shoulder score: reliability, agreement, and construct validity. *J Orthop Sports Phys Ther,* 44**,** 336-40.

MOSER, A. D., KNAUT, L. A., ZOTZ, T. G. & SCHARAN, K. O. 2012. Validity and reliability of the Portuguese version of the American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form. *Rev Bras Reumatol,* 52**,** 348-56.

NAGHDI, S., NAKHOSTIN ANSARI, N., RUSTAIE, N., AKBARI, M., EBADI, S., SENOBARI, M. & HASSON, S. 2015. Simple shoulder test and Oxford Shoulder Score: Persian translation and cross-cultural validation. *Arch Orthop Trauma Surg,* 135**,** 1707-18.

NEGAHBAN, H., BEHTASH, Z., SOHANI, S. M. & SALEHI, R. 2015. Responsiveness of two Persian-versions of shoulder outcome measures following physiotherapy intervention in patients with shoulder disorders. *Disabil Rehabil,* 37**,** 2300-4.

NETO, J. O., GESSER, R. L., STEGLICH, V., BONILAURI FERREIRA, A. P., GANDHI, M., VISSOCI, J. R. & PIETROBON, R. 2013. Validation of the Simple Shoulder Test in a Portuguese-Brazilian population. Is the latent variable structure and validation of the Simple Shoulder Test Stable across cultures? *PLoS One,* 8**,** e62890.

O'CONNOR, D. A., CHIPCHASE, L. S., TOMLINSON, J. & KRISHNAN, J. 1999. Arthroscopic subacromial decompression: responsiveness of disease-specific and health-related quality of life outcome measures. *Arthroscopy,* 15**,** 836-40.

OFFENBACHER, M., EWERT, T., SANGHA, O. & STUCKI, G. 2003. Validation of a German version of the 'Disabilities of Arm, Shoulder and Hand' questionnaire (DASH-G). *Z Rheumatol,* 62**,** 168-77.

OH, J. H., JO, K. H., KIM, W. S., GONG, H. S., HAN, S. G. & KIM, Y. H. 2009. Comparative evaluation of the measurement properties of various shoulder outcome instruments. *Am*

J Sports Med, 37**,** 1161-8.

PIITULAINEN, K., PALONEVA, J., YLINEN, J., KAUTIAINEN, H. & HAKKINEN, A. 2014. Reliability and validity of the Finnish version of the American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form, patient self-report section. *BMC Musculoskelet Disord,* 15**,** 272.

RAZMJOU, H., BEAN, A., VAN OSNABRUGGE, V., MACDERMID, J. C. & HOLTBY, R. 2006. Cross-sectional and longitudinal construct validity of two rotator cuff disease-specific outcome measures. *BMC Musculoskelet Disord,* 7**,** 26.

ROBINS, R. J., ANDERSON, M. B., ZHANG, Y., PRESSON, A. P., BURKS, R. T. & GREIS, P. E. 2017. Convergent Validity of the Patient-Reported Outcomes Measurement Information System's Physical Function Computerized Adaptive Test for the Knee and Shoulder Injury Sports Medicine Patient Population. *Arthroscopy,* 33**,** 608-616.

ROCOURT, M. H., RADLINGER, L., KALBERER, F., SANAVI, S., SCHMID, N. S., LEUNIG, M. & HERTEL, R. 2008. Evaluation of intratester and intertester reliability of the Constant-Murley shoulder assessment. *J Shoulder Elbow Surg,* 17**,** 364-9.

RODDEY, T. S., OLSON, S. L., COOK, K. F., GARTSMAN, G. M. & HANTEN, W. 2000. Comparison of the University of California-Los Angeles Shoulder Scale and the Simple Shoulder Test with the shoulder pain and disability index: single-administration reliability and validity. *Phys Ther,* 80**,** 759-68.

ROY, J. S., MACDERMID, J. C., FABER, K. J., DROSDOWECH, D. S. & ATHWAL, G. S. 2010. The simple shoulder test is responsive in assessing change following shoulder arthroplasty. *J Orthop Sports Phys Ther,* 40**,** 413-21.

RYSSTAD, T., ROE, Y., HALDORSEN, B., SVEGE, I. & STRAND, L. I. 2017. Responsiveness and minimal important change of the Norwegian version of the Disabilities of the Arm, Shoulder and Hand questionnaire (DASH) in patients with subacromial pain syndrome. *BMC Musculoskelet Disord,* 18**,** 248.

SALLAY, P. I. & REED, L. 2003. The measurement of normative American Shoulder and Elbow Surgeons scores. *J Shoulder Elbow Surg,* 12**,** 622-7.

SALOMONSSON, B. A., S.; DALEN, N.; LILLKRONA, U. 2009. The Western Ontario Shoulder Instability Index (WOSI): validity, reliability, and responsiveness retested with a Swedish translation. *Acta Orthop,* 80**,** 233-8.

SCHMITT, J. S. & DI FABIO, R. P. 2004. Reliable change and minimum important difference (MID) proportions facilitated group responsiveness comparisons using individual threshold criteria. *J Clin Epidemiol,* 57**,** 1008-18.

SKARE, O., LIAVAAG, S., REIKERAS, O., MOWINCKEL, P. & BROX, J. I. 2013. Evaluation of Oxford instability shoulder score, Western Ontario shoulder instability index and Euroqol in patients with SLAP (superior labral anterior posterior) lesions or recurrent anterior dislocations of the shoulder. *BMC Res Notes,* 6**,** 273.

STAPLES, M. P., FORBES, A., GREEN, S. & BUCHBINDER, R. 2010. Shoulder-specific

disability measures showed acceptable construct validity and responsiveness. *J Clin Epidemiol,* 63**,** 163-70.

TASHJIAN, R. Z., DELOACH, J., GREEN, A., PORUCZNIK, C. A. & POWELL, A. P. 2010. Minimal Clinically Important Differences in ASES and Simple Shoulder Test Scores After Nonoperative Treatment of Rotator Cuff Disease. *J Bone Joint Surg Am,* 92**,** 296-303.

TORRENS, C., GUIRRO, P. & SANTANA, F. 2016. The minimal clinically important difference for function and strength in patients undergoing reverse shoulder arthroplasty. *J Shoulder Elbow Surg,* 25**,** 262-8.

VAN DE WATER, A. T., SHIELDS, N., DAVIDSON, M., EVANS, M. & TAYLOR, N. F. 2014. Reliability and validity of shoulder function outcome measures in people with a proximal humeral fracture. *Disabil Rehabil,* 36**,** 1072-9.

VAN DE WATER, A. T. M., DAVIDSON, M., SHIELDS, N., EVANS, M. C. & TAYLOR, N. F. 2016. The Shoulder Function Index (SFInX): evaluation of its measurement properties in people recovering from a proximal humeral fracture. *BMC Musculoskeletal Disorders,* 17**,** 295.

VAN DER LINDE, J. A., VAN KAMPEN, D. A., VAN BEERS, L., VAN DEURZEN, D. F. P., SARIS, D. B. F. & TERWEE, C. B. 2017. The Responsiveness and Minimal Important Change of the Western Ontario Shoulder Instability Index and Oxford Shoulder Instability Score. *J Orthop Sports Phys Ther,* 47**,** 402-410.

VAN DER LINDE, J. A., WILLEMS, W. J., VAN KAMPEN, D. A., VAN BEERS, L. W., VAN DEURZEN, D. F. & TERWEE, C. B. 2014. Measurement properties of the Western Ontario Shoulder Instability index in Dutch patients with shoulder instability. *BMC Musculoskelet Disord,* 15**,** 211.

VAN KAMPEN, D. A., VAN BEERS, L. W., SCHOLTES, V. A., TERWEE, C. B. & WILLEMS, W. J. 2012. Validation of the Dutch version of the Simple Shoulder Test. *J Shoulder Elbow Surg,* 21**,** 808-14.

VAN KAMPEN, D. A., WILLEMS, W. J., VAN BEERS, L. W., CASTELEIN, R. M., SCHOLTES, V. A. & TERWEE, C. B. 2013. Determination and comparison of the smallest detectable change (SDC) and the minimal important change (MIC) of four-shoulder patient-reported outcome measures (PROMs). *J Orthop Surg Res,* 8**,** 40.

VROTSOU, K., CUELLAR, R., SILIO, F., RODRIGUEZ, M. A., GARAY, D., BUSTO, G., TRANCHO, Z. & ESCOBAR, A. 2016. Patient self-report section of the ASES questionnaire: a Spanish validation study using classical test theory and the Rasch model. *Health Qual Life Outcomes,* 14**,** 147.

WERNER, B. C., CHANG, B., NGUYEN, J. T., DINES, D. M. & GULOTTA, L. V. 2016. What Change in American Shoulder and Elbow Surgeons Score Represents a Clinically Important Change After Shoulder Arthroplasty? *Clin Orthop Relat Res,* 474**,** 2672-2681.

WIERTSEMA, S. H. D. W., P. B.; RIETBERG, M. B.; HEKMAN, K. M.; SCHOTHORST, M.; STEULTJENS, M. P.; DEKKER, J. 2014. Measurement properties of the Dutch version of the Western Ontario Shoulder Instability Index (WOSI). *J Orthop Sci,* 19**,** 242-9.

YAHIA, A., GUERMAZI, M., KHMEKHEM, M., GHROUBI, S., AYEDI, K. & ELLEUCH, M. H. 2011. Translation into Arabic and validation of the ASES index in assessment of shoulder disabilities. *Annals of Physical and Rehabilitation Medicine,* 54**,** 59-72.

YIAN, E. H., RAMAPPA, A. J., ARNEBERG, O. & GERBER, C. 2005. The Constant score in normal shoulders. *J Shoulder Elbow Surg,* 14**,** 128-33.

YUGUERO, M., HUGUET, J., GRIFFIN, S., SIRVENT, E., MARCANO, F., BALAGUER, M. & TORNER, P. 2016. Transcultural adaptation, validation and assessment of the psychometric properties of the spanish version of the Western Ontario Shoulder Instability Index questionnaire. *Rev Esp Cir Ortop Traumatol,* 60**,** 335-345.

MAB outcome measures

DUC, C., PICHONNAZ, C., BASSIN, J. P., FARRON, A., JOLLES, B. M. & AMINIAN, K. 2014. Evaluation of muscular activity duration in shoulders with rotator cuff tears using inertial sensors and electromyography. *Physiol Meas,* 35**,** 2389-400.

JOLLES, B. M., DUC, C., COLEY, B., AMINIAN, K., PICHONNAZ, C., BASSIN, J. P. & FARRON, A. 2011. Objective evaluation of shoulder function using body-fixed sensors: a new way to detect early treatment failures? *J Shoulder Elbow Surg,* 20**,** 1074-81.

KORVER, R. J., HEYLIGERS, I. C., SAMIJO, S. K. & GRIMM, B. 2014. Inertia based functional scoring of the shoulder in clinical practice. *Physiol Meas,* 35**,** 167-76.

KORVER, R. J., SENDEN, R., HEYLIGERS, I. C. & GRIMM, B. 2014. Objective outcome evaluation using inertial sensors in subacromial impingement syndrome: a five-year follow-up study. *Physiol Meas,* 35**,** 677-86.

PICHONNAZ, C., AMINIAN, K., ANCEY, C., JACCARD, H., LECUREUX, E., DUC, C., FARRON, A., JOLLES, B. M. & GLEESON, N. 2017. Heightened clinical utility of smartphone versus body-worn inertial system for shoulder function B-B score. *PLoS One,* 12**,** e0174365.

PICHONNAZ, C., DUC, C., GLEESON, N., ANCEY, C., JACCARD, H., LECUREUX, E., FARRON, A., JOLLES, B. M. & AMINIAN, K. 2015. Measurement properties of the smartphone-based B-B Score in current shoulder pathologies. *Sensors (Basel),* 15**,** 26801-17.

PICHONNAZ, C., DUC, C., JOLLES, B. M., AMINIAN, K., BASSIN, J. P. & FARRON, A. 2015. Alteration and recovery of arm usage in daily activities after rotator cuff surgery. *J Shoulder Elbow Surg,* 24**,** 1346-52.

PICHONNAZ, C., LECUREUX, E., BASSIN, J. P., DUC, C., FARRON, A., AMINIAN, K., JOLLES, B. M. & GLEESON, N. 2015. Enhancing clinically-relevant shoulder function assessment using only essential movements. *Physiol Meas,* 36**,** 547-60.

YANG, J. L., LIN, J. J., HUANG, H. Y., HUANG, T. S. & CHAO, Y. W. 2014. Shoulder physical activity, functional disability and task difficulties in patients with stiff shoulders: interpretation from RT3 accelerator. *Man Ther,* 19**,** 349-54.