MPhil: Orthopaedic Biology

Title:

Validation of a novel Patient Reported Outcome Measure for Patients with Lower Limb Osteoarthritis: Reliability, Construct Validity and Responsiveness

By:

Pg Noor Azmi Pg Dr Hj Mohammad

Student ID No: 201285545

Institute of Translational Medicine

University of Liverpool, U.K.

September 2023

Contents:

Table of Contents

Tables of Results

Tables of figures

Appendix

Appendix A

New PROM – MP20

NEW Pilot Questionnaire

Patient Name:		Date of Birth:	Today's Date:

For the following items please select the response that best describes your **level of function** on average over the **last month** *(For each item please tick one box per row)*

		Able without problems	Able but a little difficult	Able but moderately difficult	Able but very difficult	Unable
When I Need to:						
01	Stand up from a chair	☐	☐	☐	☐	☐
02	Put on footwear	☐	☐	☐	☐	☐
03	Get in and out of a car	☐	☐	☐	☐	☐
04	Walk for 10 minutes	☐	☐	☐	☐	☐
05	Go up a flight of stairs	☐	☐	☐	☐	☐
06	Go down a flight of stairs	☐	☐	☐	☐	☐
07	Carry things (e.g. shopping bag)	☐	☐	☐	☐	☐
08	Do up buttons	☐	☐	☐	☐	☐
09	Reach out for something at shoulder height	☐	☐	☐	☐	☐
10	Turn a key	☐	☐	☐	☐	☐
11	Prepare a meal	☐	☐	☐	☐	☐
12	Do my regular job or daily routine if retired	☐	☐	☐	☐	☐
13	Perform leisure or sporting activities (e.g. Dancing, bowling, gardening)	☐	☐	☐	☐	☐
14	Do Housework	☐	☐	☐	☐	☐
15	Go shopping on my own	☐	☐	☐	☐	☐

Please select the response that best describes your **level of pain while resting**

		No pain	Little pain	Moderate pain	Severe pain	Constant pain
16	Level of pain	☐	☐	☐	☐	☐

Please select how **pain** from your joints limits your **overall function**

		No Limitation	Little limitation	Moderate limitation	A lot of limitation	Completely limited
When I Need to:						
17	Use my legs	☐	☐	☐	☐	☐
18	Use my arms	☐	☐	☐	☐	☐
		No Limitation	**Little limitation**	**Moderate limitation**	**A lot of limitation**	**Completely limited**
19	How does your **general medical health (e.g. asthma)** limit your overall function?	☐	☐	☐	☐	☐
20	How does your **mood (e.g. anxiety, depression)** limit your overall function?	☐	☐	☐	☐	☐

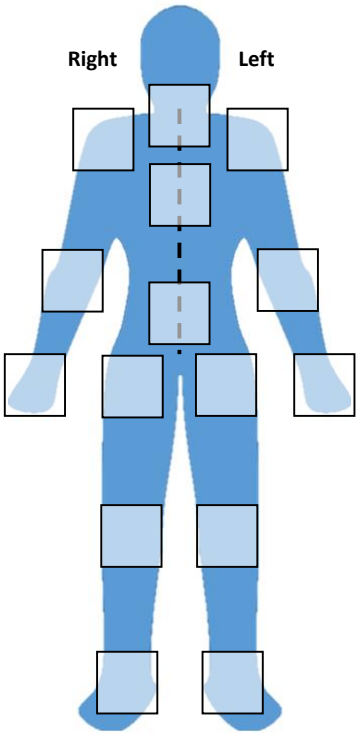
Appendix B:

Body Map of Pain (BMP) andVisual Analgoue Scale (VAS)

NEW Pilot Questionnaire

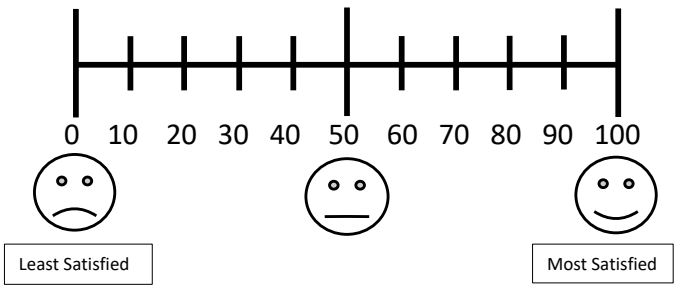
Question 21

Using a scale of 0 to 10 *(0 – no pain, 10 – worse pain),* rate the **SEVERITY of pain in your joints** by filling in the **boxes** on the picture below. **Leave unaffected joints blank**. (E.g. left knee 8, right shoulder 4)



Question 22

Using the following scale please mark with an '**X**' on the **scale line** to show how **satisfied** you are with your **overall level of function** on average over the **last month**. (0 – Least satisfied, 100 – Most satisfied)



1 Introduction:

1.1 Overview

This project is dedicated to the late Professor Simon Frostick, who originally coined the idea of developing a Holistic Patient-Reported Outcome Measuring Tool, which will allow one to measure an area or domain of interest and provide an overall view of one's physical functional status. Patient-Reported Outcome (PRO) in itself can be defined as "any report of the status of patient's health condition that comes directly from a patient, without interpretation of the patient's response by a clinician or anyone else" [1] . Therefore, patient-Reported Outcome Measures (PROMs) is a measurement tool that can independently measure a PRO. Hence PROMs are essentially a set of questions (items) that reflects a patient's chosen level of health-related quality of life. It can be very focused, like an Oxford Knee Questionnaire (OKS) that assess knee function for arthritic knee, or it can be very generic like Short Form Questionnaire (SF-12), which looks at one's quality of life in general. The purpose of this study is to validate a novel PROM that measures the Overall Physical Function of a patient with Lower Limb Osteoarthritis.

Being an Orthopaedic Surgeon, we often focus on our specialised area of interest, e.g. the knee or hip, that we can sometimes forget to view a patient 'holistically'. Therefore we thought of developing a tool that could assess the painful joint of interest and other important joints of the body as well as concomitant factors that contribute towards the overall physical function. These factors include mental and psychosocial being. Identifying these problems will significantly improve the way we customise, prioritise and deliver care. The groundwork for this project began with identifying the conceptual framework and creating a list of items (questions) through an iterative process of interviewing patients as well as health professionals before finalising the content and devising a measuring method. As a result, we developed a novel Patient Reported Outcome Measure (PROM) designed for patients with Lower Limb Osteoarthritis (OA). In HY's thesis [2], he described phase I involving primarily the development phase of a new Patient-Reported Outcome (PRO) tool. This work defined the conceptual framework and identified the health domains required for the concept. It went on to analytically quantify the items generated and also field-tested the Pilot version. The

next phase of the project is to explore the psychometric properties of this new PROM using standardised approaches that are widely accepted. This new PROM will attempt to provide a Global Functional Assessment of patients with Lower Limb OA, using just a single tool. This study revolved around field-testing this new PROM on a sample population of patients with end-stage Hip and Knee OA. We explored the three central tenets for validating a PROM, i.e. reliability, validity and responsiveness and adopted the well-established Classical Test Theory method [3] as our theoretical framework.

1.2 Background

Osteoarthritis is a condition that affects the joints, causing pain and stiffness. It is by far the most common form of joint disease, affecting people worldwide and at least 8 million people in the UK. Considering that age, obesity, and joint injury are among the most significant risk factors for OA and increasing in the population, it is expected to rise substantially with time. The socio-economic burden of OA is well documented in terms of healthcare expenditure and lost productivity [4]. The most prevalent groups of people affected by OA are mainly elderly, although not exclusively, and a large proportion of them present with multiple medical problems and pain in more than one joint. Most hip and knee OA patients usually present to an Orthopaedic surgeon in the UK after being referred from their General Practitioner. Often their symptoms are specific to one joint; however, it is not uncommon to have complex patients with multiple joint and medical problems. These complex cases pose a challenge from assessing the patient's symptomatology, their limitations in day to day physical activity and the overall impact on their quality of life. Consequently, these factors influence the decisions to intervene and the expected outcomes of the health intervention, whether the management is conservative or involves surgery.

The question is, really, how do we measure the value of the quality of care delivered to these patients that require them. As we move away from a volume-driven health care service to rewarding a value in health care delivery, PROMs are becoming central in providing better evidence that the treatment provided has delivered what it was intended to. E.g. patients undergoing TKR surgery to relieve pain and improve physical function, but to demonstrate

value, the orthopaedic surgeon must assess the result by measuring the degree of pain relief and physical function the patient experiences after surgery [5] .

The rationale for a new 'Holistic Outcome Measure.'

It is routine practice for a patient with signs and symptoms of an osteoarthritic knee to be referred to an Orthopaedic surgeon by their General Practitioners, and often they have had some form of assessment in the form of PROMs [6]. For example, a patient referred and accompanied by a low scoring Oxford Knee Score indicates a deficient functioning knee, and a degenerative radiograph that same patient will be put on the waiting list for a Total Knee Arthroplasty. However, it gets a bit complicated if that same patient were also to have pain in other joints along with a complex medical problem. The assessment becomes more complicated when you consider the state of general mental health, which we know impacts the overall outcome of the patient. Co-morbidities like Chronic Obstructive Pulmonary Disease, Diabetes, Back pain, Rheumatoid Arthritis are just some factors that play a vital role in deciding how a patient's outcome will fare following surgical intervention like joint arthroplasty. We know that current outcome measures widely used now like the Oxford Hip and Knee Scores are very good at measuring specific joint outcome measure, however, is not meant to give the overall picture. Because we already have a tool that could measure that, like the SF-12 and Euroqol EQ5D. We would like to have an outcome measure with both components, the joint-specific and the 'overall' picture. Even better if the measure for scores can be isolated to give just the combined scores of interest or combined to give the overall measure.

We can imagine this tool can be beneficial in a busy clinic, where we are seeing 20 – 25 patients with knees and hip OA. Each patient has a varying degree of medical and joint problems, and in addition, each has different levels of physical functioning. It is a real challenge to gather all the vitally important information, the physical symptoms, the limitations in daily activities, other joints affected to come up with the best treatment plan for the patient. Because we (the surgeons) are so focused on our specialised field (e.g. hip or knees), we sometimes forget to acquire information about other joints or limitations of day-to-day activities due to other medical problems. Consequently, we provide treatment for that

specific, let us say hip or knee, and hope that it may somehow have the desired improved outcome. Nevertheless, we all have cases where this expected, or some may say 'outliers' may not end up with the expected outcome.

An example would be a 70-year-old female patient who came to the clinic with severe right knee OA and a poor Oxford Knee Score(OKS) (a standard joint-specific PROM used in Orthopaedic practice). Matched with clinical findings consistent right knee OA, a Total Knee Arthroplasty is offered, and six months following surgery, a clinical review reveals that her outcome is no better. Why? Because the patient has severe Right shoulder pain from a previous injury which is still bothering her much, and back pain from degenerative spinal stenosis, which affects her use of walking aid and mobility. However, surprisingly her OKS, showed that she was doing well, giving the clinician an impression of successful operation (which it probably was) from the knee's perspective but an SF-12 Score (a generic quality of life questionnaire) which is not easily interpreted in the clinic by most clinicians reveals otherwise. The patient leaves the clinic convinced her knee is a success; however, her overall functional status remains poor. The main objective of this project was to minimise the gap between surgeon's and patient's expectations of the outcome of surgery by providing a tool that could give a surgeon a 'snapshot' of the patient's overall functional status with particular attention to other joints affected. We hoped to help surgeons/clinicians focus on the other aspects of the patient's health that could equally impact the overall quality of life and play a role in managing it. Furthermore, when we approach patients issues from a 'global perspective rather than in a 'robotic' manner, we can offer a more holistic package which patients will appreciate. We feel this is something that is still missing with our current plethora of outcome measures.

When this project was undertaken, there was no established measure of Global Functional Assessment designed for patients with Lower Limb Osteoarthritis. In the last decade, we have seen how Patient Reported Outcome Measure (PROM) has virtually revolutionised the assessment of patient's outcomes putting patient's views central to the process. The success of Orthopaedic PROMs like Oxford Hip and Knee Score (OHS/ONS), Harris Hip Scores (HHS), Knee Injury and Outcome Score (KOOS) in evaluating patient's outcomes following joint

arthroplasty has made PROMs a vital assessment tool for clinicians and 'health authorities' to measure the performances of the provision of their service. This evidence has driven the NHS to implement the PROMs project nationwide in 2009 and made it compulsory to collect PROMs data for four major surgical interventions, Total Hip Arthroplasty, Total Knee Arthroplasty, Hernia operation and Varicose Vein Surgery [7]. In Orthopaedics Surgery, however, these established PROMs we have mentioned are disease and joint-specific. Though very good at detecting changes for the specific disease or joint affected, it does not provide the clinician with the overall picture of its health status. This is important because the target population in this patient age group are elderly, and it is common to have multiple medical and joint problems. The Western Ontario and McMaster OA Index (WOMAC) comes close to assessing 'an overall function' of patients with OA [8]; however, it still focuses on only lower limb symptoms and lacks general questions such as quality of life-type items. In addition, despite the long-established presence of WOMAC, there remains confusion amongst users of how to use this tool correctly. Most orthopaedic practices overcome this issue by giving patients battery forms that contain disease-specific and generic quality of life measures, and a typical combination would be an Oxford Hip Scores with a Short-form Health Measure (SF-12). This approach of effectively giving a battery of questionnaires increases the burden to the patients and may subject the questionnaires themselves to biases, either through lack of compliance or concentration.

Hence, one of this project's goals is to develop a tool that can close the gap between these different types of PROMs (Disease-specific vs Generic) and accurately assess the Overall Functional Status of patients with OA by using patients with Lower Limb OA as a model. It is undoubtedly a challenging task, and within the development phase, it involves integrating both Joint-specific and Disease-specific items/domains to align with the intended conceptual framework to produce a more 'generalised' Disease-specific assessment tool. It requires going back to the drawing board and interviewing the study group patients to gain insight into the factors that affect their functional capacity, Activity of Daily Living (ADL) and social integration. This process later formed the Qualitative aspect of the project [2], resulting in a new concept - Holistic Functional Outcome measure, from an Orthopaedic perspective. It combines the assessment of significant joints (Upper Limb, Axial and Lower limb) and the

contribution of other factors that implies the overall outcome of a patient, e.g. General Health and Depression. We developed this new PROM specifically for patients with Lower Limb OA. We will also explore the psychometric strength and weaknesses of current available PROMs used in Orthopaedics for Lower Limb OA and compare this with the new PRO measure.

The authors hope that this novel PROM would provide surgeons with a much more meaningful overview of patients' overall functional status with OA and help both patient and clinician make better joint informed decisions.

Psychometric Evaluation: Reliability, Validity and Responsiveness

This thesis focuses on analysing the psychometric properties of a new PROM using established and widely accepted methods. The ability of a PROM to improve decision making relies on the psychometric strength of an instrument to capture the burden of disease or treatment. Reliability, Validity and Responsiveness are essential attributes to be demonstrated before using any PROM with confidence. However, the most crucial aspect is that we must understand is that evidence of reliability, validity and responsiveness falls on a continuous scale of no evaluation to a complete evaluation. Thus reliability, validity and responsiveness are continuous scales and are not dichotomous psychometric indices. So, to say an instrument is entirely reliable or valid is incorrect but more accurately described that the instrument has demonstrated strong evidence of reliability and validity. The more evidence there is that the instrument (PROM) measures the construct it is supposed to measure, the more confidence one has in it.

Reliability, validity and responsiveness are separate psychometric entities; however, some argue that responsiveness is an aspect of validity rather than a separate entity [9]. By definition, an instrument that is not reliable (lack internal consistency, test-retest) cannot be valid, and likewise, a reliable instrument may not necessarily measure what it is supposed to measure, i.e. not valid. So, one could surmise that reliability is the first prerequisite to validating a PRO measure. The third entity, responsiveness, measures the change in the burden of disease or treatment following an intervention.

1.3 Patient Reported Outcomes (PRO) and Patient Reported Outcome Measures (PROMs)

The definition of Patient Reported Outcomes (PRO) is that 'an assessment of any aspect of a person's health status that comes from the person directly, without interpretation from any other person. It should reflect the actual health state or performance' [1] [10], and Patient Reported Outcome Measures (PROMs) are primarily a measuring tool designed to collect PRO. Over the last decade, a shift towards a patient-centred approach has encouraged clinicians to be more balanced towards understanding the disease process and patient's perception of good health. Clinicians are concerned with the efficacy of treatment provided and have a vested interest in the quality of life of the patient. One way of collecting this information is using a specifically designed self-administered set of questionnaires we call patient-reported Outcome Measure (PROM) to answer a topic or question you have in mind (Health concept). More specifically, Patient-reported outcomes (PROs) are standardized measures directly reported by the patient that characterize the patient's perception of the impact of disease and treatment on health and functioning.

In contrast, patient-reported outcome measures (PROMs) are the tools used to measure patient-reported outcomes. PROs provide information that would otherwise be difficult to quantify, such as in the cases of symptom burden, social participation, and pain. PROs offer several advantages. For example, PROs are usually more feasible to implement and associated with lower costs since less health professional time is required and no specific training usually needed for them to be implemented. Finally, PROs respect the values and priorities of patients. Ultimately, most people seek treatment because of functional disability, pain, fatigue, or restrictions in social participation, which provides a solid rationale to systematically monitor these outcomes besides traditional clinical outcome measures.

Why is PROM important in Orthopaedic Surgery?

The most common surgical treatment for end-stage knee or Hip osteoarthritis is Joint Arthroplasty. Traditionally evaluation of healthcare intervention in Orthopaedic Surgery is focused on the efficiency of surgery and safety [11]. Conventional efficacy measures such as

radiographs, blood tests and revision rates are used to measure the outcomes from the surgery. However, PROMs are different in that it captures the patient's perspective on the impact of the disease and directly reports the efficacy of the treatment. They are exceptionally informative when treatment provides a satisfactory traditional clinical efficacy and safety measure, for example, excellent well aligned and balanced radiographs following Total Knee Replacement (TKR), however when asked from the patient's perspective, they offer minimal benefit. Hence PROMs can measure this vital trade-off.

Because of OA's progressive, degenerative nature, patient-reported outcome measures (PROMs) play an essential role in monitoring the course of the disease over time and the effectiveness of treatment. This is particularly the case for younger adults with OA, as the goal of management is to minimize symptoms, maximize function, and prolong the time until joint replacement surgery is required. Therefore, clinicians should use PROMs to capture the natural course of the disease, from early or mild OA to severe end-stage joint disease, and joint replacement and beyond. PROMs can help identify whether nonsurgical interventions effectively manage symptoms and may provide guidance when deciding whether a patient with OA is suitable for total joint replacement. So, for orthopaedic surgeons, PROMs are essential because:

- It is essential for research
- It can measure treatment benefits
- It can assess the efficiency, effectiveness and comparative safety of an intervention
- It can provide objective justifications for allocations of resources

1.4 Process of Developing a Patient-Reported Outcome Measure

Developing a new outcome measure involves several steps, summarized in figure 1.1 below (1). It is an iterative process that involves qualitative research and quantitative analysis where the outcome measure will undergo several iterations until the criteria for good PROMs development have been fulfilled. It can be a lengthy process and usually requires 3-5 years of development before it can be finalized for clinical use.

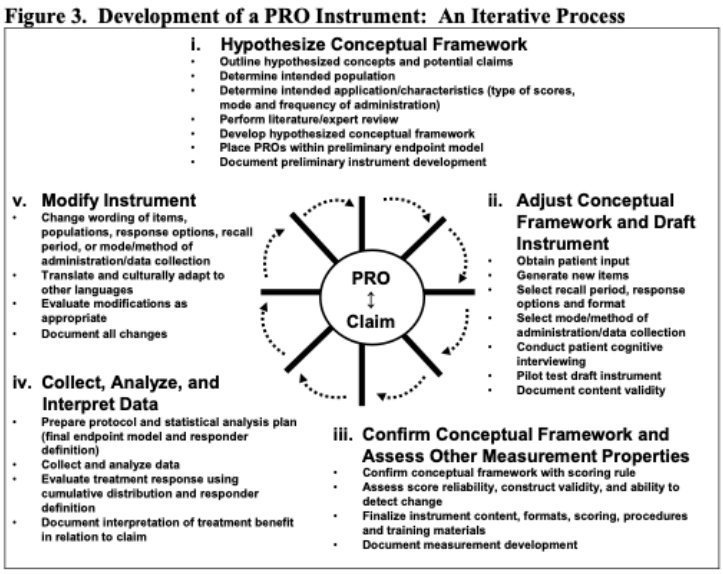


Figure 1.1 Development of a PRO Instrument (FDA recommendation)

COSMIN group, which is an initiative of an international multidisciplinary team of researchers with a background in epidemiology, psychometrics, medicine, qualitative research, and health care, who have expertise in the development and evaluation of outcome measurement instruments, have set out guidelines for outcome research methodology and has been widely accepted. A rigorous international Delphi process [12], has produced a checklist for researchers to follow in designing outcome measures or reviewing an instrument. The consensus includes the following measurement properties: internal consistency, reliability, measurement error, content validity, construct validity, responsiveness and interpretability.

Figure 1.2 Cosmin Taxonomy of relationship measurement properties

In developing an ideal Patient-Reported Outcome Measure (PROM), the most crucial factor that you have to establish is your Conceptual Framework, i.e., what are you trying to measure?

Phase I: Conceptual Framework & Item Generation

This is essentially forming your Conceptual Framework. Identify whether the 'thing' you are trying to measure is unidimensional, e.g. Lower Limb physical function, and one of the questions could be "how far you can walk in 5 mins?" Alternatively, do you wish to measure not just ability but pain and also a limitation? You may also want to consider asking about patient's satisfaction as well. You can imagine the more information you want, the more questions you will ask, the likely your PROMs will be complicated and lengthy. This increases the chances of errors and also poor compliance from the patients. Getting the balance right is vital in the planning stages, and hence the importance of the initial qualitative phase of the project cannot be underestimated.

In this project, we wanted to evaluate a patient's Global Functional Status, and hence our questions are streamlined towards physical function ability and limitations and how it affects

their overall function. Hence our conceptual Framework revolves around Overall Physical Function. I will refer you to the comprehensive qualitative work describing 'Development of a Novel PROM to assess overall physical function in a patient with Lower Limb OA: Conceptual Framework and Item Generation" (2). As a result of this work, we now have a new PROM measuring holistic function in a patient with Lower Limb Osteoarthritis.

1.5 Psychometric Analysis – Phase II

The psychometric analysis of this novel PROM forms the main bulk of this thesis to explore its measurement properties, and in this project, we used the widely accepted Classical Test Theory [13] which was ideal for the size and level of our study.

1.5.1 Classical Test Theory:

Classical test theory is a traditional quantitative approach to testing the reliability and validity of a scale based on its items. Classical test theory, also known as true score theory, assumes that each person has a true score, T, that would be obtained if there were no errors in measurement.

True scores quantify values on an attribute of interest, defined here as the underlying concept, construct, trait, or the ability of interest (the "thing" intended to be measured). As values of the true score increase, responses to items representing the same concept should also increase (i.e., there should be a monotonically increasing relationship between true scores and item scores), assuming that item responses are coded so that higher responses reflect more of the concept.

1.5.2 Item Response Theory:

This is a more modern technique, and it is a theory of testing based on the relationship between individuals' performances on a test item and the test takers' levels of performance on an overall measure of the ability that item was designed to measure [14]. Unlike simpler alternatives for creating scales and evaluating questionnaire responses, it does not assume that each item is equally difficult. This distinguishes IRT from, for instance, Likert scaling, in

which "All items are assumed to be replications of each other, or in other words items are considered to be parallel instruments". By contrast, item response theory treats the difficulty of each item as information to be incorporated in scaling items. IRT proponents claim that this method is superior to CTT, and to a certain extent, it is probably true [6]. However, the main disadvantage of IRT is that it is complex, and the prerequisite to using this method is a relatively large sample of patients to calibrate item parameters using specialised software, which is beyond the constraints of this relatively small study.

Hence in this study, we will adopt the CTT method for obvious reasons as it is more pragmatic, cost-effective and straightforward. We will also adopt the widely accepted COSMIN group recommendations to analyse PRO Measures.

Once the Pilot version of the new PROM has been finalised, it is now ready for the first phase of the validation process, which involves exploring the three central tenets, Reliability, Validity and Responsiveness.

1.5.3 Reliability

Assessment of reliability consists of determining that a scale or measurement yields reproducible and consistent results. There are two different levels of scale validation. First, reliability is used as a term to describe aspects of repeatability and stability of measurements. Any measurement or summary score, whether based upon a single item or multiple items, should yield reproducible or consistent values if used repeatedly on the same patient while the patient's condition is not known to change. Thus reliability, in this sense of repeatability, describes the differences between multiple measurements.

Secondly, for scales containing multiple items, all the items should be consistent because they should all measure the same thing. This is true if the scale is unidimensional. If, however, in multidimensional outcome measures like the WOMAC scale, the items should correlate with the sub-domain intended to measure. This form of reliability, called internal reliability or internal consistency, uses item correlations to assess the homogeneity of multi-item scales

and is, in many senses, a form of validity. From a statistical perspective, reliability is similar to variance because an unreliable measure varies between measurement occasions.

Both forms of reliability are assessed using correlation techniques. Thus, repeatability reliability is based upon the analysis of correlations between repeated measurements, where the measurements are repeated over time (test-retest reliability). Internal reliability, also often called internal consistency, is based on item-to-item and item-to-scale correlations in multi-item scales. Since these two concepts are mathematically related, estimates of the internal reliability of multi-item scales can often be used to predict the approximate value of their repeatability reliability.

Several different measures have been proposed; however, we will use the Intraclass Coefficient Correlation (ICC) in our study. Since reliability is the extent to which repeated measurements will give the same results when the true scale score remains constant, measurement concerns the level of agreement between the two scores. If a patient is in a stable condition, an instrument should yield repeatable and reproducible results if used repeatedly. This is the basis of a test-retest study, with patients who are thought to have stable disease and who are not expected to experience changes due to treatment effects. In this study, the patients are asked to complete the same questionnaires on two occasions within two weeks. The level of agreement between the occasions can be measured using the ICC, providing a measure of the instrument's reliability. In this study, we will use widely accepted standards for ICC where ICC > 0.9 are indicative of excellent reliability [15].

Internal Reliability

Also known as Internal consistency. It measures the extent to which items in the sub-scales are homogenous, thus measuring the same concept. We will be using Cronbach's alpha coefficient to determine the level of agreeability with the items and the domains it belongs to and its relationship towards the whole scale. For the domains, the target Cronbach's alpha > 0.8, although several studies have mentioned that a value of > 0.5 is adequate.

1.5.4 Validity

A measure is only as good if it measures what it is intended to measure, and there three different types of Validity, Content Validity, Criterion Validity, and Construct Validity.

Content Validity

This is the extent to which the PROMs measure the appropriate content and represents the different attributes that the construct is supposed to represent. This is usually done at the beginning of the development phase, where through focus groups and cognitive patients, the items representing the construct and sub-domains are identified and further endorsed by a group of experts where gaps are filled in. This is a vital process in the development of the PROM.

Criterion Validity

Refers to how well the tool measures to an external standard. Because when developing a PROM, typically we have no standard, hence criterion validity is usually not applicable. The measure is compared to a standard measure in a situation where it is applicable, and the consistency between the two is evaluated.

Construct Validity

Measures the extent to which the scores of an instrument scale or subscale are an adequate reflection of the dimensionality of the construct to be measured. In other words, how good does it measure what it is supposed to measure? The evaluation includes the degree to which a measure correlates to other measures that it is similar to but does not correlate to measures that it is not similar to. We will be using extensive correlation analysis of the subscales / overall scale with other PROMs scale/subscale, using Multitrait-Multimethod analysis (MTMM) as described by Fayers et al. The correlation between each item and the scale it belongs to will also be analysed using the Multittrait-Multi-Item method (MTMI). This forms the concept of Convergent and Divergent Validity. Correlation coefficient $r > 0.50$ were considered as indicator of convergent Validity, and correlations $r < 0.35$ as an indicator of divergent Validity. The hypothesis in this study is that the Physical function domains should correlate well with established WOMAC-Physical Function, Oxford Hip and Knee Scores and the SF-12 Physical Component Scores. In theory, the Pain Domain in the new PROM should correlate with the WOMAC pain scale, and the General Health domain (New PROM) should deliver a positive correlation with the SF-12 Mental Component Score.

1.5.5 Responsiveness

Responsiveness measures the ability of the instrument to detect change [16]. This can be calculated by the change between pre-operative and 6-month follow-up time points and reported as Effect Size for the mean change in terms of Cohen's , as well as Standardized Response Means (SRM).

1.6 Common Orthopaedic Patient Reported Outcome Measures

UK has long led the use of PRO in Orthopaedic Surgery, and in almost all Elective Orthopaedic Units in the country, the collection of PROMS data for patients undergoing Elective Joint Replacement has been made compulsory to monitor outcomes and for research. So much so, the UK government, via the Department of Health, has begun the initiative and made it compulsory to collect PROMs on three major elective surgery, Hernia, Varicose Vein and Joint Arthroplasty. Furthermore, in Orthopaedic Surgery, the PROMs that are routinely recorded are the Oxford Hip and Knee Scores (developed in Oxford), and occasionally for research purposes, the WOMAC or Harris Hip Scores are collected as well. These PROMs are disease-specific PROMs that focuses on the joint function mainly. It is common practice in the UK to combine it with a Generic PROM like the SF-12 or Euroqol -5D to complement the Disease-specific PROM. ALL these PROMs have undergone rigorous qualitative and quantitative psychometric analyses and have proven reliable, responsive and hence validated extensively for use in patients with Lower Limb OA.

1.6.1 Oxford Hip and Knee Scores (OHS & OKS)

OHS and OKS is a disease-specific measure consisting of 12 questions which assesses pain and function of the hip or knee in relation to different activities of daily life. Each question is answered by ticking a position on a five-point ordinal scale. Each item obtains a score of zero (worst function) to 4 (best function) giving a total ranging from 0 to 48. The Oxford scores was developed specifically to assess the outcomes of hip or knee replacements surgery and has been shown to be consistent, reproducible, valid and sensitive to change [17].

1.6.2 WOMAC (Western Ontario and McMaster Universities OA Index)

WOMAC is a well-known disease specific measure which is widely used for measuring outcome after THR and TKR. With a total of 24 items, these are grouped into 3 domains, Pain (5-items), Stiffness (2-items) and Difficulty in Function (17-items). Using a Likert scale the patients rate themselves, each item scoring between zero (worst function) to 4 (least difficulty). The scores can be summed up giving individual domain score and Total WOMAC score.

1.6.3 SF-12 (Short Form-12 Health Survey)

SF-12 is a widely used measure of general-health status which was developed to provide an alternative to SF-36. It is much shorter than the predecessor, easier to administer and has proven reliability and validity. It has a norm-base scores based on large general population in the U.S., which means it has a mean of 50 with a SD of 10. After imputing your data, the designed software will give you 2 scores, Physical Component Score (PCS) and a Mental Component Score (MCS).

1.7 The New PROM

The New Pilot PROM is based on a multi-dimensional concept and has 3 components to it. The first component is the main PROM which has 20 items and assesses 5 different domains. They are Upper Limb function, Lower Limb Function, Role Limitation, Pain and General Health. Each domain has a number of questions which can provide an individual domain score and when the scores of 5 domains are added up gives the Overall Score. The other 2 components to it are the Body Map of Pain (BMP) Score and the Visual Analogue Score (VAS) of overall satisfaction of function.

1.7.1 The main PROM – 20 items (MP20)

During the initial stages of the development of the New PROM, prior to the preliminary field testing, it had 20 items altogether hypothesized to assess 5 separate domains. They were

Lower Limb Function (LL – 6 items), Upper Limb Function (UL – 5 items), Role Limitation (RL – 4 items), Pain (Pain – 3 items) and General Health (GH - 2 items). Patient self-record following a Likert Scale which measure from zero (worst function) to 4 (best function) giving a total score between 0 to 80. However, following further reviews of the questionnaire with experts (Research Group, Consultant Orthopaedic Surgeons, Physiotherapists as well as patients input) it was realized that the questionnaire could be too generic and as a result not very sensitive to changes and intervention. In order to overcome this, we introduced 2 other adjuncts to the main questionnaire body, a Body Map diagram and a Visual Analogue Scale for satisfaction of overall function.

1.7.2 Body Map of Pain (BMP)

The concept of identifying areas that are painful in the body has been used before in relation to outcomes for joint replacements [18] [19]. However, the author has used this basic idea and modified it to provide us with more pertinent information regarding a patient's status. While most studies have used an extensive number of painful sites (up to 19) as correlation index, our Body Map is mainly focused on the major joints. It is essentially a simple diagram of a human body with empty boxes overlaying 12 main joint areas identified. They include Right and Left Hand and Wrist, Elbows, Shoulders, Cervical Spine, Thoracic Spine, Lumbar Spine, Hips, Knees and Foot and Ankle joints. Patients are then instructed to fill in the box or boxes corresponding to the painful joint and any other joint that is also painful and provide a severity score of 0 (no pain) to 10 (most painful) in each box if relevant. E.g. a patient with severe Right knee pain but also bothered with Right Ankle pain and Right shoulder pain may fill the Right knee box with the number 10 (very severe) and fill the right ankle and right shoulder with 5 and 6, respectively. By doing so, the author feels that it is possible to have a quick 'snapshot' of joints affected, whether it is localized only to a hip joint or are we dealing with a patient with multiple joint problems. It is also possible to estimate from the patient's perspective the proportion of pain coming from the index joint compared to the whole body.

1.7.3 Visual Analogue Scale (VAS) for satisfaction

The visual analogue scale for satisfaction of overall function, in principle follows the recommendations of ISAR meeting [20] of using single-item satisfaction outcome for assessing hip/knee replacements. However instead of using Likert scale we used a VAS scale and we ask the question most commonly raised in our Delphi analysis, which is how satisfied patients are. The scale is a horizontal line measuring from 0 (completely not satisfied) to 100 (very satisfied), marked with indents at intervals of 10. The line is scaled to 20cm, and the patient only needs to mark on the horizontal line how satisfied they are with the overall function in the last 4 weeks.

1.8 Goal of this Study

This study's primary goal is to explore this new PROM's psychometric properties analyzing evidence for its'' Reliability, Validity and Responsiveness. The new PROM will be tested against three other standard PROMs already in use in patients undergoing Total joint arthroplasty, the WOMAC, SF-12 and Oxford Hip/Knee Scores. We will explore the impact of adding a novel Body Map Pain and the VAS into the already multidimensional concept of the 20 items main questionnaire (MP20). In achieving this goal, we will adopt the current standards measure recommended by the Cosmin group and the Classical Test Theory methods.

1.8.1 Hypothesis

In undertaking the study of this work, our hypothesis is that this new instrument which includes the MP20, BMP and VAS, together, provides a global functional assessment of patients with Lower Limb OA. Hence, in order to prove this assumption, we need strong evidence that the new instruments are Reliable & valid. Our goals are therefore to answer these four important questions. They are as follows:

1. Is the newly develop PROM, reliable and consistent?

We will measure the test-retest reliability using Interclass Correlation Coefficient and ICC > 0.8 for the sub-scales and > 0.9 for the whole scale is deemed satisfactory. For measure of internal consistency, we will calculate Cronbach's alpha for the item to sub-scale and an alpha value of > 0.5 is considered as good.

2. Is the new PROM, measuring what is supposed to measure?

   This measure of construct validity will again use extensive correlation analysis and we will use the WOMAC subscales, SF-12 and Oxford as our comparators. Using Convergent and Divergent Validity concept, a correlation value of > 0.5 is considered having convergent validity and a correlation value of <0.35 is considered to have a divergent validity.

3. Is it the New PROM responsive to changes?

   We will analyse this by measuring the Effect Size (ES) and Standardised Response Mean (SRM) after 6 months

4. Are the added components, Body Map of Pain (BMP) and Visual Analogue Scale (VAS) for satisfaction a useful adjunct to the PROM?

   We will gather qualitative information from the research participants and also measure the correlation of the improvement in Body Map Score following intervention with improvement in outcome measure as a measure of relative validity.

1.9 Impact of Study

Firstly, it is a novel attempt at combining various domains that are important in assessing the overall functional status of a specific target population of patients with Lower Limb Osteoarthritis. We agree that combining Oxford Scores or WOMAC and SF-12 is quite a sensible approach to achieve our objectives however there are issues related to these questionnaires, especially the SF-12, which patients find confusing. Also, WOMAC questions are sometimes not relevant nowadays, especially regarding bath use, which not many people still use nowadays. Furthermore, mostly none of these questionnaires considers the function

of the upper limb, e.g. shoulders, wrist and elbows, which can offer an insight into the patient's overall function.

Secondly, the development of this new PROM is not intended to replace the current ones in practice, but we believe that the new PROM could offer a new perspective into assessing a patient with multiple joint problems. The multiple domains and the inclusion of Body Map and VAS for overall satisfaction can offer a more holistic impression of the patient's functional status and help surgeons focus treatment appropriately.

And thirdly, we also believe it could become a valuable screening tool for GP who wishes to refer a patient to secondary services. A more common scenario in an arthroplasty clinic is a referral for knee pain accompanied with a poor Oxford Score but no mention of the patient's severe lumbar back pain or shoulder pain. If this information is at hand, the referral, in theory, could be prioritized appropriately and channelled to the appropriate services.

2 Methods

2.1 Study Design

The strategy to field tests this new PROM was developed to assess Acceptability, Reliability, Validity and Responsiveness. Hence, we require two separate cohort from the target population, **Cohort A for Test-retest Reliability Study** and **Cohort B for Responsiveness Study**. Both groups require baseline data to be collected i.e. Pre-operative PROMs at initial contact and a second PROM data collected within 2 weeks for cohort A and in 6 months for cohort B. And because both groups are from the same target population i.e. end stage Hip or knee Osteoarthritis, the baseline data are combined to analyze Internal reliability (Internal consistency) and Construct Validity. The Response rate, Data quality, Ceiling and Floor effects will also be analyzed to establish its efficacy.

Baseline PROMs

The 1st set of PROMs, for both the groups will be administered at recruitment. There was a total of 4 set of PROMs to be filled in by each patient so at least 30 - 45 mins are allocated. Patients were allowed to fill in the questionnaires independently and at the end checked for missing items.

2nd set of PROMS

For the *Cohort A*, each patient was handed another set of 4 PROMs and asked to fill them in after 2 days and no longer than 14 days. There is no exact consensus as to how long the interval is between tests, however most studies quote a period of between 2 -14 days is considered to be adequate. The Oxford Hip and Knee Score Group used 2 days interval for their repeatability studies [21] and the development of WOMAC questionnaire took 7 days as their recall period [22]. Other relevant studies also follow similar time intervals for validating the reliability of the outcome measures [23] [24]. A pre-paid addressed envelope was provided to facilitate return of retest questionnaires.

For the Cohort B, participants were contacted using a postal questionnaire at 6 months post-surgery.

In order to determine the **acceptability** of the new questionnaires, patients should respond to the debriefing questions: (a) Did you need any help filling in this *questionnaire*? (c) Were there questions that you found unnecessary? If there was can you tell us why? (d) Were there any questions that you thought was not included that may have been useful? and (e) Did you feel this questionnaire have covered everything that you think is need to assess you OVERALL Function? If no can you suggest your comments for improvement?

2.2 Participants

Our participants were sampled from patients who are have been offered Total Hip or Total Knee Replacement. This should represent our target population which are patients with Lower Limb Osteoarthritis. Because it represents mainly the extreme spectrum of the disease, i.e. end stage Osteoarthritis, we will be able to analyse any ceiling effects within the study population. To ensure they are a representative heterogenous sample, all patients above the age of 18 and those who are able to fill in the questionnaires independently are eligible for recruitment. A total of 120 participants were needed following discussion with our research planning group to gain a reasonable sample size for these initial stages of psychometric evaluation.

2.3 Sample Size

There is no formal sample size estimation that could be found for evaluation of PROMs. However, the rule of thumb adopted by most studies appears to be adequate (reference) in that principally you would require a large enough sample representative of the target population. 5 to 10 participants for each item to reduce the chance of effect is recommended, and so following this recommendation we would require at least 100 patients (because maximum number of items in the new prom is 20).

This estimated sample size depends largely on the reliability of the items, as the more reliable the PRO measure is, the lesser the number of patients required to analyse its properties. However, we do not know this in advance hence it's advisable to err more than less. The estimated sample size of 100 participants should suffice in order to obtain analyses for

Internal Consistency, however the study design will split the sample size into 2 cohorts to analyse Test retest reliability and Responsiveness. In addition, we also took into consideration the estimated number of non-respondents (patients who fail to return questionnaires) in our existing NHS PROMs programme in our institution of about 60%.

Based on a further powered calculation, a minimum of 45 respondents will be needed to gain a lower bound of confidence interval of 0.55, assuming Cronbach's alpha is 0.7 with 20 items questionnaire.

Hence after discussion with our research group and biostatistician, we agreed to increase our sample population to 120 patients, with 60 patients to be recruited in each cohort. This sample is feasible within our time and financial constraints, and should be adequate to provide sufficient evidence of its psychometric properties during this early stage of field testing.

2.4 Eligibility

All NHS patients seen in our institution our elective outpatient clinic was eligible for recruitment. The criteria would be any male or female above the age of 18, who have been listed for Total Hip or Knee Replacement and are able to independently fill the questionnaire PROMs. They should also be able to provide informed consent to participate. Half of the patients would be those recruited in elective clinic after having being listed for surgery (Cohort A) and the other half would be recruited during the actual day for surgery (Cohort B).

Patients were excluded if they are having revision operation, multiple joint surgery, frail, do not speak or understand English, unable to provide consent and unable to fill the questionnaires independently.

2.5 Recruitment and Consent Procedures

Group A patients were recruited from an elective Hip and Knee Clinic led by respective Orthopaedic Consultants in our institution. Patients who have been clinically decided by their respective consultants to require a joint replacement and put on the waiting list are

approached for eligibility into the study. Group B patients were recruited on the morning of their respective surgery day. Patient having either hip or knee replacement were approached for eligibility into the study.

Patients were eligible for recruitments were identified consecutively and approached by either member of researcher team (NM and HY). A verbal explanation of the study and Patient Information Leaflet was given for patients to consider. It explains the rationale of the study, objectives, limit of involvement of participants and what is required of them. Participants are allowed as much time to discuss the with their family or healthcare professional before deciding to participate. Patients who refused consent to participate were not expected to give any reason, however if they voluntarily wish to, the details was recorded.

Patients who consented to participate were invited to provide baseline information and signed an informed consent form that has been approved by the institution's research and ethics committee. They were then invited to complete the New PROMs as well as 3 other PRO measures, The WOMAC, SF-12 form and the relevant Oxford Score Questionnaire. All patients were free to withdraw from the study at any point without giving reasons and without prejudicing future treatments.

2.6 Data Collection and Handling

Registration and Baseline Clinical Data were collected by the researcher team (NM & HY) in designated CRF and the PROM questionnaires were completed by the patients. The completed CRF and PROMs were reviewed by the NM and HY and every attempt is made to chase missing responses immediately or where possible either by telephone call or mail.

Data was entered in to Excel sheet and cross checked by NM and HY for data errors, before finally entered into SPSS version 25 for analyses. To minimize error two person would enter data where one person reads and the other inputs data into software. This is then cross-checked and the roles are swapped between the researcher to ensure 100% data cross checked.

Unreturned PROMs from participants gets another form sent to them by mail within 1-2 weeks it was due, followed by a telephone call if there are still no responses.

Missing data

Missing data could be classified into 3 categories (Little and Rubin 2002), missing completely at random, missing at random and Missing not at Random. For both cohorts, missing forms are excluded from the analysis and for missing items, patients will be contacted to complete missing data.

2.7 Instruments of measures used

Patients that have consented to participate in the study were then given 4 sets of Baseline PROMs, which includes the new PROM (MP20, BMP and VAS), WOMAC, SF-12 and Oxford Questionnaire. These comparator PROMs were used based on their existing excellent performance documented in the literature (reference) and their wide usage in the UK. It is also the combination of PROMs that our institution currently uses as part of the PROMs programme.

2.7.1 Western Ontario and McMaster Universities OA Index (WOMAC)

WOMAC is a well-known disease specific measure which is widely used for measuring outcome after THR and TKR. With a total of 24 items, these are grouped into 3 domains, Pain (5-items), Stiffness (2-items) and Difficulty in Function (17-items). Using a Likert scale the patients rate themselves, each item scoring between zero (worst function) to 4 (least difficulty). The scores can be summed up giving individual domain score and Total WOMAC score.

2.7.2 Short Form-12 Health Survey

SF-12 is a widely used measure of general-health status which was developed to provide an alternative to SF-36. It is much shorter than the predecessor, easier to administer and has

proven reliability and validity. It has a norm-base scores based on large general population in the U.S., which means it has a mean of 50 with a SD of 10. After imputing your data, the designed software (reference) will give you 2 scores, Physical Component Score (PCS) and a Mental Component Score (MCS).

2.7.3 Oxford Hip & Knee Questionnaire (OHS & OKS)

OHS and OKS is a disease-specific measure consisting of 12 questions which assesses pain and function of the hip or knee in relation to different activities of daily life. Each question is answered by ticking a position on a five-point ordinal scale. Each item obtains a score of zero (worst function) to 4 (best function) giving a total ranging from 0 to 48. The Oxford scores was developed specifically to assess the outcomes of hip or knee replacements surgery and has been shown to be consistent, reproducible, valid and sensitive to change (reference Fitzpatrick).

2.7.4 The New PROMs

20 item Main Questionnaire (MP20):

During the initial stages of the development of the New PROM, prior to the preliminary field testing, it had 20 items altogether hypothesized to assess 5 separate domains. They were Lower Limb Function (LL – 6 items), Upper Limb Function (UL – 5 items), Role Limitation (RL – 4 items), Pain (Pain – 3 items) and General Health (GH - 2 items). Patient self-record following a Likert Scale which measure from zero (worst function) to 4 (best function) giving a total score between 0 to 80. However, following further reviews of the questionnaire with experts (Research Group, Consultant Orthopaedic Surgeons, Physiotherapists as well as patients input) it was realized that the questionnaire could be too generic and as a result not very sensitive to changes and intervention. In order to overcome this, we introduced 2 other adjuncts to the main questionnaire body, a Body Map diagram and a Visual Analogue Scale for satisfaction of overall function.

Body Map of Pain (BMP):

The principle behind having a Body Map essentially is to provide a visual representation of the areas that is most affected by pain around the whole human body. We identified 12 main joints and placed an empty box over these joints. The patients will self-score any joint that they feel is bothering them with pain/discomfort and give a score from zero (no pain) to 10 (maximum pain). The main joint that is of concern is denoted as Index Joint, and we identify the score as BMP Index Joint Score.

Visual Analogoue Scale (VAS) of Satisfaction:

The visual analogue scale for satisfaction of overall function, in principle follows the recommendations of ISAR meeting [20] of using single-item satisfaction outcome for assessing hip/knee replacements. However instead of using likert scale we used a VAS scale and we ask the question most commonly raised in our Delphi analysis, which is how satisfied patients are. The scale is a horizontal line measuring from 0 (completely not satisfied) to 100 (very satisfied), marked with indents at intervals of 10. The line is scaled to 20cm, and the patient only needs to mark on the horizontal line how satisfied they are with the overall function in the last 4 weeks.

Hence the final version of the New PROM contains the following:

- 20 items (5 Domain) – Maximum Total score 80
- Body Map Diagram for pain – BMP Index Joint Score, ILR and delta ILR scores
- VAS for Overall satisfaction – Range from 0 to 100

NEW Pilot Questionnaire

Patient Name:		Date of Birth:	Today's Date:

For the following items please select the response that best describes your **level of function** on average over the **last month** *(For each item please tick one box per row)*

		Able without problems	Able but a little difficult	Able but moderately difficult	Able but very difficult	Unable
When I Need to:						
01	Stand up from a chair	☐	☐	☐	☐	☐
02	Put on footwear	☐	☐	☐	☐	☐
03	Get in and out of a car	☐	☐	☐	☐	☐
04	Walk for 10 minutes	☐	☐	☐	☐	☐
05	Go up a flight of stairs	☐	☐	☐	☐	☐
06	Go down a flight of stairs	☐	☐	☐	☐	☐
07	Carry things (e.g. shopping bag)	☐	☐	☐	☐	☐
08	Do up buttons	☐	☐	☐	☐	☐
09	Reach out for something at shoulder height	☐	☐	☐	☐	☐
10	Turn a key	☐	☐	☐	☐	☐
11	Prepare a meal	☐	☐	☐	☐	☐
12	Do my regular job or daily routine if retired	☐	☐	☐	☐	☐
13	Perform leisure or sporting activities (e.g. Dancing, bowling, gardening)	☐	☐	☐	☐	☐
14	Do Housework	☐	☐	☐	☐	☐
15	Go shopping on my own	☐	☐	☐	☐	☐

Please select the response that best describes your **level of pain while resting**

		No pain	Little pain	Moderate pain	Severe pain	Constant pain
16	Level of pain	☐	☐	☐	☐	☐

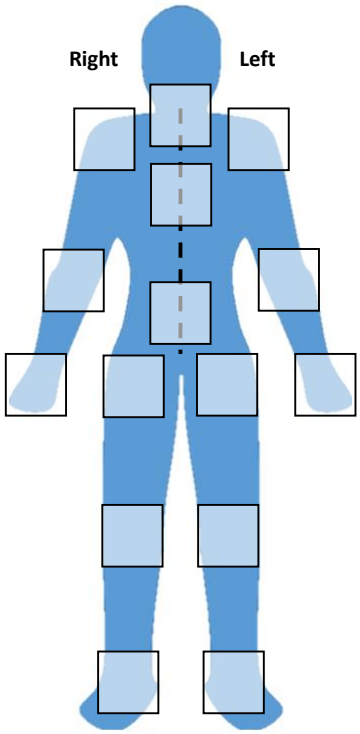
Please select how **pain** from your joints limits your **overall function**

		No Limitation	Little limitation	Moderate limitation	A lot of limitation	Completely limited
When I Need to:						
17	Use my legs	☐	☐	☐	☐	☐
18	Use my arms	☐	☐	☐	☐	☐
		No Limitation	**Little limitation**	**Moderate limitation**	**A lot of limitation**	**Completely limited**
19	How does your **general medical health (e.g. asthma)** limit your overall function?	☐	☐	☐	☐	☐
20	How does your **mood (e.g. anxiety, depression)** limit your overall function?	☐	☐	☐	☐	☐

NEW Pilot Questionnaire

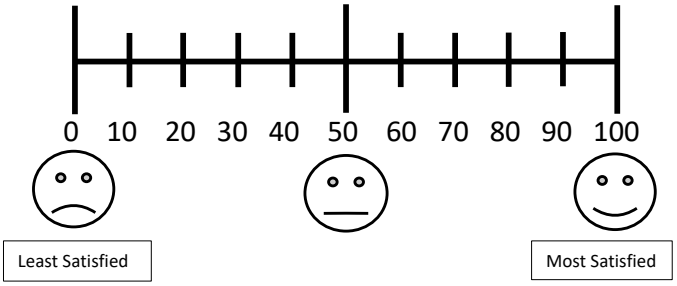
Question 21

Using a scale of 0 to 10 *(0 – no pain, 10 – worse pain),* rate the **SEVERITY of pain in your joints** by filling in the **boxes** on the picture below. **Leave unaffected joints blank**. (E.g. left knee [8], right shoulder [4])



Question 22

Using the following scale please mark with an '**X**' on the **scale line** to show how **satisfied** you are with your **overall level of function** on average over the **last month**. (0 – Least satisfied, 100 – Most satisfied)



2.8 Analysis Plan

2.8.1 Sample Characteristics

We will analyze the characteristics of sample Age, Body Mass Index and the various outcome scores based on their recruitment cohort, Group A and B. We will assess normality distribution using Shapiro-Wilk Test, visual inspection of their histograms and Q-Q plots, and skewness statistics. In the event that the outcome is mixed normal and non-normal distribution we will adopt a more conservative approach for statistical analyses.

2.8.2 Descriptive Statistics

Baseline demographics data of sample will be reported as means with Standard deviation, median with Confidence intervals, Ranges and frequencies where appropriate. Baseline outcome measure scores will be analysed between the two groups and depending on the normality distribution, average differences between the two group will be verified for significant differences. We hypothesize that the baseline data should be the same with both groups as they are both end stage Hip and Knee OA patient who are waiting for their respective operation.

2.8.3 Floor and ceiling effects

Floor and ceiling effects were calculated as the proportion (expressed as percentage) of patients showing respectively minimum and maximum scores with regards to each type of PROM. There is no consensus as to the level that should be acceptable, however most studies showed that <15% is considered adequate [25].

2.8.4 Test-retest Reliability

We will use intraclass correlation coefficient (ICC) as a measure of reliability throughout our analysis. ICC measures the extent to which patients can be distinguished from each other despite measurement error. In other words, the same patient filling the questionnaire should be able to reproduce the same outcome provided no change has occurred. It measures the

stability and hence reproducibility of an instrument over time assuming no changes has occurred. It is the most widely accepted method to assess test-retest reliability based on Cosmin's Group recommendations.

From a statistical standpoint, ICC measures the strength of agreement between repeated measurements, by assessing the proportion of the total variance, σ2 (the square of the SD), of an observation that is associated with the between-patient variability (reference). So, if we regard the error variability as 'noise' and the true value of patients' scores as the 'signal', then ICC measures the signal/noise ratio. If the ICC is large (close to 1), then the random error variability is low and a high proportion of the variance in the observations is attributable to variation between patients. The measurements are then described as having high reliability. Conversely, if the ICC is low (close to 0), the random error variability dominates and the measurements have low reliability.

There are several methods to derive ICC and here we use two-way fixed effect model with absolute agreement to derive the correlation coefficient (ICC) along with its corresponding confidence interval as recommended by most PROM validation studies (references).

A reliability coefficient of at least 0.90 is often recommended if measurements are to be used for evaluating individual patients [26], although not all PRO measures is able to achieve such a high level. For discriminating between groups of patients, as in a clinical trial, it is usually recommended that the reliability should exceed 0.70. Thus values from 0.70 to 0.90 represent 'moderate or good reliability (acceptable error)' and above 0.90 are 'high or excellent (minimal or no error)'.

2.8.5 Internal Reliability

Also known as Internal consistency. It measures the extent to which items in the sub-scales, are homogenous, thus measuring the same concept. The primary method of estimating reliability for multi-item scales, is to extensively analyze correlations provide information about the associations among different items in the scale. Internal consistency is typically indexed by Cronbach's coefficient alpha, which is estimated using a two-way fixed-effect

analysis of variance (ANOVA) that partitions the “signal” (i.e., between person variance) from the “noise” (i.e., interaction between people and responses to different items) (reference) . Alpha can also be expressed using the formula (3):

$$\text{Alpha} = (K * Rii\ )\ (1 + (K - 1) * Rii\ )$$

This alternative expression illustrates how reliability increases with the number of items (K) in a scale and the strength of the correlations among items as represented by the intraclass correlation (Rii). Rii represents the estimated reliability for a single item. Applying the formula, a scale with an intraclass correlation of 0.30 and five items will have an estimated reliability of 0.68. Thus, a PRO measurement with multi-item scales yields more precise measurement of PRO constructs than a single-item measure.

We will use Cronbach’s alpha coefficient to determine the level of agreeability with the items and the domains it belongs to as well as it’s relationship towards the measurement scale. For each of the domains the target Cronbach’s alpha > 0.7 (reference), although several other studies have mentioned that a value of > 0.3 is adequate.

2.8.6 Construct Validity

Measures the extent to which the scores of an instrument scale or subscale, are an adequate reflection of the dimensionality of the construct to be measured. In other words, how good is it measuring what it is supposed to measure. We will be analysing Spearman’s correlation of the subscales / overall scale with other PROMs scale / subscale, using Multi trait Multi method analysis (MTMM). The correlation between each item and the scale it belongs to will also be analysed using Multit trait-Multi Item method (MTMI). This forms the concept of Convergent and Divergent Validity. Correlation coefficient $r > 0.50$ were considered as indicator of convergent validity and correlations $r < 0.35$ as an indicator of divergent validity. The hypotheses in this study is that the Physical function domains should correlate well with established WOMAC-Physical Function, Oxford Hip and Knee Scores and the SF-12 Physical Component Scores. In theory the Pain Domain in the new PROM should correlate with

WOMAC pain scale, and the General Health domain (New PROM) should deliver a positive correlation with SF-12 Mental Component score.

2.8.7 Responsiveness [27]

The most widely used measures of sensitivity and responsiveness are the standardized Response mean (SRM) and the effect size (ES), which are also used for indicating clinical significance. The SRM is the ratio of the mean change to the SD of that change, and the ES is the ratio of the mean change to the SD of the initial measurement. Thus ES ignores the variation in the change, while SRM is more similar to the paired t-test (except that the t-test uses the standard error, SE, rather than the SD). The SRM is more frequently used than ES. A standardized measure of effect size (ES) was calculated using the Cohen's d. Cohen's d computes the difference in score between the baseline and the follow-up at 6 months, and then divides this difference by the baseline score standard deviation. This method takes into consideration the variability in scores, a step beyond the mean differences considered in the paired sample t-test. In interpreting Cohen's d, a small, medium, and large ES can be considered as d = 0.20, 0.50, and 0.80 respectively. The standardized response mean (SRM) is another important indicator of ES, similar to the paired t-test, but removing dependence on sample size from the equation. [27] This is computed as the mean difference between baseline and follow-up PRO scores divided by the standard deviation of difference scores, reflecting individual changes in scores. Although there is no perfect consensus, recommended guidelines for interpreting SRM values are similar to interpretation of Cohen's d.

2.8.8 Statistical Package

- For our analysis we will be using SPSS version 25 [28]

2.9 Ethics Statement & Confidentiality

2.9.1 Declaration of Helsinki and Good Clinical Practice

This study was conducted in accordance with the ethical principles that have their origin in the Declaration of Helsinki, and that are consistent with Good Clinical Practice and the applicable requirements as stated in the Research Governance Framework for Health and Social Care (2nd edition 2005). Local investigators have ensured the study is conducted in accordance with relevant regulations and with Good Clinical Practice.

2.9.2 Ethics Approvals

The protocol, informed consent form, participant information sheet and any proposed advertising material have been approved by the Research Ethics Committee (REC), and host institution.

The Chief Investigator have submitted and, obtained approval from the above parties for all substantial amendments to the original approved documents. For reference, the study IRAS number is 207639, version 2 dated 15/7/2016.

The REC has the purpose to look after the rights, well-being and dignity of patients. The REC reference number is given on the front page of this protocol. The REC that reviewed this study was the Haydock Research Ethics Committee.

2.9.3 Consent

A written version and a verbal discussion of the PIS and Informed Consent Form will be presented to the participants which details the exact nature of the study; the implications and constraints of the protocol; the known side effects and any risks involved in taking part. It will be clearly stated that the participant is free to withdraw from the study at any time for any reason without prejudice to future care, and with no obligation to give the reason for withdrawal.

Written consent will then be obtained by means of participant dated signature, and dated signature of the person who presented and obtained the informed consent. The person who obtained the consent must be suitably qualified, experienced and trained in consenting for research, and have been authorised to do so by the chief Investigator. Members of the research teams at both trusts will be involved in obtaining consent, which will mostly be a research nurse or physiotherapist. In other instances the consent will be obtained by the

investigators who are clinically qualified research students, with a medical background and who have been appropriately trained in research consent.

A copy of the signed Informed Consent Form will be given to the participants, and one copy will be kept by the research team. The original signed Consent Form will be retained in the medical notes, and a copy held in the Investigator Site File (ISF). Consent forms will be held in a secure location separately from any study data.

2.9.4 Confidentiality

A database of fully identifiable patient information will be stored on the NHS computer system of the RLBUHT and will comply with hospital information governance policies and the Data Protection Act.

Each participant will be allocated a unique study number. For analysis purposes linked anonymised data will be transferred to the computer system of the University of Liverpool. Every effort will be made to ensure that all identifiers are removed from the transferred file. The researchers undertaking this work are all trained in Good Clinical Practice (GCP) and have been trained in information governance. Only those who are suitably qualified and have honorary contracts with the RLBUHT will have access to the fully identifiable database.

The study site file containing study documentation and including the original consent forms will be stored in locked filing cabinets in a locked room with restricted access. These files will be available for inspection by any regulatory authority.

2.9.5 Audits and Inspections

The CI shall submit once a year throughout the study or on request, an Annual Progress report to the REC Committee, host organisation and Sponsor. In addition, an End of Study notification and final report will be submitted to the same parties.

The study may be monitored, or audited in accordance with the current approved protocol, International Conference of Harmonisation (ICH), GCP, relevant regulations and standard operating procedures. The monitoring plan will be developed by the CI.

2.9.6 Indemnity

The University of Liverpool has a specialist insurance policy in place, which would operate in the event of any participant suffering harm as a result of their involvement in the research. NHS indemnity operates in respect of the clinical treatment, which is provided.

3 RESULTS

3.1 Sample Characteristics

We recruited 120 patients consecutively in total from two cohort, Group A (Test Retest Cohort) and Group B (Responsiveness Cohort). 60 patients were approached from each group and nobody declined to participate, with 64 females and 56 males altogether.

A Shapiro-Wilk's test ($p < 0.05$) (Shapiro et al 1965, Razali et all 2011), and visual inspection of their histograms and Q-Q plots showed most of the outcome measure scores of the New PROMs (BMP and VAS Satisfaction) and its sub-domains (LL, UL, RL, Pain and GH) along with the Oxford Scores showed significant evidence that they were not normally distributed. For e.g. the sub-domain scores the LL domain reports a skewness of 0.504 (SE 0.309) and kurtosis -0.686 (SE=0.608) for Group B (see table 3.1 and Figure 3.1). However, there were evidence to show that he MP20 and WOMAC scores were normally distributed with skewness -0.119 (SE 0.309) and a kurtosis of -0.429 (SE=0.608) for Group B-MP20 (Table 3.1)

	Study Group	Kolmogorov-Smirnova			Shapiro-Wilk			Skewness			Kurtosis			Distribution
		Statistic	df	Sig.	Statistic	df	Sig.	Statistic	SE	z-score	Statistic	SE	z-score	
Age	Test-retest	0.143	60	0.004	0.955	60	**0.027**	-0.56	0.309	**-1.81**	-0.063	0.608	-0.10	Not Normal
	Responsiveness	0.111	60	0.066	0.937	60	**0.004**	-0.827	0.309	**-2.68**	0.177	0.608	0.29	Not Normal
BMI	Test-retest	0.11	40	.200*	0.944	40	**0.049**	0.32	0.374	**0.86**	-0.58	0.733	-0.79	Not Normal
	Responsiveness	0.09	60	.200*	0.97	60	**0.146**	0.042	0.309	**0.14**	-0.97	0.608	-1.60	Normal
LL Score Baseline	Responsiveness	0.128	60	0.015	0.948	60	**0.013**	0.504	0.309	**1.63**	-0.686	0.608	-1.13	Not Normal
	Test Retest	0.139	60	0.006	0.948	60	**0.012**	0.488	0.309	**1.58**	-0.615	0.608	-1.01	Not Normal
UL Score Baseline	Responsiveness	0.16	60	0.001	0.855	60	**0**	-1.485	0.309	**-4.81**	2.329	0.608	3.83	Not Normal
	Test Retest	0.158	60	0.001	0.893	60	**0**	-1.289	0.309	**-4.17**	1.712	0.608	2.82	Not Normal
RL Score Baseline	Responsiveness	0.14	60	0.005	0.956	60	**0.03**	0.216	0.309	**0.70**	-0.911	0.608	-1.50	Not Normal
	Test Retest	0.124	60	0.023	0.961	60	**0.05**	0.44	0.309	**1.42**	-0.52	0.608	-0.86	Not Normal
Pain Score Baseline	Responsiveness	0.138	60	0.006	0.974	60	**0.23**	-0.28	0.309	**-0.91**	0.536	0.608	0.88	Normal
	Test Retest	0.135	60	0.009	0.939	60	**0.005**	-0.742	0.309	**-2.40**	1.118	0.608	1.84	Not Normal
GH Score Baseline	Responsiveness	0.242	60	0	0.831	60	**0**	-0.725	0.309	**-2.35**	-0.649	0.608	-1.07	Not Normal
	Test Retest	0.225	60	0	0.822	60	**0**	-0.858	0.309	**-2.78**	-0.602	0.608	-0.99	Not Normal
MP20 Score Baseline	Responsiveness	0.076	60	.200*	0.985	60	**0.676**	-0.119	0.309	**-0.39**	-0.429	0.608	-0.71	Normal
	Test Retest	0.128	60	0.016	0.971	60	**0.16**	-0.387	0.309	**-1.25**	-0.177	0.608	-0.29	Normal
Body Map Pain Index Score	Responsiveness	0.212	59	0	0.857	59	**0**	-0.615	0.311	**-1.98**	-0.29	0.613	-0.47	Not Normal
	Test Retest	0.207	60	0	0.852	60	**0**	-0.881	0.309	**-2.85**	0.41	0.608	0.67	Not Normal
BMP TOTAL Score	Responsiveness	0.212	59	0	0.806	59	**0**	1.621	0.311	**5.21**	2.152	0.613	3.51	Not Normal
	Test Retest	0.243	60	0	0.724	60	**0**	2.371	0.309	**7.67**	6.437	0.608	10.59	Not Normal
VAS	Responsiveness	0.1	59	.200*	0.974	59	**0.233**	0.114	0.311	**0.37**	-0.467	0.613	-0.76	Normal
	Test Retest	0.159	60	0.001	0.929	60	**0.002**	0.305	0.309	**0.99**	-1.115	0.608	-1.83	Not Normal
WOMAC Total	Responsiveness	0.119	60	0.034	0.969	60	**0.136**	0.484	0.309	**1.57**	-0.146	0.608	-0.24	Normal
	Test Retest	0.062	60	.200*	0.983	60	**0.567**	0.182	0.309	**0.59**	-0.442	0.608	-0.73	Normal
SF12-PCS	Responsiveness	0.105	60	0.099	0.956	60	**0.031**	0.795	0.309	**2.57**	0.772	0.608	1.27	Not Normal
	Test Retest	0.129	60	0.015	0.897	60	**0**	1.393	0.309	**4.51**	2.59	0.608	4.26	Not Normal
SF12-MCS	Responsiveness	0.101	60	.200*	0.958	60	**0.036**	-0.096	0.309	**-0.31**	-1.121	0.608	-1.84	Not Normal
	Test Retest	0.083	60	.200*	0.962	60	**0.062**	0.045	0.309	**0.15**	-1.114	0.608	-1.83	Normal
Oxford Total	Responsiveness	0.12	60	0.031	0.951	60	**0.018**	0.813	0.309	**2.63**	1.022	0.608	1.68	Not Normal
	Test Retest	0.09	60	.200*	0.983	60	**0.555**	0.217	0.309	**0.70**	-0.605	0.608	-1.00	Normal

Table 3.1 Test for Normality

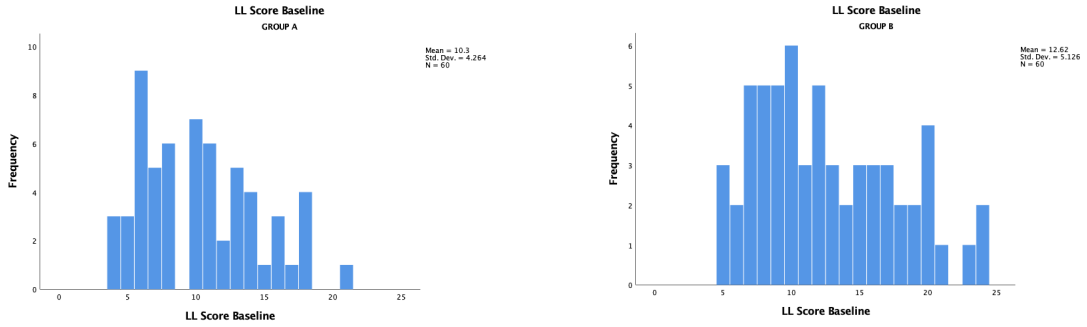


Figure 3.1 LL Domain Scores Distribution

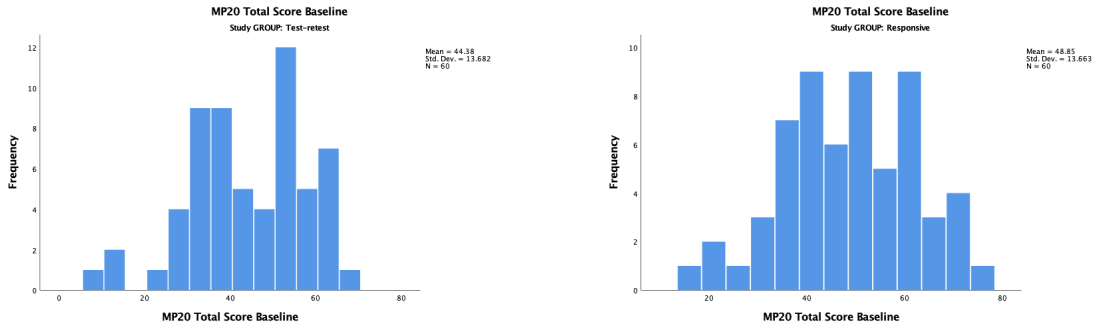
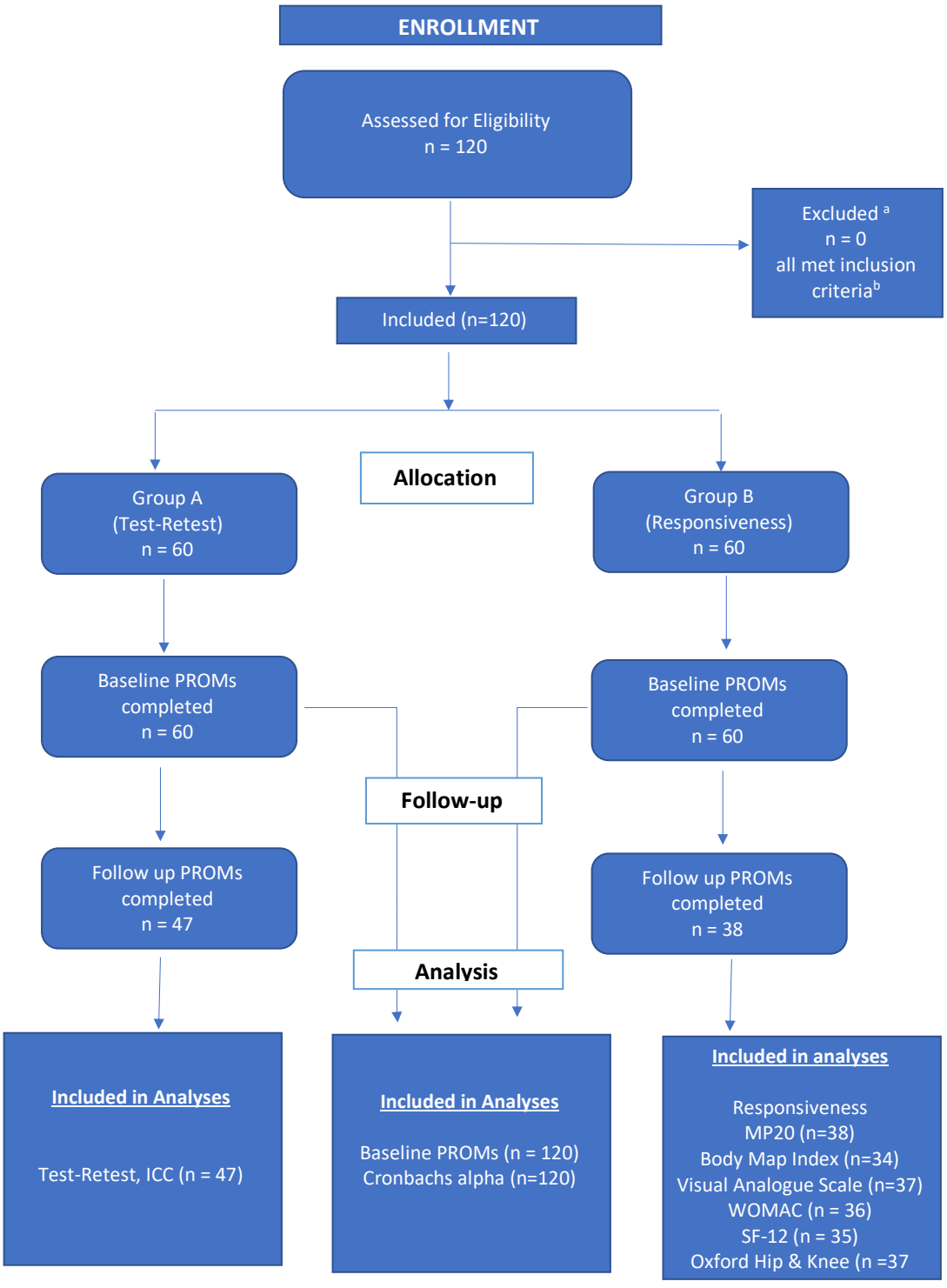


Figure 3.2 MP20 Total Scores Distribution

Apart from the outcome scores the baseline data such as age and Body Mass Index (BMI) were also not normally distributed (Table 3.1), hence we will assume that the rest of the data is mostly not normally distributed and will use non-parametric methods for most of our statistical analyses.

3.2 Consort Flowchart



a. Patients were excluded if they are having revision operation, multiple joint surgery, frail, do not speak or understand English, unable to provide consent and unable to fill the questionnaires independently.
b. The criteria would be any male or female above the age of 18, who have been listed for Total Hip or Knee Replacement and are able to independently fill the questionnaire PROMs

3.3 Descriptive Statistics

3.3.1 Baseline demographics data

We recruited a total of 120 patients consecutively from 2 cohorts, Group A (Test Retest) with 60 patients and Group B (Responsiveness) also 60 patients (see Figure 3.3 Recruitment of Study Group). In Group A, twenty-seven patients were going to have Total Hip Replacement (THR) and 33 Total Knee Replacement (TKR). The median ages were 69 (43 to 90) years old, and 71 (47 to 86) years old respectively. Independent sample t-test showed there were no significant difference in ages between those listed for TKR or TKR within Group A. In Group B, twenty-three patients had THR and 37 TKR, with median age of 68 (38 to 84) years old and 74 (47 to 86) years old respectively. There was also no significant difference in age distribution between the surgeries in this Group, and no significant difference between the two-recruitment cohort (Table 3.2).

The Body Mass Index (BMI) for both groups were collected as part of our baseline data (Table 3.3). There were 20 missing data all from Group A, and this was because these patients were recruited from Clinics and it's not routine practice for all patients to have their BMI recorded. However, all patients in Group B had their BMI recorded because they were recruited on the day of their surgery where it's a mandatory pre-requisite. There was no statistically significant difference between BMIs of patients having THR or TKR surgery and also no difference between Group A and B.

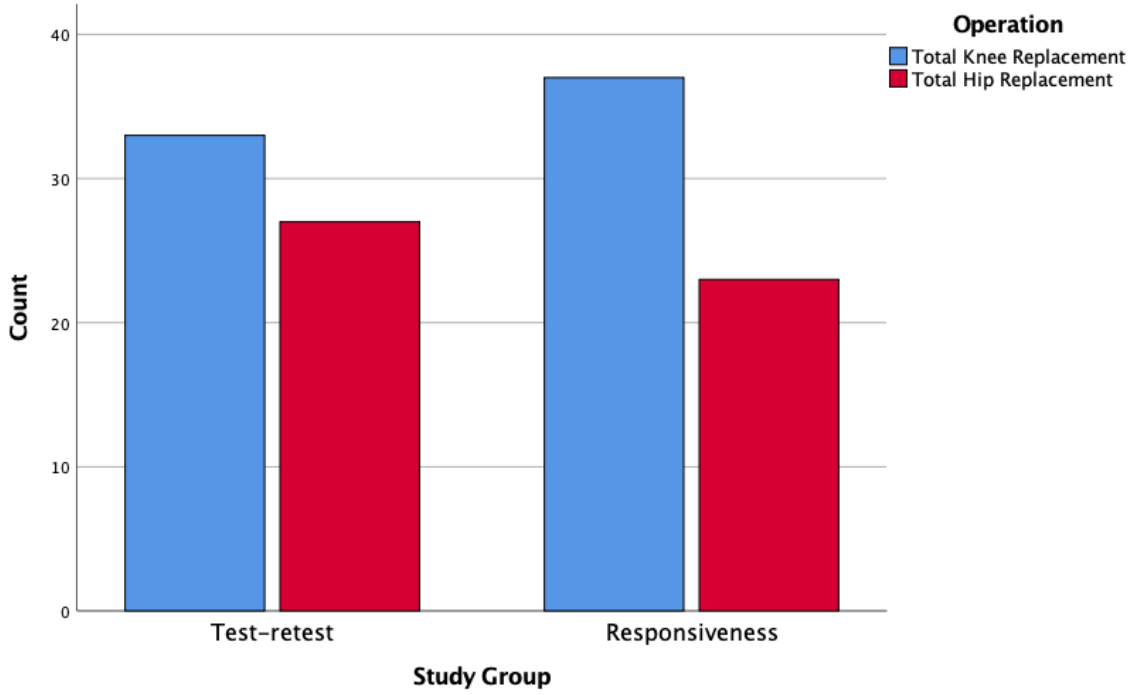


Figure 3.3 Recruitment of Study Group

Group A (Test Retest) and Group B (Responsiveness)

	GROUP A		GROUP B	
	THR	**TKR**	**THR**	**TKR**
Numbers recruited (n)	27	33	23	37
Mean Age	68.7. (sd 13.4)	69.21 (sd 8.9)	66.22 (sd 12.6)	71.65. (sd 9.7)
Median Age	69	71	68	74
Min - Max Age	43-90	47-86	38 - 84	47 - 86
p-value (Mann Whitney U test)	0.958		0.108	
	0.638			

Table 3.2. Age Distribution between Group A and B

	GROUP A		GROUP B	
	THR	TKR	THR	TKR
n	16	24	23	37
missing data	11	9	0	0
Mean BMI	28.9 (sd 6.1)	31.4 (sd 4.8)	29.55 (sd 5.6)	30.2 (sd 5.6)
Median BMI	27	30.8	29.5	30.8
Min - Max BMI	19 - 40	25 - 40	19 - 40	20 - 40
p values (Mann Whitney U test)	0.113		0.698	
	0.778			

Table 3.3 Body Mass Index (BMI) distribution of Study Population

3.3.2 Baseline New PROMs Data

We collected a complete set of baseline data for the MP20 measures with no missing data. This was possible because we were able to cross check the MP20 data during recruitment and chase up the missing data with the patients immediately. However, for the Body Map Pain (BMP) Index Joint and Visual analogue scale (VAS) for satisfaction data there was one missing data each.

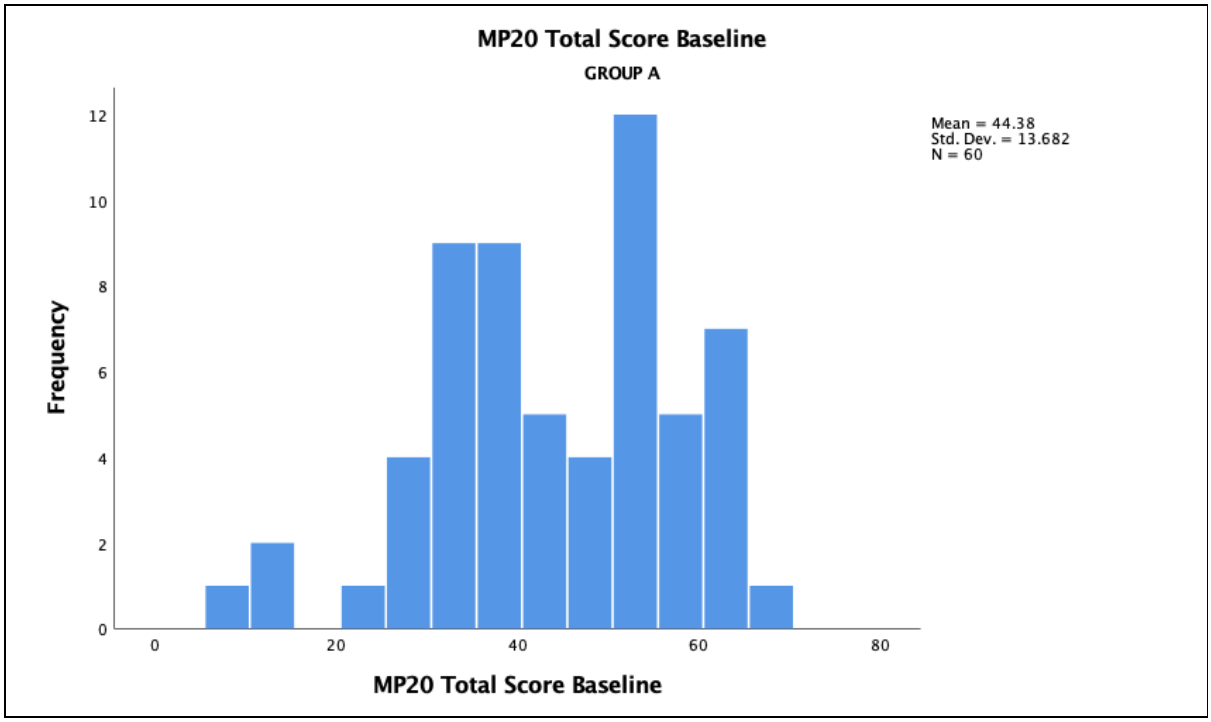
For MP20 measures, the median scores for Group A and B were 44.5 (range 8 to 70) and 49 (range of 16 to 77) respectively. The p value was 0.109 (Mann-Whitney U test) indicating no significant difference in MP20 scores between the two groups. The median scores for the rest of sub-domains are observed in Table 3.4, with minimal evidence of significant difference between the cohorts.

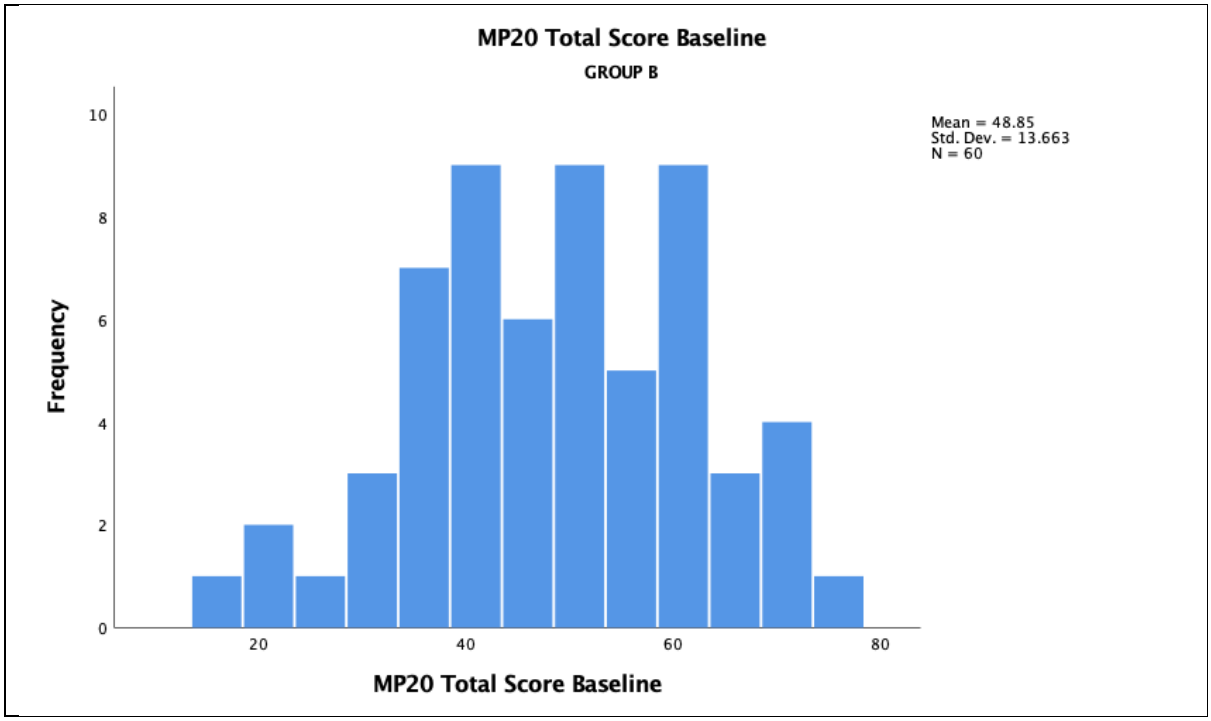
Table 3.5 shows the data for BMP Index Joint Scores and VAS for satisfaction. The median scores for BMP Index Joint were 9 (5 to 10) and 9 (6 to 10) for Group A and B respectively. The VAS for Satisfaction median scores were 37.5 (5 to 85) and 40 (0 to 90) for Group A and B respectively. Neither traits showed any significant difference between the 2 Groups.

The Baseline results for WOMAC, SF-12 and Oxford Scores are illustrated in Table 3.6, and apart from SF12-MCS scores, neither WOMAC, Oxford Scores or SF12-PCS showed any statistically significant difference between the two groups baseline data.

			LL Score Baseline	UL Score Baseline	RL Score Baseline	Pain Score Baseline	GH Score Baselne	MP20 Total Score Baseline
Group A	N	Valid	60	60	60	60	60	60
		Missing	0	0	0	0	0	0
	Mean		10.3	15.5	6.4	6.2	6.0	44.4
	Median		10	16.5	6	6	7	44.5
	Std. Deviation		4.3	4.2	3.6	2.0	2.2	13.7
	Minimum		4	1	0	0	1	8
	Maximum		21	22	15	10	8	70
Group B	N	Valid	60	60	60	60	60	60
		Missing	0	0	0	0	0	0
	Mean		12.6	16.1	7.3	6.4	6.4	48.9
	Median		12	17	7.5	6	7	49
	Std. Deviation		5.1	3.8	3.8	2.2	1.7	13.7
	Minimum		5	3	1	0	3	16
	Maximum		24	20	15	12	8	77
p value	*Mann-Whitney U Test*		0.16	0.375	0.202	0.595	0.425	**0.109**

Table 3.4 Results for New Outcome Measures

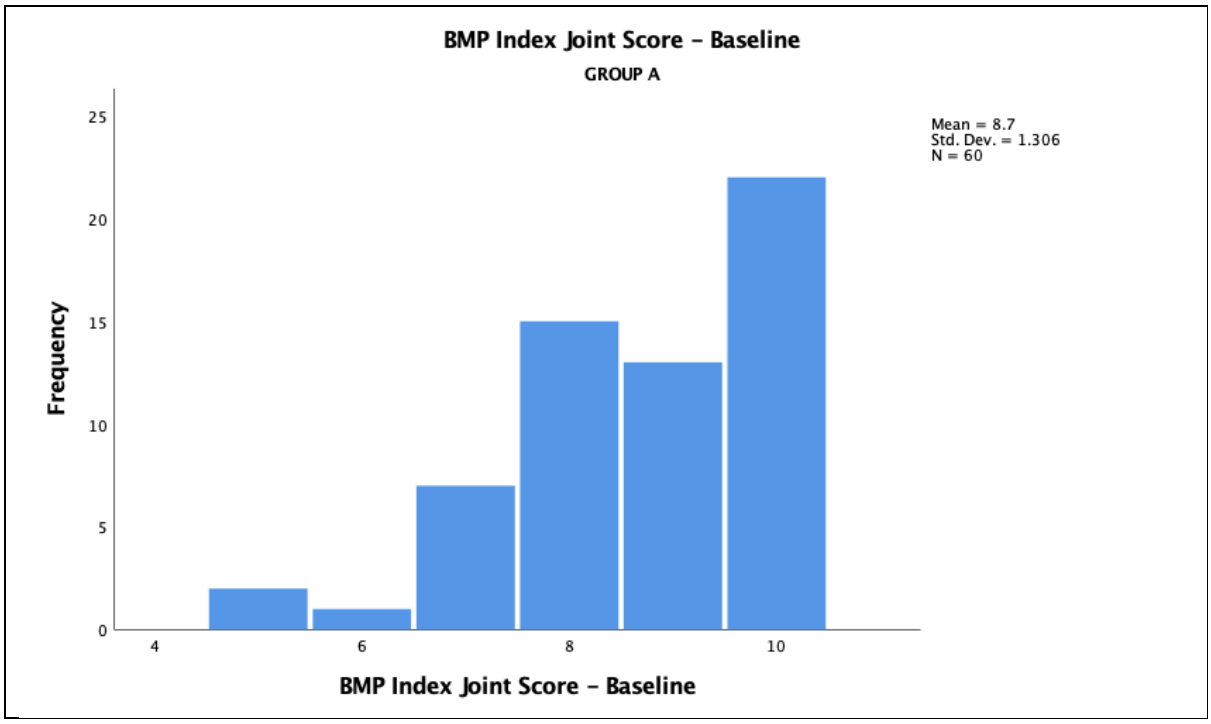

MP20 Total Score Baseline
GROUP A
Mean = 44.38
Std. Dev. = 13.682
N = 60
Frequency
0
2
4
6
8
10
12
0
20
40
60
80
MP20 Total Score Baseline

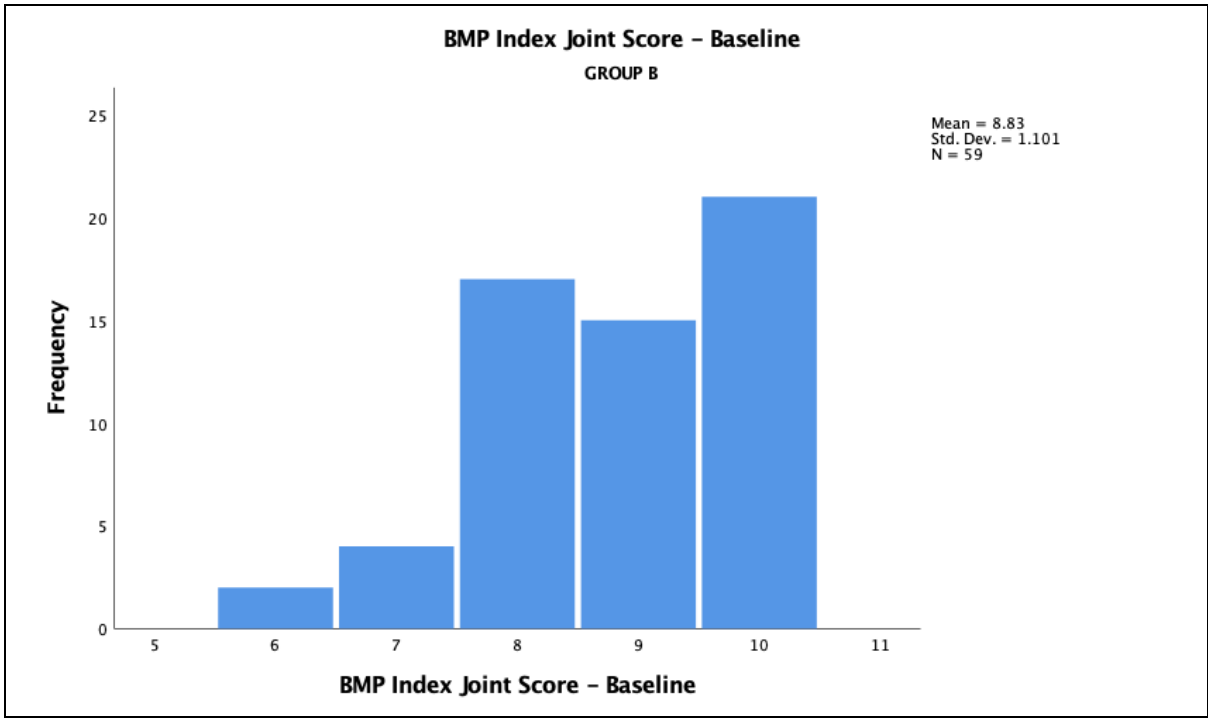

MP20 Total Score Baseline
GROUP B
Mean = 48.85
Std. Dev. = 13.663
N = 60
Frequency
0
2
4
6
8
10
20
40
60
80
MP20 Total Score Baseline

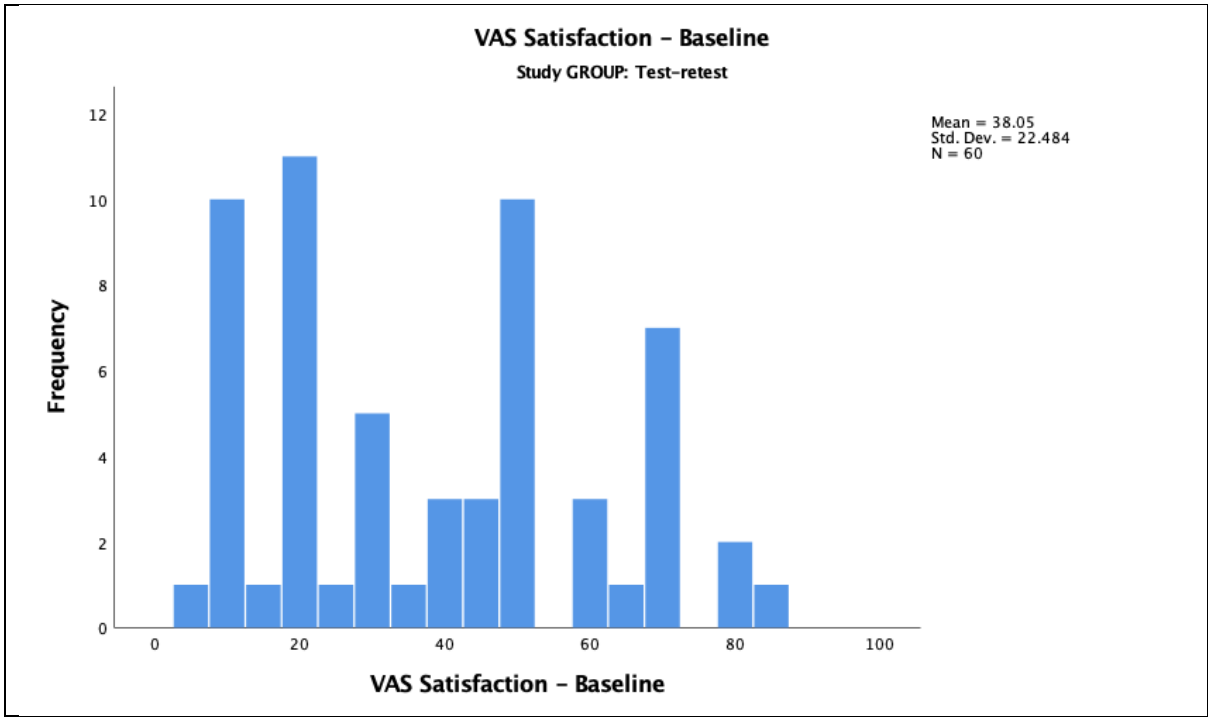
			BMP Index Joint Score - Baseline	VAS Satisfaction - Baseline
Group A	N	Valid	60	60
		Missing	0	0
	Mean		8.7	38.1
	Median		9.0	37.5
	Std. Deviation		1.3	22.5
	Minimum		5	5
	Maximum		10	85
Group B	N	Valid	59	59
		Missing	1	1
	Mean		8.8	40.5
	Median		9.0	40.0
	Std. Deviation		1.1	22.6
	Minimum		6	0
	Maximum		10	90

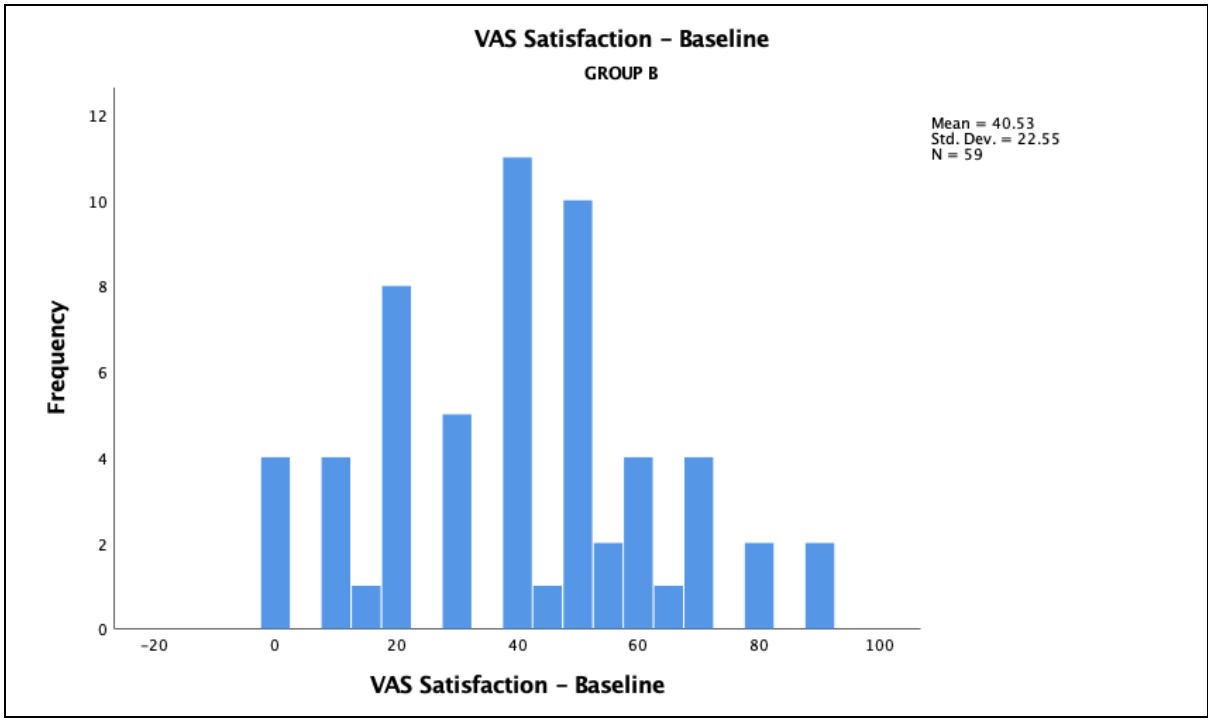
p values	*Mann-Whitney U test*	**0.739**	**0.473**

Table 3.5 Results for Body Map Pain (BMP) Index Joint Score and Visual Analogue Scale (VAS) for satisfaction Baseline measurements




BMP Index Joint Score – Baseline
GROUP B
Mean = 8.83
Std. Dev. = 1.101
N = 59
Frequency
25
20
15
10
5
0
5
6
7
8
9
10
11
BMP Index Joint Score – Baseline


VAS Satisfaction – Baseline
Study GROUP: Test-retest
Mean = 38.05
Std. Dev. = 22.484
N = 60
Frequency
12
10
8
6
4
2
0
0
20
40
60
80
100
VAS Satisfaction – Baseline



			WOMAC Total Score Baseline	PCS-SF12 score Baseline	MCS-SF12 score Baseline	Oxford Total Scores - Baseline
Test-retest	N	Valid	60	60	60	60
		Missing	0	0	0	0
	Mean		33.7	29.1	43.3	14.7
	Median		34	27.23	43.36	13.5
	Std. Deviation		16.5	6.7	13.3	7.5
	Minimum		0	19.86	20.03	0
	Maximum		69	53.99	67.19	32
Responsive	N	Valid	60	60	60	60
		Missing	0	0	0	0
	Mean		38.95	29.06	48.41	16.12
	Median		37.5	28.26	50.23	14
	Std. Deviation		17.4	7.8	11.6	8.1
	Minimum		9	16.24	27	4
	Maximum		86	54.88	67.92	44
p values	*Mann-Whitney U test*		**0.138**	**0.875**	**0.034**	**0.437**

Table 3.6 Summary results of WOMAC, SF-12 and Oxford Baseline scores

3.4 Floor and Ceiling Effects

Results of the Floor and Ceiling effect analyses demonstrates that total scores for MP20 does not record any floor or ceiling effects (Table 3.7). Upon further sub-domain analyses, the GH domain reveals a high proportion of ceiling effects on both groups of 36.7% and 41.7% for

Group A and B respectively. The rest of the domains otherwise demonstrates very low floor and ceiling proportions which are acceptable (reference) with a range from 0 to 16.7%.

The BMP analysis shows also quite a high ceiling effects for Group A and B with 36.7% and 35.6% respectively (Table 3.8). This may reflect the level of pain that the patient is at before surgery. On the other hand, the VAS demonstrates minimal floor effects Group B (5.1%) and zero floor and ceiling effect in Group A.

Upon comparative analyses, all three PROMs (MP20, WOMAC and Oxford) individual total scores performed well, with minimal floor and ceiling effects noted (Table 3.9)

			LL Score Baseline	UL Score Baseline	RL Score Baseline	Pain Score Baseline	GH Score Baselline	MP20 Total Score Baseline
Group A	N	Valid	60	60	60	60	60	60
		Missing	0	0	0	0	0	0
	n Min Scores		0.0	0.0	2.0	1.0	0.0	0
	% Floor Effect		0	0	3.3	1.7	0	0
	n Max Scores		0	8	0	0	22	0
	% Ceiling Effect		0.0	13.3	0.0	0.0	**36.7**	0
Group B	N	Valid	60	60	60	60	60	60
		Missing	0	0	0	0	0	0
	n min scores		0.0	0.0	0.0	1.0	0.0	0
	% floor effect		0	0	0	1.7	0	0
	n max scores		2	10	0	1	25	0
	% ceiling effect		3.3	16.7	0.0	1.7	**41.7**	0

Table 3.7 Results for New PROM Floor and Ceiling effects

			BMP Index joint Score	VAS for Satisfaction
Group A	N	Valid	60	60
		Missing	0	0
	n Min Scores		0.0	0.0
	% Floor Effect		0	0
	n Max Scores		22	0
	% Ceiling Effect		36.7	0.0
Group B	N	Valid	59	59
		Missing	0	0
	n min scores		0.0	3.0
	% floor effect		0	5.1
	n max scores		21	0
	% ceiling effect		35.6	0.0

Table 3.8 Result of New PROM Body Map Pain Index Joint and VAS Floor and Ceiling effects

			MP20 Total Score Baseline	WOMAC Score Baseline	Oxford Score Baseline
Group A	N	Valid	60	60	60
		Missing	0	0	0
	n Min Scores		0	1.0	0.0
	% Floor Effect		0	1.7	0
	n Max Scores		0	0	0
	% Ceiling Effect		0	0	0
Group B	N	Valid	60	60	60
		Missing	0	0	0
	n min scores		0	0.0	1.0
	% floor effect		0	0	1.7
	n max scores		0	0	0
	% ceiling effect		0	0	0

Table 3.9 Results Comparing Floor and Ceiling Effects between PROMs

3.5 Missing Data

Missing data could be classified into 3 categories (Little and Rubin 2002), Missing completely at random, missing at random and Missing not at Random. For both cohorts, missing forms are excluded from the analysis and for missing items, patients were contacted to complete missing data.

From the 60 participants recruited from each cohort, baseline data were successfully collected on all of them. This was possible because the research group was able to chase the missing data immediately during recruitment at the clinics and wards. However, in Group A the response rate for Test Retest questionnaires was only 78.3% (47 respondents) and in Group B the response rate was expectedly even lower at 61.7% (37 respondents). Attempts were made to contact the participants that did not return the questionnaires through a follow up questionnaire mailed to the addresses, however everyone failed to respond.

After cross-checking through our data, 47 complete datasets are available for Test Retest Analysis and only 35 participants dataset are available for Responsiveness analysis. For analysing Internal Reliability and Construct Validity, we assumed both cohorts at the point of recruitment are from the same population (i.e. end stage Hip and Knee OA), hence we were able to combined both cohorts baseline data giving a total of 120 complete dataset. In addition, further statistical analyses did not show any significant different between the groups baseline data, (Age and BMI), as well as most of the outcome measures.

3.6 Test-Retest Reliability Results

47 patients returned a follow up set of questionnaires within 2 weeks which gives a 78% response rate. There were a couple of missing items across all the questionnaires however a follow-up telephone call was made by the researcher (NM) and missing item scores were completed.

3.6.1 MP20

The average Interclass Correlation Coefficient (ICC) for each of the 20 items questionnaire (MP20) showed good test retest reliability with ICC results ranging from 0.62 (CI 0.32 to 0.79) to 0.88 (CI 0.78 to 0.93). The strongest correlations were items 2, 15 and 20 with ICC values of 0.87, 0.88 and 0.87 respectively and the weakest correlations were with items 10 and 18 with ICC values of 0.67 and 0.62 respectively. 70% of the items (14/20) had ICC values of > 0.75 (Very Good), 6 out of 20 items had ICC values between 0.6 to 0.75 (Good), and no items had average ICC values of < 0.6 (moderate) or > 0.9 (Excellent).

The scores for the separate domains which included domains of Lower Limb (LL), Upper Limb (UL), Role Limitation (RL), Pain (P) and General Health (GH) showed average ICC values of 0.89 (CI 0.79 to 0.94), 0.87 (CI 0.77 to 0.93), 0.88 (CI 0.77 to 0.93), 0.84 (CI 0.70 to 0.91) and 0.88 (CI 0.78 to 0.93) respectively. The domain scores demonstrate very good test retest reliability characteristics.

Finally, the MP20 total scores showed an average ICC value of 0.92 (CI 0.85 to 0.96) which indicates excellent test retest reliability characteristics (see Table 3.10).

3.6.2 Body Map Pain (BMP) Index Joint Scores and Visual Analogue Scale (VAS) for Satisfaction

The BMP index joint and VAS for satisfaction showed very good test retest results (see Table 3.10) with ICC values of 0.83 (CI 0.68 to 0.90) and 0.89 (CI 0.81 to 0.94).

3.6.3 WOMAC, SF-12 Physical Component Score (PCS) and Mental Component Score (MCS), and Oxford Scores

The average ICC values for WOMAC sub-domains for Pain, Stiffness and Activity of Daily Living (ADL) were 0.91 (CI 0.84 to 0.95), 0.88 (CI 0.79 to 0.94) and 0.91 (CI 0.84 to 0.95) respectively (see Table 3.11). The WOMAC total scores demonstrates excellent test retest reliability with average ICC values of 0.92 (CI 0.86 to 0.96).

The Generic Health-related Quality of Life Questionnaire SF-12 probably displayed the least reliable test retest results with average ICC values of SF-12 PCS and MCS of 0.59 (CI 0.28 to 0.77) and 0.88 (CI 0.78 to 0.93) respectively.

The Oxford Questionnaire demonstrated the best test retest reliability results with average ICC value of 0.95 (CI 0.91 to 0.97)

3.6.4 Summary of Test retest Reliability study

47 complete datasets were available of analysis and the average ICC values for MP20, WOMAC and Oxford Scores displayed excellent test reliability with average ICC values 0.92, 0.92 and 0.95 respectively. We consider average ICC values of > 0.90 as excellent, 0.75 to 0.90 as very good and 0.60 to 0.75 as good. Average ICC values of < 0.4 are considered poor test retest reliability.

The individual items of MP20 showed varying level of reliability, however overall average ICC values were good to very good reliability. No items scored ICC values of <0.60. The BMP Index Joint and VAS satisfaction component also displayed a very good repeatability results with average ICC values of 0.83 and 0.89 respectively. So we conclude that overall the new questionnaires performed well when compared to WOMAC and Oxford, and performed better than SF-12 questionnaires. The two items that performed poorly within MP20 was item 10 (RL domain) and 18 (Pain domain).

n = 47		Interclass Correlation Coefficient [b]	95% Confidence Interval		Interclass Correlation Coefficient [b]	95% Confidence Interval		F Test with True Value 0			
	Items / Domains	Average [c]	Lower Bound	Upper Bound	Single [a]	Lower Bound	Upper Bound	Value	df1	df2	Sig
MP20 PROMS ITEMS	Qn1	0.74	0.53	0.86	0.59	0.36	0.75	3.79	46	46	0
	Qn2	0.87	0.77	0.93	0.77	0.62	0.87	7.66	46	46	0
	Qn3	0.84	0.72	0.91	0.73	0.56	0.84	6.22	46	46	0
	Qn4	0.84	0.71	0.91	0.72	0.55	0.83	6.15	46	46	0
	Qn5	0.79	0.62	0.88	0.65	0.45	0.79	4.67	46	46	0
	Qn6	0.73	0.52	0.85	0.57	0.35	0.74	3.71	46	46	0
	Qn7	0.80	0.64	0.89	0.67	0.47	0.80	4.96	46	46	0
	Qn8	0.75	0.56	0.86	0.61	0.39	0.76	4.01	46	46	0
	Qn9	0.82	0.68	0.90	0.70	0.52	0.82	5.67	46	46	0
	Qn10	0.67	0.42	0.82	0.51	0.26	0.69	3.06	46	46	0
	Qn11	0.84	0.70	0.91	0.73	0.54	0.84	6.93	46	46	0
	Qn12	0.74	0.53	0.86	0.59	0.37	0.75	4.10	46	46	0
	Qn13	0.77	0.58	0.87	0.63	0.41	0.78	4.73	46	46	0
	Qn14	0.82	0.67	0.90	0.69	0.51	0.82	5.80	46	46	0
	Qn15	0.88	0.78	0.93	0.78	0.64	0.87	8.09	46	46	0
	Qn16	0.80	0.65	0.89	0.67	0.48	0.80	5.02	46	46	0
	Qn17	0.71	0.48	0.84	0.55	0.32	0.72	3.54	46	46	0
	Qn18	0.62	0.32	0.79	0.45	0.19	0.65	2.74	46	46	0
	Qn19	0.82	0.68	0.90	0.70	0.52	0.82	5.77	46	46	0
	Qn20	0.87	0.77	0.93	0.78	0.63	0.87	8.11	46	46	0
DOMAINS OF MP20 PROM	Lower Limb	0.89	0.79	0.94	0.79	0.66	0.88	8.50	46	46	0
	Upper Limb	0.87	0.77	0.93	0.77	0.62	0.87	8.08	46	46	0
	Role Limitation	0.88	0.77	0.93	0.78	0.63	0.88	8.94	46	46	0
	Pain Score	0.84	0.70	0.91	0.72	0.54	0.83	6.19	46	46	0
	General Health Score	0.88	0.78	0.93	0.78	0.63	0.87	8.40	46	46	0
	MP 20 Total Score	**0.92**	0.85	0.96	0.85	0.75	0.92	13.52	46	46	0
	BMP Index Joint Score	**0.83**	0.69	0.90	0.70	0.52	0.82	5.64	46	46	0
	VAS Satisfaction	**0.89**	0.81	0.94	0.81	0.68	0.89	9.20	46	46	0

Table 3.10 Summary table of ICC results

		Interclass Correlation Coefficient [b]	95% Confidence Interval		Interclass Correlation Coefficient [b]	95% Confidence Interval		F Test with True Value 0			
	Items / Domains	Average [c]	Lower Bound	Upper Bound	Single [a]	Lower Bound	Upper Bound	Value	df1	df2	Sig
WOMAC SCORES	Womac Pain	0.91	0.84	0.95	0.83	0.72	0.90	10.77	46	46	0
	Womac Stiff	0.88	0.79	0.94	0.79	0.66	0.88	8.53	46	46	0
	Womac ADL	0.91	0.84	0.95	0.84	0.73	0.91	11.09	46	46	0
	Womac Total	**0.92**	0.86	0.96	0.86	0.76	0.92	12.90	46	46	0
SF-12 SCORES	SF12 PCS	**0.59**	0.28	0.77	0.42	0.16	0.63	2.50	46	46	0.001
	SF12 MCS	**0.88**	0.78	0.93	0.79	0.64	0.88	8.79	46	46	0
OXFORD SCORES	Oxford Scores	**0.95**	0.91	0.97	0.90	0.83	0.94	18.95	46	46	0

Two-way mixed effects model where people effects are random and measures effects are fixed.

a The estimator is the same, whether the interaction effect is present or not

b Type A intraclass correlation coefficients using an absolute agreement definition

c This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise

Table 3.11 ICC results for WOMAC, SF-12 and Oxford Scores

3.7 Results for Internal Reliability

Internal reliability or also known as internal consistency is a measure how much each item is closely related to one another. We used a type of Multi item scaling analysis (a type of correlation analysis) where a measure of inter item correlation is calculated between each item measuring the same trait/domain with a Cronbach's alpha coefficient value given for each domain. An alpha value of > 0.7 is considered as good estimate internal reliability of the given trait/domain, and if we would like to see the alpha value decrease if the item was deleted indicating it's homogeneity with the given trait/domain. In this case we would keep

the item with the domain. If however the alpha value increases, it may indicate that the item does not share the same construct with the domain, i.e. it's not correlating well with the rest of the domain. In this case one would exclude the item from the domain.

3.7.1 Cronbach's alpha – estimation of internal reliability

	Qn1	Qn2	Qn3	Qn4	Qn5	Qn6	Qn7	Qn8	Qn9	Qn10	Qn11	Qn12	Qn13	Qn14	Qn15	Qn16	Qn17	Qn18	Qn19	Qn20	LL Score	UL Score	RL Score	Pain Score	GH Score	MP20 TOTAL
Qn1	1	.590**	.641**	.436**	.531**	.471**	.426**	0.16	.256**	0.151	.287**	.409**	.235**	.316**	.270**	.300**	.388**	0.162	0.046	.350**	.785**	.421**	.353**	.402**	.261**	.574**
Qn2	.590**	1	.564**	.415**	.422**	.340**	.507**	.248**	.336**	.288**	.380**	.481**	.302**	.448**	.271**	.342**	.399**	.213*	0.154	.312**	.729**	.504**	.428**	.442**	.270**	.620**
Qn3	.641**	.564**	1	.539**	.573**	.439**	.480**	.270**	.351**	.256**	.399**	.562**	.354**	.466**	.442**	.428**	.348**	.217*	0.176	.351**	.799**	.510**	.543**	.477**	.315**	.678**
Qn4	.436**	.415**	.539**	1	.571**	.468**	.539**	.218*	.281**	.280**	.430**	.483**	.431**	.520**	.446**	.321**	.514**	.188*	.238**	.283**	.737**	.485**	.569**	.441**	.289**	.654**
Qn5	.531**	.422**	.573**	.571**	1	.751**	.611**	.311**	.368**	.321**	.463**	.610**	.568**	.615**	.491**	.359**	.542**	.350**	.303**	.442**	.809**	.603**	.681**	.574**	.441**	.782**
Qn6	.471**	.340**	.439**	.468**	.751**	1	.515**	.327**	.345**	.228*	.387**	.425**	.514**	.478**	.464**	.265**	.459**	.319**	.250**	.387**	.736**	.525**	.575**	.477**	.367**	.680**
Qn7	.426**	.507**	.480**	.539**	.611**	.515**	1	.362**	.444**	.345**	.512**	.457**	.454**	.570**	.618**	.344**	.525**	.316**	.254**	.392**	.662**	.810**	.647**	.518**	.373**	.741**
Qn8	0.16	.248**	.270**	.218*	.311**	.327**	.362**	1	.669**	.735**	.598**	.395**	.288**	.439**	.358**	.260**	.330**	.408**	.358**	.436**	.329**	.716**	.442**	.423**	.448**	.577**
Qn9	.256**	.336**	.351**	.281**	.368**	.345**	.444**	.669**	1	.618**	.531**	.419**	.379**	.528**	.427**	.319**	.284**	.597**	.387**	.434**	.432**	.769**	.531**	.538**	.455**	.671**
Qn10	0.151	.288**	.256**	.280**	.321**	.228*	.345**	.735**	.618**	1	.579**	.420**	.240**	.461**	.322**	.181*	.338**	.430**	.346**	.342**	.331**	.642**	.426**	.383**	.386**	.547**
Qn11	.287**	.380**	.399**	.430**	.463**	.387**	.512**	.598**	.531**	.579**	1	.570**	.333**	.544**	.435**	.279**	.405**	.435**	.328**	.500**	.508**	.784**	.565**	.484**	.476**	.690**
Qn12	.409**	.481**	.562**	.483**	.610**	.425**	.457**	.395**	.419**	.420**	.570**	1	.448**	.633**	.486**	.362**	.408**	.337**	.250**	.460**	.625**	.576**	.765**	.509**	.413**	.737**
Qn13	.235**	.302**	.354**	.431**	.568**	.514**	.454**	.288**	.379**	.240**	.333**	.448**	1	.561**	.539**	.200*	.483**	.308**	.198*	.376**	.515**	.471**	.758**	.417**	.336**	.623**
Qn14	.316**	.448**	.466**	.520**	.615**	.478**	.570**	.439**	.528**	.461**	.544**	.633**	.561**	1	.629**	.394**	.522**	.391**	.360**	.532**	.605**	.676**	.864**	.595**	.511**	.804**
Qn15	.270**	.271**	.442**	.446**	.491**	.464**	.618**	.358**	.427**	.322**	.435**	.486**	.539**	.629**	1	.362**	.393**	.362**	.360**	.461**	.519**	.609**	.837**	.521**	.481**	.723**
Qn16	.300**	.342**	.428**	.321**	.359**	.265**	.344**	.260**	.319**	.181*	.279**	.362**	.200*	.394**	.362**	1	.348**	0.153	.189*	.416**	.446**	.369**	.413**	.806**	.360**	.524**
Qn17	.388**	.399**	.348**	.514**	.542**	.459**	.525**	.330**	.284**	.338**	.405**	.408**	.483**	.522**	.393**	.348**	1	.231*	.276**	.309**	.563**	.484**	.534**	.667**	.330**	.620**
Qn18	0.162	.213*	.217*	.188*	.350**	.319**	.316**	.408**	.597**	.430**	.435**	.337**	.308**	.391**	.362**	0.153	.231*	1	.271**	.295**	.321**	.521**	.421**	.571**	.312**	.503**
Qn19	0.046	0.154	0.176	.238**	.303**	.250**	.254**	.358**	.387**	.346**	.328**	.250**	.198*	.360**	.360**	.189*	.276**	.271**	1	.504**	.255**	.381**	.363**	.317**	.311**	.469**
Qn20	.350**	.312**	.351**	.283**	.442**	.387**	.392**	.436**	.434**	.342**	.500**	.460**	.376**	.532**	.461**	.416**	.309**	.295**	.504**	1	.461**	.546**	.552**	.488**	.893**	.662**
LL Score	.785**	.729**	.799**	.737**	.809**	.736**	.662**	.329**	.432**	.331**	.508**	.625**	.515**	.605**	.519**	.446**	.563**	.321**	.255**	.461**	1	.656**	.676**	.611**	.422**	.859**
UL Score	.421**	.504**	.510**	.485**	.603**	.525**	.810**	.716**	.769**	.642**	.784**	.576**	.471**	.676**	.609**	.369**	.484**	.521**	.381**	.546**	.656**	1	.705**	.611**	.533**	.862**
RL Sore	.353**	.428**	.543**	.569**	.681**	.575**	.647**	.442**	.531**	.426**	.565**	.765**	.758**	.864**	.837**	.413**	.534**	.421**	.363**	.552**	.676**	.705**	1	.621**	.530**	.874**
Pain Score	.402**	.442**	.477**	.441**	.574**	.477**	.518**	.423**	.538**	.383**	.484**	.509**	.417**	.595**	.521**	.806**	.667**	.571**	.317**	.488**	.611**	.611**	.621**	1	.461**	.762**
GH Score	.261**	.270**	.315**	.289**	.441**	.367**	.373**	.448**	.455**	.386**	.476**	.413**	.336**	.511**	.481**	.360**	.330**	.312**	.811**	.893**	.422**	.533**	.530**	.461**	1	.656**
MP20 TOTAL	.574**	.620**	.678**	.654**	.782**	.680**	.741**	.577**	.671**	.547**	.690**	.737**	.623**	.804**	.723**	.524**	.620**	.503**	.469**	.662**	.859**	.862**	.874**	.762**	.656**	1

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

Table 3.12 Correlation Matrix of MP20

We calculated separate Cronbach's alpha coefficient for each domain of the Lower limb, Upper limb, Role Limitation, Pain and General Health for the MP20 measure and the results are summarized in table Table 3.13.

The alpha value for LL, UL, RL, Pain and GH were 0.868, 0.85, 0.825, 0.544 and 0.712 respectively. This indicate that all the domains except for the Pain domain demonstrate good estimate of internal reliability of its corresponding items with alpha value > 0.7. The poor alpha value of Pain domain also corresponded with a poor inter-item correlation (0.15 to 0.348) and poor item-trait correlations (0.216 to 0.387).

The LL Domain item-item correlation analysis demonstrates generally good correlation coefficient values of > 0.4 apart from between item 2 and 6 which shows a value of 0.34. The corrected item-trait correlation for all the items in LL Domain were good with correlation values of 0.581 to 0.730. Cronbach's alpha value if each item were deleted all reduced

indicating evidence of its unidimensional with the LL domain. All items correlation was statistically significant and all the items supported LL domain.

In the UL Domain item 7 showed poorest correlation within the items in the domain (0.345 to 0.512) and also the lowest corrected item-trait correlation of 0.567 ($p < 0.01$). The cronbach's alpha value when item was deleted also increased to 0.88 from 0.85 showing evidence that item 7 does not sit well within the UL domain compared with the rest. The rest of items (Qn 8 to 11) on the other hand showed good evidence of correlation with values of 0.531 to 0.735 ($p<0.01$). The item-trait correlation for items 8 to 11 were also good with values in the range of 0.602 to 0.656. Looking at the correlation matrix for MP20 (Table 3.12) it appears that item 7 correlated better with items for LL Domain with values ranging from 0.426 to 0.611 ($p < 0.01$). The Item-trait correlation was also better with LL domain with correlation value of 0.662 compared to 0.576. From this analysis item 7 does not appear to support the UL Domain.

The four items in RL domain gave an alpha value of 0.825 and the alpha values decreased when the corresponding items were deleted which supports this dimension. The item-item correlation was fair-good with correlation values of 0.448 to 0.633 and corrected item-trait correlation values of 0.599 to 0.745. All the correlation values in RL domain were statistically significant with all items shows evidence of support for this domain.

The pain domain performed poorest with cronbach's alpha value of 0.544 with a maximum correlation values < 0.387 (Item 17). When we analyzed the correlation matrix further, item 16 correlated better with Item 20 on GH domain (0.416, $p < 0.01$), item 17 correlated best with item 5 of LL domain (0.542, $p < 0.01$) and item 18 correlated better with item 9 of RL domain (0.597, $p < 0.01$). There is little evidence to support the internal reliability of items in the Pain Domain.

The GH domain only had 2 items and the correlation value was 0.504 ($p < 0.01$) with alpha value of 0.712. The internal reliability evidence supports the items in the GH domain.

		Qn1	Qn2	Qn3	Qn4	Qn5	Qn6	Corrected item-trait correlation*	Cronbach's Alpha	Cronbach's Alpha if Item Deleted
LL Domain	Qn1	1	.590**	.641**	.436**	.531**	.471**	.686**	0.868	0.843
	Qn2	.590**	1	.564**	.415**	.422**	.340**	.581**		0.86
	Qn3	.641**	.564**	1	.539**	.573**	.439**	.720**		0.837
	Qn4	.436**	.415**	.539**	1	.571**	.468**	.604**		0.854
	Qn5	.531**	.422**	.573**	.571**	1	.751**	.730**		0.829
	Qn6	.471**	.340**	.439**	.468**	.751**	1	.613**		0.852

		Qn7	Qn8	Qn9	Qn10	Qn11	Corrected item-trait correlation	Cronbach's Alpha	Cronbach's Alpha if Item Deleted
UL Domain	Qn7	1	.362**	.444**	.345**	.512**	.567**	0.85	0.879
	Qn8	.362**	1	.669**	.735**	.598**	.643**		0.79
	Qn9	.444**	.669**	1	.618**	.531**	.643**		0.808
	Qn10	.345**	.735**	.618**	1	.579**	.602**		0.822
	Qn11	.512**	.598**	.531**	.579**	1	.656**		0.793

		Qn12	Qn13	Qn14	Qn15	Corrected item-trait correlation	Cronbach's Alpha	Cronbach's Alpha if Item Deleted
RL Domain	Qn12	1	.448**	.633**	.486**	.599**	0.825	0.798
	Qn13	.448**	1	.561**	.539**	.611**		0.794
	Qn14	.633**	.561**	1	.629**	.745**		0.738
	Qn15	.486**	.539**	.629**	1	.628**		0.787

		Qn16	Qn17	Qn18	Corrected item-trait correlation	Cronbach's Alpha	Cronbach's Alpha if Item Deleted
Pain Domain	Qn16	1	.348**	0.153	.336**	0.544	0.423
	Qn17	.348**	1	.231*	.387**		0.369
	Qn18	0.153	.231*	1	.216*		0.536

		Qn19	Qn20	Corrected item-trait correlation	Cronbach's Alpha	Cronbach's Alpha if Item Deleted
GH	Qn19	1	.504**	.504**	0.712	N/A
	Qn20	.504**	1	.504**		N/A

Table 3.13 Correlation matrix &Cronbach's alpha values for MP20 sub-domains

3.8 Results for Construct Validity

We analyzed the relationship of the items within each trait/domain and also with other traits/domains to see the extent of 'similarities' it has with them. In other words, if the item has strong similarity i.e. strong correlation with the trait/domain that it belongs to, then the item is said to exhibit Convergent Validity. If it does not correlate well with poor correlation values then it is assumed to demonstrate Divergent Validity. This concept of convergent and divergent validity is used extensively at item level to discern it's relationship with the domain

and forms the core principles of Multi item Multi Trait analysis (MTMI). It can also be further used at domain-scale level between different type of measures (PROMs) to analyse the validity of the scale compared with other similar measures. This method is also known as Multi Trait Multi Method Analysis (MTMM).

3.8.1 Multi trait Multi Item Analysis

Table 3.12 it gives us a general overview of correlation coefficient values of different items to different traits. We will consider further the correlations of each item in each domain.

Lower Limb Domain (LL)

Item 1 shows good correlation within items within its domain (0.436 to 0.641, $p < 0.01$) but poor correlation other domain items (0.046 to 0.388, $p < 0.01$). However, it's showing better correlations (although not as good) with item 7 and 12 with correlation values of 0.426 and 0.409 respectively ($p < 0.01$). The corrected item-trait correlation shows that Item 1 correlates best with its own domain compared to other domain (Table 3.14) with values correlation value of 0.686 ($p < 0.01$). Item 1 demonstrates Convergent and Divergent Validity.

Item 2 correlations with it's domain range from 0.340 to 0.59 ($p < 0.01$), correlating weakly with item 6. However, it showed better correlations with item 7, 12 and 14 with values of 0.507, 0.481 and 0.448 respectively ($p < 0.01$). Nevertheless, corrected item-trait correlation is still better with its own domain compared to other domain, with value of 0.581 ($p < 0.01$), Table 3.14. So we could say that Item 2 has demonstrated evidence of convergent and divergent validity.

Item 3 shows very good correlation within its domain with item-item correlation ranging from 0.439 to 0.641 ($p < 0.01$). It's item-item correlation with other domain was generally poor except for with items 7 and 12 showing correlation values of 0.48 and 0.562 respectively ($p < 0.01$). It's corrected item-trait correlation is 0.720 and is higher when compared to other domains (0.315 to 0.543, $p < 0.01$). Item 3 clearly demonstrates convergent and divergent validity.

Item 4 correlation values with LL domain ranged from 0.415 to 0.571 ($p < 0.01$). However, its correlation with item 7, 14 and 17 were even better with values of 0.539, 0.520 and 0.514 respectively ($p < 0.01$). Despite that the corrected item-trait correlation was best with it's own domain with value of 0.604 ($p < 0.01$) compared to other domains (range of 0.441 to 0.569). We can still say that item 4 demonstrate convergent and divergent validity.

Item 5 correlation values with LL domain were good ranging from 0.422 to 0.751 ($p < 0.01$). But it also well with Items 7, 12, 13 and 14 with values of 0.611, 0.610, 0.568 and 0.615 ($p < 0.01$). The corrected item-trait correlation was still better with LL domain compared to other domains demonstrating correlation values of 0.730 ($p < 0.01$). Item 5 has evidence of convergent and divergent validity.

Item 6 correlated very well with item 5 (0.751, $p <0.01$) but poorly with item 2 (0.340, $p < 0.01$). It also correlated well with 2 other items from another domain i.e. item 7 (UL) and item 13 (RL domain) having correlation values of 0.515 and 0.514 respectively ($p < 0.01$). The corrected item-trait correlation is still highest with values of 0.613 ($p < 0.01$). Item 6 has evidence of convergent and divergent validity.

Upper Limb domain (UL)

Item 7 correlated poorly with items in its domain with values of 0.362 to 0.512 ($p < 0.01$). However, it correlated better with items in LL domain (range 0.426 to 0.611) and RL domain (range 0.454 to 0.618, $p < 0.01$). The corrected item-trait correlation to UL domain (0.567, $p < 0.01$) is lower compared to LL domain (0.662, $p < 0.01$) and RL domain (0.647, $p < 0.01$). **Item 7 does not** demonstrate satisfactory convergent or divergent validity.

Item 8 correlated well with items 9, 10 and 11 with values of 0.669, 0.735 and 0.598 respectively ($p < 0.01$). It did not correlate well with item 7 with value of 0.362 ($p < 0.01$). Correlation with other items does not indicate good correlations. The corrected item-trait correlation to UL domain was 0.643 ($p < 0.01$) which is higher than the other domains (Table 3.15). Item 8 has good evidence of convergent and divergent validity.

Item 9 correlated well with items 8, 10 and 11 with values of 0.669, 0.618 and 0.531 respectively ($p < 0.01$). It also had a good correlation value with item 14 (RL domain) and item 18 (Pain domain) with correlation values of 0.528 and 0.597 respectively ($p < 0.01$). However, it did not correlate well with item 7 (0.444, $p < 0.01$) which is in its own domain. The corrected item-trait correlation to UL domain was 0.643 ($p < 0.01$), which is still higher compared to correlations with other domains. So we accept item 9 evidence of convergent and divergent validity to its domain.

Item 10 correlated very well items 8, 9 and 11 with values of 0.735, 0.618 and 0.579 respectively ($p < 0.01$). It did not correlate well with item 7 (0.345, $p < 0.01$) from its own domain and it did not correlate well with other items in other domains either. The corrected item-trait correlation to UL domain was 0.602 ($p < 0.01$) and is significantly higher than other domain. Item 10 has convergent and divergent validity demonstrated to it's domain.

Item 11 is the only item in this domain that correlated fairly well with all the items in UL domain with values ranging from 0.512 to 0.598 ($p < 0.01$). However, it also correlated well with items 12 (RL domain), 14 (RL domain) and 20 (GH domain) with values of 0.57, 0.544 and 0.50 respectively ($p < 0.01$). The corrected item-trait correlation to UL domain was 0.656 ($p < 0.01$) which is higher than other domains and thus Item 11 demonstrates evidence of convergent and divergent validity.

Role Limitation domain (RL)

Item 12 correlation with Items 13, 14 and 15 were 0.448, 0.633 and 0.486 respectively ($p < 0.01$). However it appears to also have good correlation with items 2, 3, 4 and 5 from LL domain with values of 0.481, 0.562, 0.483 and 0.610 respectively ($p < 0.01$). Another good correlation was also found with item 11 (UL domain) with value of 0.570 ($p < 0.01$). It does not appear to correlate well with the rest of RL domain, Pain and GH items. The corrected item-trait correlation to RL domain was 0.599 ($p < 0.01$). This is lower compared to item-trait correlation to LL domain which was 0.625 ($p < 0.01$). **Item 12 does not show sufficient evidence of convergent validity as it correlated better with LL domain.**

Item 13 correlations with Item 12, 14 and 15 were 0.448, 0.561 and 0.539 respectively ($p < 0.01$). However, it also correlated well with item 5 (LL) and item 6 (LL) with correlation values of 0.568 and 0.514 ($p < 0.01$). Its correlation with other items are not as impressive. The corrected item-trait correlation with RL domain was 0.611 ($p < 0.01$) which is highest compared to other domains. We accept evidence of convergent and divergent validity for Item 13.

Item 14 correlated well with items 12, 13 and 15 with correlation values of 0.633, 0.561 and 0.629 respectively ($p < 0.01$). However there was also good correlations with other items from other domain, mainly, Item 4 (LL), Item 5 (LL), Item 7 (UL), Item 9 (UL), Item 11 (UL), Item 17 (Pain) and Item 20 (GH), all of which had correlation values of > 0.5 ($p < 0.01$). The corrected Item-trait correlation was 0.745 ($p < 0.01$) which was higher than correlations with other domain. So although there is evidence to suggest that item 14 does not exhibit much of divergent validity with item-item correlation analysis, however it still correlated best with its own domain and hence we accept evidence of convergent and divergent validity.

Item 15 correlated well with items 13 and 14 with correlation values of 0.539 and 0.629 respectively ($p < 0.01$). It did not correlate as well with item 12 with correlative value of 0.486 ($p < 0.01$). For some reason it correlated quite well with item 7 with value of 0.618 ($p < 0.01$). The corrected item-trait correlation was 0.628 ($p < 0.01$), which was higher than the other correlation indices. Item 15 has evidence of convergent and divergent validity.

Pain domain

Item 16 did not correlate well with item 17 and 18 with correlation values of 0.348 ($p < 0.01$) and 0.153 ($p > 0.05$). It also correlated poorly with the rest of the domain items. Item-trait correlation recorded values of 0.336 ($p < 0.01$) which is the lowest compared to the others (Table 3.17). **Hence item 16 does not show evidence of convergent or divergent validity.**

Item 17 did not correlate well with item 16 and 18 with correlation values of 0.348 ($p < 0.01$) and 0.231 ($p < 0.05$). It correlated better with Item 4 and 5 (LL), Item 7 (UL) and Item 14 (RL)

with correlation values of 0.514, 0.542, 0.525 and 0.522 respectively ($p < 0.01$). Corrected item-trait correlation was 0.387 ($p < 0.01$) which was the lowest. **Item 17 also did not show evidence of convergent or divergent validity.**

Item 18 also correlated poorly with item 16 and 17 with correlation values of 0.153 ($p > 0.05$) and 0.231 ($p < 0.05$). It had a good correlation with item 9 (UL) with values of 0.597 ($p < 0.01$) but poor correlation with the rest of the items. Corrected Item-trait correlation was 0.216 ($p < 0.05$) which was the poorest. **Item 18 did not show evidence of convergent or divergent validity for it's domain.**

GH domain

Item 19 and 20 showed a good correlation with value of 0.504 ($p < 0.01$) and neither items showed any profound correlation with other items in other domains except of item 14 (RL) which correlated a little better with item 20, correlation value of 0.532 ($p < 0.01$). The corrected item-trait correlation for item 19 was higher when compared to the other domains, however for item 20 the correlation was better with UL and RL domain with values of 0.546 and 0.522 respectively ($p < 0.01$). **Item 19 demonstrated good convergent and divergent validity, but Item 20 showed good convergent validity but appears to correlate also with UL and RL domain.**

Summary of Inter Item Correlation Analysis

All the items in the Lower Limb domain demonstrated good evidence for convergent and divergent validity with item-trait correlation highest with its domain compared to the rest. Item 7 from Upper Limb domain clearly demonstrated that it does not sit comfortably within its domain and correlated better with Lower limb and Role Limitation items/domain. Item 12 from Role Limitation domain correlated better with Lower Limb domain and does not demonstrated good convergent or divergent validity. The rest of the items in RL domain appears to sit well in its domain. All the three pain items do not correlate well with each other and had poor item-trait correlation. For General Health only item 19 was quite consistent but item 20 appeared to be correlate better with UL and RL domain.

Domain / Trait	Item		LL Score	UL Score	RL Score	Pain Score	GH Score	Corrected Item-Trait
Lower Limb	Qn1	Correlation Coefficient	.785**	.416**	.353**	.402**	.261**	.686**
		Sig. (2-tailed)	0	0	0	0	0.004	0
		N	120	120	120	120	120	120
	Qn2	Correlation Coefficient	.729**	.501**	.428**	.442**	.270**	.581**
		Sig. (2-tailed)	0	0	0	0	0.003	0
		N	120	120	120	120	120	120
	Qn3	Correlation Coefficient	.799**	.497**	.543**	.477**	.315**	.720**
		Sig. (2-tailed)	0	0	0	0	0	0
		N	120	120	120	120	120	120
	Qn4	Correlation Coefficient	.737**	.478**	.569**	.441**	.289**	.604**
		Sig. (2-tailed)	0	0	0	0	0.001	0
		N	120	120	120	120	120	120
	Qn5	Correlation Coefficient	.809**	.605**	.681**	.574**	.441**	.730**
		Sig. (2-tailed)	0	0	0	0	0	0
		N	120	120	120	120	120	120
	Qn6	Correlation Coefficient	.736**	.530**	.575**	.477**	.367**	.613**
		Sig. (2-tailed)	0	0	0	0	0	0
		N	120	120	120	120	120	120

Table 3.14 MTMI Correlation matrix for MP20 - LL Domain

Domain / Trait	Item		LL Score	UL Score	RL Score	Pain Score	GH Score	Corrected Item-Trait
Upper Limb	Qn7	Correlation Coefficient	.662**	.813**	.647**	.518**	.373**	.567**
		Sig. (2-tailed)	0	0	0	0	0	0
		N	120	120	120	120	120	120
	Qn8	Correlation Coefficient	.329**	.716**	.442**	.423**	.448**	.643**
		Sig. (2-tailed)	0	0	0	0	0	0
		N	120	120	120	120	120	120
	Qn9	Correlation Coefficient	.432**	.769**	.531**	.538**	.455**	.643**
		Sig. (2-tailed)	0	0	0	0	0	0
		N	120	120	120	120	120	120
	Qn10	Correlation Coefficient	.331**	.642**	.426**	.383**	.386**	.602**
		Sig. (2-tailed)	0	0	0	0	0	0
		N	120	120	120	120	120	120
	Qn11	Correlation Coefficient	.508**	.785**	.565**	.484**	.476**	.656**
		Sig. (2-tailed)	0	0	0	0	0	0
		N	120	120	120	120	120	120

Table 3.15 MTMI Correlation matrix for MP20 - UL domain

Domain / Trait	Item		LL Score	UL Score	RL Score	Pain Score	GH Score	Corrected Item-Trait
Role Limitation	Qn12	Correlation Coefficient	.625**	.572**	.765**	.509**	.413**	.599**
		Sig. (2-tailed)	0	0	0	0	0	0
		N	120	120	120	120	120	120
	Qn13	Correlation Coefficient	.515**	.484**	.758**	.417**	.336**	.611**
		Sig. (2-tailed)	0	0	0	0	0	0
		N	120	120	120	120	120	120
	Qn14	Correlation Coefficient	.605**	.678**	.864**	.595**	.511**	.745**
		Sig. (2-tailed)	0	0	0	0	0	0
		N	120	120	120	120	120	120
	Qn15	Correlation Coefficient	.519**	.604**	.837**	.521**	.481**	.628**
		Sig. (2-tailed)	0	0	0	0	0	0
		N	120	120	120	120	120	120

Table 3.16 MTMI Correlation matrix for MP20 - RL domain

Domain / Trait	Item		LL Score	UL Score	RL Score	Pain Score	GH Score	Corrected Item-Trait
Pain	Qn16	Correlation Coefficient	.446**	.355**	.413**	.806**	.360**	.336**
		Sig. (2-tailed)	0	0	0	0	0	0
		N	120	120	120	120	120	120
	Qn17	Correlation Coefficient	.563**	.494**	.534**	.667**	.330**	.387**
		Sig. (2-tailed)	0	0	0	0	0	0
		N	120	120	120	120	120	120
	Qn18	Correlation Coefficient	.321**	.521**	.421**	.571**	.312**	.216*
		Sig. (2-tailed)	0	0	0	0	0.001	0
		N	120	120	120	120	120	120

Table 3.17 MTMI Correlation matrix of MP20 - Pain Domain

Domain / Trait	Item		LL Score	UL Score	RL Score	Pain Score	GH Score	Corrected Item-Trait
General Health	Qn19	Correlation Coefficient	.255**	.381**	.363**	.317**	.811**	.504**
		Sig. (2-tailed)	0.005	0	0	0	0	0.001
		N	120	120	120	120	120	120
	Qn20	Correlation Coefficient	.461**	.546**	.552**	.488**	.893**	.504**
		Sig. (2-tailed)	0	0	0	0	0	0
		N	120	120	120	120	120	120

Table 3.18 MTMI Correlation matrix for MP20 - GH domain

3.8.2 Multi Trait Multi Method Analysis (MTMM)

Next we analyzed the correlations between the domains to assess convergence and divergence characteristics, and then we will analyse their relationship with other measuring instruments (PROM) to assess further construct validity.

Table 3.19 shows the correlation between each domain scores to the Total MP20 scores and here we'd look at the Corrected domain-total scores to minimize errors arising from the scores included. We can see that overall the correlation between each domain to one another is quite good between LL, UL, RL and Pain domain with correlation values ranging from 0.611 to 0.708 ($p < 0.01$). However with GH domain, it's correlation with LL, UL, RL and Pain domain appears to be less strong with correlation values ranging from 0.422 to 0.532 ($p < 0.01$). The domains also correlated well with the total score not surprisingly with correlation values of 0.728, 0.766, 0.777, 0.697 and 0.564 for LL, UL, RL, Pain and GH domains respectively ($p < 0.01$). Note again that the GH domain's correlation is the lowest.

Spearman's rho		LL Score	UL Score	RL Score	Pain Score	GH Score	New PROM Total	Corrected Domain-Total
LL Score	Correlation Coefficient	1	.652**	.676**	.611**	.422**	.859**	.728**
	Sig. (2-tailed)	.	0	0	0	0	0	0
	N	120	120	120	120	120	120	120
UL Score	Correlation Coefficient	.656**	1	.705**	.611**	.533**	.880**	.766**
	Sig. (2-tailed)	0	.	0	0	0	0	0
	N	120	120	120	120	120	120	120
RL Score	Correlation Coefficient	.676**	.708**	1	.621**	.530**	.874**	.777**
	Sig. (2-tailed)	0	0	.	0	0	0	0
	N	120	120	120	120	120	120	120
Pain Score	Correlation Coefficient	.611**	.606**	.621**	1	.461**	.762**	.697**
	Sig. (2-tailed)	0	0	0	.	0	0	0
	N	120	120	120	120	120	120	120
GH Score	Correlation Coefficient	.422**	.532**	.530**	.461**	1	.656**	.564**
	Sig. (2-tailed)	0	0	0	0	.	0	0
	N	120	120	120	120	120	120	120

** Correlation is significant at the 0.01 level (2-tailed).

Table 3.19 Inter Domain Correlation Matrix for MP20

Spearman's rho		LL Score Baseline	UL Score Baseline	RL Score Baseline	Pain Score Baseline	GH Score Baseline	MP20 Total Score Baseline	BMP Index Joint Score - Baseline	VAS Satisfaction - Baseline	Pain Score WOM Baseline	Stiffnes Score WOM Baseline	ADL Score WOM Baseline	WOMAC Total Score Baseline	PCS-SF12 score Baseline	MCS-SF12 score Baseline	Oxford Total Scores - Baseline
LL Score Baseline	Correlation Coefficient	1	.656**	.676**	.611**	.422**	.859**	-.340**	.522**	.611**	.558**	.749**	.734**	.420**	.469**	.718**
	Sig. (2-tailed)	.	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	N	120	120	120	120	120	120	119	119	120	120	120	120	120	120	120
UL Score Baseline	Correlation Coefficient	.656**	1	.705**	.611**	.533**	.862**	-.383**	.456**	.552**	.464**	.628**	.628**	.393**	.428**	.652**
	Sig. (2-tailed)	0	.	0	0	0	0	0	0	0	0	0	0	0	0	0
	N	120	120	120	120	120	120	119	119	120	120	120	120	120	120	120
RL Score Baseline	Correlation Coefficient	.676**	.705**	1	.621**	.530**	.874**	-.404**	.480**	.557**	.410**	.702**	.679**	.488**	.490**	.744**
	Sig. (2-tailed)	0	0	.	0	0	0	0	0	0	0	0	0	0	0	0
	N	120	120	120	120	120	120	119	119	120	120	120	120	120	120	120
Pain Score Baseline	Correlation Coefficient	.611**	.611**	.621**	1	.461**	.762**	-.469**	.372**	.702**	.537**	.706**	.722**	.371**	.439**	.711**
	Sig. (2-tailed)	0	0	0	.	0	0	0	0	0	0	0	0	0	0	0
	N	120	120	120	120	120	120	119	119	120	120	120	120	120	120	120
GH Score Baseline	Correlation Coefficient	.422**	.533**	.530**	.461**	1	.656**	-.305**	.381**	.409**	.285**	.427**	.426**	0.159	.565**	.463**
	Sig. (2-tailed)	0	0	0	0	.	0	0.001	0	0	0.002	0	0	0.084	0	0
	N	120	120	120	120	120	120	119	119	120	120	120	120	120	120	120
MP20 Total Score Baseline	Correlation Coefficient	.859**	.862**	.874**	.762**	.656**	1	-.437**	.546**	.665**	.541**	.776**	.767**	.474**	.550**	.797**
	Sig. (2-tailed)	0	0	0	0	0	.	0	0	0	0	0	0	0	0	0
	N	120	120	120	120	120	120	119	119	120	120	120	120	120	120	120
BMP Index Joint Score - Baseline	Correlation Coefficient	-.340**	-.383**	-.404**	-.469**	-.305**	-.437**	1	-.304**	-.497**	-.476**	-.477**	-.504**	-0.044	-.313**	-.520**
	Sig. (2-tailed)	0	0	0	0	0.001	0	.	0.001	0	0	0	0	0.638	0.001	0
	N	119	119	119	119	119	119	119	118	119	119	119	119	119	119	119
VAS Satisfaction - Baseline	Correlation Coefficient	.522**	.456**	.480**	.372**	.381**	.546**	-.304**	1	.439**	.348**	.463**	.471**	.328**	.407**	.513**
	Sig. (2-tailed)	0	0	0	0	0	0	0.001	.	0	0	0	0	0	0	0
	N	119	119	119	119	119	119	118	119	119	119	119	119	119	119	119
Pain Score WOM Baseline	Correlation Coefficient	.611**	.552**	.557**	.702**	.409**	.665**	-.497**	.439**	1	.736**	.844**	.905**	.368**	.436**	.767**
	Sig. (2-tailed)	0	0	0	0	0	0	0	0	.	0	0	0	0	0	0
	N	120	120	120	120	120	120	119	119	120	120	120	120	120	120	120
Stiffnes Score WOM Baseline	Correlation Coefficient	.558**	.464**	.410**	.537**	.285**	.541**	-.476**	.348**	.736**	1	.694**	.758**	.311**	.322**	.666**
	Sig. (2-tailed)	0	0	0	0	0.002	0	0	0	0	.	0	0	0.001	0	0
	N	120	120	120	120	120	120	119	119	120	120	120	120	120	120	120
ADL Score WOM Baseline	Correlation Coefficient	.749**	.628**	.702**	.706**	.427**	.776**	-.477**	.463**	.844**	.694**	1	.988**	.428**	.535**	.856**
	Sig. (2-tailed)	0	0	0	0	0	0	0	0	0	0	.	0	0	0	0
	N	120	120	120	120	120	120	119	119	120	120	120	120	120	120	120
WOMAC Total Score Baseline	Correlation Coefficient	.734**	.628**	.679**	.722**	.426**	.767**	-.504**	.471**	.905**	.758**	.988**	1	.418**	.518**	.863**
	Sig. (2-tailed)	0	0	0	0	0	0	0	0	0	0	0	.	0	0	0
	N	120	120	120	120	120	120	119	119	120	120	120	120	120	120	120
PCS-SF12 score Baseline	Correlation Coefficient	.420**	.393**	.488**	.371**	0.159	.474**	-0.044	.328**	.368**	.311**	.428**	.418**	1	-0.007	.490**
	Sig. (2-tailed)	0	0	0	0	0.084	0	0.638	0	0	0.001	0	0	.	0.942	0
	N	120	120	120	120	120	120	119	119	120	120	120	120	120	120	120
MCS-SF12 score Baseline	Correlation Coefficient	.469**	.428**	.490**	.439**	.565**	.550**	-.313**	.407**	.436**	.322**	.535**	.518**	-0.007	1	.498**
	Sig. (2-tailed)	0	0	0	0	0	0	0.001	0	0	0	0	0	0.942	.	0
	N	120	120	120	120	120	120	119	119	120	120	120	120	120	120	120
Oxford Total Scores - Baseline	Correlation Coefficient	.718**	.652**	.744**	.711**	.463**	.797**	-.520**	.513**	.767**	.666**	.856**	.863**	.490**	.498**	1
	Sig. (2-tailed)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.
	N	120	120	120	120	120	120	119	119	120	120	120	120	120	120	120

Table 3.20 Multiple Correlation Matrix between different Measures

We will now analyze the results of correlations between the new PROMs domain versus other measures (Table 3.21).

The LL domain correlated well Oxford Scores with correlation value of 0.718 ($p < 0.01$) which is what we we'd expected but correlated even better with WOMAC-ADL domain and WOMAC-Total score with values of 0.749 and 0.734 respectively ($p < 0.01$). The LL domain did not correlate well with any of the SF-12 components (PCS=0.420, MCS=0.469).

The UL domain did not get as good correlation as LL had with Oxford and WOMAC, with values of 0.652 and 0.628 ($p < 0.01$), but this was expected as it is supposed to assess Upper limb function.

RL domain demonstrated good correlation with WOMAC-ADL, WOMAC-TOTAL and Oxford Scores with values of 0.702, 0.679 and 0.744 respectively ($p < 0.01$).

Pain domain correlated well with WOMAC-PAIN with correlation value of 0.702 ($p < 0.01$) and it also correlated well with WOMAC-TOTAL and Oxford Scores.
GH domain did not show any strong correlation with any of the existing measures, except for SF12-MCS (mental health component) with a good correlation value of 0.565 ($p < 0.01$). This indicate that GH domain is distinctively different from the rest and is more similar to a generic general health theme than a specific physical symptom type of question.

The MP20 score which is the total score of the five domains summed together was correlated against the other measures and the results were quite good across the board. MP20 correlated best with Oxford Scores with values of 0.797 ($p < 0.01$) and also quite good with WOMAC-TOTAL (0.767, $p < 0.01$).

The Body Map Pain Index Joint Score was a pain score (0 to 10) given by patients for their Index joint, and their baseline correlation measures were analyzed. The correlation was negative values because more severe pain was given a value closer to 10 whereas outcome measure scores higher for better function. The correlation coefficient was found to be modest to good with WOMAC-Pain domain with value of -0.497 ($p < 0.01$).

The VAS for satisfaction of joint function also was found to correlate modestly with WOMAC and Oxford Scores with correlation values of 0.471 and 0.513 respectively ($p < 0.01$). This is good because we did not expect either VAS or BMP scores to correlate that well with the existing outcome measure, as we know they are measuring a different health concept. VAS for satisfaction was meant to measure level of satisfaction one has with their joint function and it is a very subjective question, but a very important question that we ask all our patients following their joint replacement. However it is encouraging to see that it did not differ too much away so as to measure a completely unrelated trait.

Finally we looked at the correlation between WOMAC and Oxford (Table 3.22) and as expected it showed very good correlation value of 0.863 ($p < 0.01$) indicating very good evidence that these PROMs are measuring similar concepts.

Summary of Multi Trait Multi Method Analysis (MTMM)

The five domains of MP20 showed overall good correlation with the total score (corrected domain-total correlation), however the GH domain displayed weakest correlation with the total score as well as with other domains. This demonstrates that the general health questions are clearly covering a different dimensions concept compared to concept covered by LL, UL, RL and Pain which are more closely related to physical functioning. And hence provide more evidence for the multi-dimensional concept of this new PROM.

LL domain showed good evidence of construct validity with high correlation with Oxford and WOMAC scores, and so did Role Limitation items. The Pain items had good correlation with WOMAC pain and with Oxford Scores. The 2 other components to the new PROM, BMP and VAS was indeed an attempt to make the new PROM more 'holistic' giving opportunity for patients to report other functional issues and general health that was not covered yet. The results of it's correlation analysis demonstrated that both components correlated modestly with the existing standard questionnaires. This again provides more evidence that MP20 when combined with BMP and VAS provide more information, and not just the same information about the patients overall functional status.

NEW PROMS Domains		Pain Score WOMAC Baseline	Stiffnes Score WOMAC Baseline	ADL Score WOMAC Baseline	WOMAC Total Score Baseline	PCS-SF12 score Baseline	MCS-SF12 score Baseline	Oxford Total Scores - Baseline
LL Score Baseline	Correlation Coefficient	.611**	.558**	**.749****	**.734****	.420**	.469**	**.718****
	Sig. (2-tailed)	0	0	0	0	0	0	0
	N	120	120	120	120	120	120	120
UL Score Baseline	Correlation Coefficient	.552**	.464**	.628**	.628**	.393**	.428**	.652**
	Sig. (2-tailed)	0	0	0	0	0	0	0
	N	120	120	120	120	120	120	120
RL Score Baseline	Correlation Coefficient	.557**	.410**	.702**	.679**	.488**	.490**	**.744****
	Sig. (2-tailed)	0	0	0	0	0	0	0
	N	120	120	120	120	120	120	120
Pain Score Baseline	Correlation Coefficient	**.702****	.537**	.706**	**.722****	.371**	.439**	**.711****
	Sig. (2-tailed)	0	0	0	0	0	0	0
	N	120	120	120	120	120	120	120
GH Score Baselne	Correlation Coefficient	.409**	.285**	.427**	.426**	0.159	.565**	.463**
	Sig. (2-tailed)	0	0.002	0	0	0.084	0	0
	N	120	120	120	120	120	120	120
MP20 Total Score Baseline	Correlation Coefficient	.665**	.541**	**.776****	**.767****	.474**	.550**	**.797****
	Sig. (2-tailed)	0	0	0	0	0	0	0
	N	120	120	120	120	120	120	120
BMP Index Joint Score - Baseline	Correlation Coefficient	-.497**	-.476**	-.477**	-.504**	-0.044	-.313**	-.520**
	Sig. (2-tailed)	0	0	0	0	0.638	0.001	0
	N	119	119	119	119	119	119	119
VAS Satisfaction - Baseline	Correlation Coefficient	.439**	.348**	.463**	.471**	.328**	.407**	.513**
	Sig. (2-tailed)	0	0	0	0	0	0	0
	N	119	119	119	119	119	119	119

Table 3.21 Multi Trait Multi Method (MTMM) Correlation Matrix

OTHER PROMS / Domains		Pain Score WOMAC Baseline	Stiffnes Score WOMAC Baseline	ADL Score WOMAC Baseline	WOMAC Total Score Baseline	PCS-SF12 score Baseline	MCS-SF12 score Baseline	Oxford Total Scores - Baseline
Pain Score WOM Baseline	Correlation Coefficient	1	.736**	.844**	.905**	.368**	.436**	.767**
	Sig. (2-tailed)	.	0	0	0	0	0	0
	N	120	120	120	120	120	120	120
Stiffnes Score WOM Baseline	Correlation Coefficient	.736**	1	.694**	.758**	.311**	.322**	.666**
	Sig. (2-tailed)	0	.	0	0	0.001	0	0
	N	120	120	120	120	120	120	120
ADL Score WOM Baseline	Correlation Coefficient	.844**	.694**	1	.988**	.428**	.535**	.856**
	Sig. (2-tailed)	0	0	.	0	0	0	0
	N	120	120	120	120	120	120	120
WOMAC Total Score Baseline	Correlation Coefficient	.905**	.758**	.988**	1	.418**	.518**	**.863****
	Sig. (2-tailed)	0	0	0	.	0	0	0
	N	120	120	120	120	120	120	120
PCS-SF12 score Baseline	Correlation Coefficient	.368**	.311**	.428**	.418**	1	-0.007	.490**
	Sig. (2-tailed)	0	0.001	0	0	.	0.942	0
	N	120	120	120	120	120	120	120
MCS-SF12 score Baseline	Correlation Coefficient	.436**	.322**	.535**	.518**	-0.007	1	.498**
	Sig. (2-tailed)	0	0	0	0	0.942	.	0
	N	120	120	120	120	120	120	120
Oxford Total Scores - Baseline	Correlation Coefficient	.767**	.666**	.856**	.863**	.490**	.498**	1
	Sig. (2-tailed)	0	0	0	0	0	0	.
	N	120	120	120	120	120	120	120

Table 3.22 MTMM Correlation Matrix for Other PROMs

3.9 Responsiveness Results

Out of the sixty participants recruited in this cohort, thirty-eight participants responded to the follow up PROMs at 6 months post-surgery which gives a response rate of 63.3 %. This rate of response is similar to the PROM programme in our institution.

Missing Items

From the 38 datasets available, there was no missing items in MP20 questionnaires, however there were 4 missing data in Body Map Index Joint score (1 baseline and 3 Post op), one missing values of VAS (at baseline), 2 missing values in WOMAC scales (at postop), three missing values of SF12 (at postop) and one missing Oxford Score values (at postop). The proportion of missing items for MP20, BMP Index Score, VAS, WOMAC, SF12 and Oxford Scores were, 0%, 5.3%, 1.3%, 2.6%, 3.9% and 1.3% respectively.

We then analyzed the paired data individually (Table 3.24) and identified the pattern of direction following treatment. Positive rank is the outcome when the outcome measure is higher following treatment indicating that the patient has improved. Negative rank would mean that the patient has not improve and would be concerning. Ties indicate that there has been no difference in the outcome following treatment.

Next we analyzed statistically the standardized difference between these paired means by calculating their Effect Size (ES) and Standardised Ratio Mean (SRM). We calculated Effect size by measuring the standardized difference between the paired means divided by the standard deviation of the original mean. This gives the Cohen's d value of ES. We calculated the Standardised Ratio Mean (SRM) by calculating the mean difference between the paired sample and dividing it by the standard deviation of the mean difference. To interpret the ES values we used cohen's reference value of 0.2, 0.5 and 0.8 to indicate small, moderate and large effect size. There is no perfect consensus literature however it's generally agreed that we use same reference values for SRM as well. We used a paired sample t-test to derive the figures which are summarized in table Table 3.25.

We analysed the mean difference of the paired sample and negative values indicate the direction of scores from a smaller score indicating a poorer outcome to a larger score indicating that the participant has improved. This is generally the expected trend following treatment. In the event that the mean difference is positive it means the outcome following surgical intervention is now worst, and if the mean difference is zero, there is no difference in outcome.

New PRO measure

General Health (GH)

Within the domains of the MP20 PROMs, the General Health domain showed the weakest responsiveness with equal number of patients (11 each) displaying positive and negative ranks and sixteen patients recorded no difference in their GH domains post operatively. The mean difference was 0.16 (CI – 0.35 to 0.67) and the **cohen's d value was 0.1 with Standard Ratio Mean (SRM) of 0.1**. However, the difference is not statistically significant with p-value of 0.53.

Lower Limb (LL)

The LL domain performed best with 33 positive ranks and only 2 negative ranks and only 3 showing no difference in LL scores (Table 3.24). The mean difference between the baseline and postoperative scores was -5.63 (CI -7.39 to -3.88), and the **cohen's d value was -1.13 and the SRM was -1.06** (p-value < 0.01), indicating a large effect size.

Upper Limb (UL)

The UL domain recorded 26 positive ranks and 6 negative ranks with 6 ties (Table 3.24). The mean difference was -1.74 (CI -2.87 to -0.6), and the cohen's d was -0.49 and SRM was 0.5 (p-value < 0.01) indicating a moderate effect size.

Role Limitation (RL)

RL reported 32 positive ranks and only 4 negative ranks with 2 ties. The mean differences was -4.13 (CI -5.52 to -2.74) and **cohen's d was -1.05 with SRM value of -0.98** (p value < 0.01), indicating a large effect size.

Pain

Pain domain recorded 30 positive ranks, 6 negative ranks and 2 ties. The mean difference was -2.24 (CI -3.03 to -2.45) and **cohens d was -1.07 with SRM of -0.93** (p values < 0.01). This indicate a large effect size for the pain domain.

MP20

The MP20 total score also performed well with 34 positive ranks (better outcome) and 4 negative ranks (poorer outcome). The mean difference between baseline and postoperative scores was -13.16 (CI -17.54 to -8.87) and the **Cohen's d value was -1.01 with SRM of -0.99**, indicating a large effect size.

BMP and VAS

For the other 2 components of the new PRO measure we'd expect the direction of scores for BMP index joint should be negative as you'd expect the index joint score to be smaller following treatment. Conversely for VAS of satisfaction, the score is expected to be larger

postoperatively, i.e. positive ranks. There were only 34 patients that filled in the BMP scores (4 missing) and all showed negative ranks i.e. improvement in their index joint pain score following surgery. The mean difference in the BMP index joint score between baseline and postoperatively was 6.59 (CI 5.72 to 7.46) and the **Cohen's d value was 5.87 with SRM of 2.63** (p value < 0.01). This demonstrate significantly large effect size of the BMP index joint score. The VAS scores also reported well with 31 positive ranks, 3 negative ranks, and 3 felt no difference in level of satisfaction of joint function post-surgery. The mean difference in the VAS score was -30.30 (CI -41.21 to -19.38) with **Cohen's d value of -1.24 and SRM value of -0.93** (p value 0.01), indicating significantly large effect size.

WOMAC, SF12 and Oxford Scores

We compared these results of New PRO measure with our 'control' measures and found the results we comparable. Two items were missing in the WOMAC scores from the 38 respondents however all gave a positive rank with WOMAC total scores (Table 3.24). The mean difference for WOMAC total scores was -30.49 (CI-36.29 to -24.68) and the cohen's d value **was -1.65 with SRM of -1.78**, indicating a significantly large effect size (Table 3.25). The WOMAC domains also showed large effect sizes with the Pain, Stiffness and ADL recording 35, 30 and 35 positive ranks and SRM values of -1.29, -1.35 and -1.69 respectively.

There were 3 missing items in the SF12 scores and are reported as Mental component scores (MCS) and Physical Component scores (PCS). MCS performed poorly reporting only 19 positive ranks, 15 negative ranks and 1 tie. The mean difference of MCS scores was -1.67 (CI -6.47 to 3.13) and Cohen's value of -0.14 and SRM of -0.12. However, the differences were not statistically significant with p value of 0.485. The PCS performed much better reporting 27 positive ranks, 7 negative ranks and no ties. The mean difference of PCS scores was -9.56 (CI -13.29 to -5.84) with **Cohen's d value of -1.15 and SRM -0.88**, indicating a large effect size.

And finally the Oxford scores performed very well with 34 positive ranks (better outcome) and 3 negative ranks (poorer outcome) no ties. Their mean differences between the baseline and post operative scores was -15.73 (CI -19.28 to -12.18) and the Cohen's d value was **-1.80 and SRM -1.48**, demonstrating a large effect size.

Summary of Responsiveness study

All the PRO measures demonstrated large effect size with Cohen's d values of > 0.80 with exception of SF12 MCS which reported a cohen's d value of **-0.14 and SRM of -0.12**. The most responsive PRO measure was the Oxford score with cohen's d value of **-1.80 and SRM of -1.48**. WOMAC total scores is next with cohen's d value of **-1.65 and SRM of -1.78**. MP20 demonstrated a slightly lesser level of responsiveness compared to Oxford and WOMAC with cohen's d value of **was -1.01 with SRM of -0.99.** SF12-PCS also reported a similar effect size to MP20 with cohen's d value **of -1.15 and SRM -0.88.** These results reflect the construct of each type of PRO measure where it was clear that a more disease specific joint PROM like Oxford and WOMAC demonstrated a higher level responsiveness where reporting physical symptoms are much more specific and less ambiguous. The PRO measure like MP20 and SF12 PCS showed lesser degree of responsiveness given the fact that these measures are aimed at a more overall measure of physical function.

At domain/component level the most responsive was the BMP index joint score with cohen's d value of **5.87 and SRM of 2.63.** The VAS of satisfaction demonstrated good responsiveness with **Cohen's d value of -1.24 and SRM value of -0.93.** LL, RL and Pain domain also demonstrated large effect size (d > 0.8) albeit lesser level of responsiveness with cohen's d value of **-1.13, -1.05 and -1.07 respectively with SRM values of -1.06, -0.98 and -0.93 respectively.** This is expectedly not as high as WOMAC domains of Pain, Stiffness and Role limitation which reported **cohen's d value of -8.36, -1.36 and -1.57** SRM values of -1.29, -1.35 and -1.69 respectively. The UL domain of MP20 was moderately responsive with cohens d value of -0.49 and SRM was 0.5. The domain with the least evidence of responsiveness was the GH domain which showed very low Effect size however the results were not statistically significant. So only LL, UL, RL and Pain domain showed adequate evidence for level of responsiveness, but GH domain did not display any evidence of responsiveness. This may reflect the difficulty of wording the questions that aimed to gather information about one's general health. Even the SF12-MCS did not demonstrate good evidence of responsiveness and this may reflect the similarity this PRO measure has with GH domain.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	LL Score Baseline	13.18	38	4.98	0.81
	LL Score Post	18.82	38	4.61	0.75
Pair 2	UL Score Baseline	16.61	38	3.52	0.57
	UL Score Post	18.34	38	3.03	0.49
Pair 3	RL Score Baseline	7.47	38	3.95	0.64
	RL Score Post	11.61	38	4.02	0.65
Pair 4	Pain Score Baseline	7.03	38	2.09	0.34
	Pain Score Post	9.26	38	2.37	0.38
Pair 5	GH Score Baselline	6.71	38	1.51	0.24
	GH Score Post	6.55	38	1.77	0.29
Pair 6	MP20 Total Score Baseline	51.00	38	13.06	2.12
	MP20 Total Score - Post	64.16	38	14.00	2.27
Pair 7	BMP Index Joint Score - Baseline	8.79	34	1.12	0.19
	Index Jt Score Post	2.21	34	2.07	0.36
Pair 8	VAS Satisfaction - Baseline	41.51	37	24.52	4.03
	VAS Satisfaction Post	71.81	37	23.44	3.85
Pair 9	Pain Score WOM Baseline	8.74	36	4.21	0.70
	Pain Score WOMAC Post	43.94	36	28.14	4.69
Pair 10	Stiffnes Score WOM Baseline	3.00	36	1.66	0.28
	Stiffnes Score WOMAC Post	5.25	36	1.86	0.31
Pair 11	ADL Score WOM Baseline	27.81	36	13.70	2.28
	ADL Score WOMAC Post	49.36	36	13.31	2.22
Pair 12	WOMAC Total Score Baseline	39.54	36	18.49	3.08
	WOMAC Total Score - Post	70.03	36	17.67	2.94
Pair 13	PCS-SF12 score Baseline	29.60	35	8.32	1.41
	PCS score Post	39.16	35	10.85	1.83
Pair 14	MCS-SF12 score Baseline	49.59	35	11.62	1.96
	MCS score Post	51.25	35	10.79	1.82
Pair 15	Oxford Total Scores - Baseline	17.51	37	8.76	1.44
	Oxford Total Scores Postop	33.24	37	9.16	1.51

Table 3.23 Paired sample statistics for Responsiveness Cohort (Group B)

		N	Mean Rank	Sum of Ranks
LL Score Post - LL Score Baseline	Negative Ranks	2	14.75	29.5
	Positive Ranks	33	18.2	600.5
	Ties	3		
	Total	38		
UL Score Post - UL Score Baseline	Negative Ranks	6	18.33	110
	Positive Ranks	26	16.08	418
	Ties	6		
	Total	38		
RL Score Post - RL Score Baseline	Negative Ranks	4	10.13	40.5
	Positive Ranks	32	19.55	625.5
	Ties	2		
	Total	38		
Pain Score Post - Pain Score Baseline	Negative Ranks	6	10.92	65.5
	Positive Ranks	30	20.02	600.5
	Ties	2		
	Total	38		
GH Score Post - GH Score Baselne	Negative Ranks	11	13.14	144.5
	Positive Ranks	11	9.86	108.5
	Ties	16		
	Total	38		
MP20 Total Score - Post - MP20 Total Score Baseline	Negative Ranks	4	16.63	66.5
	Positive Ranks	34	19.84	674.5
	Ties	0		
	Total	38		
Index Jt Score Post - BMP Index Joint Score - Baseline	Negative Ranks	34	17.5	595
	Positive Ranks	0	0	0
	Ties	0		
	Total	34		
VAS Satisfaction Post - VAS Satisfaction - Baseline	Negative Ranks	3	17.83	53.5
	Positive Ranks	31	17.47	541.5
	Ties	3		
	Total	37		
Pain Score WOMAC Post - Pain Score WOM Baseline	Negative Ranks	1	3	3
	Positive Ranks	35	18.94	663
	Ties	0		
	Total	36		
Stiffnes Score WOMAC Post - Stiffnes Score WOM Baseline	Negative Ranks	1	4.5	4.5
	Positive Ranks	30	16.38	491.5
	Ties	5		
	Total	36		
ADL Score WOMAC Post - ADL Score WOM Baseline	Negative Ranks	0	0	0
	Positive Ranks	35	18	630
	Ties	1		
	Total	36		
WOMAC Total Score - Post - WOMAC Total Score Baseline	Negative Ranks	0	0	0
	Positive Ranks	36	18.5	666
	Ties	0		
	Total	36		
PCS score Post - PCS-SF12 score Baseline	Negative Ranks	7	9.86	69
	Positive Ranks	27	19.48	526
	Ties	1		
	Total	35		
MCS score Post - MCS-SF12 score Baseline	Negative Ranks	15	16.77	251.5
	Positive Ranks	19	18.08	343.5
	Ties	1		
	Total	35		
Oxford Total Scores Postop - Oxford Total Scores - Baseline	Negative Ranks	3	5.17	15.5
	Positive Ranks	34	20.22	687.5
	Ties	0		
	Total	37		

Table 3.24 Ranking between paired scores

Negative ranks indicate that the post op score is smaller than baseline i.e. expected in BMP

Positive ranks indicate the post op score being bigger than baseline i.e. expected in MP20, VAS, WOMAC and Oxford

Ties indicated no difference in baseline and postop scores.

		Paired Differences						t	df	Sig. (2-tailed)	Effect Size	SRM
		Mean	Std. Deviation	SD of baseline	Std. Error Mean	95% Confidence Interval of the Difference						Standardised Response Mean
						Lower	Upper					
Pair 1	LL Score Baseline - LL Score Post	-5.63	5.33	4.98	0.87	-7.39	-3.88	-6.51	37	0	-1.13	**-1.06**
Pair 2	UL Score Baseline - UL Score Post	-1.74	3.45	3.52	0.56	-2.87	-0.60	-3.11	37	0.004	-0.49	**-0.50**
Pair 3	RL Score Baseline - RL Score Post	-4.13	4.23	3.95	0.69	-5.52	-2.74	-6.02	37	0	-1.05	**-0.98**
Pair 4	Pain Score Baseline - Pain Score Post	-2.24	2.41	2.09	0.39	-3.03	-1.45	-5.72	37	0	-1.07	**-0.93**
Pair 5	GH Score Baseline - GH Score Post	0.16	1.55	1.51	0.25	-0.35	0.67	0.63	37	0.534	0.10	**0.10**
Pair 6	MP20 Total Score Baseline - MP20 Total Score - Post	-13.16	13.32	13.06	2.16	-17.54	-8.78	-6.09	37	0	**-1.01**	**-0.99**
Pair 7	BMP Index Joint Score - Baseline - Index Jt Score Post	6.59	2.49	1.12	0.43	5.72	7.46	15.44	33	0	**5.87**	**2.65**
Pair 8	VAS Satisfaction - Baseline - VAS Satisfaction Post	-30.30	32.74	24.52	5.38	-41.21	-19.38	-5.63	36	0	**-1.24**	**-0.93**
Pair 9	Pain Score WOM Baseline - Pain Score WOMAC Post	-35.21	27.33	4.21	4.55	-44.45	-25.96	-7.73	35	0	-8.36	**-1.29**
Pair 10	Stiffnes Score WOM Baseline - Stiffnes Score WOMAC Post	-2.25	1.66	1.66	0.28	-2.81	-1.69	-8.12	35	0	-1.36	**-1.35**
Pair 11	ADL Score WOM Baseline - ADL Score WOMAC Post	-21.56	12.76	13.70	2.13	-25.87	-17.24	-10.13	35	0	-1.57	**-1.69**
Pair 12	WOMAC Total Score Baseline - WOMAC Total Score - Post	-30.49	17.16	18.49	2.86	-36.29	-24.68	-10.66	35	0	**-1.65**	**-1.78**
Pair 13	PCS-SF12 score Baseline - PCS score Post	-9.56	10.85	8.32	1.83	-13.29	-5.84	-5.21	34	0	-1.15	**-0.88**
Pair 14	MCS-SF12 score Baseline - MCS score Post	-1.67	13.98	11.62	2.36	-6.47	3.13	-0.71	34	0.485	-0.14	**-0.12**
Pair 15	Oxford Total Scores - Baseline - Oxford Total Scores Postop	-15.73	10.66	8.76	1.75	-19.28	-12.18	-8.98	36	0	**-1.80**	**-1.48**

Table 3.25 Summary results of Effect Size (Cohen's d) and Standardised Response Mean (SRM)

Summary of Psychometric Analyses

Bringing together all the results we have so far analysed Test-Retest Reliability, Internal Reliability, Construct Validity and Responsiveness of the new PRO measure MP20, BMP and VAS for satisfaction of joint function.

Test Retest Reliability

The test retest study showed MP20 performed with excellent ICC values of > 0.9. The items that performed poorest were items 10 and 18 albeit still having ICC values of > 0.6.

Internal Reliability

For evidence of internal reliability, the LL, UL and RL had cronbach's alpha value of > 0.8, with most items except for item 7 causing a reduction in cronbach's alpha value when the item

was deleted. This demonstrated good evidence that items correlate well with the domain it represents. For item 7 cronbach's alpha increased when it was deleted indicating lack of homogeneity with the domain it was suppose to represent (UL domain). The Pain domain, cronbach's alpha value was only 0.544 and with item-trait correlation values of < 0.4, all 3 items in this domain does not appear to be homogenous. The GH domain showed cronbach's value of > 0.7 which supports evidence of internal reliability.

Construct Validity

MTMI analyses demonstrated that item 7 (UL domain) did not correlate well with UL domain, but infact correlated better with LL and RL items/domains. Item 12 (RL domain) also correlated better with LL domain and did not have evidence of convergent and divergent validity with it's domain. All items in Pain domain does not correlate well with each other, and item 20 from GH domain appeared to correlate better with UL and RL domain.

MTMM analyses

The five domains of MP20 showed overall good correlation with the total score (corrected domain-total correlation), however the GH domain displayed weakest correlation with the total score as well as with other domains. This demonstrates that the general health questions are clearly covering a different dimensions concept compared to concept covered by LL, UL, RL and Pain which are more closely related to physical functioning. And hence provide further evidence for the multi-dimensional concept of this new PROM.

LL domain showed good evidence of construct validity with high correlation with Oxford and WOMAC scores, and so did Role Limitation items. The Pain items had good correlation with WOMAC pain and with Oxford Scores. The 2 other components to the new PROM, BMP and VAS was indeed an attempt to make the new PROM more 'holistic' giving opportunity for patients to report other functional issues and general health that was not covered yet. The results of it's correlation analysis demonstrated that both components correlated modestly with the existing standard questionnaires. This again provides more evidence that MP20 when combined with BMP and VAS provide more information, and not just the same information about the patients overall functional status.

Responsiveness

The MP20 PRO measure and three of its domains i.e. LL, RL, and Pain showed large effect sizes with cohen's d value > 0.8 (Table 3.25). The Upper limb (UL) domain demonstrated moderate effect size with cohen's d value of -0.49 ($p < 0.01$) and the GH domain showed the least responsive quality with cohen's d value of -0.1, however the result was not shown to be statistically significant with p value of 0.5. The BMP Index Joint score and VAS for satisfaction demonstrated very large effect sizes with cohen's d value of 5.87 and -1.24 respectively. The Standardised Ratio Mean (SRM) for both these domains also reported large Effect size > 0.8. The WOMAC total scores, Oxford Scores and SF12-PCS also showed large effect sizes with cohen's d and SRM values of > 0.8.

4 DISCUSSIONS

4.1 Test Retest Study Outcome

More than 90% of the items in MP20 reported average ICC values of > 0.7, with exceptions for items 10 and 18, which had ICC values of 0.67 and 0.62, respectively, albeit still considered good reliability levels. Item 10 was part of an upper limb domain question that asks about a patient's limitation when turning a key, and item 18 was part of a Pain domain which asked about how the pain from your joint limits overall function when using their arms. Lack of test-retest reliability, or in this case, reduced level of reliability, can be a simple indication of measurement difficulties arising either from the items or scales under investigation or from the nature of the target population. It is clear that both these items revolve around upper limb context, and with the target population being patients with hip and knee OA, it is likely an indication that the question may not be suitable for the target population. Poor construct of the item can also be a problem, especially with item 18, which was a double-barreled question asking about overall function when using the arms and legs simultaneously. Patients were likely to be confused about which one he/she needs to report. We feel item 10, on the other hand, although it demonstrates a reduced level of test-retest reliability, is still valid as it highlights an essential function of the upper limb and hence should be retained.

The scores of the five separate domains demonstrated good reliability with average ICC values > 0.8, and the MP20 total scores showed excellent test-retest reliability with an ICC value of 0.92. The BMP index joint score and VAS of satisfaction of overall joint function also performed well with very good level of ICC values. Hence collectively, MP20 PRO measures satisfy the test-retest criterion, and deletion of item 18 will likely improve the overall test-retest result.

4.2 Internal Reliability Outcome

We used Cronbach's alpha as a measure of internal consistency of items to the domain it belonged to, along with inter-item and item-trait correlation analysis. The alpha values of LL, UL, RL showed alpha values of > 0.8, indicating overall good consistency within its items (Item 1 to 15), except for item 7 of the Upper Limb domain. Item 7 asked about the ability of a patient to carry things (e.g. shopping bag), and although it was meant to give information about upper limb function, the item construct does not appear to support it. The Cronbach's alpha value increased when item 7 was deleted from the UL domain, and further correlation analysis revealed item 7 correlated better with Lower limb and Role Limitation domain. Hence item 7 is probably more suited to be a feature of LL or RL domain rather than UL domain.

The pain domain had a poor alpha value of 0.544, with all the items correlating poorly with each other and towards its own trait (corrected item-trait correlation). Although alpha value did reduce as items were deleted, this could be an indication of both insufficient items in this domain to describe the pain characteristics as well as a problem with construct definition.

For example, item 16 considered the assessment of current pain while resting, but it depends whether the patient had painkillers yet or not, and also pain could vary at different times of the day. We know that pain is one trait that is most often difficult to measure due to multiple factors that could be affecting it. Item 17 had a slightly better alpha value as it asked about how the pain from the joints limit overall function when using the legs, but again the issue of double-barrel question and poor item construct makes the characteristic of this item inconsistent with the trait. A better-phrased item would ask 'How pain from your joint limits your overall function'. Item 18 has similar issues with 17, but this question is probably

confusing, as we said earlier in test-retest analyses. A better-worded item would probably be 'How pain from your joints limit your upper limb function'.

The General Health domain had an alpha value of > 0.7, demonstrating quite a reasonable level of internal consistency between the items and the domain it represents.

Following internal reliability analyses, we would suggest moving item 7 to LL or RL domain and either considering rephrasing all the three items in the Pain domain or even perhaps deleting the three items completely from MP20.

4.3 Construct Validity Outcome

MTMI analysis

All Items in MP20 except items 7, 12, 16, 17, 18 and 20, demonstrated good evidence of convergent and divergent validity. We have already seen how item 7 from the UL domain did not sit well in the UL domain and correlated better with LL and RL domains. Item 12, which asks about a patient's ability when they need to do a regular job or daily routine if retired, appeared to correlate better with Lower Limb domain and did not demonstrate good evidence of divergent validity within the Role Limitation domain. Further analyses revealed that the difference in the level of correlation between RL and LL was relatively marginal, i.e. 0.599 and 0.625 respectively, and the content structure of this item also fits the trait, i.e. Role Limitation. In this case, we would still suggest keeping item 12 within the RL domain.

All the three pain items (16, 17 and 18) did not correlate well with each other and had poor item-trait correlation. We have already discussed this earlier in our internal consistency discussion and perhaps restructuring the content of the items in order to maintain the Pain dimension as one of its multi-dimensional PRO measures. The alternative will be deleting the three items completely, which will improve the performance of the PRO measure at the expense of the ability of the PROM to assess the Pain dimension of patients.

Item 20 asks about how mood (e.g. anxiety and depression) limit the overall function, and it appeared to show a marginally better correlation with UL and RL domain. Further analyses revealed that the actual difference in item-trait correlation across all the domains is relatively small, between 0.461 to 0.552 ($p < 0.01$). This indicates that item 20 does not possess evidence of divergent validity on its own. However, as we will see later on in MTMM analysis, together with item 19, it forms a good dimension. So, we suggest keeping the GH items as it is.

MTMM analysis

The five domains of MP20 showed an overall good correlation with the total score (corrected domain-total correlation); however, the GH domain demonstrated a significantly weaker correlation with other domains. This shows that the general health (GH) items are covering the concept of a different dimension compared to the concept covered by LL, UL, RL and Pain, which are more closely related to physical symptoms and functioning. Thus provide more evidence for the multi-dimensional concept of this new PROM.

LL domain showed good evidence of construct validity with a high correlation with Oxford and WOMAC scores, and so did Role Limitation items. The Pain items had a good correlation with WOMAC pain and Oxford Scores and hence good evidence of construct validity; however, we know from previous analysis that the Pain domain has got insufficient evidence of internal reliability. The two other components to the new PROM, BMP and VAS, were an attempt to make the new PROM more 'holistic', allowing patients to report other functional issues and general health that was not covered yet. The results of its correlation analysis demonstrated that both components correlated modestly with the existing standard questionnaires. This again provides more evidence that MP20, when combined with BMP and VAS, provide more information, and not just the same information about the patients overall functional status.

So, from the MTMI and MTMM analyses, we conclude that item 7 would better serve the RL domain, and we would rephrase the items in the pain domain to be less obscure and remove double-barrel questions. This would allow the PROM to maintain its Pain domain and allow

assessment of the pain dimension. Despite that, all the hypothetically related domains still showed good correlation with other existing PROM, e.g. correlation of MP20-Pain domain with WOMAC-Pain domain was 0.702 ($p < 0.01$) and correlation of MP20-LL with Oxford Scores is 0.718 ($p < 0.01$), which shows good evidence of construct validity. And finally, the MP20 total score showed an encouragingly good correlation with Oxford and WOMAC (table Table 3.21), but not as good as the correlation between Oxford and WOMAC (Table 3.22), which fits in nicely as we do not expect the correlation to be as good based on the multi-dimension construct of the new PROM.

4.4 Responsiveness Outcome

The MP20 demonstrated a large effect size, although not as high as the WOMAC and Oxford scores. This reflects the construct of each type of PRO Measure where it was expected that the more disease-specific joint PROM like Oxford and WOMAC would demonstrate a higher level responsiveness where reporting physical symptoms are much more specific and less ambiguous. The PRO measure like MP20 and SF12 PCS showed a lesser degree of responsiveness given the fact that these measures are aimed at a more overall measure of physical function.

At domain/component level, the BMP index joint score and the VAS of satisfaction demonstrated good responsiveness, and so did the LL, RL and Pain demonstrating a large effect size ($d > 0.8$). The UL domain of MP20 was moderately responsive, but the domain with the least evidence of responsiveness was the GH domain which showed a very small Effect size; however, the results were not statistically significant. So only LL, UL, RL and Pain domains showed adequate evidence for the level of responsiveness, but the GH domain did not display any evidence of responsiveness. This may reflect the difficulty of constructing an item that aimed to gather information about one's general health. Even the SF12-MCS did not demonstrate good evidence of responsiveness, and this may reflect the similarity this PRO measure has with the GH domain.

4.5 Is there sufficient evidence for Validity?

In order to decide if we have sufficient evidence to support validity of this new PRO measure, we shall review again the hypothesis to see if we've answered the questions we set out in the beginning of the study.

Question 1. Is the newly develop PROM, reliable and consistent?
The test retest study have clearly shown that the MP20 along with BMP index joint score and VAS of satisfaction of overall function are reliable. Internal reliability study also demonstrated adequate cronbach's alpha value for all the MP20 domain even for the pain domain (alpha value > 0.5). So yes the new PROM is reliable and consistent.

Question 2. Is the new PROM, measuring what is supposed to measure?
Following extensive correlation analysis, we conclude that, firstly item 7 needs to transferred from Upper Limb domain to Role Limitation domain. Secondly, we will keep pain domain but the items will need reconstructing to eliminate double-barrel questions and avoid vagueness. And finally, the domains, and PROMs have satisfied adequate convergent and divergent validity to satisfy evidence of construct validity.

Question 3. Is it the New PROM responsive to changes?
The MP20, BMP index joint score and VAS demonstrated large effect sizes and hence has good evidence of responsiveness

Question 4. Are the added components, Body Map of Pain (BMP) and Visual Analogue Scale (VAS) for satisfaction a useful adjunct to the PROM?

In short we feel the answer is yes however several factors have to be born in mind. Firstly, BMP provides a quick snapshot of the patients pain distribution and a measure of severity. We have analyzed the Index Joint Score test retest reliability, construct validity (via MTMM analysis) and it's responsiveness and found that it provides satisfactory evidence. However we did not analyze the psychometric properties of other joints (excluding index joint) and

hence we cannot conclude the clinemetric properties of the rest of BMP. Secondly, VAS of satisfaction of Overall Function is quite a broad and multidimensional trait and will be quite complicated to measure. As a consequence measuring this broad trait with just a single method using VAS obviously has limitations. Although test retest and responsiveness study results provided satisfactory evidence, the construct validity testing (MTMI & MTMM analysis) did not show very good correlation with existing PRO Measures used. This could be a good thing, the fact that the whole idea of the study is to measure an Overall Function, hence we shouldn't expect the correlation to be too good. The next question would be how can we be sure it's measuring what it is suppose to measure, i.e. do we have sufficient evidence of construct validity for it's use. Well assuming that MP20 itself is able give us overall function, a correlation of > 0.5 is fair indication that VAS does satisfy some evidence of construct validity.

We will gather qualitative information from the research participants and also measure the correlation of the improvement in Body Map Score following intervention with improvement in outcome measure as a measure of relative validity.

4.5.1 Body Map Pain and Visual Analogue Scale for Satisfaction

BMP and VAS for satisfaction of overall function were additional features to the 20 item PROM (MP20). BMP provides both an overall visual representation as well as a severity index of pain coming from the major joints of the human body. It is meant to simplify communication of pain from a patient to the clinician by putting a severity score between zero to ten and marking the score over relevant joints on a sketch of a human body provided. Patients are encouraged to put scores over not just the hip and knee joint but also onto other joints which are painful and troublesome. It is a quick way to ascertaining 'geographically' where the most pain are, and other locations of pain that the patient is suffering. It is a 'holistic' way of easily finding out a patient's pain experiences, in a snapshot. Although we have not completed a full psychometric analysis of the BMP, both test retest reliability and

responsiveness study of the index joint score (i.e. the main hip or knee pain) demonstrated satisfactory evidence. Feedback given from patients are that the BMP was easy to understand and allowed patient to report other pain from different joints that the patient is experiencing which may have been overlooked during clinical consultation. As we have found out from our study, pain is a trait which is most difficult to measure due multiple factors influencing the outcome and hence the BMP was designed. We have not found this feature in any type of PROM used in Orthopaedic surgery and is potentially a very useful tool to understand more about a patients experience of pain.

Visual Analogue Scale (VAS) is a form of measure where patients are asked to mark on a standardized scale, for e.g. a ruler, where they would put their measure of health. In this study, we'd like to measure the level of satisfaction of overall function that one has before and after surgery. This again is a difficult trait to measure because level of satisfaction can be due to various factors, success of the operation, polyarthritis, experience of patients during rehabilitation etc.

4.6 Recommended Amendments to new PROM

Based on the psychometric analysis and qualitative feedback from our study population, and research group several changes to the new PROM were recommended. Firstly item 7 will be incorporated into RL domain as it appeared to fit better following internal reliability testing and extensive correlation analysis. The pain domain is to be maintained but restructured to avoid ambiguity and double-barrel questions. The Item will now be phrased as follows

Item 16. Please select your level of pain on average over the last month, with or without analgesia.

Item 17. Please select how pain from hip or knee limit your overall function

Item 18. Please select how pain from using your arms limit your overall function

Although the GH domain was not very responsive, so was the SF12-MCS and both the results were not statistically significant. However it is felt the content of the items were suitable and fulfilled content validity criteria for measuring general health traits.

4.7 Potential use for the New PROM

We can see various ways that this new PROM may have positive impact when used concurrently with current outcome tool measures already in used in patients with Lower limb Osteoarthritis especially those who are undergoing joint arthroplasty. From the clinical perspective they can be:-

- As a screening tool for clinicians, especially General Practitioners to identify those patients who are likely going to be straight forward candidates for Joint arthroplasty or those which requires more 'thought' prior to committing to surgery. E.g. Patients with Knee OA with degenerative spine or Rotator cuff arthropathy.
- Good for risk assessment and managing patients' expectations, Identifying 'high risk' group. i.e. potentially patients who may not have as good outcome, due to multiple factors.
- More patient centred approach. Allows patient to convey more patient's perspective, highlighting issues which may have been under-addressed, e.g. poor upper limb functioning in patients with multiple joint diseases.

From a research standpoint, we will begin to look at patient's outcome in a more holistic view and assess not just the joint of interest but able to evaluate overall index of physical function.

4.8 Limitations of this Study

4.8.1 Power of the Study

- *Numbers of participants*

  The strength of this study can be improved by recruiting more participants. Power calculation done in the beginning of the study showed a minimum of 45 participants required to gain a decent Confidence interval and statistically significant results. However, with such a high attrition rate in these study group population, we were only able to achieve the satisfactory number in the test-retest group, but in the responsiveness group the figures fell to 38 patients. This may affect our degree of

confidence by the evidence presented in the responsiveness group. Such a high attrition rate is quite common in such study population because our study group normally presents elderly population. In addition, the burden of filling in so many forms adds to the fatiguability of participants to continue the study.

4.8.2 Skewed Study Population

The selection of patients for this study only included patients who are at the extreme end of the disease (awaiting joint replacement) and is not really a reflection of the whole disease population with Osteoarthritis of the lower limb. In other words, the study population is skewed. This was also highlighted during the descriptive analyses. Hence although this new PROM shows a promising validation result, assessing the minimal clinical differences can be challenging and it's not necessarily about the inability to assess them but rather the complexity and potential biases introduced by the skewed distribution of data.

Elaborating further, in a skewed population, the majority of participants will fall within a particular range, while the minority represents the extreme values. This can lead to a non-representative sample that doesn't reflect the broader population, which in this case are patients with lower limb osteoarthritis. This may further lead to inaccurate results and potentially erroneous conclusions about minimal clinical differences. Related particularly to this study is the impact on effect size. The magnitude of differences in the responsiveness cohort may have been influenced by the skewness of the data, resulting in larger effect size that may not represent the clinical significance of an intervention. It can also be argued that due to this particular skewed population, you have observed statistically significant difference but these differences may not be clinically relevant. This will result in decisions and interventions made based on statistical significance rather than clinical importance.

Skewed data can also make it challenging to interpret the clinical relevance of a statistical findings. A small change in a skewed variable may have a disproportionately large impact on certain individual which may not be clinically meaningful.

Overcoming the challenges

To address these challenges the next phase of the study should also include patients of varying spectrum of severity. This is likely to involve a much larger number of participants and recruitment entry points will need to be wider and not focused only on Orthopaedic Clinics. This will enable us to employ a more robust statistical methods like 'Items Response Theory' which are less sensitive to the distributional assumptions of data and allow better administration of the questionnaires using platform such as Computer Adaptive Test (CAT). These methods are already being used and is continually evolving in large Health Related quality of life platforms such as PROMIS score [29] and Versus Arthritis Musculoskeletal Health Questionnaire (MSK HQ) [30] . Both these new concepts are related, but in essence CAT is a method of test administration that uses principles of Items Response Theory (IRT) to adaptively select test items based on a test and calibrate test items, regardless of whether the test is administered adaptively or traditionally. Both CAT and IRT play essential roles in improving precision and efficiency of outcome measures assessment.

4.8.3 Feedback from patients

- The pain questions (16,17 & 18) and the general health question (Qns 19 & 20) appears to be the most challenging questions in terms of accuracy of reflection. Its genuinely accepted that these are the most complex domain to describe and hence should require more engagement from patients in order to iterate it further. However, these are beyond the scope of such small study, nevertheless it has certainly highlighted the fact that Overall Health Status is a complex concept and that requires not just Health care professionals input but more importantly the patients themselves. More qualitative data on this aspect would have provided better insight to move research further.

5 CONCLUSION

Creating a new outcome tool that looks at the overall functional status of a patient is a complex task of identifying the multiple domains that contribute towards a holistic outcome measure of a patient's physical function. The overarching conceptual framework here is the

Overall Physical Function and involves a variety of domains. The major domains are Physical (upper limb, lower limb and spine) and whilst others include activities in which patients are involved in a day to day functioning. One domain that has proven to be most challenging to measure has been pain, reflecting the subjective nature of this domain entity, which will continue to trouble PROM developers for years to come. Despite the complexity, it is possible to come up with the best fit multi-domain outcome measuring tool to assess the Overall Functional Status of a patient, but we believe it can be addressed using the 3 component approach. These three components that make up the overall functional tool are the MP20, i.e. the Main PROM 20 item questionnaire, the Body Map Pain, and the Visual Analogue Scale of Satisfaction.

An MP20 score alone can give one an overall physical functioning score for a patient, which in itself may be sufficient. It constitutes 5 of the most important domains of overall physical function, Lower limb physical function, Upper limb physical function, Role Limitation, Pain and General Health. However, it is still at the end of the day a number, though it is a valuable index of functioning for research purposes; clinically, we believe it does not contribute much to a patient. Hence, we recommend using the MP20 PROM alongside the BMP and VAS of Satisfaction. Together it fulfils the 'holistic' requirement of a patient's assessment of overall physical function and mutually beneficial to clinicians and patients.

6 REFERENCES

[1] F. a. D. Administration, "Guidance for Industry on Patient Reported Outcome Measures: Use in Medicinal Product Development to Support Labelling Claims.," *Federal Register,* vol. 74(235), pp. 65132-65133, 2009.

[2] M. Yakob, Development and Validation of a Patient Reported Outcome Measure Assessing Global Physical Function in Patient's suffering from Lower Limb Osteoarthritis, MPhil Thesis, 2020.

[3] P. Fayers and R. Hays, "Assessing qualty of Life in Clinical Trials," Oxford University Press, 2011, pp. 3-54.

[4] P. Jaume and R. Alba, "Socio-economic coxts of osteoarthritis: A systematic review of cost-of-illness studies," *Seminars in Arthritis and Rheumatism,* vol. 44, no. 5, pp. 531-541, 2015.

[5] D. Ayers and K. Bozic, "The Importance of Outcome Measurement in Orthopaedics," *Orthopaedic Healthcare Worldwide,* pp. 3409-3411, 2013.

[6] A. Lundgren-Nilsson, A. Dencker, A. Palstam, G. Person, M. C. Horton, R. Escorpizo, A. Kucukdeveci, S. Kutlay, A. Elhan, G. Stucki, A. Tennat and P. Conaghan, "Patient-reported outcome measures in osteoarthritis: a systematic search and review of their use and pscyhometric properties," *RMD Open,* vol. 4, pp. doi:10.1136/rmdopen-2018-000715, 2018.

[7] N. England, "digital.nhs.uk," 2009. [Online]. Available: https://digital.nhs.uk/data-and-information/publications/statistical/patient-reported-outcome-measures-proms-in-england.

[8] N. Bellamy, W. Buchanan, C. Goldsmith, J. Campbell and L. Stitt, "Validation Study of WOMAC: A Health Status Instrument for measuring Clinically Important Patient Relevant Outcomes in Antirheumatic Drug Therapy in Patients with Osteoarthritis of the Hip or Knee," *The Journal of Rheumatology,* vol. 15, no. 12, pp. 1833-1840, 1988.

[9] A. Sikorskii and P. Noble, "Statistical Considerations in Psychometric Validation of Outcome Measures," *Clinical Orthopaedics and Related Research,* vol. 471, pp. 3489-3495, 2013.

[10] T. Weldring and S. Smith, "Patient-Reported Outcomes (PROs) and Patient-Reported Outcome Measures (PROMs)," *Health Services Insights,* vol. 6, no. doi: 10.4137/HSI.S11093, pp. 61-61, 2013.

[11] D. Ayers and K. Bozic, "The Importance of Outcome Measurement in Orthopaedics," *Clinical Orthopaedics and Related Research,* vol. 471, pp. 3409-3411, 2013.

[12] L. Mokkink, C. Terwee, D. Patrick, J. Alonso, P. Stratford, D. Knol, L. Bouter and H. de Vet, "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,* pp. DOI 10.1007/s11136-010-9606-8, (2010) 19:539–549.

[13] P. Fayers and R. Hays, "Assessing Quality of Life in Clinical trials (2nd Edition)," *Chapter 1,* pp. 3 -53, 2011.

[14] J. Cappelleri, J. Lundy and R. Hays, "Overview of Classical Test Theory and Item Response Theory for Quantitative Assessment of Items in Developing Patient-Reported Outcome Measures," *Clin Ther,* vol. 36, no. 5, pp. 648-662, 2014.

[15] T. Koo and M. Li, "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research," *Journal of Chiropractic Medicine,* vol. 15, pp. 155-163, 2016.

[16] G. Norman, K. Wyrwich and D. Patrick, "The mathematical relationship among different forms of responsiveness coefficients," *Qual Life Res,* vol. 16, no. DOI: 10.1007/s11136-007-9180-x, pp. 815-822, 2007.

[17] D. Murray, R. Fitzpatrick, K. Rogers, H. Pandit, D. Beard, A. Carr and J. Dawson, "The use of the Oxford Hip and Knee Scores," *The Journal of Bone and Joint Surgery (Br),* vol. 89, no. B, pp. 1010-14, 2007.

[18] A. Dave, F. Selzer, F. Selzer, Losina and J. Collins , "Is There an Association Between Whole-body Pain With Osteoarthritis-related Knee Pain, Pain Catastrophizing, and Mental Health?," *Clinical Orthopaedics and Related Research,* pp. DOI 10.1007/s11999-015-4575-4, (2015) 473:3894–3902.

[19] D. Southerst, P. Cote, M. Stupar and P. Stern, "THE RELIABILITY OF BODY PAIN DIAGRAMS IN THE QUANTITATIVE MEASUREMENT OF PAIN DISTRIBUTION AND LOCATION IN PATIENTS WITH MUSCULOSKELETAL PAIN:A SYSTEMATIC REVIEW," *Journal of Manipulative and Physiological Therapeutics,* Vols. Volume 36, Number 7, September 2013.

[20] O. Rolfson, E. Bohm, P. Franklin, J. Dawson, J. Dunn and S. Lubekke, "Patient-reported outcome measures in arthroplasty registries Report of the Patient-Reported Outcome Measures Working Group of the Inter- national Society of Arthroplasty Registries Part

II. Recommendations for selection, administration, and analysis," *Acta Orthopedica,* vol. 87 (eSuppl 362): , p. 9–23 9 , 2016.

[21] D. J. B. H. M. K. H. C. J. A. J. P. J. Dawson, "Development of a patient-reported outcome measure of activity and participation (the OKS- APQ) to supplement the Oxford knee score," *THE BONE & JOINT JOURNAL ,* Vols. VOL. 96-B, No. 3, MARCH , pp. 332-338, 2014 .

[22] N. Bellamy, W. Buchannan, C. Goldsmith, J. Campbell and L. Stitt, "Validation Study of WOMAC," *The Journal of Rheumatology,* no. 15-12, pp. 1833-1840, 1988.

[23] E. Roos, H. Roos, L. Lohmander, C. Ekdahl and B. Beynnon, "Knee Injury and Osteoarthritis Outcome Score (KOOS): Development of a Self-Administered Outcome Measure," *Journal of Orthopaedic Sports and Physical Therapy,* vol. Volume 78 Number 2 , pp. 88-95, August 1998.

[24] F. Impellizerri , A. Mannion, M. Leunig, M. Bizzini and F. Naal, "Comparison of the Reliability, Responsiveness, and Construct Validity of 4 Different Questionnaires for Evaluating Outcomes after Total Knee Arthroplasty," *Journal of Arthroplasty,* vol. Vol. 26 No. 6 , pp. 861-868, September 2011 .

[25] T. W. W. F. Wyrwich KW, "Further evidence supporting an SEM-based criterion for identify- ing meaningful intra-individual changes in health-related quality of life," *J Clin Epidemiology,* vol. 52, p. 861, 1999.

[26] J. Nunnally and I. Bernstein, "The Assessement of Reliability. Psychometric Theory.," 1994, pp. 248-292.

[27] W. K. P. D. Norman GR, "The mathematical relationship among different forms of responsiveness coefficients," *Qual Life Res,* vol. 16, p. 815, 2007.

[28] I. C. R. 2017., *IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY:IBM Corp.*

[29] "PROMIS Assessment Center Website," [Online]. Available: https://www.assessmentcenter.net/..

[30] "Versus Arthritis Musculoskeletal Health Questionnaire (MSK-HQ)," [Online]. Available: https://innovation.ox.ac.uk.