HEALTH OUTCOME MEASURES IN AN AGING POPULATION: VALIDITY, RELIABILITY AND INTERPRETABILITY

by

Erin M. Macri

B.Sc., Simon Fraser University, 1997

M.P.T., University of British Columbia, 2006

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

(Experimental Medicine)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

May 2013

Abstract

In Canada, older adults currently represent a record high proportion of about 15% of the population. Associated with aging is the increased prevalence of multiple morbidity, resulting in widely varied and complex health statuses among our aged. Identifying effective strategies to promote healthy aging and reduce comorbidity hinges on the ability to accurately measure health outcomes. This requires the use of valid and reliable instruments with associated reference statistics to enhance interpretability of test scores.

In Chapter 2, I present a validity study of a patient-reported outcome measure, the Patellofemoral Pain and Osteoarthritis Outcome Scale (PFOOS). The PFOOS is designed to evaluate symptoms relating to patellofemoral (PF) pain and osteoarthritis. I recruited 54 adults aged 18+ with peri- or retro-patellar pain for ≥ 3 months, rated ≥3/10 on a numeric pain scale, aggravated by PF-loading tasks (e.g. squats). People with diffuse knee pain, history of total knee or hip replacement, or severe knee trauma in the past year were excluded. Recruitment was done through adverts to staff & students at an Australian University. Participants completed paper & online versions of the PFOOS, Anterior Knee Pain Scale and SF-36, and repeated the PFOOS in 1-2 weeks. Analysis included internal consistency (Cronbach's α), test-retest & alternate forms reliability (ICC), and construct validation (hypothesis testing). ICCs were ≥ 0.79, Cronbach's α ≥ 0.61. The PFOOS performed largely as hypothesized. Overall, the PFOOS demonstrated good validity & reliability in this sample.

In Chapter 3, I report results of a cross-sectional study aimed to develop normative data for the de Morton Mobility Index (DEMMI). The DEMMI is a performance-based instrument that measures mobility across a spectrum from bed bound to functional independent mobility. A sample of 183 healthy, community-dwelling adults age 60+ were recruited. Mean DEMMI scores varied by age category, by living arrangement (independent vs. assisted living), and by use of mobility aid ($p<0.05$). Scores did not differ by sex ($p=0.49$) or falls history ($p=0.21$). Reference intervals were provided for individual and group comparison, to facilitate use of the DEMMI across the mobility spectrum in clinical and research settings.

Preface

Chapter 2 is based on work conducted at the University of Queensland, Australia. The PFOOS instrument was developed by Drs. Kay Crossley, Ewa Roos, and Sallie Cowan. Ethics for this study was obtained at the University of Queensland in Australia (project #2012000025). The developers of the instrument were responsible for generating the items in conjunction with input from a panel of experts (patients, physiotherapists, orthopaedic surgeons, rheumatologists, sports physicians, and researchers). I was responsible for completing participant recruitment, screening interested individuals for eligibility, obtaining informed written consent, administering all patient-reported outcome instruments, collecting and entering all data, and statistical analysis. To date, there have been no other publications of this research (manuscript writing in progress).

Chapter 3 is based on collaborative work conducted at the University of British Columbia as well as Monash University in Melbourne, Australia. This study was approved for Canadian data collection by the Clinical Research Ethics Board (CREB Certificate #H10-02748) as well as the Vancouver Coastal Health Authority. Australian ethics was approved by the Monash University Human Research Ethics Committee (project #CF07/3954- 2007001870). A version of chapter 3 has been published: Macri EM, Lewis JA, Khan KM, Ashe MC, de Morton NA. *The de Morton Mobility Index: Normative Data for a Clinically Useful Mobility Instrument*. Journal of Aging Research, Vol. 2012, Article ID 353252, 7 pages, 2012. doi:10.1155/2012/353252. I critically reviewed the DEMMI publications beginning with the first one in 2008 onwards as well as key papers related to mobility instruments and their measurement properties. As a graduate student working with Dr. Khan and Dr. de Morton on her Churchill Fellowship in Canada, I was the local (Vancouver) lead for logistics and participant recruitment (with Dr. de Morton). I scheduled appointments, ensured adherence to measurement protocol; attended all site visits; administered the DEMMI instrument to all participants; worked with and supervised a team in recruitment (screening for eligibility, obtaining written informed consent) and administration of study paperwork; and completed all data entry and analysis. As lead author, I wrote the manuscript and critically reviewed revisions. Dr. de Morton developed and validated the instrument, designed the current study, and led/oversaw all aspects of the projects in both Australia and Canada. Ms. Lewis collected data in Australia, conducted assessments, and

assisted with manuscript writing. Dr. Khan and Dr. Ashe provided input to the Canadian arm of the study in design and methodology, statistical analysis and manuscript writing.

Table of Contents

List of Tables

List of Figures

List of Symbols and Abbreviations

α	alpha
s_x	standard deviation of a collection of scores
μ_i	theoretical 'true' score of an underlying health construct for an individual
x_i	score obtained for an underlying health construct for an individual
ADL(s)	activities of daily living
AKPS	Anterior Knee Pain Scale (also called the Kujala Scale)
AL	Assisted Living
ANOVA	analysis of variance
APACHE II	Acute Physiology and Chronic Health Evaluation II
BP	Bodily Pain (SF-36 subscale)
COSMIN	COnsensus-based Standards for the selection of health status Measurement INstruments
DEMMI	de Morton Mobility Index
GH	General Health (SF-36 subscale)
HRQOL	Health-related quality of life
ICC	intraclass correlation coefficient
IQ	intelligence quotient
KOOS	Knee injury and Osteoarthritis Outcome Scale
MCID	minimal clinically important difference
MCS	Mental Component Summary (SF-36 subscore of mental items)
MD	minimum difference
MDC(x)	minimal detectable change (with or without confidence interval specified, i.e. MDC / MDC_{90} / MDC_{95})
MH	Mental Health (SF-36 subscale)
PA	posterior-anterior
PCS	Physical Component Summary (SF-36 subscore of physical items)
PF	Physical Function (SF-36 subscale)
PFOOS	Patellofemoral Pain and Osteoarthritis Outcome Scale
PFPS	Patellofemoral pain syndrome
PSFS	Patient-Specific Functional Scale

PROM	Patient-reported outcome measure
QOL	quality of life
r	correlation coefficient (e.g. Pearson's r)
RE	Role Emotional (SF-36 subscale)
RP	Role Physical (SF-36 subscale)
SD	standard deviation
SDC	smallest detectable change
SEM	standard error of measurement
SF	Social Function (SF-36 subscale)
SF-36	Short Form (36) Health Survey
TUG	Timed 'Up & Go' test
UBC	University of British Columbia, Canada
UM	University of Melbourne, Australia
UQ	University of Queensland, Australia
VT	Vitality (SF-36 subscale)
WOMAC	Western Ontario McMaster Osteoarthritis Index

Acknowledgements

As I come to the end of this chapter of my life, I am grateful to have this space to take pause and reflect on the many colleagues, friends, and family members who provided me with the support, love and inspiration so necessary to my arrival here.

Academically, I would like to thank Dr. Karim Khan for seeing my potential, introducing me to this academic world with such patience and dedication, and for offering so many pivotal opportunities for my professional and personal growth. I am also grateful for Dr. Kay Crossley for opening a new path of research for me, and for providing me with the guidance, resources and networking to inspire me towards completion of this degree and onward into the next phase (and thanks to Maddie for the laughter and play through it all). Thank you also to Dr. Jen Davis for setting the bar so high in terms of critical thinking and attention to detail, and thank you for so graciously agreeing to continue to work with me as my path forged new directions throughout my studies. Immense thanks also to Dr. Natalie de Morton, for your wisdom, kindness, and mentorship through my earliest days of research. Inviting me to work with you during your fellowship provided me with a nourishing and engaging environment in which to learn. Another important influence in my academic experience includes Dr. Maureen Ashe, thank you for the opportunity to work with your research team, and for generously offering your time, feedback, and commitment to my learning trajectory.

Equally important in providing a positive experience throughout this degree are my colleagues, fellow students and lab mates. Special thank you to Vanessa Young, Vina Tan, Douglas Race, Kerry Mellifont, Naomi Beck, Janneke van Leeuwen, Adam Culvenor, Natalie Collins, Steph Filbay, Jessica Lewis, and Anna Chudyk. You are all generous and kind and your contributions to my work are immense and will always be remembered.

Without financial support this thesis would absolutely not have been possible. Thank you to the UBC team responsible for offering me financial support through the Canadian Institute of Health Research's Vancouver Integrated Study of Aging. The PFOOS study was made possible through the Australian International Endeavour Award Research Fellowship. The DEMMI study was

made possible for me, indirectly, through funding for Dr. de Morton’s Vincent Fairfax Churchill Fellowship award.

It has taken until the third degree to have the opportunity to thank Colleen and Mike Macri, my mom and dad, on paper. This display of gratitude in no way matches that which I feel for you both. There are no words to describe my deep love and appreciation. You have been the pillars of my life, without you I would not have these wings.

And finally, Jaia. Some might argue you came on board at the most inopportune of times, mere weeks before I started this degree. Thank you for allowing me this love affair of learning. Thank you for giving me a reason to seek balance, health, evolution and vitality in my life. Bulumusu.

To the late Dr. Thomas L. Richardson

Chapter 1: Introduction – Measuring Age-Related Health Outcomes

"Exhilaration may come from recognition that the goal of a vigorous long life may be an attainable one"

~James Fries

1.1 Aging and comorbidity

The number of Canadians aged 65 and older is currently estimated at five million[1]. This represents a record high of almost 15% of the national population[1], substantially higher than the worldwide proportion of 8%[2]. The association between morbidity and aging is well documented in the literature[3-8]. Chronic conditions such as arthritis (see Figure 1.1), cardiovascular disease or cognitive decline, and the concurrence of multiple comorbidities, become more prevalent and disabling with age. The result is a diverse and complex spectrum of health statuses among older adults that can result in loss of ability to perform functional tasks, participate in meaningful activities, or live autonomously[5, 6, 9-12]. In addition to the physical burden of living with chronic comorbidity come high economic costs at both individual and societal levels[13-18]. Older adults in Canada represent approximately 40% of acute hospital admissions, and use about 70% higher resources when hospitalized, compared to their younger counterparts[19]. In 1998, annual direct health care expenses were estimated at $4.4 billion for arthritis alone[20], just one of the many prevalent chronic illnesses, and the highest cause of disability in women and second highest in men[21]. In 2010, estimates of total direct and indirect costs for arthritis approached $33 billion annually[21].

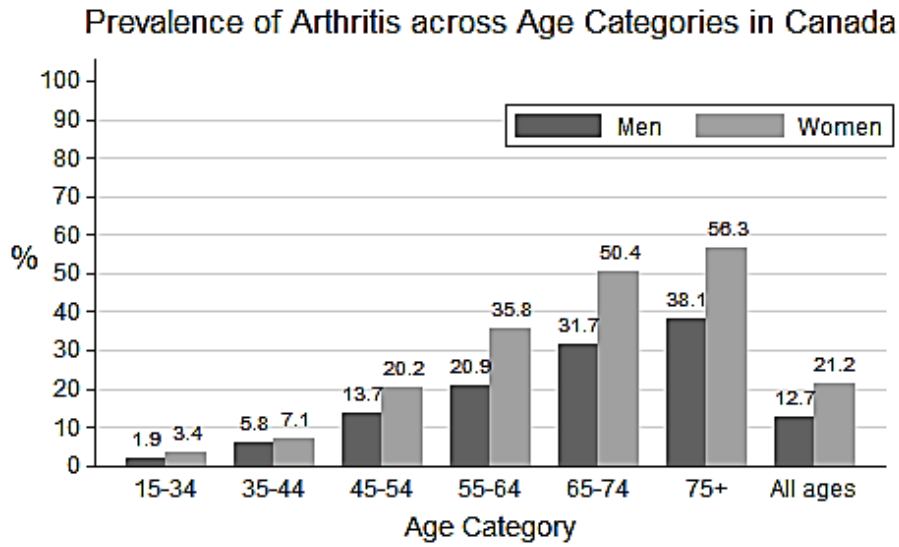

Figure 1.1 Prevalence of arthritis in Canada, graph derived from data in Canadian Community Health Survey 2011[22]

It is well established in the literature that risk factor modification (i.e. through medical intervention or preventive lifestyle modification) can alter chronic disease trajectories, resulting in lower morbidity and mortality[11, 23-26]. However, measuring morbidity and health status is a challenging and often daunting task that has been the crux of considerable confusion in the literature[27]. What exactly should be measured? What is the best way to measure it? And finally, how closely does a measurement approximate the "truth" such that inferences and decisions can be made with confidence? The task of identifying effective strategies to promote healthy aging and reduce chronic comorbidity hinges on the ability to accurately measure 'health outcomes'.

1.2 Health outcomes, constructs, and health outcome measures

A 'health outcome' is the effect that a process, such as disease progression or medical treatment, has on health status (or factors influencing health status, called health determinants)[28-30]. Health status is influenced by a complex interaction of health determinants (contextual factors such as disease processes, age, sex, health behaviours, or socioeconomic status) and interventions (medical care or social programs)[31]. Some health outcomes are relatively easy to quantify. Examples include the occurrence of an event (e.g. hospital admission, cancer diagnosis, death) or the observation of a biological or physiological output (e.g. range of motion, height, serum cholesterol levels). However, many health outcomes are less tangible. Consider outcomes such as depression, pain, or quality of life. These more abstract health attributes are called 'constructs' and they can be challenging to define and heavily reliant on theory to describe[32, 33]. Quantifying or measuring these states of health cannot be done solely through rater observation because they involve patient perspective[34, 35].

'Health outcome measures' are the measurements one obtains in an attempt to quantify, describe, or understand some aspect of health status. At either an individual or a group level, health outcome measurements can be used to: (i) better understand the natural course of a disease (such as arthritis) or physiological process (such as aging); or (ii) to evaluate the effects of a deliberate health intervention. Many definitions of health outcomes refer to the latter, since often it is the effects of targeted interventions that are of primary interest in health care and research settings[28-30, 33].

1.2.1 Measuring mobility

'Mobility' is a health attribute, or construct, that is commonly affected by age-related chronic comorbidity and is therefore an important construct to measure in both clinical and research settings. Mobility is also a construct that illustrates some of the challenges in defining and subsequently measuring health outcomes. The term has many different meanings. For example, a literature search of the keyword 'mobility' using Medline (Ovid SP) reveals over 100,000 publications. Definitions of 'mobility' among these journal articles include, but are not limited to: range of motion[36]; joint laxity or hypermobility[37]; walking ability[38]; using a mechanical lift to put a patient into a chair[39]; or a spectrum ranging from passive range of motion to walking[39]. To further complicate the issue, the construct of mobility can be further broken down into other attributes such as balance, gait speed or functional capacity[40]. Clearly if one wishes to measure mobility, this construct must first be carefully defined prior to developing an instrument for measuring it.

Both studies in my thesis involve health outcome instruments that are developed to include some measure of mobility. The definition for 'mobility' in this document is that developed by the World Health Organization[41]: "moving by changing body position or location or by transferring from one place to another, by carrying, moving or manipulating objects, by walking, running or climbing, and by using various forms of transportation".

1.3 Features of health outcome instruments

There is a tremendous variety of types of health outcome instruments. This includes: laboratory instruments (e.g. to conduct assays of biological samples such as blood or tissue biopsies); diagnostic imaging (e.g. magnetic resonance or X-ray); orthopaedic instruments (e.g. tape measures or dynamometers); and paper-based instruments (e.g. questionnaires). The focus of this thesis is to explore the latter category, investigating instruments where each item is presented and scored in a written format. Selecting an instrument will depend on the construct of interest; how that construct is best measured; and the population of interest within the context (or purpose) of the clinical or research question.

1.3.1 Dimensionality

A health outcome instrument can measure one or more 'dimensions', also called constructs, health attributes or domains. Scales that target a single construct improve one's ability to draw inferences from test scores and thus contributes to validity[33, 42, 43]. The concept of unidimensionality can appear murky when considering that many constructs in health care (such as mobility), while unidimensional, can be further divided into several other unidimensional constructs. For example, the Short Form 36 (SF-36) is a health-related quality of life (HRQOL) instrument that is divided into eight subscales, each representing a different aspect of quality of life[44]. Streiner and Norman suggest the goal is not to achieve the smallest unit of unidimensionality, but to ensure the scale is "unidimensional enough"[32].

Often, one wishes to know about health outcomes that span several dimensions. For example, one may wish to know how a therapeutic intervention affects a person's mobility, but might also want to know how it then affects their ability to participate in daily activities, how their pain has changed, or how their overall quality of life has been affected. An improvement in mobility may not carry over into an improvement in quality of life if certain contextual factors have been missed[35]. For example, an outcome measure that captures an improvement in an individual's ability to walk the length of a hallway may not reflect improved quality of life at home if she is still house-bound due to inability to negotiate stairs in her house.

If a single instrument targets several dimensions, it is most appropriate to divide the instrument into subscales, each representing a unique dimension. This approach maximizes interpretability within each domain - important information can be missed if a total score is used instead of separated subscores[33, 45].

1.3.2 Performance-based vs. patient-reported outcomes

We can divide paper-based health outcome instruments into two categories: performance-based and patient-reported[34]. Performance-based instruments involve direct observation by a rater (i.e. the therapist or researcher). The de Morton Mobility Index (DEMMI – see Appendix E) is an example of a performance-based instrument further described in Chapter 3. Patient-reported outcome measures (PROMs), on the other hand, refer to measures that cannot be directly observed by an examiner, and involve an individual's self-reported perceptions, opinions,

feelings, or experiences[46]. The Patellofemoral Pain and Osteoarthritis Outcome Scale (PFOOS – see Appendix A) is an example of a PROM further described in Chapter 2.

1.3.3 Generic, population-specific or patient-specific outcome measures

Health outcome instruments are designed to glean information defined by a clinical or research question. To answer the question, instruments can target (i) one or more diseases or conditions; (ii) one or more body regions/systems; or (iii) one or more individuals[34].

At the broadest level, generic instruments are designed to compare health states across multiple populations, such as across a range of diseases. A common example of this would be the SF-36[44]. The DEMMI[47] is also a generic instrument, in this case assessing mobility in older adults whose mobility may be limited by a variety of conditions or factors.

Instruments can also be designed to target a more narrowly defined population. This can include targeting a given condition or disease, a given system, or both[34, 48]. A condition-specific instrument targets a defined disease. An example is the Western Ontario McMaster Osteoarthritis Index (WOMAC), which targets osteoarthritis of the knees or hips[34, 49]. A system-specific instrument targets a body region. For example, the Anterior Knee Pain Scale (AKPS) targets the knee but may be used across a variety of conditions affecting the knee [50, 51]. Often, an instrument is designed to be both condition- and system-specific, such as the Knee injury and Osteoarthritis Outcome Scale (KOOS)[52].

At the narrowest level, instruments can be designed to target a single person. For example, with the Patient-Specific Functional Scale (PSFS) individuals generate their own list of up to five activities that they currently feel they are having trouble with as a result of their given condition or situation[53]. Another example is the Geriatric Quality of Life Questionnaire, in which an individual is presented with a list of 24 items and they select the eight they feel are most bothersome to them[54].

Each level of measurement (generic, population, or individual) has particular strengths and weaknesses (see Table 1.1). With regards to research, Streiner and Norman[32] and others[55] recommend that a generic scale be administered as well as a population-specific instrument in order to obtain the benefits they each offer in terms of generalizability and sensitivity to change.

Table 1.1 Comparing levels of measurement (generic, population-specific, person-specific)

Level	Definition	Example	Pros[32, 48]	Cons[32, 48]
Generic	Multiple body regions, multiple conditions	SF-36[44]	i) Can compare across many health states; ii) Often more data available re: validity/ reliability; iii) Particularly useful for informing policy and resource allocation decision-making	i) May miss relevant information for a specific condition; ii) May have irrelevant information for a specific condition iii) May show smaller effect sizes for a specific condition
Population-specific	Condition-specific: Multiple regions One condition System-specific: One body region Multiple conditions	WOMAC[49] KOOS[52] AKPS[50]	i) May be more responsive than generic scales in specific population ii) Potentially better content validity for a given population	i) Limited generalizability or comparability across conditions ii) May miss secondary effects important to the individual due to narrower focus
Person-specific	Individual self-selects items of importance/ relevance	PFPS[53] GQOL[54]	i) All items are relevant and meaningful to the individual; ii) No items are irrelevant iii) Helpful in clinical settings for specific goal setting and evaluation	i) Cannot compare across individuals, let alone populations ii) Hard to establish validity & reliability iii) Limited research applications

1.4 Purposes of health outcome instruments

Ultimately, health outcome instruments are used to assess health status[34]. Within this overarching goal, instruments have three broad purposes: evaluative, discriminative and predictive[32, 33]. For example, a mobility instrument can be administered for the purpose of determining an individual's mobility at a single time point ('evaluative'). This measure could serve as a baseline against which future evaluations can be compared. An instrument may additionally provide cut-points to help differentiate between groups with different attributes, thus serving a 'discriminative' role. For example, a person whose mobility score is below a certain cut-point might be identified as being in need of therapeutic intervention compared to someone whose score is within a healthy range. Finally, an instrument could be used for 'predictive' purposes, such as when a person's mobility score identifies them as being at high risk for future falls. Instruments can be designed and developed for any or all of these purposes from the outset, or they can be developed for one purpose and subsequently validated for additional purposes[32].

1.5 Measurement properties of health outcome instruments

Mokkink et al.[56] define a 'measurement property' as "a feature of a measurement instrument that reflects the quality of the measurement instrument". These 'features' can broadly be described in three categories: validity, reliability and responsiveness[47, 52, 56-58], each of which will be defined below (Sections 1.6 to 1.8). Two additional concepts, interpretability and usability, are also described (Sections 1.9 and 1.10). Understanding the nuances of measurement properties and their relevant statistical tests is no small feat[59], owing to the fact that these "deceptively simple"[32] concepts are, in fact, substantially complicated[32-34, 56]. In health care, the cost of drawing incorrect conclusions from an instrument can be devastating[60], highlighting the importance of understanding measurement properties both clinically and scientifically.

It is important to note that all measurement properties are *context specific*. It is not sufficient to say that an instrument is valid and reliable. It must be described as being valid and reliable *for a specific population and purpose*[32]. The degree to which scores from an instrument are valid and reliable determines the confidence with which one can draw inferences and conclusions about a given scientific or clinical inquiry and specific population.

Measurement properties will be described in this document using Mokkink et al's taxonomy (see Figure 1.2)[56, 61, 62]. Their guideline, called the COSMIN checklist, was developed for PROMs, however the authors recommend the checklist for use with all health outcome instruments[56]. Terms synonymous to 'measurement properties' include 'psychometrics' and 'clinimetrics'[63, 64], and the history and some inherent challenges regarding these different terms can be reviewed elsewhere[32, 34, 64-66].

1.6 Validity

Most simply, validity is defined as the extent to which an instrument measures the construct it is intended to measure[33, 56]. Three different aspects of validity are commonly defined: content, criterion, and construct validation[61] (see Figure 1.2)[32, 56, 67, 68].

Figure 1.2 The COSMIN categories of measurement properties. *Reprinted with permission*[1].

[1] Reprinted from Journal of Clinical Epidemiology, Volume 63, Mokkink LB et al., "*The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes*", pp 737-45, 2010, with permission from Elsevier.

1.6.1 Criterion validation

Criterion validation refers to how well an instrument performs compared to a gold standard instrument[32, 33, 62, 63]. Objective criterion measures rarely exist for measuring abstract health constructs like mobility, pain, or quality of life[32-34, 40, 62, 63, 69, 70]. If they do exist, there are several scenarios where the development of a new instrument is justified and criterion validation should be undertaken. First, an instrument historically accepted as the gold standard may be shown on closer inspection to have inadequate validity and reliability[32, 33]. A new instrument might also be developed if the criterion instrument is cost-prohibitive, and a new and less expensive technology becomes available. A third indication for criterion validation is when a shortened version of a previously validated instrument is developed. For example, the Short Form 12 (SF-12) is a shortened version of the SF-36. This new shorter scale has undergone criterion validation, with the SF-36 behaving as the 'gold standard', to compare performance of the two scales[71]. Appropriate statistics for criterion validation include correlations or area under the receiver-operator curve[32, 61].

1.6.2 Content validation

Content validation can be undertaken when a criterion measure does not exist[32, 56, 62]. 'Content' refers to both the depth and breadth covered by all items of an instrument[32]. In other words, how well does it capture *all* aspects of the construct of interest, and does it capture *only* aspects of that construct, or are there missing or irrelevant items[32]? If an instrument appears to measure the construct of interest, it can be said to have 'face validity'[32, 56].

There is no statistic for confirming content validity; rather, it is a matter of expert opinion[33, 62, 63]. Ideally, expert opinion should include expert clinicians and researchers but should also include members of the relevant patient population[32, 72]. For example, a hypothetical group of researchers might be interested in measuring mobility in older adults. They might develop a draft of an instrument that contains four questions they feel adequately measure mobility:

1. How far are you able to walk?
2. How much difficulty do you have getting in and out of the bath?
3. How much difficulty do you have climbing stairs?
4. How much difficulty do you have getting up from a chair?

The items are reviewed by a group of expert clinicians, who decide that walking distance is not as important as walking speed, since they know their patients typically report trouble when they can no longer cross the road before the traffic lights turn red[73, 74]. The question therefore becomes:

1. How fast can you walk?

The team might then agree that these items are suitable for assessing mobility in older adults. However, if a patient group is consulted, they might find out that gait speed is not important to this population, but rather pain during walking is. Secondly, contemporary homes may not have a bath tub, or people prefer to shower – this item may not be relevant to a large portion of the population. And finally, getting up from a chair with arm rests is achievable, but rising from the toilet is of greater concern to this population. These important considerations might be missed without patient input.

1.6.3 Construct validation

Construct validation defines how well an instrument measures an abstract concept, or construct[33]. Constructs are often theoretical in nature (see Section 1.2)[33], and these theories and assumptions must be tested in order to confirm validity. Contemporary views of validity are that all forms of validation contribute to construct validity[32, 67, 68]. The COSMIN definition, however, is somewhat narrower, and includes hypothesis testing, structural validation, and cross-cultural validation[56, 61, 62].

Hypothesis testing

If an instrument possesses construct validation, then outcomes should be consistent with *a priori* beliefs or hypotheses of how scores should behave[32]. Hypothesis testing is used to establish 'convergent', 'divergent', and 'known groups' validation.

'Convergent' validation is confirmed when scores for two instruments measuring a similar construct behave in similar ways. This is measured using correlation coefficients such as Pearson's r or Spearman's rho[45, 63]. Correlations between instruments that are expected to converge should range between 0.4 – 0.8[32]. For example, a group that reports high perceived functional abilities might also demonstrate high levels of mobility on a performance-based instrument. This was demonstrated by Davenport et al.[75], who reported a moderately high

Pearson's correlation of 0.69 between the performance-based DEMMI to the self-report scores obtained in the Lower Extremity Functional Scale (LEFS).

'Divergent' validation (or 'discriminant' validation) is established when correlations between unrelated constructs would be expected to be low[32, 33, 62, 63]. The same group with high mobility might demonstrate vastly differing IQ scores amongst them, as the two constructs are not related. Importantly, correlations for both convergent and divergent validation do not need to achieve statistical significance; it is only important that the overall pattern of correlations reflects the *a priori* hypotheses[32].

'Known groups' validation establishes whether an instrument can differentiate between groups that differ in some meaningful way. Known groups validation (also called 'extreme groups' or 'discriminative' validation - not to be confused with discriminant validation above) involves comparing mean scores (for example using a t-test) of the two groups[32, 33]. For example, older adults who ambulate with a mobility aid (such as a cane or walker) have demonstrated lower mobility scores compared to those ambulating without an aid[75].

Structural validation

Structural validation involves the evaluation of an instrument's dimensionality[33]. This is commonly evaluated through factor analysis[32, 61], which can confirm dimensionality, identify an irrelevant item in a scale, or assist with breaking a larger scale into appropriate subscales (such as pain, function and quality of life). However, since many constructs in health care are, by nature, multidimensional, Streiner and Norman remind us that the goal of structural validation is to ensure a scale is "unidimensional enough" to draw meaningful inferences[32].

Cross-cultural validation

Once instruments have been developed and validity established for a given population or purpose, the instrument can next be validated on a new population or for a new purpose, thus improving the generalizability of the instrument[33]. Alternatively, it is not uncommon to translate instruments into different languages[76] or to culturally adapt them in some other way. For example, a mobility questionnaire enquiring about use of mobility aids would use words like "cane" or "walker" for use in North American populations, however would substitute words like "stick" or "frame" for use in other English speaking countries like Australia. When substantial

changes to the original instrument occur, the instrument must be re-evaluated for validity and reliability in its new form[56, 61].

One very common cultural shift that has occurred quite ubiquitously in that past decade or so is the use of computers and internet. In Canada, internet use is at 80% for individuals 16 and over, and is even common among older adults – 51% in ages 65 to 74, and 27% of those 75 or older[77]. As such, many PROMs are now being administered electronically. Benefits of electronically administered forms include: (i) reduced administrative costs; (ii) patient preference/acceptance; (iii) reduced risk of data entry errors; (iv) reductions in missing data; and (v) further reaching catchment area and/or reduced participant burden (e.g. out of town administration)[34, 78, 79]. Limitations include: (i) requirement of computer literacy and internet access; and (ii) less control over timing and order of test administration[34, 78, 79]. A large meta-analysis of 46 publications comparing electronic and paper-and-pencil formats of 278 scales concluded that scales in these two formats are grossly equivalent[78, 79]. It is therefore acceptable to administer in either format, though format equivalence should be evaluated in cases where substantial changes have been made between formats (e.g. changes in wording of items or instructions, response options, or number of questions visible on one sheet/screen)[78, 79]. While included under the umbrella of 'validity' here, format equivalence involves both validation and reliability. Validation could include cognitive interview techniques on a small sample to ensure items are being interpreted the same way in both formats. Reliability testing would be similar to assessing test-retest reliability (see Section 1.7.2).

1.7 Reliability

Reliability describes "the extent to which a measurement is consistent and free from error"[32, 33]. Importantly, a reliable test is not necessarily valid[33] (see Figure 1.3c). Reliability is subcategorized into: internal consistency, reliability (relative measures), and measurement error[61](see Figure 1.2).

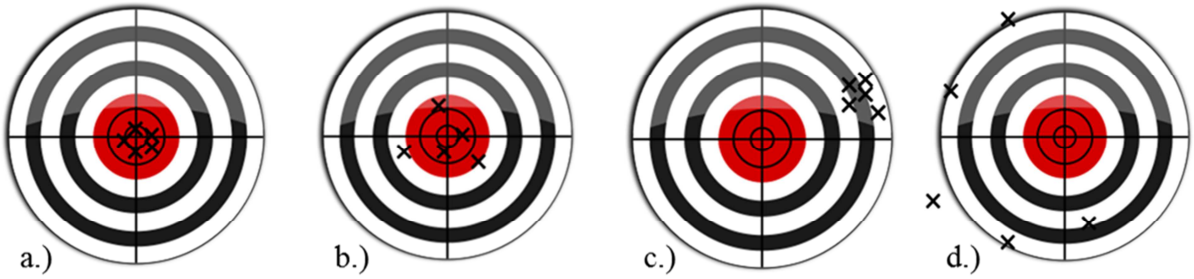

Figure 1.3 Validity vs. reliability

a.) valid and reliable; b.) valid, not as reliable; c.) reliable but not valid; d.) neither reliable nor valid

1.7.1 Internal consistency

Internal consistency is "the degree of interrelatedness among items"[62]. This means that scores for items of a unidimensional instrument should change in a similar direction and magnitude if the construct being tapped by the instrument changes. Importantly, this does not mean that high internal consistency confirms unidimensionality[80] – a physically fit individual might answer two items in a similar way even though they represent different dimensions such as function and quality of life.

Cronbach's α is the most widely used statistic to assess internal consistency[33, 80-82]. This statistic ranges between zero (0) and one (1), where one represents perfect internal consistency; $\alpha \geq 0.70$ or 0.80 reflects adequate internal consistency[32, 79, 80]. Cronbach's α will be higher when a unidimensional instrument has high internal consistency. However, Cronbach's α also varies for other reasons. In addition to item intercorrelation, Cronbach's α will vary as a function of the variability of the sample, with a more heterogeneous sample obtaining a higher α[42, 80]. Also, as the number of items within an instrument increases, so too does α[32, 42, 80]. Finally, Cronbach's α will decrease as the number of dimensions increases[42, 80]. Cortina[80] described an instrument with two dimensions, moderate correlation (r=50) and 12 items that still had an adequate α of 0.78. The same instrument but with 18 items would have an α increased to 0.85[80]. This also demonstrates that α can still be higher than 0.70 even if it is not unidimensional.

It should be noted that Cronbach's α is not an independently meaningful statistic. Rather than prematurely judging an instrument on Cronbach's α alone, it is advisable to engage in appreciating the nuances underlying the statistic[42]. One must evaluate dimensionality of the instrument; look carefully at all items; consider whether there are enough items to have adequate

scope (content validity) without creating item redundancy; and consider the variability of the sample. A low Cronbach's α likely indicates low internal consistency or multidimensionality. However, at the other end of the spectrum a high alpha may indicate item redundancy or a very heterogeneous sample. Therefore, very high alphas (>0.9) should also be interpreted with caution[34].

1.7.2 Reliability

This subdomain refers to the "degree to which the measurement is free from measurement error"[56]. If an individual has not changed, then the score also should not change between subsequent test administrations[32, 33]. This can be assessed with 'test-retest reliability', 'intrarater reliability' or 'interrater reliability'.

Simple correlations such as Pearson's r have been used to calculate reliability coefficients[33, 47, 63]. However, covariance does not account for systematic differences between test administrations[32-34, 79, 83-85]. A more appropriate statistic to use is the intraclass correlation coefficient (ICC)[59, 61, 78, 79, 84, 85]. ICC represents the variance of interest (e.g. variance of test scores) as a percentage of the total variance (i.e. variance of interest plus error variance)[32, 33, 59, 84]. It can be written as

$$reliability = \frac{subject\ variance}{subject\ variance + error\ variance} \qquad (1)$$

ICC theoretically ranges from zero (0) to one (1), with one representing perfect reliability. An ICC value of ≥ 0.70 is generally regarded as acceptable[32, 57, 79, 83]. McGraw and Wong[85] describe ten types of ICC that fall within three models. Which one to use depends on the study question and methods, the intended clinical use, and whether or not one wishes to evaluate systematic differences (i.e. absolute agreement) between test administrations [32, 33, 59, 78, 79, 83, 85].

Like internal consistency, relative measures are context specific and reflect the population, methods and circumstances specific to a study – they are not static characteristics of an instrument itself. For example, in test-retest reliability, issues that might affect a score include participant fatigue (lower scores over time), learning effects (systematically improving scores

over time), memory effects (remembering a score from last time or forgetting something relevant like a previous episode of pain or a fall), motivation or competitive behaviours[33]. Factors in rater reliability studies include standardization of tester training and methods, plus the effects of memory, learning and fatigue of the tester(s). Study methods should consider potential sources of variability and design studies accordingly (e.g. optimizing inter-test intervals)[33, 83]. However, some contextual factors cannot be as easily controlled through study design. For example, some variables are inherently unstable to measure or control (such as mood)[33]. Also, as in the case of internal consistency, the statistic includes between-subjects variability[84], and therefore even if variability from trial to trial is very small, a heterogeneous sample will have high between-subjects variability and thus the statistic will be higher[32].

1.7.3 Measurement error

Mokkink et al.[56] define measurement error as "the systematic and random error of a patient's score that is not attributed to true changes in the construct to be measured". Quantifying measurement error is an essential component of determining reliability. ICC is a unitless statistic, and used alone provides only the proportion of variance that is attributable to true variance[33, 84, 86]. Measurement error is reported in the same unit as the health outcome instrument, and provides an estimate of reliability that is considered to be more of a "fixed characteristic" of an instrument than the ICC[84]. ICC should always be reported together with measurement error to maximize interpretability[33, 84].

The statistic of choice for instrument measurement error is the *standard error of measurement* (SEM). The SEM is most simply defined as the standard deviation of measurement error[33]. Therefore, there is a 68% chance that an individual's 'true' score will be within 1 SEM of a test score, and 95% confidence intervals[83, 87] can be added to an individual test score with

$$\mu_i = x_i \pm 1.96\text{*SEM} \qquad (2)$$

where μ_i is the theoretical 'true' score; and x_i is the score for one test administration.

As with other measurements of reliability (e.g. ICC) there are many ways to calculate SEM[87]. A common method cited is

$$\text{SEM} = s_x\sqrt{1 - ICC} \qquad (3)$$

where s_x is the standard deviation of the observed scores[32, 33, 63, 84, 87]. Fortunately, Stratford[87] illustrates that SEM is a highly stable measurement across several different calculation methods and any sample size, supporting the notion that SEM is a relatively fixed characteristic of an instrument[84]. One of the easiest methods to measure SEM is to report the 'root mean square error' ($\sqrt{MSE}$) following a repeated measures analysis of variance (ANOVA) which is calculated prior to determining the ICC.

Once calculated, SEM should be compared to the expected range of scores for the instrument, and the expected changes in scores, to consider whether an instrument will be suitable for a given purpose[32]. For example, the SEM (and more importantly, its extension, minimal detectable change - see Section 1.9.1) should be smaller than the minimal amount of change deemed to be important for an instrument[83].

1.8 Responsiveness

Responsiveness refers to the ability to "detect change over time in the construct to be measured"[56, 62]. While responsiveness is commonly presented as a concept separate from validity and reliability, it actually pertains to the validity of *change* scores as opposed to *cross-sectional* validity, or validity of a score at a single time point[32, 56, 88, 89]. Given this definition, one can refer to the same subcategories of criterion, content, and construct validation as described above (see Section 1.6), with the caveat that hypotheses should refer to those of *change* scores rather than a single score[62]. Determining responsiveness permits a scale to be used for evaluative purposes[33], facilitating its use in longitudinal or interventional trials.

Some researchers define responsiveness as the ability to detect 'meaningful' or 'significant' change[45, 83, 90, 91] while others define it as the ability to detect 'any' change[56, 62, 92, 93]. The COSMIN checklist[62] employs the latter definition ('any' change) in order to not muddle the concept of *detecting* change with the *importance* of the detected change (which is an issue of

interpretability – see Section 1.9.2). Statistics selected for confirming responsiveness, therefore, depends on which of the two definitions one uses. Based on the COSMIN definition[56], correlations should be used to compare change scores of the instrument in question to those of another known scale[61, 62].

1.9 Interpretability

Establishing validity, reliability and responsiveness gives clinicians and researchers a certain level of confidence as to how well a health outcome instrument is measuring what it is intended to measure. An additional important step in instrument development is to provide qualitative meaning to an outcome score[56].

1.9.1 Minimal detectable change

Minimal detectable change (MDC) is sometimes also referred to as 'smallest detectable change' (SDC) or 'minimal difference' (MD)[84, 86]. This statistic refers to the smallest change in score that exceeds measurement error and can therefore be interpreted as representing true change in an individual[33, 47, 63, 86, 94]. MDC is based on the standard error of measure (SEM) and is typically calculated as MDC_{90} or MDC_{95}[47, 75, 95], indicating whether the statistic is designed to overcome error at 90% or 95% confidence [87, 96]:

$$MDC_{90} = 1.65 * \sqrt{2} * SEM \quad (4)$$

$$MDC_{95} = 1.96 * \sqrt{2} * SEM \quad (5)$$

MDC_{95} is recommended in situations where the outcome will guide higher-risk decision making such as whether surgery is indicated[95]; in many other clinical settings MDC_{90} is suggested [47, 75, 83, 87, 96]. In situations where group changes are being evaluated (such as in clinical trials), the MDC_{group} is calculated with MDC divided by $\sqrt{n}$ [83, 97, 98].

1.9.2 Minimal clinically important difference

While the MDC is important in determining whether the amount of change observed exceeds measurement error, the ultimate question is not simply "has the patient improved?" Rather, it is "has the patient improved in a meaningful way?" The definition of 'important' depends on the lens through which we assess a change[33]. The definition also depends on the purpose of the

study: does an intervention ‘improve balance in frail older adults’ compared to ‘reduce incidence of falls’[33].

Minimal clinically important difference (MCID) is the statistic that reflects the relevance of a change score. It can be measured using either an ‘anchor-based method’ (also called a ‘criterion-based method’) or a ‘distribution-based method’ (also called a ‘normative-based method’)[32, 33, 47]. An example of an anchor-based method would be to evaluate difference in mean change scores between two groups. For example, in health care, patient perspective is valued in establishing MCID, so one could use a ‘global change scale’ to help determine MCID. A global change scale is usually a single question that asks a participant to subjectively rate change over a specified period of time or following some intervention. The responses are normally on a Likert scale. We therefore evaluate the differences in change scores between a group that feels “no better” following intervention vs. a group that feels “a little better” or “a lot better” [62, 86, 93]. An example of a distribution-based method is that developed by Norman et al. following a systematic review in which they determined that MCID consistently approached half of a standard deviation[99].

Having established the MCID, it is now appropriate to compare this to the MDC previously calculated (see Section 1.9.1). The MCID must be larger than the MDC (i.e. for individual comparison, or MDC_{group} for group comparisons) in order for it to represent both real and important change.

1.9.3 Floor/ceiling effects

Floor and ceiling effects can limit interpretability of a health outcome instrument. A floor effect is said to occur when a large portion of a sample scores the lowest possible score on an instrument; and a ceiling effect occurs at the other extreme end of the scale. Up to about 15% at either end of the scale is reportedly acceptable[100, 101]. If floor or ceiling effects exist, the instrument will have limited ability to detect changes in individuals at the ends of a scale. In other words, someone scoring at the high end of a scale will not score any higher even if they experience true improvement in the construct being measured[33]. This can limit responsiveness of the scale.

While floor and ceiling effects do have the potential to limit validity of an instrument, these statistics should be taken in context. For example, if an instrument is measuring pain as a feature of a certain disease or condition, and it is common for a large portion of individuals with this condition to be asymptomatic, then scores will bottom out in this portion of the sample. However, this is not a limitation of the instrument, per se, but a characteristic of the underlying construct. Similarly, if a sample of very healthy, happy people completed a quality of life scale such as the SF-36, it would be expected (and perfectly appropriate) that a ceiling effect would emerge from this sample (indeed, ceiling effects of as high as 56% have been reported[100]). If a health instrument has been rigorously developed, and demonstrates good content validity, floor and ceiling effects might be expected for some groups. Importantly, one must clarify the purpose of the instrument: if the instrument is designed to evaluate change over time, then the sample scores of interest should not demonstrate floor or ceiling effects.

1.9.4 Normative / referent data

A score or measure from any health instrument is meaningless without a frame of reference for comparison. References can be anchor-based or distribution-based. An anchor (or 'criterion') reference is generally a set score that has some inherent meaning in terms of the health outcome of interest[33]. For example, a person who has sufficient mobility to walk 1.2 metres per second will be capable of crossing most streets before the traffic light changes[73]. A distribution-based (or 'normative') reference, on the other hand, compares a score to those of a reference population[33, 61]. Using the same example, 'normal' walking speed is typically between 1.2 – 1.4 m/s[73].

While the term 'normative' data remains in common use today, a more appropriate term is 'reference' data[102]. The term 'reference' implies that any population of interest can be used for comparison. Second, the fact that someone scores beyond two standard deviations of 'normal' does not necessarily imply that they are 'abnormal'. Finally, the term 'normal' was historically interpreted as meaning 'the absence of disease'[5]. Today, it is well recognized that older adults display considerable heterogeneity within the spectrum of non-disease states[103]. The term 'reference' reduces the potential for misinterpretation and allows some flexibility in identifying a relevant reference population.

Reference intervals can be presented for the purposes of individual comparison or for group comparison. Reference data are generally presented as ranges or intervals within which most individuals in a group of interest would fall (as opposed to criterion references which are more often cut-points). Reference intervals provide a range of scores usually representing the middle 95% (or 90%) of a reference population, and can be used for scores that are normally or non-normally distributed[102]. Confidence intervals around group means can be used when scores are normally distributed. For comparisons at a group level (i.e. comparing mean scores of a group to the reference population of interest), confidence intervals should be based on standard error (rather than standard deviation).

1.10 Usability

Having established validity, reliability, responsiveness, and interpretability, one question remains. Will the instrument be adopted by clinicians and researchers?

Usability should be a consideration throughout the instrument development process, from the moment of inception, in order to ensure that the instrument will be attractive to its intended audience, thus maximizing uptake. There are several features of usability. For example, how feasible is the instrument in a clinical vs. research setting? How much space and equipment is required for testing, and what are the associated costs of test administration? Who can administer the testing and how much training is required? How long does the test take to complete, and are there any associated issues such as patient fatigue or other perceived burden? Is it safe to administer to patients or is there any risk associated with the tests? Can a clinician complete the testing within the usual time allotted with a patient, or will it interfere with treatment time? Does the researcher believe it to be a valid and relevant tool for their population of interest?

It is recommended that stakeholders are consulted in the very early stages of instrument development to optimize usability (and validity as previously discussed), and that the process of development be iterative such that items can be modified as the developers learn more about the instrument through research and consultation. Usability will be optimized through a two-direction model of communication consistent with current theories in knowledge translation[104].

1.11 Study objectives

The overall objective of my thesis work was to evaluate the measurement properties of two new instruments designed to assess health outcomes associated with age-related comorbidities.

1.11.1 Objectives of study #1

The primary objective of study #1 was to evaluate the validity and reliability of a new patient-reported outcome scale, the Patellofemoral Pain and Osteoarthritis Outcome Scale (PFOOS). This scale was developed for assessing health outcomes in five different domains (pain, symptoms, activities of daily living, activities of sport and recreation, and quality of life) relating to pain and osteoarthritis at the patellofemoral joint.

1.11.2 Objectives of study #2

The primary objective of study #2 was to assess mobility of community-dwelling older adults using a performance-based health outcome instrument, the de Morton Mobility Index (DEMMI), and to contribute to the interpretability of the DEMMI by developing reference scores for healthy men and women over 60 years old.

Chapter 2: The Patellofemoral Pain and Osteoarthritis Outcome Scale: Validity, Reliability and Interpretability

2.1 Introduction

Knee osteoarthritis (OA) is one of the five chronic conditions (including stroke, depression, hip fracture and heart disease) responsible for the most physical disability of any other diseases in older community-dwelling adults[6]. Radiographic knee OA (i.e. signs of degenerative changes on X-ray) affects almost 40% of North Americans, and one third of these individuals are symptomatic[20].

Traditionally considered one joint, the 'knee' is more accurately described as a 'joint complex' made up of multiple compartments: the medial and lateral tibiofemoral, and patellofemoral compartments. Historically, interest in the knee has often focused on the tibiofemoral joint. The importance of this distinction will become evident in the ensuing paragraphs.

Knee OA is typically diagnosed based on radiographic changes and clinical presentation (signs and symptoms). Radiographic knee OA is commonly assessed with a posterior-anterior (PA) view X-ray[105], and degenerative changes include joint space narrowing, osteophyte formation, subchondral sclerosis and subchondral cyst formation[106, 107]. Classifying the severity of radiographic changes is done using grading systems such as those developed by Kellgren & Lawrence[106] or the Osteoarthritis Research Society International (OARSI)[105]. Clinical features, on the other hand, include localized knee pain; brief morning stiffness; functional limitations; crepitus; bony enlargement; and restricted movement[107]. Peat et al.[108, 109] caution that these classic clinical features of knee OA tend to be more accurate in cases of advanced OA, and also tend to reflect clinical features of tibiofemoral OA more so than patellofemoral. They argue that clinical guidelines are needed to identify both earlier stages of knee OA as well as patellofemoral OA.

2.1.1 Patellofemoral osteoarthritis

Both clinical and scientific investigations of knee OA have traditionally focused on the tibiofemoral joints, and this has resulted in vast under-recognition of both the prevalence and clinical relevance of OA of the patellofemoral joint. This may be due in part because the patellofemoral joint cannot be adequately examined with a standard PA X-ray, and instead

requires alternate views such as skyline or lateral[108]. In addition, patellofemoral OA tends to present clinically quite differently from tibiofemoral OA, and even in its advanced stages the clinical signs of patellofemoral OA are not consistent with the “classic signs of knee OA”[108].

Interest in OA of the patellofemoral joint has emerged in recent years and the scientific knowledge gleaned suggests that patellofemoral OA is more prevalent than previously thought[110-116]. In a study of 777 adults over 50 years old who reported knee pain, 531 (68%) had evidence of radiographic knee OA[117]. Importantly, the patellofemoral joint was involved in 94% of those with radiographic knee OA: 314 had combined tibiofemoral/patellofemoral OA (59% of OA sample, or 41% of the full knee pain sample); 186 had isolated patellofemoral OA (35% of OA sample, 24% of knee pain sample); and only 31 had isolated tibiofemoral OA (6% of OA sample, 4% of knee pain sample)[117]. This high prevalence of patellofemoral joint involvement may be in stark contrast to current diagnostic patterns in clinical practice. A retrospective review of 57,555 general practice medical charts in the UK revealed 1782 knee-related consultations (though admittedly this included people aged 15 or older); of these, 303 were coded as disorders of the patellofemoral joint, and only 13 of these (<0.05%) were diagnosed as patellofemoral OA[118].

In addition to its prevalence, patellofemoral OA is clinically important. The presence of isolated patellofemoral OA predicts progression to generalized (multi-compartment) OA[119, 120]. Also, patellofemoral OA has greater association with knee OA symptoms than tibiofemoral OA[113-115]. Duncan et al.[113] reported a significant correlation between self-reported symptoms (pain, stiffness and function using the WOMAC) and radiographic severity of isolated patellofemoral OA including those classified as having ‘mild’ patellofemoral OA. This suggests that even people with early isolated patellofemoral OA experience symptoms of pain and reduced function, with the most difficult tasks including descending stairs, getting in and out of the bath, and getting in and out of a car[113]. Symptoms associated with isolated patellofemoral OA are at least as severe as with either isolated tibiofemoral or multi-compartment knee OA; in cases of mild patellofemoral OA, pain and function may be worse than isolated tibiofemoral OA[121].

While patellofemoral OA is still poorly understood relative to tibiofemoral OA, there are signs and symptoms that currently define patellofemoral OA. Importantly, many of these symptoms

are common to other causes of anterior knee pain[122]. Signs and symptoms associated with patellofemoral OA include anterior knee pain that is made worse by tasks that load the patellofemoral joint. For example, pain and limited function may occur with ascending and descending stairs; getting in or out of a bath; getting in or out of a car; or rising from sitting[113]. Stiffness may occur after resting (sitting or lying) for prolonged periods or first thing in the morning[113]. With moderate to severe patellofemoral OA, clinical signs can include swelling, valgus deformity, quadriceps weakness, and pain on compression of the patellofemoral joint[108].

2.1.2 Justification for a new patient-reported outcome measure (PROM)

Observation-based measurements such as diagnostic imaging (X-ray) or clinical tests (effusion, range of motion) are insufficient for evaluating the patient-relevant symptoms of patellofemoral OA. Additionally, in the case of early isolated patellofemoral OA, clinical signs are lacking, and Peat et al.[108] recommend that early patellofemoral OA be evaluated through self-reported symptoms more so than clinical signs. This is consistent with recent overall trends and recommendations in healthcare towards including PROMs as a component of comprehensive health assessment[32, 123, 124].

When searching for a suitable PROM, it is important to consider whether the PROM is suitable for use with the specific research question and target population; is sufficiently valid and reliable to enable correct interpretation of study results; and is feasible for the study of interest (e.g. is not cumbersome to complete or score)[32]. Without meeting these conditions, clinicians and researchers may be left with difficulties in interpreting study results and limited ability to generalize findings across studies.

Currently there is no gold standard condition-specific PROM for assessing patellofemoral pain or OA, which can render it difficult to select a PROM when designing trials. This is reflected in the vast array of questionnaires that have been used in studies to date[50, 55, 69, 70, 125-136]. A recent systematic review of 37 knee-related outcome measures identified only one scale of ‘sufficient quality’ that targeted anterior knee symptoms, Kujala’s Anterior Knee Pain Scale[50, 137], though this was not designed to evaluate patellofemoral OA. Many researchers investigating patellofemoral pain or OA have also opted to use generic instruments[55, 69, 129, 133, 138] (see Section 1.3.3 to review generic vs. condition-specific instruments), or knee-

related scales[113, 139, 140]. Presumably the justification for selecting a knee-specific scale is because validation studies have included patellofemoral disorders, however it should be noted that these sub-samples were not sufficiently powered to establish validity or reliability for this population[58, 101, 141].

Given the lack of an existing suitable PROM for patellofemoral pain and OA, Crossley, with co-investigators Cowan and Roos, undertook the task of developing a new PROM to target this important clinical population (unpublished).

2.1.3 The Patellofemoral Pain and Osteoarthritis Outcome Scale (PFOOS)

PFOOS instrument developers used the Knee injury and Osteoarthritis Outcome Scale (KOOS) as a template for the new scale[52]. The KOOS is a patient-reported outcome instrument designed for use in long-term evaluation of post-traumatic knee pain and knee OA[52, 142, 143]. The KOOS consists of five subscales: Pain, Symptoms, Function in Daily Living, Function in Sport and Recreation, and Quality of Life. It has been used in studies of various patellofemoral disorders, but has not been formally validated for use in a patellofemoral OA population[144-148].

The PFOOS instrument developers kept all items of the KOOS and added new items based on current knowledge about patellofemoral OA. This was done in consultation with relevant content experts[149] that attended the Second International Patellofemoral Pain Retreat held in Belgium in 2011 (thus a sample of convenience). This included surveys provided to orthopaedic surgeons (n=3), rheumatologists (n=1), sports physicians (n=2), medical doctors (n=1), physical therapists (n=7) and researchers who are active in the field of patellofemoral pain (n=14). Importantly, patients (n=44) were also contacted directly for initial content input. The surveys consisted of open ended questions, and all recommended items were included in the initial draft of the PFOOS. Twenty patients then completed the instrument to pilot the items. This sample size is consistent with recommended sizes (eight to 15) aimed to ‘sample to redundancy’ [32]. Cognitive debriefing followed, in which the patients were asked specific questions about how they interpreted and understood different aspects of the instrument. Items were modified or removed following this feedback process to create the preliminary PFOOS (see Appendix A).

2.1.4 Study objective

The purpose of this study was to evaluate the validity, reliability and interpretability of the preliminary version of the PFOOS.

2.2 Methods

2.2.1 Study design

This was a validation study with methods designed and results reported to satisfy the COSMIN guidelines for evaluation of measurement properties of patient reported outcome instruments[61]. In addition, where possible I incorporated the reporting recommendations of the CONSORT guideline's 'STARD Initiative'[150] ("Standards for Reporting of Diagnostic Accuracy") since the CONSORT guidelines[151] do not currently offer recommendations specific to the reporting of measurement properties of non-diagnostic health outcome instruments.

This study represented Phase 2 in the development of the instrument (Phase 1 being item generation, see Section 2.1.3). Validation (beyond content validation, which was established in Phase 1) was achieved through hypothesis testing (convergent, divergent, and known groups validation) and reliability was examined through evaluating internal consistency, test-retest reliability, and measurement error. Alternate forms reliability (online vs. paper-and-pencil) was also evaluated through a nested cross-over study design. Item reduction was considered based on item performance. Minimal detectable change (MDC) and minimal clinically important difference (MCID) was estimated and potential for floor and ceiling effects evaluated to enhance interpretability.

2.2.2 Participants

Eligible participants were adults aged 18 or older and living with chronic anterior knee pain. Chronic anterior knee pain was defined as:

- Peri-patellar or retro-patellar pain
- Pain rated at least 3/10 on a numeric pain rating scale
- Aggravated by tasks known to load the patellofemoral joint: squatting/crouching, ascending or descending stairs, rising from sitting, running
- Pain of at least three months duration

Exclusion criteria included diffuse or generalized knee pain; history of total knee or total hip replacement; or any severe trauma to the target knee in the previous year, such as meniscal injury or surgery.

2.2.3 Recruitment

Recruitment strategies for this convenience sample were primarily targeted at University of Queensland (UQ) students and staff. Posters were displayed throughout the St. Lucia campus of UQ. Advertisements were included in the "Volunteers" section of the "UQ Update", a weekly online newsletter for UQ staff. Mass e-mails were sent out to students and staff within the UQ School of Health and Rehabilitation Sciences. Recruitment was also accomplished through external advertising through a local Brisbane newspaper and through similar university recruitment strategies at the University of Melbourne (UM).

Interested individuals contacted a study coordinator or research assistant by telephone or e-mail. Screening was conducted via telephone (very rarely, e-mail was used for screening). I then mailed study packages to eligible individuals who provided verbal approval to participate in the study. Within the study package was written information about the study and a consent form to participate. Individuals were encouraged to contact the study coordinator with any questions prior to giving free and informed written consent.

2.2.4 Ethics

Ethics for this study was obtained through the University of Queensland Medical Research Ethics Committee, project number 2012000025.

2.2.5 Outcome measures

The primary outcome of interest was the PFOOS. Participants also completed the Kujala Anterior Knee Pain Scale (AKPS)[50] and the Short Form 36 (SF-36)[44], as well as some general questions (see Appendix D).

Patellofemoral Pain and Osteoarthritis Outcome Scale (PFOOS)

The PFOOS was the primary outcome in this validation study. The preliminary version of the scale consists of five subscales: Symptoms including stiffness (containing eight items), Pain (19 items), Function in Activities of Daily Living (ADL) (17 items), Function in Sports and

Recreation (Sport/Rec) (five items), and Quality of Life (QOL) (five items). Individuals were asked to answer the questions with regards to their symptoms during the past week. The response scheme was similar throughout the subscales, with five graded response options in an adjectival scale ranging from zero to four, with zero representing 'no problem'. Adjectival scales are unipolar in nature, meaning an individual chooses options ranging from none of the item characteristic to maximal item characteristic. This is in contrast to a Likert scale, which is bipolar and ranges from maximally endorsing an item to maximally rejecting an item[32]. Each subscale is summed, expressed as a percentage of the subscale, then reported as [100 – score], so that each subscale ranges from zero – 100, with 100 representing no disability and 0 representing maximum disability. Each subscale, while ordinal in nature, is treated statistically as an interval scale, based on recommendations that severe bias is not introduced unless score distributions are severely skewed[32, 152]. See Appendix A for the full preliminary questionnaire.

Anterior Knee Pain Scale (AKPS)

The AKPS[50] is commonly used to evaluate patellofemoral disorders[69, 153-157] and is currently one of the only acceptable quality PROMs for anterior knee pain available[69, 137]. Also known as the Kujala Scale in recognition of the scale's developer, the tool consists of 13 items. Response options are specific to each question and weighted such that summing all item scores provides a total score out of a possible maximum of 100. The scale, like the PFOOS, is treated statistically as an interval level scale. The highest score of 100 represents no disability, and 0 represents maximum disability. Strengths of the scale include its common use in the literature; ease of use (short); established test-retest reliability and internal consistency with patellofemoral pain syndrome (PFPS)[69, 129, 131] and patellofemoral instability[55]; established known groups validation in separating those with or without PFPS[50] and in separating single vs. recurrent dislocations[55]; and convergent validation where self-reported improvements correlated with AKPS scores[69]. Limitations of the scale include: time frame for symptoms not specified; not a unidimensional scale; no patient input in scale development (i.e. content validity); use of 'jargon' language in several items ("atrophy", "flexion deficiency" and "subluxation"); only three or four response options for seven of the 13 items (loss of discriminative information); arbitrary weighting of response options; and concerns of narrow content width[55, 131]. See Appendix B for the full questionnaire.

Short Form (36 item), version 2 (SF-36v2)

The Australian English version of the SF-36 version 2 questionnaire was used for this study[44, 158]. This is a generic HRQOL instrument that asks individuals to consider the past four weeks when answering the 36 items. The scale is divided into eight subscales: general health (GH); physical function (PF); role physical (RP); bodily pain (BP); vitality (VT); social function (SF); role emotional (RE); and mental health (MH). In addition, the subscales are combined and weighted to form two larger subscales, the Physical Component Summary (PCS) and Mental Component Summary (MCS). The response options vary by question from three to six options, and are adjectival in nature. For scoring purposes, initial scores (ranging from zero to a maximum of six, depending on the item) are converted into a score out of 100, and are treated as interval level subscales. Like the PFOOS and AKPS, the highest score of 100 represents no disability, and 0 represents maximum disability. Scores for each subscale are then obtained through use of proprietary software that offers both a raw score and a norm-based score considering sex and age. Strengths of the SF-36 include extensive studies of validity and reliability across a spectrum of clinical populations, including well over 1000 publications using the SF-36[44]; and the benefits of being a generic scale that offers the ability to compare scores across different populations. Limitations include the fact that generic scales tend to be less responsive than population-specific instruments. The SF-36 is used extensively in the literature, has well-documented validity and reliability across multiple populations, and is a generic scale thus enhancing generalizability across studies and populations[32, 55]. See Appendix C for the full questionnaire.

2.2.6 Test administration

Questionnaires were mailed to all study participants at baseline. At their convenience, they completed a pencil-and-paper version of the PFOOS, AKPS[50], and SF-36[44, 158], along with some demographic questions (age, sex, height, weight, knee pain history, current knee-related limitations, surgical history, physical activity, and format preference i.e. online vs. paper). Within 48 hours, a subset of participants completed the same three questionnaires in an online format using Survey Monkey[159]. All participants then completed the PFOOS within 1-2 weeks of baseline. At this time point, participants completed either the paper-and-pencil version, the online version, or both versions. For consistency, the order of the questionnaires in the paper version was the same as the order of the questionnaires online. However, given participants were completing the questionnaires from home, there was no explicit control over the order in which

they completed the forms. In order to enhance participant adherence, reminders were provided to participants via telephone and/or e-mail if completed questionnaires were not returned within the expected time frame (i.e. 1-2 weeks for paper versions, 3-5 days for online versions - see Sections 2.2.12 and 2.2.14 for more details about time intervals for questionnaire completion).

2.2.7 Statistical methods

Descriptive statistics, including central tendency and distributions, were presented graphically and in tables. Exploratory statistics included evaluating PFOOS scores by age and sex. All questionnaires (PFOOS, AKPS, and SF-36) were treated statistically as interval level scales. All statistical analysis was done using Stata Intercooled 12.0 (StataCorp, Texas, USA).

2.2.8 Sample size

I aimed to enroll 50 in this study based on recommended sample sizes for this type of study[32, 83]. This sample size is adequately powered for determining MCID, MDC, floor and ceiling effects, and reliability (test-retest) estimates[79, 83, 160]. More specifically, with a sample size of 43, Coons et al.[79], using Walter's sample size estimation methods[160], determined that a study has 80% power to assert that reliability will exceed 0.70 with 95% confidence, if the true population reliability coefficient is 0.85. Importantly, Cicchetti[161, 162] also argues that increasing sample size beyond 50 is unnecessarily costly given the lack of clinically meaningful improvement in precision that a larger sample size would offer.

2.2.9 Item reduction

The process of item reduction was based on the completed baseline questionnaires. I evaluated the paper-and-pencil format separately from the online format. Following this, I combined the two formats and performed a third evaluation. To consider an item for deletion (or modification), I evaluated all of the following possible indicators of an unsuitable item:

- Endorsement of more than 50% on the "no problem" response option
- Endorsement of more than 95% or less than 5% on any single response option
- Mean item score of less than 1

I defined performance of an item to be inadequate, and therefore a candidate for deletion, if all of these indicators were present. To accommodate items performing at the cusp of adequate performance (i.e. for further evaluation), I chose to keep items performing within a 15% window of these criteria. Therefore, for an item to be deleted, it must have at least 3 of the following:

endorsement of >58% ‘no problem’; endorsement of >96% or <4% of any item; and mean score <0.85. Additional considerations included missing items (and reasons for missing), feedback from participants regarding any item or response option, clinical considerations, as well as changes to internal consistency statistics (see Section 2.2.13) with an item removed.

2.2.10 Scoring and missing data

At baseline, missing items were quantified and described. Data missing were characterized as either: (i) an item left blank (whether unintentional or not); or (ii) an item where the ‘not applicable’ (N/A) box was selected. Missing responses for each item were reported for the entire sample as number of occurrences and as a percentage. The total number of missing responses (i.e. across all five subscales combined) was also reported, and expressed as a percentage of the total number of items for the entire PFOOS scale. Less than 5% missing was considered acceptable[163-165].

Missing items were handled using ‘person mean imputation’ for all valid subscales as has been done by the KOOS and is recommended by others (further described below)[32, 163, 165-170]. To confirm the appropriateness of this method, I quantified the amount of missing data and also looked for patterns of missingness among all variables. I then looked for a possible mechanism of missingness by creating a new dichotomous variable for each item (missing vs. non-missing) and evaluating correlations between these new variables and other variables (sex, age, pain severity, BMI, laterality of affected knee, traumatic vs. insidious onset, history of surgery, physical activity level, and access to a computer).

For scoring, when marks were placed outside a box, the closest box was chosen as the item response. For situations where two boxes were selected, the box representing the more severe answer was selected. For scoring purposes there was a set limit to the number of missing items allowable for each subscale for each individual, based on the KOOS guidelines for determining if a subscale is valid[166]. These guidelines were updated in 2012. I therefore used both the original and updated methods and did a sensitivity analysis to compare the two approaches. Using the original method, an individual’s score could be calculated (i.e. considered valid) for each subscale provided no more than two items were missing. Using the new method, a score could be calculated for each subscale provided at least half of the items within the subscale were

answered, consistent with methods recommended elsewhere[58, 171]. Since each subscale functions independently, all valid subscale scores were reported.

To score each valid subscale, all responses were first summed. In the case of missing responses, person mean imputation was employed: the average of all remaining complete subscale items was imputed into the missing response(s). The raw score was then converted to a standardized score by dividing it by the total maximum possible score for that subscale (i.e. number of items times four) and multiplying by 100 (i.e. expressed as a percentage). Finally, this number was transformed by subtracting it from 100, giving a score out of 100 where 100 represents no problem and zero represents extreme problems.

Missing data were left in an uncleaned state (i.e. missing) for the purposes of item reduction. Also, mean subscores and standard deviations were compared in a raw state (complete case analysis, where any missing item renders that individual's score invalid)[170, 172] and in a cleaned state (person mean imputation and remove invalid subscales) to look for potential bias. For all other evaluations of validity and reliability, the cleaned and imputed data were used for analysis. Central tendencies and distributions for the PFOOS subscores were presented as a full sample and were also reported by sex and by age category. For age category, participants aged 40 years or less were compared to participants aged 50 years or greater to more fully elucidate any possible difference with age (also it is more likely that a younger sample is presenting with patellofemoral pain syndrome vs. patellofemoral OA, which would be more probable in an older sample).

2.2.11 Floor and ceiling effects

The potential for floor or ceiling effects was evaluated by calculating the percentage of participants achieving scores equaling zero for each subscale (floor) and the percentage achieving scores of 100 for each subscale (ceiling)[33, 100, 101]. Up to 15% of participants scoring at either scale extreme was considered acceptable[100].

2.2.12 Alternate forms reliability

Alternate forms reliability was completed by comparing completion of the PFOOS in two different formats, online (Survey Monkey) and pencil-and-paper. Participants completed the two versions within 48 hours at baseline, and again 1-2 weeks later. This time interval is most likely to capture stable symptoms across test administrations[32, 173].

A Bland-Altman plot[174] was done initially to evaluate the potential for outliers and to confirm the assumption of homoscedasticity. Following this, repeated measures ANOVA was used to assess for any systematic differences between the two formats (using $\alpha < 0.05$ for assessing significance). Mean differences between the two formats, for each subscale, was reported as a percentage (%) of each subscale range[78, 79]. For thorough format comparison, mean differences were also compared for paper vs. paper test-retest and online vs. online test-retest. Finally, ICC(3,1) for absolute agreement (sometimes written ICC3 (A,1)) was evaluated for the two formats, and again for paper vs. paper and online vs. online test-retest. Comparing in these three ways (online vs. paper, paper vs. paper, and online vs. online) is recommended to evaluate equivalence between the two methods, but also to determine if the online format has better reliability than pencil-and-paper format[78, 79]. For example, if online methods reduce missing data, there exists potential for it to perform superiorly to the traditional pencil-and-paper format[79].

2.2.13 Internal consistency

Internal consistency was evaluated using Cronbach's alpha (α), reported together with average inter-item correlation[42, 82]. Cronbach's α will ideally fall between 0.7 and 0.9[32]. Less than 0.7 may call into question the unidimensionality of the subscale; more than 0.9 may suggest item redundancy. Mean inter-item correlation should be moderate, ideally at least 0.25[32].

2.2.14 Test-retest reliability

Test-retest reliability was assessed on completion of the PFOOS at two time points by each participant, 1-2 weeks apart[32, 57, 83]. This time interval is optimal since with a longer time period the underlying condition and symptoms are more likely to have changed; and with a very short interval there may be bias due to the influence of memory of the first test, or fatigue, on the second test administration[32, 173].

Prior to assessing test-retest reliability, alternate forms reliability was assessed to determine if the two formats are equivalent. If equivalent, test-retest reliability was to be assessed on all participants; if not equivalent, this statistic would be evaluated for each format separately.
A Bland-Altman plot[174] was done initially to evaluate the potential for outliers and to confirm homoscedasticity. Following this, repeated measures ANOVA was used to assess for any systematic differences between the two time points. Finally, ICC3(A,1) was used to evaluate test-retest reliability[85, 175]. ICCs of at least 0.7 were considered to represent adequate reliability[57, 83].

Measurement error was reported with the reliability coefficient as the standard error of measurement (SEM). SEM should be relatively small in comparison to the range of scores expected in a target (i.e. clinical) population, as well as small in comparison to the expected difference scores[32]. Therefore, SEM values of less than 10 points (i.e. 10% of scale width) were considered acceptable.

2.2.15 Interpretability

Minimal detectable change was calculated at 90% confidence (MDC_{90}) [33, 86, 87, 95]. This statistic is appropriate for individuals. In consideration for situations where changes are being evaluated as sample means, the group MDC_{90} was also reported by dividing the individual MDC_{90} by $\sqrt{n}$ [83, 97, 98].

The MCID was calculated using Norman's distribution-based method[99]. It was calculated as half of a standard deviation of the baseline subscores.

2.2.16 Construct validation: convergent and divergent validation

The following *a priori* hypotheses were posited regarding the relationship between PFOOS scores, the AKPS, and the SF-36v2:

a. (H1): All PFOOS subscales will correlate with the AKPS more so than the SF-36 physical subscales (Physical Component Summary [PCS], PF [physical function], RP [role physical], BP [bodily pain] and GH [general health]); and in turn more so than the SF-36 mental

subscales (Mental Component Summary [MCS], VT [vitality], SF [social function], RE [role emotional] and MH [mental health])

b. (H2): Correlations between the PFOOS and the AKPS will be higher in the ADL, QOL and SPR subscales than the Symptoms and Pain subscales
c. (H3): Correlations between the PFOOS and pain intensity (Numeric Pain Rating Scale) at baseline will be higher in the Pain subscale than the remaining subscales
d. (H4): Correlations between the PFOOS and pain severity (on a five-point adjectival scale) at baseline will be higher in the Pain subscale than the remaining subscales.

Pearson's correlation coefficient was used to evaluate convergent and divergent validation.

2.2.17 Construct validation: known groups validation

A general question asked participants "How would you rate your knee pain?" and offered five possible response options (no problem, mild, moderate, somewhat severe, and severe). A second question used the Numeric Pain Rating Scale (NPRS) and asked "How would you rate your level of knee pain?" with 11 response options ranging from zero "No pain" to 10 "Worst Imaginable Pain". The following *a priori* hypotheses were posited:

a. (H5): Participants who subjectively rate symptoms as 'moderate', 'somewhat severe', or 'severe' on the first general pain question will have lower PFOOS scores than those who rate their pain as "no problem' or 'mild'.
b. (H6) Participants rating their pain on the NPRS between six – 10 will have lower PFOOS scores than those who rate their pain between zero – five.

Welch's two-sample t-test were used for known groups validation, with $p<0.05$ considered statistically significant.

2.3 Results

2.3.1 Sample description

Study enrollment occurred from February to September, 2012. Eighty-three individuals were screened for eligibility, and of those 29 were excluded, leaving 54 who participated in the study (see Figure 2.1 for flow chart). A description of the participants included in the analysis is presented in Tables 2.1 and 2.2.

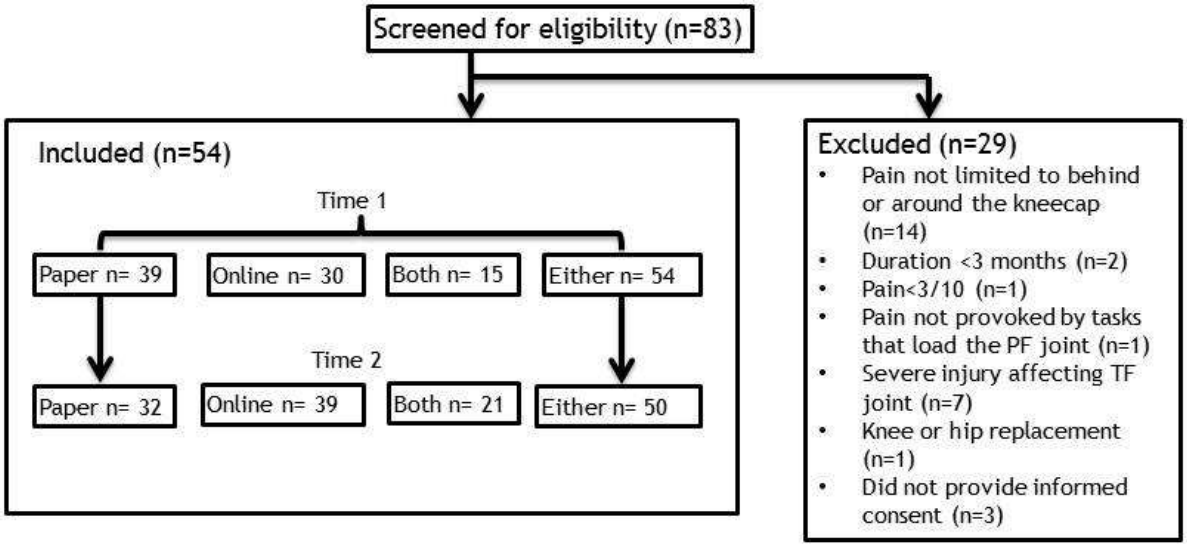

Figure 2.1 Flow diagram for recruitment. Note 39 participants completed paper format at Time 1, and 30 completed online format (some of whom completed both formats, n=15). This is the same for Time 2 as indicated.

Table 2.1 Descriptive statistics (frequency table) (N=52*)

Feature	n	%
Women	34	63
Men	20	37
Traumatic onset	18	35
Insidious onset	34	65
Surgery on affected knee	3	6
Bilateral pain	30	58
Unilateral pain	22	42
Age ≥40	28	53

*Missing data on 2 participants for all variables except sex, where N=54

Table 2.2 Descriptive statistics of study participants (N=52*)

Feature	Mean(SD)	Range
Age mean(SD)	42.6 (13.3)	19 – 66
Body Mass Index mean(SD)	26.6 (4.8)	19.7 – 37.6
Pain rating (0 – 10)	4.7 (1.5)	2 – 8
Duration of knee problems (months)	75.8 (84.8)	3 – 300
Duration of current episode (months)	53.9 (75.4)	0.5 – 300

*Missing data for these variables on 2 participants

Of the 54 participants, all participants completed at least one format (online and/or paper) of all three questionnaires at baseline. Fifty completed at least one format of the PFOOS at retest.

2.3.2 Item reduction

Results for item performance are presented in Tables 2.3 through 2.7. Since there were no items where a single response option had greater than 95% endorsement, this indicator is not presented in the tables. Also, under missing items, only N/A or 'not completed' is reported. However, administrative error rendered eight participants' paper-and-pencil scores invalid in three subscales (Pain, ADL, and Sports/Rec). Two of those eight participants had also completed an online form. Therefore, for those three subscales, there were 48 completed scales. For QOL, 1 person missed the entire subscale on paper, which was positioned on the back of the last page of the questionnaire, leaving 53 completed QOL subscales.

Symptoms subscale: I deleted items S4 and S5 based on their performance (see Table 2.3).

Pain subscale: I deleted item P7 based on performance (see Table 2.4). Items P3, P5, P9 and P8 performed questionably but were kept since they performed within the 15% window of my defined cut-points. Despite the high number of missing responses for item NP14, I kept this item based on the clinical importance of the task "hopping/jumping" for athletes (non-athletes can choose N/A without invalidating the subscale).

Function, Activities of Daily Living (ADL) subscale: I elected to delete items A6 and A17 based on performance (see Table 2.5). Items A4, A8, A9, A10, A11, A12, A13 and A14 performed within the 15% window of my established cut-points and were therefore kept.

Function, Sports & Recreation (SPR) subscale: All five items were kept in this subscale (see Table 2.6). With the exception of a high percentage of missing responses here, performance on each item was very good. It was expected that some items would not apply to non- athletes or those who do not participate in active recreation activities for lifestyle choices rather than due to their anterior knee symptoms.

Quality of Life (QOL) subscale: All items were kept in this subscale based on performance (see Table 2.7).

Table 2.3 Item performance for Symptoms, combined format (N = 54)

ITEM	Response options with <5% endorsement	>50% respond "No Problem"	Mean score <1	Missing responses (% of that item)	Decision
S1	4	-	-	1 incomplete (2%)	-
S2	-	-	-	-	-
S3	3,4	-	1.0	-	-
S4	2,3,4	83.3%	0.3	-	DELETE
S5	3,4	70.37%	0.5	-	DELETE
S6	4	-	-	-	-
S7	3,4	-	-	-	-
NS8	3,4	-	-	-	-

Table 2.4 Item performance for Pain, combined format (N = 48*)

ITEM	Response options with <5% endorsement	>50% respond "No Problem"	Mean score <1	Missing responses (% of that item)	Decision
P1	0	-	-	-	-
P2	-	-	-	2 N/A (4%)	-
P3	3,4	-	0.7	-	-
P4	4	-	-	-	-
P5	3,4	-	0.7	1 N/A (2%)	-
P9	3,4	55.32%	0.6	1 incomplete (2%)	-
P7	2,4	63.83%	0.6	1 N/A (2%)	DELETE
P8	3,4	52.08%	0.7	-	-
NP9	4	-	-	-	-
NP10	0	-	-	1 N/A (2%)	-
NP11	0	-	-	-	-
NP12	4	-	-	1 incomplete, 1 N/A (4%)	-
P13	-	-	-	-	-
NP14	-	-	-	6 N/A (13%)	-
NP15	-	-	-	2 N/A (4%)	-
NP16	-	-	-	2 N/A (4%)	-
NP17	-	-	-	-	-
NP18	-	-	-	1 incomplete (2%)	-
NP19	2	-	-	2 incomplete (4%)	-

*Six subscales invalid due to administrative error

Table 2.5 Item performance for Function, Activities of Daily Living (ADL), combined format (N = 48*)

ITEM	Response options with <5% endorsement	>50% respond "No Problem"	Mean score <1	Missing responses (% of that item)	Decision
A1	4	-	-	-	-
A2	-	-	-	-	-
A3	4	-	-	-	-
A4	3,4	56.25%	0.6	-	-
A5	4	-	-	-	-
A6	3,4	60.42%	0.6	-	DELETE
A7	4	-	-	-	-
A8	3,4	52.08%	0.6	-	-
A9	3,4	57.45%	0.6	1 N/A (2%)	-
A10	3,4	-	0.8	-	-
A11	3,4	58.33%	0.6	-	-
A12	4	56.52%	0.7	1 incomplete, 1 N/A (4%)	-
A13	3,4	57.14%	0.6	1 incomplete, 12 N/A (27%)	-
A14	3,4	50.00%	0.7	-	-
A15	3,4	-	0.9	1 incomplete (2%)	-
A16	4	-	-	4 N/A (8%)	-
A17	3,4	60.42%	0.5	-	DELETE

*Six subscales invalid due to administrative error

Table 2.6 Item performance for Function, Sports & Recreation (SPR), combined format (N = 48*)

ITEM	Response options with <5% endorsement	>50% respond "No Problem"	Mean score <1	Missing responses (% of that item)	Decision
SP1	0	-	-	3 N/A (6%)	-
SP2	-	-	-	3 N/A (6%)	-
SP3	-	-	-	2 incomplete, 10 N/A (25%)	-
SP4	-	-	-	3 N/A (6%)	-
SP5	-	-	-	3 N/A (6%)	-

*Six subscales invalid due to administrative error

Table 2.7 Item performance for Quality of Life (QOL), combined format (N = 53*)

ITEM	Response options with <5% endorsement	>50% respond "No Problem"	Mean score <1	Missing responses* (% of that item)	Decision
P1	0,1	-	-	1 incomplete (2%)	-
P2	-	-	-	1 incomplete (2%)	-
P3	4	-	-	1 incomplete (2%)	-
P4	0,4	-	-	1 incomplete (2%)	-
P5	-	-	-	1 incomplete (2%)	-

*One participant left this entire subscale blank

2.3.3 Scoring and missing data

Missing responses for each preliminary item are reported above in Tables 2.3 through 2.7. In Table 2.8, missing responses for the reduced scale are presented by subscale and by format. At least one item was missing in 14 out of 39 participants who completed the paper-and-pencil format at baseline (36% of subsample) and in 15 out of 30 participants who completed the online format (50% of subsample). As anticipated, there were more missing responses in the paper-and-pencil format than the online format, however, missing responses did occur online, something that Survey Monkey is supposed to prevent by not allowing a responder to move to the next item until they have selected a response. For the full PFOOS scale, participants either missed or selected N/A on 3% of responses on paper; 2% of responses online; and 3% of responses in the combined format (latter not shown). For each subscale, the percentage of missing items was less than 5% except Sport/Recreation which was 13% on paper; 6% online (Table 2.8) and 10% combined (latter not shown). No patterns emerged among the missing points, and correlations between missingness and other variables showed no to low-moderate correlations (maximum 0.28), confirming the data were likely 'missing at random'.

Invalid subscales, comparing both methods, are reported in Table 2.9. There were no invalid subscales by either scoring method for the Symptoms subscale. For Pain, two participant scores were invalid using the pencil-and-paper format and original scoring method only. For ADL, one participant score was invalid using the online format and original scoring method only. For Sport/Recreation, four participant scores were invalid using the pencil-and-paper format, regardless of the scoring method. For Quality of Life, one participant score was invalid using the pencil-and-paper format, using both scoring methods. The only differences between the two methods, therefore, were in regards to two Pain subscales (paper format) and one ADL subscale (online format).

Table 2.8 Missing items (incomplete or N/A) at baseline, by subscale and format

Subscale (# items)	n	Incomplete (n)	N/A(n)	% of responses missing
Symptoms (6)				
Paper	39	S1 (1)	-	0%
Online	30	-	-	0%
Pain (18)				
Paper	31*	NP12 (1) NP14(1) NP15(1) NP19(3)	P2 (2) NP10 (1) NP12 (1) NP14 (3) NP15 (1) NP16 (1)	3%
Online	30	P4(1) P9 (1) NP18(1) NP19(1)	P5 (1) NP14(3) NP15(2) NP16(1)	2%
Function, ADLs (15)				
Paper	31*	A16(1)	A7 (1) A13 (6) A16(1)	2%
Online	30	A12 (1) A13 (1) A15(1)	A9 (1) A13 (7) A16(3)	3%
Sport/Recreation (5)				
Paper	31*	SP2(1) SP3(2) SP4 (1)	SP1(2) SP2(2) SP3(6) SP4(3) SP5(3)	13%
Online	30	-	SP1(1) SP2(2) SP3(6)	6%
Quality of Life (5)				
Paper	39	Q1 (1) Q2 (1) Q3 (1) Q4 (1) Q5 (1)	-	3%
Online	30	-	-	0%
Full Scale (49)		**Total**	**Total**	
Paper	39	17	33	3%
Online	30	7	27	2%

*Due to administrative error, 3 subscales in paper format were invalid for 8 participants and were therefore excluded from analysis

A sensitivity analysis revealed differences in means between scoring methods of no more than 0.8%, and Welch's two-sampled t-tests revealed the differences were not statistically significant. Therefore, all subsequent analysis was performed and reported using the newer scoring method (i.e. where up to 50% of items can be missing in a given subscale and still included for analysis). Importantly, after the invalid subscales were removed, the percentage of missing items in Sport/Recreation (which had been 13% on paper, 6% online, and 10% combined format) was reduced to below a preferred cut-point of 5%.

Table 2.9 Invalid subscales at baseline, by format (substantial missingness is in bold)

Subscale (# items)	Number of missing items (n)	Invalid with original method	Invalid with updated method
Symptoms (6)		**>2 items missing**	**>3 items missing**
Paper	1 missing (1)	-	-
Online	-	-	-
Pain (18)		**>3 items missing**	**>9 items missing**
Paper	1 missing (4) 2 missing (1) **4 missing (1)** **5 missing (1)**	2 invalid	-
Online	1 missing (7) 2 missing (2)	-	-
Function, ADLs (15)		**>3 items missing**	**>7 items missing**
Paper	1 missing (7) 2 missing (1)	-	-
Online	1 missing (6) 2 missing (2) **4 missing (1)**	1 invalid	-
Sport/Recreation (5)		**>2 items missing**	**>2 items missing**
Paper	1 missing (4) **3 missing (2)** **5 missing (2)**	4 invalid	4 invalid
Online	1 missing (3) 2 missing (3)	-	-
Quality of Life (5)		**2+ items missing**	**2+ items missing**
Paper	**5 missing (1)**	1 invalid	1 invalid
Online	-	-	-

At retest, there were four missing cases, therefore only 50 of the 54 participants completed a retest questionnaire. Of the completed pencil-and-paper format questionnaires, there were three invalid Sports/ Recreation subscales. Online, there two invalid Sports/Recreation subscales.

Using complete case analysis resulted in subscores based on a reduced sample of 53 (Symptoms); 35 (Pain); 32 (ADL); 37 (Sport/Rec); and 53 (QOL) participants. This can be compared with the larger sample size using the person mean imputation method of 54

(Symptoms); 48 (Pain); 48 (ADL); 45 (Sport/Rec); and 53 (QOL). The differences in subscale means for the two methods of handling missing data differed by no more than 2.3%, and Welch's two-sampled t-tests revealed that these differences were not statistically significant.

Mean scores and standard deviations for the baseline PFOOS subscales that were valid are reported in Table 2.10. In total, 39 participants completed the paper-and-pencil format, while 30 completed the online format. Any difference in scores by format here should not be compared directly (e.g. for reliability) since it is two different sub-samples for each format, with some sample overlap (see Section 2.3.5 for more accurate comparison of formats). Figure 2.2 provides a visual representation of the PFOOS subscore distribution using histograms. Distribution was approximately normal with the exception of Sports/Recreation.

Table 2.10 PFOOS baseline scores, by format

PFOOS Subscale	**Paper-and-pencil Mean(SD)** ***n***	**Online* Mean(SD)** ***n***	**Combined formats[#] Mean(SD)** ***n***
Symptoms	66.2 (14.1) *n=39*	58.5 (16.5) *n=30*	63.3 (15.5) *n=54*
Pain^	59.4 (14.9) *n=31*	56.6 (16.5) *n=30*	56.9 (16.0) *n=48*
ADL[+]	74.8 (15.3) *n =31*	75.4 (18.4) *n=30*	74.8 (17.6) *n=48*
Sport/ Recreation	49.1 (29.4) *n = 27*	43.8 (26.2) *n=30*	44.7 (27.9) *n=45*
Quality of Life	47.6 (16.1) *n=38*	48.7 (17.5) *n=30*	47.0 (16.5) *n=53*

**NB: for a direct comparison of paper vs. online scores (i.e. reporting only scores of participants who completed both formats), see Section 2.3.5*

[#]Combined by adding online scores wherever paper score not available (39 paper, 15 online)

^With 2 subscales removed (invalid by original scoring method), combined mean remains 56.9, SD becomes 16.3 (n=46). No statistically significant difference.

[+]With 1 subscale removed (invalid by original scoring method), combined mean increases to 75.6, SD becomes 16.9 (n=47). No statistically significant difference.

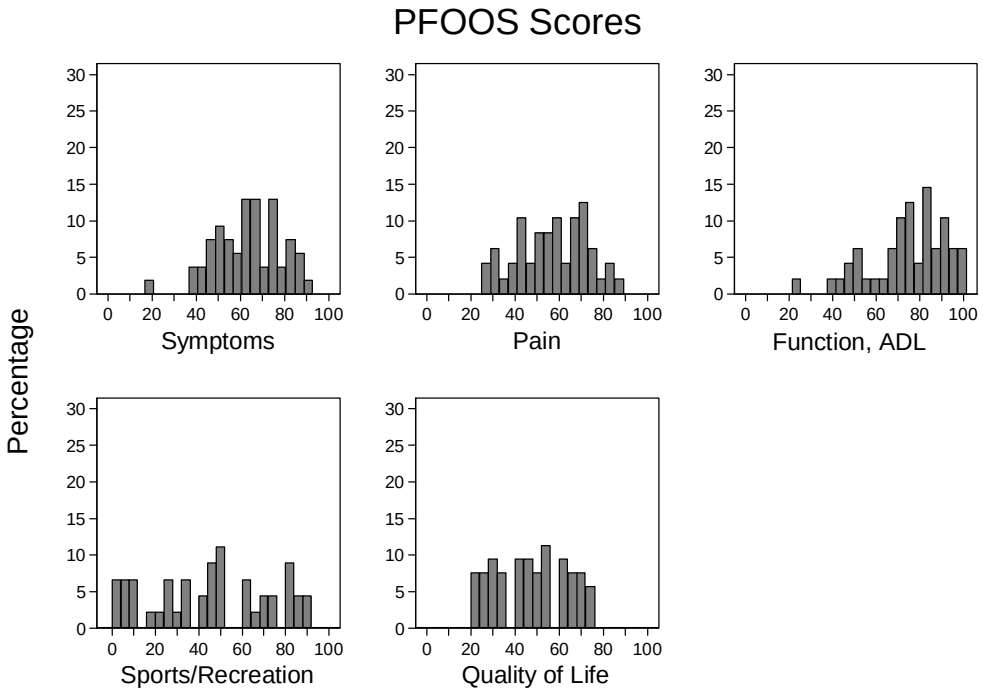

Figure 2.2 Histogram for each subscale of the PFOOS

PFOOS subscores are presented by sex and age category in Table 2.11. Unadjusted two-tailed Welch's t-tests revealed no statistical difference by sex or age. Score means and distributions for the AKPS and SF-36 subscales are presented in Table 2.12.

Table 2.11 PFOOS baseline scores, by sex and by age

PFOOS	**Men Mean (SD) *n***	**Women Mean (SD) *n***	**p**	**Age ≤40 Mean (SD) *n***	**Age ≥50 Mean (SD) *n***	**p**
Symptoms	63.5 (14.9) *n=20*	63.1 (16.1) *n=34*	0.92	65.0 (17.6) *n=25*	58.7 (10.8) *n=19*	0.15
Pain^	58.1 (15.3) *n=18*	56.2 (16.6) *n=30*	0.70	58.7 (17.3) *n=22*	53.7 (14.4) *n=17*	0.33
ADL^{+}	76.9 (17.2) *n=18*	73.6 (18.0) *n=30*	0.53	78.5 (17.1) *n=22*	67.5 (17.5) *n=17*	0.06
Sport/ Recreation	49.4 (30.9) *n=18*	41.6 (25.8) *n=27*	0.39	53.5 (22.3) *n=21*	39.0 (28.4) *n=16*	0.10
Quality of Life	49.5 (14.6) *n=20*	45.4 (17.6) *n=33*	0.37	49.6 (15.0) *n=24*	44.2 (17.2) *n=19*	0.29

Table 2.12 AKPS and SF-36 baseline scores, by format

Instrument	Paper-and-pencil Mean(SD) *n=39*	Online Mean(SD) *n=30, 15**	Combined formats[#] Mean(SD) *N=54*
AKPS	68.0 (12.6)	61.2 (13.3)	65.9 (12.6)
SF-36 v2			
PCS[^]	46.5 (7.6)	48.9 (6.7)	47.1 (7.4
MCS[^]	53.2 (9.6)	50.5 (8.9)	52.4 (9.4)
PF[^]	73.9 (14.1)	78.3 (18.5)	75.1 (15.4)
RP[^]	73.2 (36.0)	76.7 (41.7)	74.2 (37.3)
BP[^]	58.3 (16.0)	65.2 (20.4)	60.2 (17.4)
GH[^]	68.8 (22.6)	67.56 (21.4)	68.4 (22.1)
VT[^]	59.1 (15.3)	54.2 (14.5)	57.8 (15.1)
SF[^]	86.2 (17.9)	85.8 (21.1)	86.1 (18.6)
RE[^]	85.7 (26.2)	82.2 (37.5)	84.7 (29.4)
MH[^]	77.6 (15.4)	75.0 (9.8)	76.9 (14.0)

*n=30 for AKPS. For SF-36, n=15 since resource priority for use of proprietary software went to paper format (30 questionnaires were completed, but of those 15 had paper versions and therefore only the paper version was scored)
[^]PCS = Physical Component Summary; MCS=Mental Component Summary; PF=Physical Function; RP =Role Physical; BP =Bodily Pain; GH =General Health; VT =Vitality; SF =Social Function; RE =Role Emotional; MH =Mental Health

2.3.4 Floor and ceiling effects

There were no floor or ceiling effects in any of the PFOOS subscales. In the Sport/Recreation subscale, three participants scored zero (7% of sample). There were no other scores of zero or 100 in any subscale. For the AKPS there were also no floor or ceiling effects, with no scores of zero or 100. Ceiling effects were present in the SF-36 for three subscales (Role Physical 59.3%; Social Function 53.7%; and Role Emotional 70.3%). There were no floor effects in any SF-36 subscale, though Role Physical was close with 12.9% scoring zero.

2.3.5 Alternate forms reliability

Comparison of paper-and-pencil format to online format of the PFOOS is presented in Table 2.13. Mean difference between the two formats ranged by subscale from 0.3% to 2.7%. Repeated measures ANOVA of each subscale revealed that these differences were all non-significant. The absolute ICC reliability coefficients were all ≥ 0.80.

In comparing alternate formats (Table 2.13) to paper only formats (Table 2.14), ICCs were lower for paper only in three of the subscales (Symptoms, Pain and ADL) and higher for two subscales

(Sport/Rec, QOL). In addition, the mean differences between paper only test administrations was larger in four of the subscales (all except Sports/Recreation) compared to alternate forms. With online only formats (Table 2.15), ICCs were higher for all five subscales compared to paper only; and higher in four subscales compared to alternate forms.

These results confirmed paper and online format equivalence. Therefore, all remaining analyses (internal consistency, hypothesis testing, test-retest reliability, measurement error, MDC and MCID) were completed using the larger sample size that combined both formats.

Table 2.13 Summary of paper vs. online versions of the PFOOS

Subscale	n	Paper Mean (SD)	Online Mean (SD)	Mean Difference (%)	ANOVA p-value	ICCA(3,1) [95%CI]	SEM
Symptoms	31	64.5 (12.9)	64.8 (14.7)	0.3%	0.86	0.80 [0.63,0.90]	6.2
Pain	26	60.6 (14.2)	61.3 (15.2)	0.7%	0.57	0.92 [0.83,0.96]	4.2
Function, ADLs	26	76.0 (14.8)	77.8 (12.6)	1.8%	0.10	0.91 [0.80,0.96]	4.0
Function, Sports/ Recreation	26	52.6 (26.2)	50.3 (26.8)	2.3%	0.46	0.83 [0.65,0.92]	11.1
Quality of Life	31	50.0 (15.2)	52.7 (18.0)	2.7%	0.12	0.83 [0.68,0.91]	6.7

Table 2.14 Summary of paper vs. paper versions of the PFOOS

Subscale	n	Baseline Mean (SD)	Retest Mean (SD)	Mean Difference (% of scale range)	ANOVA p-value	ICC3(A,1) [95%CI]	SEM
Symptoms	31	65.6 (13.6)	68.0 (13.3)	2.4%	0.23	0.70 [0.46,0.84]	7.4
Pain	27	59.2 (15.0)	62.1 (13.6)	2.9%	0.08	0.82 [0.64,0.91]	5.9
Function, ADLs	27	74.2 (15.1)	76.5 (11.3)	2.3%	0.22	0.75 [0.52,0.88]	6.6
Function, Sports/ Recreation	21	46.1 (29.1)	47.6 (26.2)	1.5%	0.56	0.91 [0.80,0.96]	8.2
Quality of Life	30	46.7 (16.0)	50.3 (15.5)	3.6%	0.01	0.87 [0.71,0.94]	5.2

Table 2.15 Summary of online vs. online versions of the PFOOS

Subscale	n	Baseline Mean (SD)	Retest Mean (SD)	Mean Difference (% of scale range)	ANOVA p-value	ICC3(A,1) [95%CI]	SEM
Symptoms	23	60.0 (14.7)	60.4 (16.8)	0.4%	0.78	0.88 [0.74,0.95]	5.5
Pain	23	55.5 (17.8)	58.8 (16.0)	3.3%	0.01	0.92 [0.79,0.97]	4.2
Function, ADLs	23	74.6 (18.2)	74.9 (16.9)	0.3%	0.89	0.82 [0.62,0.92]	7.6
Function, Sports/ Recreation	21	45.2 (28.6)	49.0 (32.3)	3.8%	0.09	0.94 [0.86,0.98]	6.9
Quality of Life	24	47.3 (18.4)	50.0 (18.7)	2.7%	0.25	0.89 [0.75,0.95]	6.3

2.3.6 Internal consistency

Internal consistency is reported in Table 2.16. Three of the five subscales fell within the ideal range of 0.7 – 0.9 (Pain, Sports/Recreation and QOL). The Symptoms subscale fell below at 0.61, with inter-item covariance slightly low at 0.24. The ADL subscale was higher than ideal at 0.94, though average inter-item covariance was acceptable.

Table 2.16 Internal consistency for PFOOS scores at baseline – combined formats (N=54)

Subscale	n	Number of items	Average inter-item covariance	Cronbach's α
Symptoms	54	6	0.24	0.61
Pain	48	18	0.36	0.89
Function, ADLs	48	15	0.47	0.94
Function, Sports/Recreation	45	5	1.11	0.90
Quality of Life	53	5	0.34	0.78

2.3.7 Test-retest reliability

Reliability coefficients (ICC3(A,1)) and respective standard errors of measurement (SEM) are reported in Table 2.17. Mean differences from baseline to retest range from 0.4% to 3.9%. Unadjusted p-values of less than 0.05 were noted for Pain (p=0.01) and QOL (p=0.00) with differences between baseline and retest scores for these two groups of 3.3% and 3.9%, respectively. Reliability coefficients were all acceptable, ranging from 0.79 to 0.91. Measurement error (SEM) was also acceptable, ranging from 5.9 to 8.5 points.

Table 2.17 Summary of test-retest reliability and associated measures for PFOOS

Subscale	n	Baseline Mean (SD)	Retest Mean (SD)	Mean Difference (% scale range)	ANOVA p-value	ICC3(A,1) [95%CI]	SEM	MDC_{90}	$MDC_{90} \div \sqrt{n}$	MCID
Symptoms	49	64.6 (14.2)	65.0 (15.8)	0.4%	0.77	0.79 [0.66,0.88]	6.5	15.2	2.2	7.1
Pain	43	57.0 (16.5)	60.3 (15.1)	3.3%	0.01	0.87 [0.74,0.93]	5.9	13.8	2.1	8.3
Function, ADLs	43	74.3 (17.1)	75.2 (14.3)	0.9%	0.57	0.79 [0.64,0.88]	7.8	18.2	2.8	8.6
Function, Sports/ Recreation	38	47.1 (28.4)	48.4 (29.1)	1.3%	0.52	0.91 [0.84,0.95]	8.5	19.8	3.2	14.2
Quality of Life	48	46.5 (16.9)	50.4 (17.2)	3.9%	0.00	0.87 [0.73,0.93]	6.1	14.2	2.0	8.5

2.3.8 Interpretability

MDC at 90% confidence (MDC_{90}) is reported in Table 2.17. MDC_{90} is reported at the individual level as well as at the group level ($MDC_{90}/\sqrt{n}$). MCID using Norman's distribution method[99] of half a standard deviation is also reported in Table 2.17.

2.3.9 Construct validation: convergent and divergent validation

Pearson correlation coefficients for all hypotheses are reported in Table 2.18. The following outlines results for each *a priori* hypothesis:

a. (H1): All PFOOS subscales will correlate with the AKPS more so than the SF-36 physical subscales (Physical Component Summary [PCS], PF [physical function], RP [role physical], BP [bodily pain] and GH [general health]); and in turn more so than the SF-36 mental subscales (Mental Component Summary [MCS], VT [vitality], SF [social function], RE [role emotional] and MH [mental health]).

   Confirmed with notable exceptions. PF presented with the largest correlations across all five PFOOS subscales than all other SF-36 subscales. RP did not correlate with the PFOOS as expected. VT correlated better than expected.

b. (H2): Correlations between the PFOOS and the AKPS will be higher in the ADL, QOL and SPR subscales than the Symptoms and Pain subscales.

   Not confirmed. Correlations were similar across subscales, and ranged from 0.59 to 0.71, with Pain having the highest correlation.

c. (H3): Correlations between the PFOOS and pain intensity (Numeric Pain Rating Scale) at baseline will be higher in the Pain subscale than the remaining subscales.
   Partially confirmed. Pain and ADL subscales were the highest, with negative correlations of -0.48 and -0.49, respectively.
d. (H4): Correlations between the PFOOS and pain severity (on a five-point adjectival scale) at baseline will be higher in the Pain subscale than the ramaining subscales.
   Not confirmed. QOL was the highest correlation at -0.55, and ADL was the second highest at -0.50. Pain was third at -0.44.

Overall, each PFOOS subscale, taken individually, correlated as predicted with the AKPS and SF-36 subscales. However, when comparing among the PFOOS subscales (e.g. correlation of one subscale to AKPS compared to another correlation of a subscale to AKPS), the predicted relationships did not hold as well (see Table 2.18).

Table 2.18 Correlations between PFOOS subscales, AKPS, SF-36, and two pain variables (NPRS and Pain)

Subscale	AKPS	PCS	MCS	PF	RP	BP	GH	VT	SF	RE	MH	NPRS	Pain
Symptoms	0.61	0.13	-0.11	0.25	-0.07	0.12	0.00	0.14	-0.13	-0.14	0.01	-0.37	-0.40
Pain	0.71	0.18	0.10	0.46	-0.06	0.16	0.18	0.26	0.09	0.06	0.18	-0.48	-0.44
ADLs	0.59	0.30	0.03	0.54	0.00	0.27	0.22	0.21	0.05	0.08	0.11	-0.49	-0.50
Sport/ Recreation	0.65	0.18	0.09	0.44	-0.10	0.10	0.27	0.26	0.16	0.04	0.10	-0.21	-0.30
Quality of Life	0.66	0.36	-0.04	0.45	0.00	0.30	0.31	0.16	0.16	-0.04	0.04	-0.37	-0.55

Legend: AKPS - Anterior Knee Pain Scale; PCS - Physical Component Summary; MCS – Mental Component Summary; PF – Physical Function; RP – Role Physical; BP – Bodily Pain; GH – General Health; VT – Vitality; SF – Social Function; RE – Role Emotional; MH – Mental Health; NPRS – numeric pain rating scale.

2.3.10 Construct validation: known groups validation

Mean differences between milder pain (n=21) vs. moderate to severe pain (n=30) sub-samples are reported in Table 2.19 under the column labeled “Pain severity”, with p-values for Welch’s two-sample t-test reported in the next column. Mean differences between two groups with lower (n=36) vs. higher (n=16) pain levels on a NPRS are reported in the column labeled “NPRS”. As expected, all five PFOOS subscale scores were lower in the group that reported higher pain (for both pain variables), and these differences were statistically significant in four subscales (all except Sport/Recreation).

Table 2.19 Known groups – comparing PFOOS subscales between higher and lower reported pain levels.

Subscale	Pain severity* Mean difference	p	NPRS^ Mean difference	p
Symptoms	8.9	0.02	12.9	0.01
Pain	13.1	0.00	14.7	0.00
Function, ADLs	17.4	0.00	20.1	0.00
Function, Sports/ Recreation	12.2	0.08	9.2	0.27
Quality of Life	14.3	0.00	13.3	0.00

*Pain: compared milder scores (0, 1) to moderate/severe scores (2, 3, 4)
^NPRS (Numeric Pain Rating Scale): compared pain ratings 0 – 5 to pain ratings 6 – 10
NB: Mean differences were unanimously lower for the higher pain subgroup vs. lower.

2.4 Discussion

This is the first study to evaluate the validity and reliability of a preliminary version of the PFOOS, a new PROM intended to evaluate pain, symptoms, function and QOL in people with anterior knee pain or patellofemoral OA. This is an essential step in the development of the PFOOS, as it provides insight regarding the extent to which meaningful inferences can be drawn from PFOOS scores about an individual or a group.

2.4.1 Alternate forms reliability

The online version of the PFOOS performed equivalently to the paper version, with absolute agreement ICC3(A,1) of at least 0.80 and non-significant mean differences of no more than 2.7%. These findings are consistent with previous findings comparing alternate forms. In a meta-analysis of 46 publications comparing alternate formats of 278 scales, reliability coefficients were > 0.75 in 94% of the studies, and mean differences were within +/- 5% in 93% of the studies.

With equivalence established between the two formats, it is worthwhile to consider whether the online format may, in fact, be superior to the paper-and-pencil format. For example, missing data and subsequent subscale validity was less of a problem with the online format in this study. Using the original method, seven subscales were deemed invalid (due to missing items) using paper-and-pencil compared to only one in the online format. Using the new method, paper-and-pencil had five invalid subscales, and none using the online format. This reduction in missing data and subsequent increase in the number of valid subscales may in part explain why test-retest reliability coefficients for the online format (Table 2.15) were higher than for the paper format (Table 2.14) across all five subscales. Additional benefits of online administration might also

include reduced risk of data entry errors (again potentially improving reliability), and convenience for patients (e.g. complete forms at their leisure; potentially less travel time)[34, 78, 79]. In this study sample, access to the internet was reportedly unanimous (except one individual who left that question incomplete), and 14 expressed preference for the online format, while 36 had no preference and only four preferred the paper-and-pencil format.

2.4.2 Internal consistency

The high α of 0.94 for the ADL subscale could be in part explained by a relatively high number of items (n=15). It is also possible that some items were redundant. For example, items A3 ("Rising from sitting"), A7 ("Getting in/out of car"), A10 ("Rising from bed") and A15 ("Getting on/off toilet") may all be tapping the same functional task, transferring from a seated to a standing position. In addition, eight items in the ADL subscale (A4, A8, A9, A10, A11, A12, A13, A14) performed on the cusp (within 15%) of pre-determined cut-points for item reduction, primarily due to lack of spread of scores throughout the response options. Therefore, it is likely that several items are contributing to the high internal consistency of this subscale.

In contrast to ADL, the Symptoms subscale had $\alpha = 0.61$ and inter-item covariance $r = 0.24$, suggesting that the Symptoms subscale does not have adequate internal consistency. Looking at the individual items that comprise this subscale, it is possible that the low α is a reflection of lack of unidimensionality. To further investigate this subscale, I evaluated item-partial correlation (correlation of each item to the subscale with the item removed), which Streiner and Norman indicate should be at least 0.20[32]. I also looked at how Cronbach's α changed with each item removed (if an item fits well with the subscale, α should be lower with the item removed)[32]. Item S2 ("Do you feel grinding, hear clicking or any other type of noise when your knee moves?") appeared to be problematic within this subscale, with an item-partial correlation of 0.11. Also, with this item removed, Cronbach's α increased to 0.68, indicating the item was creating statistical problems for the scale.

Having said this, a question about crepitus at the knee is important for both clinicians and researchers who want richness of clinical information. In fact, looking at each question in this subscale (swelling, crepitus, catching/locking, stiffness in the morning / after resting / after exercise), it is likely that not all symptoms will correlate well within an anterior knee pain population. It has been established in the literature that symptoms vary considerably with

anterior knee pain and knee OA depending in part on which compartment is affected and how severe the radiographic findings are[107-109, 113, 122]. Therefore, the question here becomes one of priority: is it more important for a scale to have high internal consistency or is it more important for the scale to have adequate breadth and depth of content? Streiner and Norman[32] argue it is "better to sacrifice internal consistency for content validity" (p. 253), since the purpose of a PROM is inferential in nature. Patellofemoral symptoms are, by nature, heterogeneous. Therefore, umbrella terms like "symptoms" are likely to have lower internal consistency. Based on this reasoning, item S2 should be kept in the scale, though it is recommended that users of this subscale consider individual items carefully rather than relying exclusively on the overall subscore.

2.4.3 Test-retest reliability

The PFOOS demonstrated adequate test-retest reliability, with absolute ICC3(A,1) values ranging from 0.79 to 0.91. These findings compare to the test-retest reliability coefficients of the AKPS, which have ranged from 0.81 – 0.95 in other studies of anterior knee pain or instability[55, 69, 129, 131]. Test-retest reliability coefficients of the SF-36 in a patellofemoral instability cohort ranged from 0.47 to 0.77 with the exception of Role Emotional (RE) which was 0.04. Coefficients for the KOOS (though slightly different samples, i.e. awaiting arthroscopy or reconstruction for ACL or meniscal injuries, or knee OA) ranged from 0.6 to 0.95[52, 72, 143, 176].

Two subscales, Pain and QOL, had p-values less than 0.05 on the repeated measures ANOVA, suggesting the possibility of systematic error between the two time points. However, mean differences from baseline to retest for these two subscales were small at 3.3 and 3.9, respectively, both well under the difference deemed to be clinically important (MCID 8.3 and 8.5 respectively). While these changes may represent real physiological change, they could also represent error such as bias (e.g. memory and learning effects) or random chance.

2.4.4 Construct validation: convergent and divergent validation

Hypotheses regarding performance of the PFOOS subscales relative to the AKPS relative to the SF-36 were largely confirmed in this study. Moderate correlations existed among most subscales where a correlation was expected, consistent with expectations[32, 52, 57, 131]. Correlations > 0.70 would raise concerns that the new PROM is so highly related to an existing scale that it is

redundant and therefore not necessary[32]. Hypotheses regarding relationships among the SF-36 subscales were only partially confirmed. Specifically, I had expected correlations with PFOOS subscales to SF-36 physical subscales (PCS, PF, RP, BP, GH) to be greater than mental subscales (MCS, VT, SF, RE, MH). Instead, the correlations were highest with PF > BP, GH, VT > RP, SF, MH, RE. In looking at the specific items in the SF-36 that make up each subscale, these results are not unreasonable. Subscales RP (role physical) and BP (bodily pain) were comprised of many questions about how their pain or physical limitations were affecting their work or productivity. Since our sample was recruited from a work place (mainly students and staff), it is not surprising that they reported being able to work. Therefore, the absence of a positive correlation with PFOOS subscales and the RP, and a lower relationship with BP than expected, is understandable given the sample used in this study. It is noted that authors investigating the KOOS have previously hypothesized lower correlations with RP than the other physical health subscales[142, 176], contrary to the current study's hypotheses. Future studies should consider the demographics of the likely participants when considering of hypotheses.

2.4.5 Interpretability

MDC_{90} values in this study ranged from 13.8 – 19.8 for individual measures and from 2.0 – 3.2 for group measures(see Table 2.17). These findings are larger than those of the AKPS in other studies, with MDC_{95} values that ranged from 7 – 14 in several anterior knee pain cohorts[69, 129, 131]. In a cohort with knee OA, MDC_{95} values for the KOOS were in a similar range to the present study, ranging from 13.4 – 21.1[177]. However, among various other cohorts including younger individuals and athletes with traumatic knee injuries, values were typically smaller, with KOOS values ranging from 5 – 12[72, 143]. Importantly, MDC estimates are reflections not only of the instrument itself, but of the individuals being assessed, so comparisons should be done with caution (the KOOS studies were not done with a patellofemoral pain and OA cohort; and one anterior knee pain study excluded knee OA)[129].

MCID values in the present study ranged from 7.1 – 14.2. These findings compare to estimates for the AKPS of 8 – 10 in an anterior knee pain cohort[69]. MCID values have not been evaluated for the KOOS[72].

While MCID was larger than $MDC_{90group}$ for all five PFOOS subscales, it was smaller than the MDC_{90} values which are for individual comparison. Therefore, at an individual level, the amount

of change that must be overcome to be confident that it is not due to error is larger than the average amount of change that is deemed clinically meaningful. This could be problematic in situations where small but clinically meaningful individual changes are expected. Having stated this, it is important to note that the MCID in this study was calculated using a distribution based method, and is therefore only a statistical estimate of "clinical importance", something arguably best determined through empirical methods. Further, it is noted that using Norman's method, it is impossible to estimate an MCID that is greater than MDC_{90}. This is because MDC_{90} is based on SEM, and with a reliability coefficient of at least 0.75, SEM will always be at least ½ a standard deviation (i.e. MCID and SEM will converge)[99], so MDC_{90} will consistently be 2.3 times larger than Norman's MCID estimate (see equation #4, also Sections 1.9.1, 1.9.2) . Importantly, Norman's method is recommended by the authors[178] as a 'starting point', as they recognize high variability across the many types of MCID estimation methods[179-181]. For purposes of individual comparison, future studies could re-evaluate MCID for the PFOOS employing methods such as an anchor-based method or calculation of a reliability change index: this must be done in concert with a clinical intervention trial[32, 86, 182].

2.4.6 Usability

The PFOOS took respondents approximately 15 minutes to complete. The preliminary version contained 54 items and in this study was reduced to 49 items. Depending on the context in which this questionnaire is administered, this may or may not be acceptable to the responder. For example, in a clinical trial, a 15 minute time period may be feasible, whereas in a clinical setting some clinicians or patients may find this more burdensome.

Future studies with a broader spectrum of clinical severities may reveal that the PFOOS can be further reduced in size, and while reducing responder burden this may also enhance the scale's measurement properties (i.e. internal consistency). Alternatively, consideration could be given to developing a short form of this scale. The entire scale in its current state could be used to follow people longitudinally where it is expected that patellofemoral OA may progress to generalized OA – in this situation, the KOOS items that did not perform as well in this study may do better with a cohort that develops tibiofemoral symptoms. The short form scale could target patellofemoral-specific items that would be appropriate for cross-sectional studies or shorter duration prospective studies. This would reduce responder burden while also enabling items to be removed that may be less appropriate for an isolated patellofemoral cohort.

2.4.7 Limitations

There are several limitations noted in the present study. First, study methods are based on assumptions that the PFOOS scale can be treated statistically like an interval scale, despite technically being ordinal in nature. While this is true, Portney and Watkins[33] argue that statistics such as ICC calculations can be applied without distortion to ordinal level data provided the intervals between response options are assumed to be equal (p. 561). Further, Gaito[152] argues that numbers, by their nature, are intervals, and so long as data follow a normal distribution, then treating the numbers statistically as interval is appropriate. In the case of the PFOOS, results did show approximately normal distributions (see Figure 2.2). Therefore, it is appropriate that the PFOOS, like many other PROMs (including AKPS and SF-36) be treated statistically as an interval scale[32, 152].

The second limitation is that unidimensionality has not been explicitly evaluated in this study. However, the PFOOS has been created with five domains with the goal of each being unidimensional. This is in contrast to the AKPS, which is multidimensional in nature, thus limiting the ability to draw inferences from the AKPS score. Assessing unidimensionality, a component of structural validity, requires a larger sample size in order to conduct either factor analysis or Rasch principle components analysis. Depending on the confidence level desired, sample sizes for Rasch analysis vary from 64 – 243[47, 66, 183]. Having said this, evaluation of PROMs is an ongoing process done across multiple studies over time, and it is common to begin evaluation of an instrument with sample sizes adequate for reliability testing as an important early step[57, 66]. It would not be a good use of resources to begin evaluation with a large sample size when items are still being assessed (and possibly modified) at an individual level for wording, acceptability, and reliability. Having established adequate validity and reliability at this stage of the PFOOS development, it is now appropriate to recruit a larger sample size for conducting analysis of unidimensionality.

A third limitation is that responsiveness was not assessed in this study. Assessing responsiveness requires collection of data over a time period where change (worsening or improvement) would be expected, and is frequently collected during a clinical trial. Again, it is appropriate to assess the PFOOS in an iterative process, and having established adequate validity and reliability for the PFOOS, it is now appropriate to include the PFOOS in an intervention trial. An additional

benefit of such a trial would be the ability to calculate MCID for the PFOOS subscales using alternative anchor-based methods.

Chapter 3: The de Morton Mobility Index (DEMMI): Reference Data for a Performance-Based Mobility Instrument

3.1 Introduction

Mobility is an important marker and predictor of physical abilities, independence, morbidity and mortality in older adults[12, 184-188]. Loss of mobility can result in a decline in one's ability to complete daily activities[5], which may render individuals more reliant on caregivers such as family members or community health workers, to meet their basic needs. Further decline, or lack of adequate caregiver supports, can result in loss of ability to live independently in the community, with subsequent transition to assisted living or nursing home. Such functional decline can also increase risk of injury (e.g. as a result of a fall) and can increase hospital admissions[9]. Up to 40% of nursing home admissions are reportedly related to falls[9, 10], and decreased mobility is an important predictor of falls[189-191]. Determining mobility status is therefore an important component of any medical or health assessment for older adults, regardless of whether an individual is apparently healthy, acutely ill, or living with chronic comorbidity.

3.1.1 Selecting an appropriate mobility instrument

As with any health outcome measure, an instrument designed to measure mobility in older adults should demonstrate adequate validity and reliability for the target population. Also, given the diverse functional abilities and complex health statuses of older adults, a clinically relevant mobility measure would span the spectrum of functionally relevant mobility tasks such that an individual can be assessed across time points from hospitalization, through rehabilitation and toward full recovery. This would prevent a clinician from having to change instruments as the individual recovers (or declines), which is an important consideration given the challenges of comparing scores between different instruments[192].

Recent systematic reviews have described a plethora of mobility instruments for older adults[45, 63, 193]. Many of these instruments have limitations: (i) they are often designed for narrowly defined populations (e.g. specific medical conditions, specific age ranges), which can result in floor- or ceiling-effects; (ii) have not undergone rigorous validation and reliability testing; (iii) lack adequate validity, reliability or responsiveness; (iv) lack estimates or guidelines for

interpretability (such as MCID or normative references); or (v) have limited usability due to unreasonable equipment or cost requirements, or substantial time requirements to complete the test[45, 63, 193].

Following comprehensive review of existing mobility instruments, de Morton concluded that a mobility instrument designed to evaluate older adults across a broad spectrum of abilities, of sufficient scientific rigour, and convenient to administer, did not exist[45]. In response to these findings, she developed the de Morton Mobility Index (DEMMI)[47].

3.1.2 The de Morton Mobility Index (DEMMI)

The DEMMI is a unidimensional performance-based instrument developed using item response theory. It consists of 15 hierarchical items that span from bed mobility tasks through to high level dynamic balance tasks, in order of increasing difficulty (see Appendix E). On completion of the test, a raw ordinal score is converted (through Rasch analysis) to an interval score out of 100, with a higher score representing greater mobility.

Validity

The DEMMI was originally developed and validated in an acutely hospitalized older adult population[47]. Content validation was achieved during development through input from researchers, clinicians and patients, with pilot testing on patients and clinicians to assist with item refinement and reduction (details published elsewhere)[47]. Construct validation was achieved through structural validation (Rasch analysis) as well as hypothesis testing. Convergent validation was confirmed with good correlations between the DEMMI and two existing mobility instruments, the HABAM (Hierarchical Assessment of Balance and Mobility)[194] (r= 0.91) and the Barthel Index[195] (r=0.68). Divergent validation was shown with poor correlations between the DEMMI and measures of non-mobility constructs, the Mini Mental State Exam[196] (r=0.24), APACHE II[197] (r=0.07), and the Charlson Comorbidity Index[198] (r= -0.04). Known groups validation was confirmed through t-tests comparing a group being discharged home compared to a group being referred for additional rehabilitation ($p = 0.03$).

Following its initial validation study[47], the DEMMI was further validated across samples representing a broad continuum of functional abilities: healthy community-dwelling older adults[75], community-dwellers who require additional care from family or hired home care services[43], those in transitional care[91], subacute individuals[199], older adults with hip or knee OA[76], and another acute medical sample[200]. In each study, evidence supported convergent, divergent and known groups validation (with the exception of the OA cohort, where divergent validation was not tested[76]). Cross-cultural validation was also completed through evaluating a translated version of the DEMMI into Dutch[76].

Reliability

The DEMMI has demonstrated good inter-rater reliability across multiple populations, with Pearson's r of 0.87[199] and 0.94[47] and an ICC of 0.85[76]. The SEM has varied with values of 2.9[76], 4.1[47], and 5.5[199].

Responsiveness

Responsiveness was assessed in different populations using an Effect Size Index calculation and Guyatt's Responsiveness Index. Responsiveness was shown to be comparable to the modified Barthel Index and HABAM in acute medical patients[200]; as good as the Timed Up & Go[40], 6 metre walk test[90], step test[201], and Clinical Test of Sensory Organization and Balance[202] in subacute patients[199]; and more responsive than the modified Barthel Index in transitional care patients[91].

Usability

The DEMMI is easy to use and requires very little training to administer: a short video is sufficient to prepare a clinician or researcher to administer the test (*www.demmi.org.au*). It requires minimal equipment that is found in most clinical settings or homes: a chair, a bed, a timer, and a short walk-way such as a corridor. It can be administered in less than 9 minutes in hospital settings. Overall, these features make this instrument convenient and attractive for uptake in various clinical settings[47].

Interpretability and normative data

Throughout studies of various older adult populations, the MDC_{90} for the DEMMI has ranged from 6.7 to 12.7[47, 76, 199]. The MCID using anchor-based methods has ranged from 9.4 to 14[47, 91, 200], and using a distribution method has ranged from 8.4 – 12[43, 47, 91, 199, 200].

An important next step for improving interpretability of this instrument is the development of reference intervals (or "normative data"). Normative data are important for both clinicians and researchers. Applied to individuals, they provide information about expected mobility levels for sex and age-matched community dwelling peers. Applied to groups, normative data can enhance the ability to draw inferences about sample means.

Just as reference values are critical for the usefulness of measuring blood pressure, having accurate benchmarks for quantifying mobility is essential for treatment, evaluation and goal setting.

3.1.3 Study objective

The purpose of this study was to develop reference scores for the DEMMI for community-dwelling men and women over 60 years old.

3.2 Methods

3.2.1 Study design

This was a cross-sectional observational study.

3.2.2 Participants

The study used a convenience sample of community-dwelling adults aged 60 years and older in two large cities: Vancouver, Canada and Melbourne, Australia. 'Community-dwelling' refered to those living in a house, apartment or assisted living (AL). Those community-dwelling individuals living in a house or apartment were considered to be 'independent'. Those community-dwelling individuals living in AL facilities were considered to be 'semi-independent'. This is consistent with current definitions of AL in British Columbia, Canada[203] and its equivalent in Victoria, Australia, called a retirement village[204]. In both regions, AL provides accommodation and some services that are distinct from the services provided in

residential care/nursing home environments, and requires that residents are able to make decisions and otherwise live independently. I presented this data both as a complete sample (i.e. 'community-dwelling'), as well as separated into 'independent' (living independently in a house or apartment) and 'semi-independent' (living in AL).

Prior to participation, interested individuals were screened for eligibility to ensure they had no clinical conditions that might affect their mobility, including neuromuscular, orthopaedic, or cardiovascular impairments. All participants were required to speak English and were screened for cognitive limitations that would preclude the provision of informed consent. The cognitive screening included three orientation questions (full name, name of location, current date), as well as one question, *"Have you been diagnosed with dementia?"*. Where the ability to read was limited, the consent form was read aloud. All participants provided free and informed written consent.

3.2.3 Ethics

Ethics approval was obtained in Vancouver from the Clinical Research Ethics Board at the University of British Columbia and Vancouver Coastal Health Authority (CREB Certificate #H10-02748). In Melbourne, approval was from the Monash University Human Research Ethics Committee (project #CF07/3954- 2007001870).

3.2.4 Recruitment

In Vancouver, recruitment strategies included advertisements in local newspapers and on bulletin boards in seniors' centres, as well as word-of-mouth advertising by local fitness instructors. In Melbourne, residents of a retirement village and members of a Returned and Services League (RSL) were invited to participate through flyer distribution. Interested individuals contacted researchers by telephone to arrange an appointment for screening and participation or signed up at a specific site on the day of the assessments.

3.2.5 Data collection

In Vancouver, assessments took place at eight different sites, including one AL residence, community centres, seniors' activity centres, and onsite at the Centre for Hip Health and Mobility, University of British Columbia. In Melbourne, assessments took place at two sites, one AL facility (i.e. retirement village) and one RSL centre. Each assessment lasted 40 – 60 minutes.

The primary outcome of interest was the DEMMI. Demographic information collected included age, sex, use of mobility aids, living situation and medical comorbidities/history.

Research assistants or physiotherapists screened interested individuals for eligibility, obtained written informed consent, conducted interviews and administered questionnaires. A physiotherapist (E.M.) administered the DEMMI in Vancouver; in Melbourne, the DEMMI was administered by either a physiotherapist or a physiotherapy undergraduate honors student. The developer of the DEMMI, Dr. de Morton, trained all persons who administered the test by demonstrating test administration. Test administrators also watched a 30-minute instructional DVD (video available through *www.demmi.org.au*). Dr. de Morton was present during data collection in both countries to ensure procedural consistency.

3.2.6 Statistical analysis

Exploratory data analysis included graphical exploration of the data, with descriptive statistics presented in tables. Visual inspection of univariate data as well as a scatter plot of the DEMMI by age and sex with overlying LOWESS ('locally weighted scatterplot smoothing') was used to guide analysis. The DEMMI scores were explored by age category (60-69, 70-79, 80-89, 90+), sex, falls history, living situation, country of residence, and use of mobility aid (e.g. cane, walker). Welch's two-sample t-tests were used to compare DEMMI scores by dichotomous variables (e.g. sex) to account for heteroscedasticity. For comparing DEMMI scores by age category, ANOVA was done followed by post-estimation pairwise comparisons. The potential for floor or ceiling effects was investigated by calculating the percentage of participants achieving scores of zero or 100.

Reference intervals were constructed using empirical centiles (5^{th}, 50^{th}, and 95^{th}) for individual reference[102]. Means and 95% confidence intervals were presented for group reference. All statistical analysis was done using Stata Intercooled 12.0 (StataCorp, Texas, USA).

3.3 Results

Initially, 208 individuals were screened for eligibility, and 23 were excluded (see Figure 3.1 for flow chart). Two participants also had missing data for age, and were therefore also excluded, leaving 183 participants for analysis.

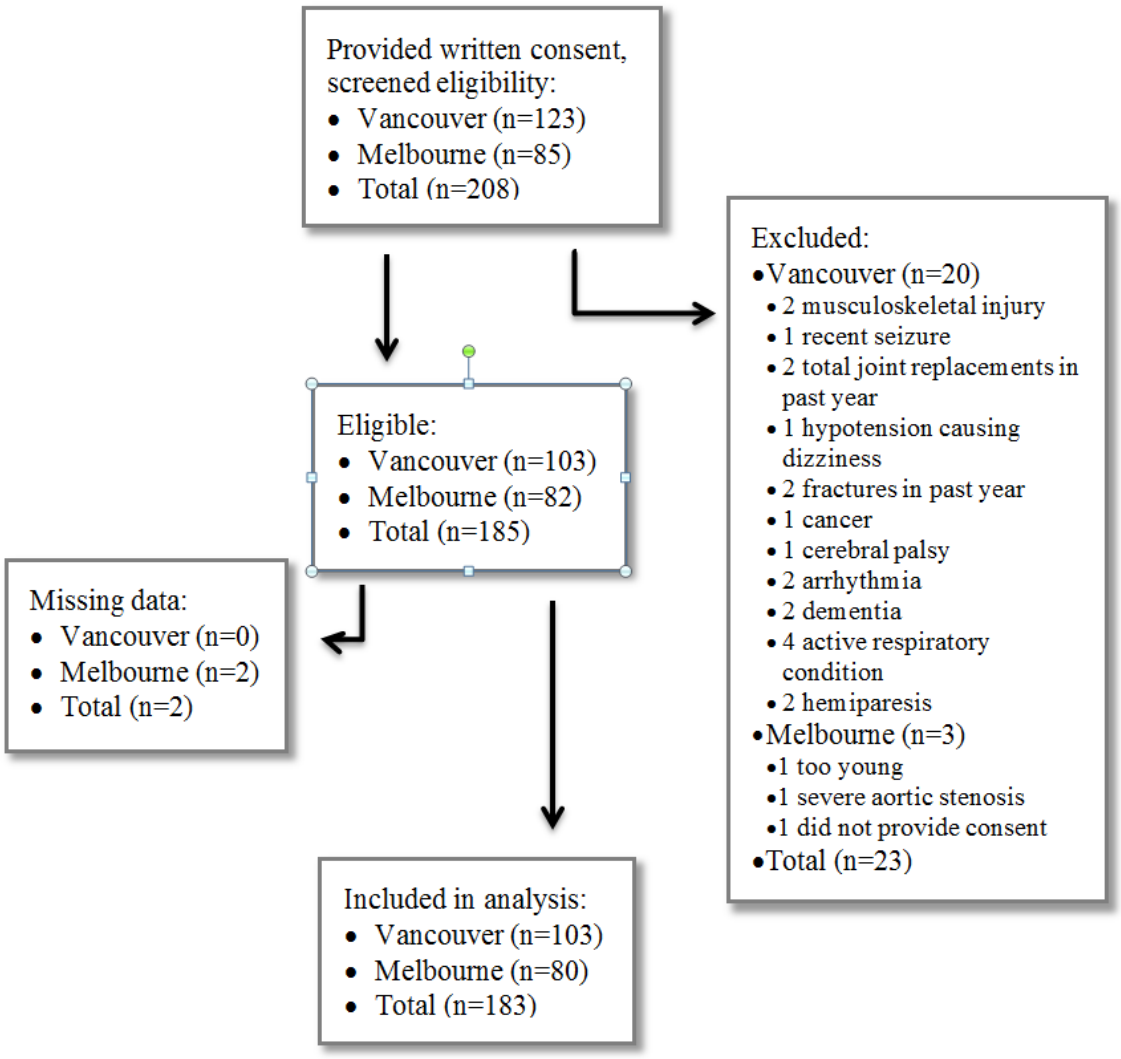

Figure 3.1 Flow diagram

A description of the 183 participants included in the analysis is presented in Table 3.1 (full sample) and Table 3.2 (by site). The majority of the participants lived in Vancouver (n=103, 56%) with the remainder in Melbourne (n=80, 44%). Over half the participants were in their 70s and approximately three quarters were women. Twenty one percent of participants reported at least one fall during the past year.

Table 3.1 Description of study participants (N=183)

	n	%
Vancouver, Canada	103	56
Melbourne, Australia	80	44
Age mean(SD)	74.6(6.7)	
60 – 69	43	23
70 – 79	96	52
80 – 89	43	23
90+	1	1
Women	136	74
≥ 1 fall in past year	8	21
Used mobility aid	28	15
Lived in house or apartment#	120	66
Assisted living / retirement village#	62	34
Lived alone	92	50

Data incomplete for 1 participant, therefore n=182

Table 3.2 Breakdown of study participants by site, number (percent)

	Melbourne		Vancouver							
	Site1[1] n=61	Site2[2] n=21	Site3[3] n=15	Site4[4] n=22	Site5[4] n=1	Site6[4] n=8	Site7[5] n=5	Site8[5] n=21	Site9[5] n=10	Site10[6] n=21
Age mean (SD)	76.8[^] (5.5)	76.9 (7.5)	78.3 (6.8)	72.1 (5.8)	75	70.8 (6.4)	66.7 (5.3)	74.5 (7.6)	73.2 (7.0)	70.1 (4.3)
60 – 69	6 (10%)	4 (19%)	1 (7%)	5 (23%)	-	4 (50%)	3 (60%)	6 (29%)	4 (40%)	10 (48%)
70 – 79	37 (63%)	8 (38%)	8 (53%)	14 (64%)	1	3 (38%)	2 (40%)	10 (48%)	3 (30%)	10 (48%)
80 – 89	16 (27%)	8 (38%)	6 (40%)	3 (14%)	-	1 (13%)	-	5 (24%)	3 (30%)	1 (5%)
90+	-	1 (5%)	-	-	-	-	-	-	-	-
Women n(%)	41 (67%)	12 (57%)	10 (67%)	20 (90%)	1	8 (100%)	2 (40%)	17 (81%)	8 (80%)	18 (86%)
≥ 1 fall in past year	18 (30%)	4 (19%)	4 (27%)	3 (14%)	-	1 (13%)	1 (20%)	3 (14%)	1 (10%)	3 (14%)
Used mobility aid	19 (31%)	5 (24%)	1 (7%)	-	-	-	-	2 (10%)	-	1 (5%)
Lived in house or apartment	5[#] (8%)	20 (95%)	7 (47%)	22 (100%)	1	8 (100%)	5 (100%)	21 (100%)	10 (100%)	21 (100%)
Assisted living	54 (90%)	1 (5%)	8 (53%)	-	-	-	-	-	-	-
Lived alone	29 (48%)	12 (57%)	7 (47%)	12 (55%)	1	1 (13%)	1 (20%)	14 (67%)	6 (60%)	10 (48%)

[1]Melbourne AL; [2]Melbourne RSL; [3]Vancouver AL; [4]Vancouver seniors centre; [5]Vancouver all ages community centre; [6]Vancouver research centre
^Age missing for 2 participants
Data missing for 1 participant

3.3.1 Normative data: DEMMI scores by age, sex and living situation

DEMMI scores are presented visually in Figure 3.2 and Figure 3.3. Reference data are provided for group evaluation as means and confidence intervals (Table 3.3) and for individual evaluation as medians with 5th and 95th percentiles (Table 3.4). The second column of each table provides scores for the full sample (bottom row, "Total" includes all ages). The next column labeled "Independent" provides scores for those living fully independently in the community. Scores for men and women living independently are provided separately in the adjacent columns. Finally, a column labeled "Semi-Independent" includes both men and women living in AL.

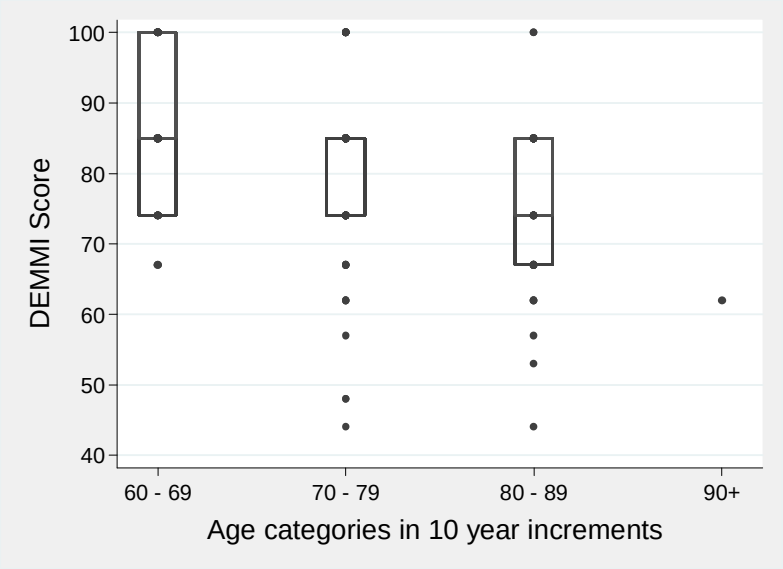

Figure 3.2 Box plot of DEMMI scores by age category with interquartile ranges (nb in age 70-79, median is the same as the 75th percentile; also, the oldest age category has only one observation)

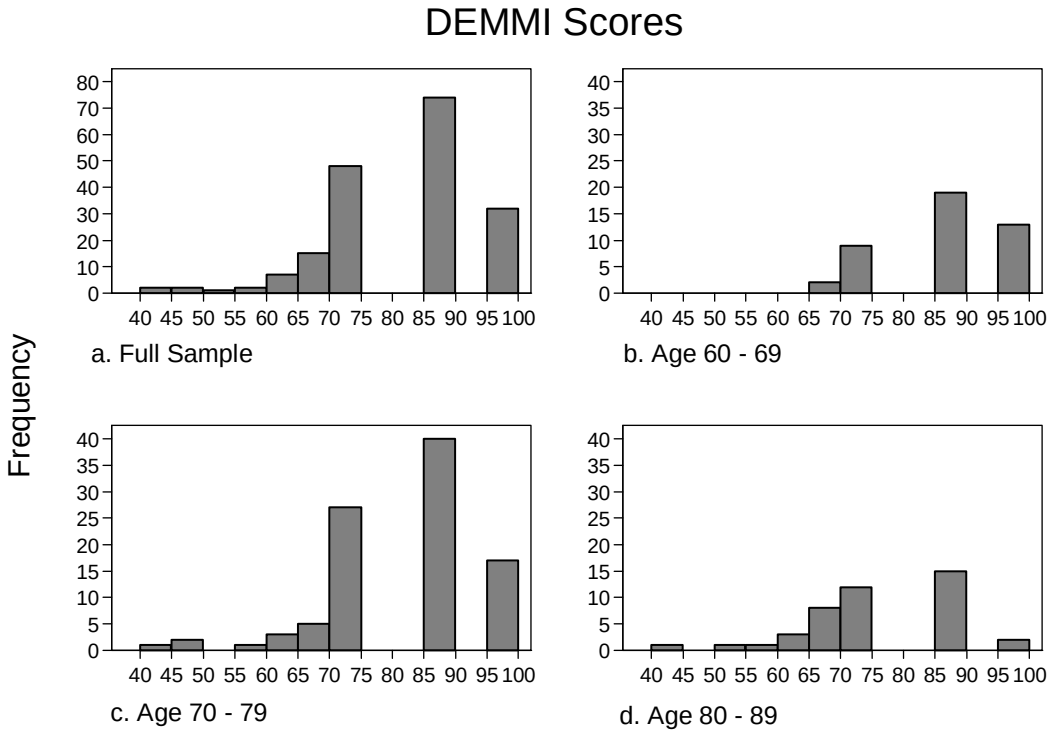

Figure 3.3 Histograms showing DEMMI scores for a) total sample, and b) through d) by age category (note category 90+ contained only 1 participant, therefore not represented graphically)

Table 3.3 DEMMI reference data for group comparisons: mean, 95% confidence interval

Age Category	Overall Mean (95%CI) n	Independent	Independent Women	Independent Men	Semi-Independent
60 – 69	**86.4** (83.2, 89.6) n=43	**86.2** (82.9, 89.6) n=38	**87.2** (83.3, 91.1) n=30	**82.8** (75.4, 90.1) n=8	**87.4** (70.4, 100) n=5
70 – 79	**81.4** (78.9, 83.9) n=96	**84.3** (81.5, 87.1) n=57	**83.4** (80.2, 86.6) n=47	**88.4** (82.1, 94.6) n=10	**77.2** (72.9, 81.6) n=39
80 – 89	**75.3** (71.8, 78.9) n=43	**77.9** (73.3, 82. 5) n=24	**78.6** (73.3, 83.8) n=18	**76.0** (63.10, 88.9) n=6	**71.9** (65.9, 77.9) n=18
90+	**62** n=1	**62** n=1	-	**62** n=1	-
Total	**81.0** (79.3, 82.8) n=183	**83.5** (82.5, 85.4) n=120	**83.7** (81.4, 85.9) n=95	**82.6** (77.9, 87.2) n=25	**76.5** (73.1, 79.9) n=62

Mean DEMMI scores were progressively lower across older age categories for all comparisons with the exception of "Independent Men" which had the highest score in the age 70-79 category (see Table 3.3).

Table 3.4 DEMMI reference intervals for individual comparison, median (5th,95th percentiles). Please refer to Table 3.3 for subsample sizes.

Age Category	Overall median (p5,p95)	Independent	Independent Women	Independent Men	Semi-Independent
60 - 69	**85** (74,100)	**85** (74, 100)	**85** (74,100)	**85** (74,100)	**85** (67, 100)
70 – 79	**85** (62,100)	**85** (67, 100)	**85** (67,100)	**85** (74,100)	**74** (48, 100)
80 - 89	**74** (57,85)	**74** (62, 100)	**85** (57,100)	**74** (67,100)	**74** (44, 85)
90+	**62***	**62***	-	**62***	-
Total	**85** (62,100)	**85** (67, 100)	**85** (67,100)	**85** (67,100)	**74** (48, 100)

**only 1 observation*

Median scores were largely consistent across all columns for a given age category (see Table 3.4). The exceptions were a lower score for "Semi-Independent" in the age 70-79 row; a lower score for "Semi-Independent" in the "Total" row; and a higher score in the "Independent Women" in the age 80-89 row. The 5th percentile scores tended to be progressively lower across older age categories for all comparisons. The 95th percentile was almost always 100, with the

exception of 85 for both “Overall” and “Semi-Independent” in the age 80-89 row. A mild ceiling effect was revealed with 18% of the study sample scoring the maximum of 100.

3.3.2 Comparisons of DEMMI scores by key variables

Differences in mean DEMMI scores by age category were statistically significant on all pairwise comparisons, with lower scores for older age categories (see Table 3.5). There were no statistically significant differences by sex (p=0.49) or falls history (p=0.21). Scores were statistically different by city (Vancouver scores higher than Melbourne, p=0.00), living situation (independent dwellers had higher scores than AL, p=0.00) and mobility aid (those using a gait aid had lower scores, p=0.00). The vast majority of people living in AL used mobility aids (only three did not). Also, the majority of people using mobility aids lived in Melbourne (only four lived in Vancouver, one of whom also lived in AL). Also, a significant age difference existed by city (Melbourne mean age 76.8 vs. Vancouver 72.8, p=0.00) as well as by living situation (AL mean age 77.2 vs. independent 73.1, p=0.00). The proportion of individuals who resided in AL in this study also differed by city (Melbourne, 68% vs. Vancouver, 8%, p=0.00).

Table 3.5 Comparison of DEMMI scores by variables of interest

Variable	**DEMMI Mean(SD) n**	**DEMMI Mean(SD) n**	**p-value**
Age: 70-79 vs. 60-69	81.4(12.3) *n=96*	86.4(10.4) *n=43*	0.02
Age: 80-89 vs. 60-69	75.3 (11.5) *n=43*	86.4(10.4) *n=43*	0.00
Age: 80-89 vs. 70-79	75.3 (11.5) *n=43*	81.4 (12.3) *n=96*	0.01
Sex: female vs. male	80.7 (12.6) *n=136*	82.1 (11.4) *n=47*	0.49
Falls history: yes vs. no	78.3 (16.1) *n=8*	81.8 (11.0) *n=175*	0.21
Semi-independent vs. independent living	76.5 (13.4) *n=120*	83.5 (11.0) *n=62*	0.00
City: Vancouver vs. Melbourne	85.1 (10.0) *n=103*	75.8 (12.9) *n=80*	0.00
Mobility aid use: yes vs. no	66.2 (12.4) *n=28*	83.7 (10.2) *n=155*	0.00

3.4 Discussion

This study generated normative data for the DEMMI mobility instrument for men and women over 60 years old who live independently or semi-independently in the community.

These normative data can be used either for group or individual reference. For example, in a study, normative data can enhance the ability to draw inferences about sample means. An example would be looking at the mean for a sample of independent community dwelling older adults who report knee pain. Let's say their study mean score at baseline was 78.6. Looking at Table 3.3, in the bottom row ("Total") of column 3 ("Independent"), this score is below the lower limit of the 95% confidence interval of 82.5. This suggests that the sample may have lower mobility than their healthy peers, and may therefore be appropriate for a clinical intervention trial.

More commonly, reference data is referred to on an individual basis, such as when a clinician wishes to know how their patient compares to age- and sex-matched peers. For example, a 78-year-old woman may visit a clinician with concerns about increasing difficulties managing things around her home. An initial assessment might reveal a DEMMI score of, say, 62. Looking at Table 3.4, the reference intervals in row 3 ("Age 70–79"), column 4 ("Independent Women") show healthy independent community-dwelling women of her age score a median of 85, with 95 percent scoring above 67. The clinician would then use the results of this assessment to engage in meaningful education and goal setting with the patient. Over the course of treatment, the DEMMI would be periodically re-administered to evaluate the patient's response to treatment and to guide ongoing treatment planning. An end goal for treatment could be set in consultation with the patient that either targeted a certain score (e.g. 74) or identified an item the patient would like to achieve (e.g. pick a pen up off the floor).

In a previous study, de Morton et al.[200] revealed that older adults with acute medical hospital admissions were discharged home with mean DEMMI scores of 60. The current study reveals higher mean scores than this for community dwellers aged 60+ (see Table 3.3). However, it is conceivable that those recovering from acute illness and hospitalization would demonstrate lower mobility at the time of discharge, and that mobility would improve with ongoing recovery and rehabilitation ("rehab potential" is an important factor in discharge planning). Therefore, the

current study supports previous findings, and demonstrates that means will likely differ between healthy community dwellers and those recovering from acute illness. The reference data obtained in this study could be used in the hospital setting to facilitate physiotherapy discharge planning and to support decision making for funding inpatient and community-based rehabilitation services after acute hospital discharge.

3.4.1 Limitations

I note limitations with this study. First, this was a relatively small sample (N = 183), just 25% of whom were men. Previous studies have developed normative data for health instruments from sample populations as small as 32[205], though most commonly have over 100 participants[206-208] and some have over 1000[209, 210]. A larger sample would provide greater representation across both sexes and across all age groups (especially those ≥80) and therefore better confidence in the reference value estimates. Further, my study sample was one of convenience, and therefore brings into question how well this sample represents the target population. Finally, the current study demonstrated 18% of community-dwelling older adults scored the maximum possible score of 100. This is slightly higher than limits suggested as indicating a ceiling effect[100, 101]. Recognizing this, the DEMMI was designed to measure across a broad spectrum of abilities (acutely hospitalized to healthy community dwellers) with a targeted use for clinical settings – documenting improvements in those who already have excellent health is not a primary goal of the DEMMI[211]. During the development of the DEMMI, one of the hardest preliminary items (standing on one leg with eyes closed)[47] was removed as there were no participants who could complete this item in an acutely hospitalized older population, and this item hence negatively impacted on some of the measurement properties of the DEMMI. The present study does not provide convincing evidence to suggest that the inclusion of this item is warranted.

Chapter 4: Conclusion

This thesis work contributes to the literature in the areas of health outcomes measurement and healthy aging. Aging and its association with chronic comorbidity results in widely varied and complex health statuses among older adults. Identifying effective strategies to promote healthy aging and reduce comorbidity relies critically on the ability to accurately measure health outcomes. This requires the use of valid and reliable instruments that have associated reference values to give meaning to test scores. This thesis provides evidence of adequate measurement properties in one new instrument and provides reference values for a second. Both instruments measure health outcomes known to vary with advancing age/comorbidity.

4.1.1 Study #1: PFOOS

In chapter 2, I provided evidence of validity and reliability of the PFOOS in assessing symptoms in an anterior knee pain and patellofemoral OA population. This is relevant to healthy aging because OA is a substantially disabling condition that increases in prevalence with age[22]. Recent evidence suggests that OA of the patellofemoral joint may precede generalized knee OA[119, 120] and is more highly associated with symptoms of knee OA than the tibiofemoral joint[113-115]. The PFOOS is the first instrument developed to assess anterior knee pain and patellofemoral OA symptoms across five different domains (symptoms, pain, activities of daily living, sports/recreation, and quality of life).

Having established validity and reliability of this instrument, it is now appropriate to undertake further evaluation of the PFOOS. First, a study with a larger sample size and a broader spectrum of symptom severity would enable assessment of unidimensionality, further assessment of internal consistency, and further assessment of known groups validation, for example by comparing mean scores of groups of varying OA severity; or comparing patellofemoral pain to patellofemoral OA. A clinical trial should also be undertaken to evaluate the instrument's responsiveness. Future studies should evaluate the PFOOS across multiple populations including patellofemoral pain, isolated patellofemoral OA, multi-compartment OA, and the general population for evaluation of measurement properties, including collection of reference (normative) data. Because the PFOOS contains items from the KOOS, it has the potential to assess both patellofemoral and tibiofemoral symptoms. This is in contrast to the AKPS, which

was designed for anterior knee pain only and does not specifically assess OA or the tibiofemoral joint. The PFOOS therefore may be more appropriate for longitudinal studies where early OA, isolated to the patellofemoral joint, is expected to progress to generalized knee OA. Future studies should therefore assess the validity of the PFOOS longitudinally as knee OA progresses. A final consideration for future studies would be the development of a short form of the PFOOS. Because several items in the PFOOS did not perform well in the current study, there may be some statistical challenges in regards to using the PFOOS with a sample of high functioning people (i.e. working) with isolated anterior knee pain. A short form PFOOS might perform statistically better in this type of cohort in a cross-sectional study or short intervention trial where progression to the tibiofemoral joint is not expected.

4.1.2 Study #2: DEMMI

In Chapter 3, I developed normative data for the DEMMI and assessed mobility of healthy, community-dwelling older adults. This is relevant to healthy aging because mobility is a vital component of health and can decrease with age and associated comorbidity, potentially leading to loss of functional independence, inability to remain living autonomously in the community, and increased mortality[184-188]. Providing reference data for healthy community-dwelling older adults for the DEMMI is important for clinicians and researchers because it gives test administrators the ability to compare individual or group scores with known scores of an age- and sex-matched population. Its potential applications are broad: annual screening for early signs of clinically relevant mobility decline; therapeutic goal setting; evaluation of changes in mobility during rehabilitation or clinical trial; and discharge planning.

It is recommended that future studies be undertaken to develop reference intervals with larger samples of men, and a greater representation of both men and women aged ≥ 80. Reference data should also be developed for other populations where the DEMMI might be used. For example, reference intervals should be developed for acutely hospitalized people to enhance the ability to use DEMMI scores for discharge planning. Alternatively, for specific conditions, such as lower extremity arthritis or people with a history of falls, reference data could enhance comparison with relevant clinical populations. Finally, while cross-cultural validation has begun with a Dutch translation of the DEMMI[76], efforts should be made to develop and validate multiple translations for test administration, which may also benefit from the collection of reference data in different world regions where the DEMMI would be used.

4.1.3 Strengths and limitations of instrument types

While the two instruments evaluated in this thesis are not meant to be directly compared, it is worth considering the respective strengths and limitations of these types of instruments in general.

Performance-based instruments such as the DEMMI offer some benefits not achievable with a self-report test. For example, the test administrator has the opportunity to witness a task's performance first hand. Between the range of 'unable' and 'able' lie many possible contributors to an individual's performance. Watching someone perform a task gives a richness of information beyond the test score alone. A clinician may be able to begin problem solving by watching the quality of the person's movements and their willingness to move – is the movement limited by apparent stiffness, weakness, unsteadiness, improper use of a gait aid, or apparent fear of movement? These details can assist with goal setting and treatment planning. Additional benefits of a performance-based test are that bias introduced from the person being tested is avoided (details of self-report bias are described below).

On the other hand, limitations of a performance-based instrument are that human resources and training are required to administer the test. Also, who administers the test can potentially influence the test results. Bias could be a problem, for example, if a clinician is conducting subsequent test administrations and is hopeful that their treatment has helped someone improve – they may be more likely to 'see' or 'report' an improvement where there may not be any. Alternatively, they may rate someone's performance lower if the person rated just prior was exceptionally high functioning or if the test administrator tends to work with very high functioning people. Additionally, a performance-based instrument misses the important contribution of patient-perspective. Finally, observing a task in a test environment is not the same as how that task might be performed in the individual's natural environment. Performance can be affected by nervousness of the individual being tested, or key relevant features of the individual's environment may be missed (e.g. multitasking while performing a task).

A self-report instrument like the PFOOS offers benefits that are not obtainable through a performance-based test. First, resources are minimized since a trained test administrator is not required. Second, asking the individual about their personal experience or perspective

incorporates important details that would be missed by simply watching someone complete a task. For example, just because they are able to stand up from a chair in a test setting does not necessarily translate to doing this freely at home – pain, fatigue, self-confidence or family dynamics can affect whether or not the individual actually performs this task at home, and if so, how much difficulty they have with the task. Finally, self-report can assist with goal setting and treatment planning in that it can identify features that are important and relevant to the patient.

Limitations of self-report instrument, in addition to what has already been discussed above, pertain primarily to the numerous sources of bias inherent in self-report[32]. Some of this bias can be reduced through careful test development (e.g. avoiding double-barreled or leading questions). Many sources of bias still remain, however. Accurate responses by self-report have considerable cognitive demands, such as understanding and interpreting a question, and remembering current and previous health states. In addition, their level of investment in the test or treatment determines how much effort they put into giving accurate answers. Social desirability, needing to be heard, an overall self-impression regarding a given construct, or wanting to please the test administrator can all affect individual responses.

Reconciling the respective strengths and limitations of these types of instruments can be challenging. To a certain extent, sources of error can be reduced through careful test development and meticulous evaluation of measurement properties of an instrument. However, ultimately there is no one type of test that is superior in obtaining accurate scores. It is therefore recommended that constructs should be assessed by both performance-based and self-report where possible. Further, self-report should be assessed with both a condition-specific and generic instrument where possible and appropriate.

Therefore, in conclusion, the DEMMI and PFOOS have both been shown to be valid and reliable instruments for measuring their respective constructs and populations, and normative data have been provided for the DEMMI. The DEMMI is applicable to a range of conditions affecting mobility, and the PFOOS addresses health constructs relevant to anterior knee pain and patellofemoral OA. The DEMMI targets older adults, while the PFOOS captures a wider age span. The two instruments could therefore serve as complimentary health outcome instruments for certain populations. For example, in a group of people aged 60+ with patellofemoral OA, a

thorough assessment could include both the DEMMI and the PFOOS, and including the SF-36 would enable comparison across other conditions and populations.

References

1. Statistics Canada. *The Canadian Population in 2011: Age and Sex (Analytical Document)*. 2011 Accessed 20 Dec 2012; Available from: http://www12.statcan.gc.ca/census-recensement/2011/as-sa/98-311-x/98-311-x2011001-eng.cfm

2. Central Intelligence Agency. *The World Factbook*. 2012 Accessed 20 December 2012; Available from: https://www.cia.gov/library/publications/the-world-factbook/geos/xx.html

3. Inouye, S.K., S. Studenski, M.E. Tinetti and G.A. Kuchel, *Geriatric syndromes: clinical, research, and policy implications of a core geriatric concept.* J Am Geriatr Soc, 2007. 55(5): p. 780-91

4. Fries, J.F., *The Compression of Morbidity.* Milbank Mem Fund Q, 1983. 61(3): p. 397 - 419

5. Rowe, J.W. and R.L. Kahn, *Successful Aging.* Gerontologist, 1997. 37(4): p. 433 - 40

6. Guccione, A.A., D.T. Felson, J.J. Anderson, J.M. Anthony, Y. Zhang, P.W. Wilson, et al., *The effects of specific medical conditions on the functional limitations of elders in the Framingham Study.* Am J Public Health, 1994. 84(3): p. 351-8

7. Gruenberg, E.M., *The Failures of Success.* Milbank Mem Fund Q, Health and Society, 1977. 55(1): p. 3 - 24

8. Kadam, U.T., F.G. Schellevis, M. Lewis, D.A.W.M. van der Windt, H.C.W. de Vet, L.M. Bouter, et al., *Does age modify the relationship between morbidity severity and physical health in English and Dutch family practice populations?* Qual Life Res, 2009. 18(2): p. 209-20

9. Kellog International Work Group, *The prevention of falls in later life: a report of the Kellogg International Work Group on the Prevention of Falls by the Elderly.* Dan Med Bull, 1987. 34(S4): p. 1 - 24

10. Smallegan, M., *How families decide on nursing home admission.* Geriatric Consulting, 1983. 1: p. 21 – 4

11. Vita, A.J., R.B. Terry, H.B. Hubert and J.F. Fries, *Aging, health risks, and cumulative disability* N Engl J Med, 1998. 338(15): p. 1035–41

12. Guralnik, J.M., L.P. Fried and M.E. Salive, *Disability as a Public Health Outcome in the Aging Population.* Annu Rev Public Health, 1996. 17: p. 25 - 46

13. Loza, E., J.M. Lopez-Gomez, L. Abasolo, J. Maese, L. Carmona, E. Batlle-Gualda, et al., *Economic burden of knee and hip osteoarthritis in Spain.* Arthritis & Rheumatism, 2009. 61(2): p. 158-65

14. Gupta, S., G.A. Hawker, A. Laporte, R. Croxford and P.C. Coyte, *The economic burden of disabling hip and knee osteoarthritis (OA) from the perspective of individuals living with this condition.* Rheumatology, 2005. 44(12): p. 1531-7

15. Leardini, G., F. Salaffi, R. Caporali, B. Canesi, L. Rovati and R. Montanelli, *Direct and indirect costs of osteoarthritis of the knee.* Clin Exp Rheumatol, 2004. 22(6): p. 699-706

16. Valtorta, N.K. and B. Hanratty, *Socioeconomic variation in the financial consequences of ill health for older people with chronic diseases: A systematic review.* Maturitas, 2013. 74(4): p. 313-33

17. Zhu, C.W., M. Sano, S.H. Ferris, P.J. Whitehouse, M.B. Patterson and P.S. Aisen, *Health-Related Resource Use and Costs in Elderly Adults with and without Mild Cognitive Impairment.* J Am Geriatr Soc, 2013. 61(3): p. 396-402

18. Stevens, J.A., P.S. Corso, E.A. Finkelstein and T.R. Miller, *The costs of fatal and non-fatal falls among older adults.* Inj Prev, 2006. 12(5): p. 290-5

19. Canadian Institute for Health Information, *Health Care in Canada, 2011: A Focus on Seniors and Aging*, 2011

20. Public Health Agency of Canada. *Arthritis in Canada: An Ongoing Challenge*. Accessed 20 December 2012; Available from: http://www.phac-aspc.gc.ca/publicat/ac/index-eng.php

21. The Arthritis Society. *Facts and Figures*. Accessed 20 December 2012; Available from: http://www.arthritis.ca

22. Statistics Canada. *Percentage diagnosed with arthritis, by age group and sex, household population aged 15 or older, Canada, 2011 – description*. 2011. Accessed November 15, 2012]; Available from: http://www.statcan.gc.ca/pub/82-625-x/2012001/article/c-g/desc/11657-02-desc-eng.htm

23. Chakravarty, E.F., H.B. Hubert, E. Krishnan, B.B. Bruce, V.B. Lingala and J.F. Fries, *Lifestyle Risk Factors Predict Disability and Death in Healthy Aging Adults.* Am J Med, 2012. 125: p. 190-197

24. Chakravarty, E.F., H.B. Hubert, V.B. Lingala, E. Zatarain and J.F. Fries, *Long distance running and knee osteoarthritis. A prospective study.* Am J Prev Med, 2008. 35(2): p. 133–8

25. Willcox, B., *Successful aging: is there hope?* . CMAJ, 2012. 184(18): p. 1973

26. Andersen, S.L., P. Sebastiani, D.A. Dworkis, L. Feldman and T.T. Perls, *Health Span Approximates Life Span Among Many Supercentenarians: Compression of Morbidity at the Approximate Limit of Life Span.* J Gerontol A Biol Sci Med Sci, 2012. 67A(4): p. 395–405

27. Freedman, V.A., E. Crimmins, R.F. Schoeni, B.C. Spillman, H. Aykan, E. Kramarow, et al., *Resolving Inconsistencies in Trends in Old-Age Disability: Report from a Technical Working Group.* Demography 2004. 41(3): p. 417 - 41

28. World Health Organization. Accessed 21 November 2012; Available from: http://www.who.int/hia/about/glos/en/index1.html

29. Australian Government Department of Health and Aging. Accessed 21 November, 2012; Available from: http://www.health.gov.au/internet/main/publishing.nsf/content/health-pbs-general-pubs-pharmpac-glossary-glossary.htm

30. Wluka, A.E., F. Hanna, M. Davies-Tuck, Y. Wang, R.J. Bell, S.R. Davis, et al., *Bone marrow lesions predict increase in knee cartilage defects and loss of cartilage volume in middle-aged women without knee pain over 2 years.* Ann Rheum Dis, 2009. 68(6): p. 850-5

31. Schneider, M., in *The setting of health research priorities in South Africa.* 2001, Burden of Disease Research Unit Tygerberg, South Africa. p. 45

32. Streiner, D.L. and G.R. Norman, *Health Measurement Scales: a practical guide to their development and use.* Fourth ed. 2008, Oxford: Oxford University Press

33. Portney, L.G. and M.P. Watkins, *Foundations of Clinical Research: Applications to Practice*. Third ed. 2009, Upper Saddle River, New Jersey: Pearson Eduction, Inc

34. Poolman, R.W., M.F. Swiontkowski, J.C.T. Fairbank, E.H. Schemitsch, S. Sprague and H.C.W. de Vet, *Outcome instruments: rationale for their use.* J Bone Joint Surg Am, 2009. 91 Suppl 3: p. 41-9

35. Wilson, I.B. and P.D. Cleary, *Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes.* JAMA, 1995. 273(1): p. 59-65

36. Drężewska, M., *Hip joint mobility in dancers. Pilot study.* Ortop Traumatol Rehabil, 2012. 14(5): p. 1-10

37. Klaassens, M., E. Reinstein, Y. Hilhorst-Hofstee, J.J. Schrander, F. Malfait, H. Staal, et al., *Ehlers-Danlos arthrochalasia type (VIIA-B)--expanding the phenotype: from prenatal life through adulthood.* Clinical Genetics, 2012. 82(2): p. 121-30

38. Harvey, L.A., R. Adams, J. Chu, J. Batty and D. Barratt, *A comparison of patients' and physiotherapists' expectations about walking post spinal cord injury: a longitudinal cohort study.* Spinal Cord, 2012. 50(7): p. 548-52

39. Nusdorfer, L. and A. Jeffs *How mobility can shorten stay, improve outcomes.* Hosp Case Manag 2012. 20(9): p. 141-3

40. Podsiadlo, D. and S. Richardson, *The timed "Up & Go": a test of basic functional mobility for frail elderly persons.* J Am Geriatr Soc, 1991. 39(2): p. 142-8

41. World Health Organisation, *International Classification of Functioning, Disability and Health*, 2001: Geneva, Switzerland

42. Streiner, D.L., *Starting at the beginning: an introduction to coefficient alpha and internal consistency.* Statistical developments and applications, 2003. 80(1): p. 99-103

43. de Morton, N.A., C. Meyer, K.J. Moore, B. Dow, C. Jones and K. Hill, *Validation of the de Morton Mobility Index (DEMMI) with older community care recipients.* Australas J Ageing., 2011. 30(4): p. 220-5

44. Ware, J.E., *SF-36 health survey update.* Spine, 2000. 25: p. 3130-39

45. de Morton, N.A., D.J. Berlowitz and J.L. Keating, *A systematic review of mobility instruments and their measurement properties for older acute medical patients* Health Qual Life Outcomes, 2008. 6: p. 44

46. Dawson, J., H. Doll, R. Fitzpatrick, C. Jenkinson and A.J. Carr, *The routine use of patient reported outcome measures in healthcare settings.* BMJ, 2010. 340: p. c186

47. de Morton, N.A., M. Davidson and J.L. Keating, *The de Morton Mobility Index (DEMMI): an essential health index for an ageing world.* Health Qual Life Outcomes, 2008. 6: p. 63

48. Marra, C.A., J.C. Woolcott, J.A. Kopec, K. Shojania, R. Offer, J.E. Brazier, et al., *A comparison of generic, indirect utility measures (the HUI2, HUI3, SF-6D, and the EQ-5D) and disease-specific instruments (the RAQoL and the HAQ) in rheumatoid arthritis.* Soc Sci Med, 2005. 60(7): p. 1571-82

49. Bellamy, N., *WOMAC Osteoarthritis Index. A user's guide.* 1995, London: University of Western Ontario

50. Kujala, U.M., L.H. Jaakkola, S.K. Koskinen, S. Taimela, M. Hurme and O. Nelimarkka, *Scoring of patellofemoral disorders.* Arthroscopy, 1993. 9(2): p. 159-63

51. Garratt, A.M., S. Brealey, W.J. Gillespie and D.T. Team, *Patient-assessed health instruments for the knee: a structured review.* Rheumatology, 2004. 43(11): p. 1414-23

52. Roos, E.M., H.P. Roos, L.S. Lohmander, C. Ekdahl and B.D. Beynnon, *Knee Injury and Osteoarthritis Outcome Score (KOOS)--development of a self-administered outcome measure.* J Orthop Sports Phys Ther, 1998. 28(2): p. 88-96

53. Stratford, P., C. Gill, M. Westaway and J. Binkley, *Assessing disability and change on individual patients: a report of a patient specific measure.* Physiother Can, 1995. 47: p. 258-63

54. Guyatt, G.H., D.J. Eagle, B. Sackett, A. Willan, L. Griffith, W. McIlroy, et al., *Measuring quality of life in the frail elderly* J Clin Epidemiol, 1993. 46(12): p. 1433-44

55. Paxton, E.W., D.C. Fithian, M.L. Stone and P. Silva, *The reliability and validity of knee-specific and general health instruments in assessing acute patellar dislocation outcomes.* Am J Sports Med, 2003. 31(4): p. 487-92

56. Mokkink, L.B., C.B. Terwee, D.L. Patrick, J. Alonso, P.W. Stratford, D.L. Knol, et al., *The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes.* J Clin Epidemiol, 2010. 63: p. 737-45

57. Thorborg, K., P. Holmich, R. Christensen, J. Petersen and E.M. Roos, *The Copenhagen Hip and Groin Outcome Score (HAGOS): development and validation according to the COSMIN checklist.* Br J Sports Med, 2011. 45(6): p. 478-91

58. Marx, R.G., E.C. Jones, A.A. Allen, D.W. Altchek, S.J. O'Brien, S.A. Rodeo, et al., *Reliability, validity, and responsiveness of four knee outcome scales for athletic patients.* J Bone Joint Surg Am, 2001. 83-A(10): p. 1459-69

59. Shrout, P.E. and J.L. Fleiss, *Intraclass Correlations: Uses in Assessing Rater Reliability.* Psychol Bull, 1979. 86(2): p. 420 – 8

60. St. John's Telegram, *Misdiagnosis knocked her off her feet, patient tells cancer inquiry*, in *Victoria Times Colonist* 2008: Victoria

61. Mokkink, L.B., C.B. Terwee, D.L. Patrick, J. Alonso, P.W. Stratford, D.L. Knol, et al., *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, 2010. 19: p. 539-49

62. Mokkink, L.B., C.B. Terwee, D.L. Knol, P.W. Stratford, J. Alonso, D.L. Patrick, et al., *The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content.* BMC Med Res Methodol, 2010. 10: p. 22

63. Davenport, S.J., S. Paynter and N. de Morton, *What instruments have been used to assess the mobility of community-dwelling older adults?* Phys Ther Rev, 2008. 13(5): p. 345-354

64. Streiner, D.L., *Clinimetrics vs. psychometrics: an unnecessary distinction.* J Clin Epidemiol, 2003. 56: p. 1142-5

65. Feinstein, A.R., *Clinimetrics.* 1987: Yale University Press

66. Comins, J., J. Brodersen, M. Krogsgaard and N. Beyer, *Rasch analysis of the Knee injury and Osteoarthritis Outcome Score (KOOS): a statistical re-evaluation.* Scand J Med Sci Sports, 2008. 18(3): p. 336-45

67. Guion, R.M., *On Trinitarian Doctrines of Validity.* Prof Psychol, 1980. 11(3): p. 385 - 98

68. Messick, M., *Test Validity and the Ethics of Assessment.* American Psychologist, 1980. 35(11): p. 1012 – 27

69. Crossley, K.M., K.L. Bennell, S.M. Cowan and S. Green, *Analysis of outcome measures for persons with patellofemoral pain: which are reliable and valid?* Arch Phys Med Rehabil, 2004. 85(5): p. 815-822

70. Smith, T.O., L. Davies, M.-L. O'Driscoll and S.T. Donnell, *An evaluation of the clinical tests and outcome measures used to assess patellar instability.*The Knee, 2008. 15: p. 255 - 62

71. Schofield, M.J., *Validity of the SF-12 Compared with the SF-36 Health Survey in Pilot Studies of the Australian Longitudinal Study on Women's Health.* J Health Psychol, 1998. 3(2): p. 259 – 71

72. Collins, N.J., D. Misra, D.T. Felson, K.M. Crossley and E.M. Roos, *Measures of knee function: International Knee Documentation Committee (IKDC) Subjective Knee Evaluation Form, Knee Injury and Osteoarthritis Outcome Score (KOOS), Knee Injury and Osteoarthritis Outcome Score Physical Function Short Form (KOOS-PS), Knee Outcome Survey Activities of Daily Living Scale (KOS-ADL), Lysholm Knee Scoring Scale, Oxford Knee Score (OKS), Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC), Activity Rating Scale (ARS), and Tegner Activity Score (TAS).* Arthritis Care Res, 2011. 63(S11): p. S208-28

73. Fritz, S. and M. Lusardi, *White Paper: "Walking Speed: the Sixth Vital Sign".* J Geriatr Phys Ther, 2009. 32(2): p. 2 - 5

74. Stanaway, F.F., D. Gnjidic, F.M. Blyth, D.G. Le Couteur, V. Naganathan, L. Waite, et al., *How fast does the Grim Reaper walk? Receiver operating characteristics curve analysis in healthy men aged 70 and over.* BMJ, 2011. 343: p. d7679

75. Davenport, S.J. and N.A. de Morton, *Clinimetric properties of the de Morton Mobility Index in healthy, community-dwelling older adults.* Arch Phys Med Rehabil, 2011. 92(1): p. 51-8

76. Jans, M.P., V.C. Slootweg, C.R. Boot, N.A. de Morton, G. van der Sluis and N.L. van Meeteren, *Reproducibility and validity of the Dutch translation of the de Morton Mobility Index (DEMMI) used by physiotherapists in older patients with knee or hip osteoarthritis.* Arch Phys Med Rehabil, 2011. 92(11): p. 1892-9

77. Statistics Canada. *Individual Internet use and E-commerce*. 2010 Accessed 21 December 2012; Available from: www.statcan.gc.ca/daily-quotidien/111012a-eng.htm

78. Gwaltney, C.J., A.L. Shields and S. Shiffman, *Equivalence of electronic and paper-and-pencil administration of patient-reported outcome measures: a meta-analytic review.* Value Health, 2008. 11(2): p. 322-33

79. Coons, S.J., C.J. Gwaltney, R.D. Hays, J.J. Lundy, J.A. Sloan, D.A. Revicki, et al., *Recommendations on evidence needed to support measurement equivalence between electronic and paper-based patient-reported outcome (PRO) measures: ISPOR ePRO Good Research Practices Task Force report.* Value Health, 2009. 12(4): p. 419-29

80. Cortina, J.M., *What is coefficient alpha? An examination of theory and applications.* J Appl Psychol, 1993. 78(1): p. 98-104

81. Cronbach, L.J., *Coefficient alpha and the internal structure of tests.* Psychometrika, 1951. 16: p. 297-334

82. Streiner, D.L., *Being inconsistent about consistency: when coefficient alpha does and doesn't matter.* J Pers Assess, 2003. 80(3): p. 217-22

83. Terwee, C.B., S.D.M. Bot, M.R. de Boer, D.A.W.M. van der Windt, D.L. Knol, J. Dekker, et al., *Quality criteria were proposed for measurement properties of health status questionnaires.* J Clin Epidemiol, 2007. 60(1): p. 34-42

84. Weir, J.P., *Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM.* J Strength Cond Res, 2005. 19(1): p. 231-40

85. McGraw, K.O. and S.P. Wong, *Forming Inferences About Some Intraclass Correlation Coefficients.* Psychol Methods, 1996. 1(1): p. 30 - 46

86. Terwee, C.B., L.D. Roorda, D.L. Knol, M.R. De Boer and H.C.W. de Vet, *Linking measurement error to minimal important change of patient-reported outcomes.* J Clin Epidemiol, 2009. 62: p. 1062-7

87. Stratford, P.W., *Getting more from the literature: estimating the standard error of measurement from reliability studies.* Physiother Can, 2004. 56(1): p. 27-30

88. Angst, F., *The new COSMIN guidelines confront traditional concepts of responsiveness.* BMC Med Res Methodol 2011. 11: p. 152

89. Guyatt, G.H., R.A. Deyo, M. Charlson, M.N. Levine and A. Mitchell, *Responsiveness and validity in health status measurement: a clarification.* J Clin Epidemiol, 1989. 42(5): p. 403-8

90. Salbach, N.M., N.E. Mayo, J. Higgins, S. Ahmed, L.E. Finch and C.L. Richards, *Responsiveness and predictability of gait speed and other disability measures in acute stroke.* Arch Phys Med Rehabil 2001. 82: p. 1204-12

91. de Morton, N.A., N.K. Brusco, L. Wood, K. Lawler and N.F. Taylor, *The de Morton Mobility Index (DEMMI) provides a valid method for measuring and monitoring the mobility of patients making the transition from hospital to the community: an observational study.* J Physiother, 2011. 57(2): p. 109-16

92. Beaton, D.E., *Understanding the relevance of measured change through studies of responsiveness.* Spine, 2000. 25(24): p. 3192-9

93. Cowan, S.M., K.L. Bennell, K.M. Crossley, P.W. Hodges and J. McConnell, *Physical therapy alters recruitment of the vasti in patellofemoral pain syndrome.* Med Sci Sports Exerc, 2002. 34(12): p. 1879-85

94. de Vet, H.C., C.B. Terwee, R.W. Ostelo, H. Beckerman, D.L. Knol and L.M. Bouter, *Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change.* Health Qual Life Outcomes, 2006. 4: p. 54

95. Donoghue, D., R. Physiotherapy, g. Older People and E.K. Stokes, *How much change is true change? The minimum detectable change of the Berg Balance Scale in elderly people.* J Rehabil Med, 2009. 41(5): p. 343-6

96. Stratford, P.W., J. Binkley, P. Solomon, E. Finch, C. Gill and J. Moreland, *Defining the minimum level of detectable change for the Roland-Morris questionnaire.* Phys Ther, 1996. 76(4): p. 359-65; discussion 366-8

97. de Vet, H.C.W., L.M. Bouter, P.D. Bezemer and A.J. Beurskens, *Reproducibility and responsiveness of evaluative outcome measures. Theoretical considerations illustrated by an empirical example.* Int J Technol Assess Health Care, 2001. 17: p. 479 - 87

98. de Boer, M.R., H.C.W. de Vet, C.B. Terwee, A.C. Moll, H.J.M. Volker-Dieben and G.H.M.B. van Rens, *Change to the subscales of two vision-related quality of life questionnaires are proposed.* J Clin Epidemiol, 2005. 58: p. 1260 – 8

99. Norman, G.R., J.A. Sloan and K.W. Wyrwich, *Interpretation of changes on health related qualith of life. The remarkable universality of half a standard deviation.* Med Care, 2003. 41: p. 582-92

100. McHorney, C. and A. Tarlov, *Individual-patient monitoring in clinical practice: are available health status surveys adequate?* Qual Life Res, 1995. 4(4): p. 293-307

101. Barber-Westin, S.D., F.R. Noyes and J.W. McCloskey, *Rigorous statistical reliability, validity, and responsiveness testing of the Cincinnati knee rating system in 350 subjects with uninjured, injured, or anterior cruciate ligament-reconstructed knees.* Am J Sports Med, 1999. 27(4): p. 402-16

102. Wright, E.M. and P. Royston, *Calculating reference intervals for laboratory measurements.* Stat Methods Med Res, 1999. 8(2): p. 93-112

103. Rowe, J.W. and R.L. Kahn, *Human Aging: Usual and Successful.* Science, 1987. 237(4811): p. 143 - 9

104. Canadian Institutes of Health Research. *About Knowledge Translation*. 2012. Accessed 13 July 2012; Available from: http://www.cihr-irsc.gc.ca/e/29418.html

105. Altman, R.D. and G.E. Gold, *Atlas of individual radiographic features in osteoarthritis, revised.* Osteoarthritis Cartilage, 2007. 15(Suppl A): p. A1-56

106. Kellgren, J.H. and J.S. Lawrence, *Radiological assessment of osteo-arthrosis.* Ann Rheum Dis, 1957. 16(4): p. 494-502

107. Zhang, W., Doherty, M., Peat, G., Bierma-Zeinstra, S.M.A., Arden, N.K., Bresnihan, B., et al. *EULAR Evidence Based Recommendations for the Diagnosis of Knee Osteoarthritis.* Ann Rheum Dis, 2010. 69: p. 483-9

108. Peat, G., R.C. Duncan, L.R.J. Wood, E. Thomas and S. Muller, *Clinical features of symptomatic patellofemoral joint osteoarthritis.* Arthritis Res Ther, 2012. 14(2): p. R63

109. Peat, G., E. Thomas, R. Duncan, L. Wood, E. Hay and P. Croft, *Clinical classification criteria for knee osteoarthritis: performance in the general population and primary care.* Ann Rheum Dis, 2006. 65(10): p. 1363-7

110. Hinman, R.S. and K.M. Crossley, *Patellofemoral joint osteoarthritis: an important subgroup of knee osteoarthritis.* Rheumatology, 2007. 46(7): p. 1057-62

111. Kalichman, L., Y. Zhang, J. Niu, J. Goggins, D. Gale, D.T. Felson, and D. Hunter, *The association between patellar alignment and patellofemoral joint osteoarthritis features--an MRI study.* Rheumatology, 2007. 46(8): p. 1303-8

112. Crossley, K., G. Marino, M. Macilquham, A. Schache and R. Hinman, *Can patellar tape reduce the patellar malalignment and pain associated with patellofemoral osteoarthritis?* Arth Rheumatism, 2009. 61(12): p. 1719-2

113. Duncan, R., G. Peat, E. Thomas, L. Wood, E. Hay and P. Croft, *Does isolated patellofemoral osteoarthritis matter?* Osteoarthritis Cartilage, 2009. 17: p. 1151-5

114. Kornaat, P.R., J.L. Bloem, R.Y.T. Ceulemans, N. Riyazi, F.R. Rosendaal, R.G. Nelissen, et al., *Osteoarthritis of the knee: association between clinical features and MR imaging findings.* Radiology, 2006. 239(3): p. 811-7

115. Hunter, D.J., L. March and P.N. Sambrook, *The association of cartilage volume with knee pain.* Osteoarthritis Cartilage, 2003. 11(10): p. 725-9

116. Neuman, P., I. Kostogiannis, T. Friden, H. Roos, L.E. Dahlberg and M. Englund, *Patellofemoral osteoarthritis 15 years after anterior cruciate ligament injury--a prospective cohort study.* Osteoarthritis Cartilage, 2009. 17(3): p. 284-90

117. Duncan, R.C., E.M. Hay, S. J. and C. P.R., *Prevalence of radiographic osteoarthritis - it all depends on your point of view.* Rheumatology, 2006. 45(6): p. 757-60

118. Wood, L., S. Muller and G. Peat, *The epidemiology of patellofemoral disorders in adulthood: a review of routine general practice morbidity recording.* Prim Health Care Res Dev, 2011. 12(2): p. 157-64

119. Mazzuca, S.A., K.D. Brandt, B.P. Katz, Y. Ding, K.A. Lane and K.A. Buckwalter, *Risk factors for progression of tibiofemoral osteoarthritis: an analysis based on fluoroscopically standardised knee radiography.* Ann Rheum Dis, 2006. 65: p. 515-9

120. Duncan, R., G. Peat, E. Thomas, E.M. Hay and P. Croft, *Incidence, progression and sequence of development of radiographic knee osteoarthritis in a symptomatic population.* Ann Rheum Dis, 2011. 70(11): p. 1944-8

121. Duncan, R., G. Peat, E. Thomas, L. Wood, E. Hay and P. Croft, *How do pain and function vary with compartmental distribution and severity of radiographic knee osteoarthritis?* Rheumatology, 2008. 47(11): p. 1704-7

122. Crossley, K., S. Cowan and J. McConnell, *Anterior Knee Pain*, in *Clinical Sports Medicine*, P. Bruckner and K. Khan, Editors. 2012, McGraw-Hill: Sidney

123. Rothman, M.L., P. Beltran, J.C. Cappelleri, J. Lipscomb, and B. Teschendorf, *Patient-reported outcomes: conceptual issues.* Value Health, 2007. 10 Suppl 2: p. S66-75

124. McGrail, K., S. Bryan and J. Davis, *Let's all go to the PROM: the case for routine patient-reported outcome measurement in Canadian healthcare.* Healthc Pap, 2011. 11(4): p. 8-18; discussion 55-8

125. Harrison, E. and Q. H, *Analysis of outcome measures used in the study of patellofemoral pain syndrome.* Physiother Can, 1995. 47(4): p. 264-72

126. Harrison, E., D. Magee and H. Quinney, *Development of a clinical tool and patient questionnaire for evaluation of patellofemoral pain syndrome patients.* Clin J Sport Med, 1996. 6(3): p. 163-70

127. Selfe, J. and L. Harper, *Four Outcome Measures for Patellofemoral Joint Problems. Part 1. Development and validity.* Physiother, 2001. 87(10): p. 516-2

128. Laprade, J.A. and E.G. Culham, *A self-administered pain severity scale for patellofemoral pain syndrome.* Clin Rehabil, 2002. 16(7): p. 780-8

129. Watson, C.J., M. Propps, J. Ratner, D.L. Zeigler, P. Horton and S.S. Smith, *Reliability and responsiveness of the lower extremity functional scale and the anterior knee pain scale in patients with anterior knee pain.* J Orthop Sports Phys Ther, 2005. 35(3): p. 136-46

130. Saltzman, C.L., J.A. Goulet, R.T. McClellan, L.A. Schneider and L.S. Matthews, *Results of treatment of displaced patellar fractures by partial patellectomy.* J Bone Joint Surg Am, 1990. 72(9): p. 1279-85

131. Bennell, K., S. Bartam, K. Crossley and S. Green, *Outcome measures in patellofemoral pain syndrome: test retest reliability and inter-relationships.* Phys Ther Sport 2000. 1: p. 32-41

132. Flandry, F., J.P. Hunt, G.C. Terry and J.C. Hughston, *Analysis of subjective knee complaints using visual analog scales.* Am J Sports Med, 1991. 19(2): p. 112-8

133. van Linschoten, R., M. van Middelkoop, E.M. Heintjes, S.M.A. Bierma-Zeinstra, J.A.N. Verhaar and B.W. Koes, *Patellofemoral Pain Syndrome and Exercise Therapy*, in *General Practice* 2012, Erasmus Universiteit Rotterdam: Rotterdam

134. Smith, T.O., L. Davies, R. Chester, A. Clark and S.T. Donell, *Clinical outcomes of rehabilitation for patients following lateral patellar dislocation: a systematic review.* Physiotherapy, 2010. 96(4): p. 269-81

135. Chesworth, B.M., E. Culham, G.E. Tata and M. Peat, *Validation of outcome measures in patients with patellofemoral syndrome.* J Orthop Sports Phys Ther, 1989. 10(8): p. 302-8

136. Fulkerson, J.P., G.J. Becker, J.A. Meaney, M. Miranda and M.A. Folcik, *Anteromedial tibial tubercle transfer without bone graft.* Am J Sports Med, 1990. 18(5): p. 490-7

137. Howe, T.E., L.J. Dawson, G. Syme, L. Duncan and J. Reid, *Evaluation of outcome measures for use in clinical practice for adults with musculoskeletal conditions of the knee: a systematic review.* Man Ther, 2012. 17(2): p. 100-18

138. Warden, S., R. Hinman, M. Watson, K. Avin, A. Bialocerkowski and K. Crossley, *Patellar taping and bracing for the treatment of chronic knee pain: a systematic review and meta-analysis.* Arth Rheumatism, 2008. 59(1): p. 73-83

139. Rathleff, M.S., E.M. Roos, J.L. Olesen and S. Rasmussen, *Early intervention for adolescents with patellofemoral pain syndrome--a pragmatic cluster randomised controlled trial.* BMC Musculoskelet Disord, 2012. 13: p. 9

140. Piva, S.R., G.K. Fitzgerald, J.J. Irrgang, J.M. Fritz, S. Wisniewski, G.T. McGinty, et al., *Associates of physical function and pain in patients with patellofemoral pain syndrome.* Arch Phys Med Rehabil, 2009. 90(2): p. 285-95

141. Lysholm, J. and J. Gillquist, *Evaluation of knee ligament surgery results with special emphasis on use of a scoring scale.* Am J Sports Med, 1982. 10(3): p. 150-4

142. Roos, E.M. and S. Toksvig-Larsen, *Knee injury and Osteoarthritis Outcome Score (KOOS) - validation and comparison to the WOMAC in total knee replacement.* Health Qual Life Outcomes, 2003. 1: p. 17

143. Salavati, M., B. Akhbari, F. Mohammadi, M. Mazaheri and M. Khorrami, *Knee injury and Osteoarthritis Outcome Score (KOOS); reliability and validity in competitive athletes after anterior cruciate ligament reconstruction.* Osteoarthritis Cartilage, 2011. 19(4): p. 406-10

144. Vanlauwe, J.J.E., T. Claes, D. Van Assche, J. Bellemans and F.P. Luyten, *Characterized chondrocyte implantation in the patellofemoral joint: an up to 4-year follow-up of a prospective cohort of 38 patients.* Am J Sports Med, 2012. 40(8): p. 1799-807

145. Larsson, S., M. Englund, A. Struglics and L.S. Lohmander, *The association between changes in synovial fluid levels of ARGS-aggrecan fragments, progression of radiographic osteoarthritis and self-reported outcomes: a cohort study.* Osteoarthritis Cartilage, 2012. 20(5): p. 388-95

146. Thaunat, M., C. Bessiere, N. Pujol, P. Boisrenoult and P. Beaufils, *Recession wedge trochleoplasty as an additional procedure in the surgical treatment of patellar instability with major trochlear dysplasia: early results.* Orthop Traumatol Surg Res, 2011. 97(8): p. 833-45

147. Bonner, K.F., W. Daner and J.Q. Yao, *2-year postoperative evaluation of a patient with a symptomatic full-thickness patellar cartilage defect repaired with particulated juvenile cartilage tissue.* J Knee Surg, 2010. 23(2): p. 109-14

148. Lygre, S.H.L., B. Espehaug, L.I. Havelin, S.E. Vollset and O. Furnes, *Does patella resurfacing really matter? Pain and function in 972 patients after primary total knee arthroplasty.* Acta Orthop, 2010. 81(1): p. 99-107

149. Crossley, K.M., *Personal Communication*, 2012

150. Bossuyt, P.M., J.B. Reitsma, D.E. Bruns, C.A. Gatsonis, P.P. Glasziou, L.M. Irwig, et al. *Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative.* Radiology, 2003. 226(1): p. 24-8

151. Moher, D., K.F. Schulz, and D. Altman, *The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials.* JAMA, 2001. 285(15): p. 1987-91

152. Gaito, J., *Measurement scales and statistics: resurgence of an old misconception.* Psychol Bull, 1980. 87(3): p. 564 - 7

153. Petri, M., E. Liodakis, M. Hofmeister, F.J. Despang, M. Maier, P. Balcarek, et al., *Operative vs conservative treatment of traumatic patellar dislocation: results of a prospective randomized controlled clinical trial.* Arch Orthop Trauma Surg, 2013. 133(2): p. 209-13

154. Kuru, T., A. Yaliman and E.E. Dereli, *Comparison of efficiency of Kinesio taping and electrical stimulation in patients with patellofemoral pain syndrome.* Acta Orthop Traumatol Turc, 2012. 46(5): p. 385-92

155. Pattyn, E., N. Mahieu, J. Selfe, P. Verdonk, A. Steyaert and E. Witvrouw, *What predicts functional outcome after treatment for patellofemoral pain?* Med Sci Sports Exerc, 2012. 44(10): p. 1827-33

156. Frye, J.L., L.N. Ramey and J.M. Hart, *The effects of exercise on decreasing pain and increasing function in patients with patellofemoral pain syndrome: a systematic review.* Sports Health, 2012. 4(3): p. 205-10

157. Lake, D.A. and N.H. Wofford, *Effect of therapeutic modalities on patients with patellofemoral pain syndrome: a systematic review.* Sports Health, 2011. 3(2): p. 182-9

158. Ware, J.E., M. Kosinski, J.B. Bjorner, D.M. Turner-Bowker, B. Gandek and M.E. Maruish, *SF-36v2® Health Survey: Administration guide for clinical trial investigators.* 2008, Lincoln, RI: Quality Metric Incorporated

159. Survey Monkey. Available from: http://www.surveymonkey.com

160. Walter, S.D., M. Eliasziw and A. Donner, *Sample size and optimal designs for reliability studies.* Stat Med, 1998. 17(1): p. 101-10

161. Cicchetti, D.V., *Sample size requirements for increasing the precision of reliability estimates: problems and proposed solutions.* J Clin Exp Neuropsychol, 1999. 21(4): p. 567 - 70

162. Cicchetti, D.V., *The precision of reliability and validity estimates re-visited: distinguishing between clinical and statistical significance of sample size requirements (Methodological Commentary).* J Clin Exp Neuropsychol, 2001. 23(5): p. 695 - 700

163. Sijtsma, K. and L.A. van der Ark, *Investigation and treatment of missing item scores in test and questionnaire data.* Multivariate Behav Res, 2003. 38(4): p. 505-28

164. Hardouin, J.-B., R. Conroy and V. Sebille, *Imputation by the mean score should be avoided when validating a Patient Reported Outcomes questionnaire by a Rasch model in presence of informative missing data.* BMC Med Res Methodol, 2011. 11: p. 105

165. Chavance, M., *Handling missing items in quality of life studies.* Commun Stat Theory Methods, 2004. 33(6): p. 1371-83

166. Roos, E. *The Knee Injury and Osteoarthritis Outcome Score.* Accessed 15 November, 2012; Available from: http://www.koos.nu

167. Huisman, M., *Imputation of missing item responses: some simple techniques.* Qual Quant, 2000. 34: p. 331-51

168. Peyre, H., A. Leplege and J. Coste, *Missing data methods for dealing with missing items in quality of life questionnaires. A comparison by simulation of personal mean score, full information maximum likelihood, multiple imputation, and hot deck techniques applied to the SF-36 in the French 2003 decennial health survey.* Qual Life Res, 2011. 20(2): p. 287-300

169. Streiner, D.L., *Personal Communication*, 2013

170. Eekhout, I., M.R. De Boer, J.W.R. Twisk, H.C.W. De Vet and M.W. Heymans, *Missing data: a systematic review of how they are reported and handled (Brief Report).* Epidemiol, 2012. 23: p. 729 - 32

171. Fayers, P. and D. Machin, *Quality of Life: The Assessment, Analysis and Interpretation of Patient-reported Outcomes. Second Edition*, 2007, John Wiley & Sons: West Sussex

172. Fielding, S., G. Maclennan, J.A. Cook and C.R. Ramsay, *A review of RCTs in four medical journals to assess the use of imputation to overcome missing data in quality of life outcomes.* Trials, 2008. 9: p. 51

173. Marx, R.G., A. Menezes, L. Horovitz, E.C. Jones and R.F. Warren, *A comparison of two time intervals for test-retest reliability of health status instruments.* J Clin Epidemiol, 2003. 56(8): p. 730-5

174. Bland, J.M. and D.G. Altman, *Statistical methods for assessing agreement between two methods of clinical measurement.* Lancet, 1986. 1: p. 307 - 10

175. Visintainer, P., *personal communication*, 2012

176. Roos, E.M., H.P. Roos, C. Ekdahl and L.S. Lohmander, *Knee injury and Osteoarthritis Outcome Score (KOOS)--validation of a Swedish version.* Scand J Med Sci Sports, 1998. 8(6): p. 439-48

177. Ornetti, P., S. Parratte, L. Gossec, C. Tavernier, J.N. Argenson, E.M. Roos, et al., *Cross-cultural adaptation and validation of the French version of the Knee injury and Osteoarthritis Outcome Score (KOOS) in knee osteoarthritis patients.* Osteoarthritis Cartilage, 2008. 16(4): p. 423-8

178. Norman, G.R., J.A. Sloan and K.W. Wyrwich, *The truly remarkable universality of half a standard deviation: confirmation through another look.* Expert Rev, 2004. 4(5): p. 581-5

179. Cook, C.E., *Clinimetrics Corner: The Minimal Clinically Important Change Score (MCID): A Necessary Pretense.* J Manual Manipulative Ther, 2008. 16(4): p. E82-3

180. Wells, G., D. Beaton, B. Shea, M. Boers, L. Simon, V. Strand, et al., *Minimal clinically important differences: review of methods.* J Rheumatol, 2001. 28(2): p. 406-412

181. Farivar, S.S., H. Liu and R.D. Hays, *Half standard deviation estimate of the minimally important difference in HRQOL scores?* Expert Rev, 2004. 4(5): p. 515-23

182. Wise, E.A., *Methods for analyzing psychotherapy outcomes: a review of clinical significance, reliable change, and recommendations for future directions.* J Pers Assess, 2004. 82(1): p. 50-9

183. Walton, D. and J.M. Elliott, *A higher-order analysis supports use of the 11-item version of the tampa scale for kinesiophobia in people with neck pain.* Phys Ther, 2013. 93(1): p. 60-8

184. Warren, M.D. and R. Knight, *Mortality in relation to the functional capacities of people with disabilities living at home.* J Epidemiol Community Health, 1982. 36: p. 220-3

185. Corti, M.C., J.M. Guralnik, M.E. Salive and J.D. Sorkin, *Serum Albumin Level and Physical Disability as Predictors of Mortality in Older Persons.* JAMA, 1994. 272: p. 1036 – 42

186. Donaldson, L.J., D.G. Clayton and M. Clarke, *The elderly in residential care: mortality in relation to functional capacity.* J Epidemiol Community Health, 1980. 34: p. 96 – 101

187. Studenski, S., S. Perera, D. Wallace, J. Chandler, P. Duncan, E. Rooney, et al., *Physical performance measures in the clinical setting.* J Am Geriatr Soc, 2003. 51(3): p. 314-22

188. Dumurgier, J., A. Elbaz, P. Ducimetiere, B. Tavernier, A. Alperovitch and C. Tzourio, *Slow walking speed and cardiovascular death in well functioning older adults: prospective cohort study.* BMJ. Online First, 2009. 339

189. Shumway-Cook, A., S. Brauer and M. Woollacott, *Predicting the probability for falls in community-dwelling older adults using the Timed Up & Go Test.* Phys Ther, 2000. 80: p. 896-903

190. Tromp, A.M., S.M. Pluijm, J.H. Smit, D.J. Deeg, L.M. Bouter and P. Lips, *Fall-risk screening test: a prospective study on predictors for falls in community-dwelling elderly.* J Clin Epidemiol, 2001. 54(8): p. 837-844

191. Sherrington, C., S.R. Lord, J.C.T. Close, E. Barraclough, M. Taylor, S. O'Rourke, et al., *Development of a Tool for Prediction of Falls in Rehabilitation Settings (Predict_First): A Prospective Cohort Study.* J Rehabil Med, 2010. 42: p. 482-8

192. Dorans, N.J., *Linking scores from multiple health outcome instruments.* Qual Life Res, 2007. 16(Suppl 1): p. 85-94

193. Pollock, C.L., J.J. Eng and S.J. Garland, *Clinical measurement of walking balance in people post stroke: a systematic review.* Clin Rehabil 2011 25(8): p. 693-708

194. MacKnight, C. and K. Rockwood, *Rasch analysis of the hierarchical assessment of balance and mobility (HABAM).* J Clin Epidemiol, 2000. 53(12): p. 1242-7

195. Mahoney, F.I. and D.W. Barthel, *Functional Evaluation: The Barthel Index.* Md State Med J, 1965. 14: p. 61-5

196. Folstein, M.F., S.E. Folstein and P.R. McHugh, *"Mini-mental state". A practical method for grading the cognitive state of patients for the clinician.* J Psychiatr Res, 1975. 12(3): p. 189-98

197. Knaus, W.A., E.A. Draper, D.P. Wagner and J.E. Zimmerman, *APACHE II: a severity of disease classification system.* Crit Care Med, 1985. 13(10): p. 818-29

198. Charlson, M.E., P. Pompei, K.L. Ales and C.R. MacKenzie, *A new method of classifying prognostic comorbidity in longitudinal studies: development and validation.* J Chronic Dis, 1987. 40(5): p. 373-83

199. de Morton, N.A. and K. Lane, *Validity and reliability of the de Morton Mobility Index in the subacute hospital setting in a geriatric evaluation and management population.* J Rehabil Med, 2010. 42(10): p. 956-61

200. de Morton, N.A., M. Davidson and J.L. Keating, *Validity, responsiveness and the minimal clinically important difference for the de Morton Mobility Index (DEMMI) in an older acute medical population.* BMC Geriatrics, 2010. 10: p. 72

201. Hill, K.D., J. Bernhardt, A.M. McGann, D. Maltese and D. Berkovits, *A new test of dynamic standing balance for stroke patients: reliability, validity, and comparison with healthy elderly.* Physiother Can, 1996. 48(257-62)

202. Cohen, H., C.A. Blatchly and L.L. Gombash, *A study of the clinical test of sensory interaction and balance.* Phys Ther, 1993. 73(6): p. 346-51; discussion 351-4

203. Government of British Columbia Canada, *Community Care and Assisted Living Act, Part 1, Paragraph 1*, 2002

204. Government of Victoria Australia, *Retirement Villages Act, Number 126. Version 064. Section 3*, 1986: Victoria, Australia

205. Cohen, H., L. Heaton, S. Congdon and H. Jenkins, *Changes in sensory organization test scores with age.* Age Ageing, 1996. 25: p. 39 - 44

206. Blake, C., M. Codd and Y. O'Meara, *The short form 36 (SF-36) health survey: normative data for the Irish population.* Ir J Med Sci, 2000. 169: p. 195-200

207. Bohannon, R., *Comfortable and maximum walking speed of adults aged 20-79 years: reference values and determinants.* Age Ageing, 1997. 26: p. 15 - 9

208. Vereeck, L., F. Wuyts, S. Truijen and P. Van de Heyning, *Clinical assessment of balance: normative data, and gender and age effects.* Int J Audiol, 2008. 47(2): p. 67-75

209. Era, P., P. Sainio, S. Koskinen, P. Haavisto, M. Vaara and A. Aromaa, *Postural balance in a random sample of 7,979 subjects aged 30 years and over.* Gerontology, 2006. 52(4): p. 204-13

210. Macfarlane, D.J., K.L. Chou, Y.H. Cheng and I. Chi, *Validity and normative data for thirty-second chair stand test in elderly community-dwelling Hong Kong Chinese.* Am J Human Biol, 2006. 18(3): p. 418-21

211. Bindman, A.B., D. Keane and N. Lurie, *Measuring health changes among severely ill patients: the floor phenomenon.* Med Care, 1990. 28: p. 1142-52

Appendices

Appendix A PFOOS Study: The Patellofemoral Pain and Osteoarthritis Outcome Scale

PFOOS Version 1.3

Instructions

This survey asks for your view about your knee. This information will help us keep track of how you feel about your knee and how well you are able to do your usual activities. Answer every question by ticking the appropriate box, only one box for each question. If you are unsure about how to answer a question, please give the best answer you can.

Symptoms

These questions should be answered thinking of your knee symptoms during the **last week**.

S1 Do you have swelling in your knee?

Never	Rarely	Sometimes	Often	Always
☐	☐	☐	☐	☐

S2 Do you feel grinding, hear clicking or any other type of noise when your knee moves?

Never	Rarely	Sometimes	Often	Always
☐	☐	☐	☐	☐

S3 Does your knee catch or hang up when moving?

Never	Rarely	Sometimes	Often	Always
☐	☐	☐	☐	☐

S4 Can you straighten your knee fully?

Always	Often	Sometimes	Rarely	Never
☐	☐	☐	☐	☐

S5 Can you bend your knee fully?

Always	Often	Sometimes	Rarely	Never
☐	☐	☐	☐	☐

Stiffness

The following questions concern the amount of joint stiffness you have experienced during the **last week** in your knee. Stiffness is a sensation of restriction or slowness in the ease with which you move your knee joint.

S6 How severe is your knee joint stiffness after first wakening in the morning?

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

S7 How severe is your knee stiffness after sitting, lying or resting **later in the day**?

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

NS8 How severe is your knee stiffness after exercise?

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

Pain

P1 How often do you experience knee pain?

Never	Monthly	Weekly	Daily	Always
☐	☐	☐	☐	☐

What amount of knee pain have you experienced the **last week** during the following activities?

- If you haven't done this activity because of fear of pain or on medical advice, please tick "EXTREME"
- If you are unsure about an item, please give the best answer you can
- If you do not do an activity for reasons other than pain or medical advice, tick "N/A"

P2 Twisting/pivoting on your knee

N/A	None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐	☐

P3 Straightening knee fully

N/A	None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐	☐

P4 Bending knee fully

N/A	None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐	☐

P5 Walking on flat surface

N/A	None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐	☐

P9 Standing upright

N/A	None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐	☐

P7 At night while in bed

N/A	None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐	☐

P8 Sitting or lying

N/A	None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐	☐

NP9 Rising from sitting (getting out of the car)

N/A	None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐	☐

NP10 Kneeling

N/A	None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐	☐

NP11 Squatting

N/A	None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐	☐

NP12 Heavy household activities (including carrying and lifting)

N/A	None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐	☐

P13	Going up or down stairs					
	N/A	None	Mild	Moderate	Severe	Extreme
	☐	☐	☐	☐	☐	☐
NP14	Hopping/jumping					
	N/A	None	Mild	Moderate	Severe	Extreme
	☐	☐	☐	☐	☐	☐
NP15	Running/jogging (including prolonged)					
	N/A	None	Mild	Moderate	Severe	Extreme
	☐	☐	☐	☐	☐	☐
NP16	After sport and recreational activities					
	N/A	None	Mild	Moderate	Severe	Extreme
	☐	☐	☐	☐	☐	☐
NP17	How often do you experience knee pain after stopping activity?					
	N/A	Never	Monthly	Weekly	Daily	Always
	☐	☐	☐	☐	☐	☐
NP18	How often does pain limit your activity?					
	N/A	Never	Monthly	Weekly	Daily	Always
	☐	☐	☐	☐	☐	☐
NP19	How often do you experience pain with cold weather?					
	N/A	Never	Monthly	Weekly	Daily	Always
	☐	☐	☐	☐	☐	☐

Function, daily living

The following questions concern your physical function. By this we mean your ability to move around and to look after yourself. For each of the following activities please indicate the degree of difficulty you have experienced in the **last week** due to your knee.

- If you haven't done this activity because of fear of pain or on medical advice, please tick "EXTREME"
- If you are unsure about an item, please give the best answer you can
- If you do not do an activity for reasons other than pain or medical advice, tick "N/A"

A1	Descending stairs					
	N/A ☐	None ☐	Mild ☐	Moderate ☐	Severe ☐	Extreme ☐
A2	Ascending stairs					
	N/A ☐	None ☐	Mild ☐	Moderate ☐	Severe ☐	Extreme ☐
A3	Rising from sitting					
	N/A ☐	None ☐	Mild ☐	Moderate ☐	Severe ☐	Extreme ☐
A4	Standing					
	N/A ☐	None ☐	Mild ☐	Moderate ☐	Severe ☐	Extreme ☐
A5	Bending to floor/pick up an object					
	N/A ☐	None ☐	Mild ☐	Moderate ☐	Severe ☐	Extreme ☐
A6	Walking on flat surface					
	N/A ☐	None ☐	Mild ☐	Moderate ☐	Severe ☐	Extreme ☐
A7	Getting in/out of car					
	N/A ☐	None ☐	Mild ☐	Moderate ☐	Severe ☐	Extreme ☐
A8	Going shopping					
	N/A ☐	None ☐	Mild ☐	Moderate ☐	Severe ☐	Extreme ☐
A9	Putting on socks/stockings					
	N/A ☐	None ☐	Mild ☐	Moderate ☐	Severe ☐	Extreme ☐
A10	Rising from bed					
	N/A ☐	None ☐	Mild ☐	Moderate ☐	Severe ☐	Extreme ☐
A11	Taking off socks/stockings					
	N/A ☐	None ☐	Mild ☐	Moderate ☐	Severe ☐	Extreme ☐

A12	Lying in bed (turning over, maintaining knee position)					
	N/A ☐	None ☐	Mild ☐	Moderate ☐	Severe ☐	Extreme ☐
A13	Getting in/out of bath					
	N/A ☐	None ☐	Mild ☐	Moderate ☐	Severe ☐	Extreme ☐
A14	Sitting					
	N/A ☐	None ☐	Mild ☐	Moderate ☐	Severe ☐	Extreme ☐
A15	Getting on/off toilet					
	N/A ☐	None ☐	Mild ☐	Moderate ☐	Severe ☐	Extreme ☐
A16	Heavy domestic duties (moving heavy boxes, scrubbing floors, etc.)					
	N/A ☐	None ☐	Mild ☐	Moderate ☐	Severe ☐	Extreme ☐
A17	Light domestic duties (cooking, dusting, etc.)					
	N/A ☐	None ☐	Mild ☐	Moderate ☐	Severe ☐	Extreme ☐

Function, sports and recreational activities

The following questions concern your physical function when being active on a higher level. The questions should be answered thinking of what degree of difficulty you have experienced during the **last week** due to your knee.

- If you haven't done this activity because of fear of pain or on medical advice, please tick "EXTREME"
- If you are unsure about an item, please give the best answer you can
- If you do not do an activity for reasons other than pain or medical advice, tick "N/A"

SP1	Squatting					
	N/A ☐	None ☐	Mild ☐	Moderate ☐	Severe ☐	Extreme ☐
SP2	Running					
	N/A ☐	None ☐	Mild ☐	Moderate ☐	Severe ☐	Extreme ☐
SP3	Jumping					
	N/A ☐	None ☐	Mild ☐	Moderate ☐	Severe ☐	Extreme ☐
SP4	Twisting/pivoting on your injured knee					
	N/A ☐	None ☐	Mild ☐	Moderate ☐	Severe ☐	Extreme ☐
SP5	Kneeling					
	N/A ☐	None ☐	Mild ☐	Moderate ☐	Severe ☐	Extreme ☐

Quality of Life

Q1	How often are you aware of your knee problem?				
	Never	Monthly	Weekly	Daily	Constantly
	☐	☐	☐	☐	☐
Q2	Have you modified your life style to avoid potentially damaging activities to your knee?				
	Not at all	Mildly	Moderately	Severely	Totally
	☐	☐	☐	☐	☐
Q3	How much are you troubled with lack of confidence in your knee?				
	Not at all	Mildly	Moderately	Severely	Totally
	☐	☐	☐	☐	☐
Q4	In general, how much difficulty do you have with your knee?				
	None	Mild	Moderate	Severe	Extreme
	☐	☐	☐	☐	☐
NQ5	Have you modified your sport or recreational activities due to your knee pain?				
	Not at all	Mildly	Moderately	Severely	Totally
	☐	☐	☐	☐	☐

Appendix B PFOOS Study: Anterior Knee Pain Scale[2]

For each question, circle the letter which best corresponds to your symptoms:

1. Limp

[a] none
[b] slight or periodic
[c] constant

2. Weight-bearing

[a] full weight-bearing without pain
[b] weight-bearing possible, but painful
[c] weight bearing impossible

3. Walking

[a] unlimited
[b] more than 2 km
[c] 1-2 km
[d] unable

4. Stairs

[a] no difficulty
[b] slight pain when descending
[c] pain when both ascending and descending
[d] unable

5. Squatting

[a] no difficulty
[b] repeated squatting painful
[c] painful each time
[d] possible with partial weight-bearing
[e] unable

6. Running

[a] no difficulty
[b] pain after more than 2 km
[c] slight pain from start
[d] severe pain
[e] unable

7. Jumping

[a] no difficulty
[b] slight difficulty
[c] constant pain
[d] unable

[2] Reprinted from Arthroscopy, Volume 9, Kujala U.M. et al., "*Scoring of patellofemoral disorders*", pp. 159-63, 1993 with permission from Elsevier.

8. Prolonged sitting with knees flexed

[a] no difficulty
[b] pain after exercise
[c] constant pain
[d] pain forces to extend knee temporarily
[e] unable

9. Pain

[a] none
[b] slight and occasional
[c] interferes with sleep
[d] occasionally severe
[e] constant and severe

10. Swelling

[a] none
[b] after exertion
[c] after daily exercise
[d] every evening
[e] constant

11. Abnormal kneecap movements (subluxation-kneecap going out)

[a] none
[b] occasionally in sports activities
[c] occasionally in daily activities
[d] at least one documented dislocation
[e] more than two dislocations

12. Atrophy of thigh

[a] none
[b] slight
[c] severe

13. Flexion deficiency

[a] none
[b] slight
[c] severe

Appendix C PFOOS Study: SF-36 v.2[3]

Your Health and Well-Being

This questionnaire asks for your views about your health. This information will help keep track of how you feel and how well you are able to do your usual activities. *Thank you for completing this survey!*

For each of the following questions, please mark an ☒ in the one box that best describes your answer.

1. In general, would you say your health is:

Excellent	Very good	Good	Fair	Poor
☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

2. <u>Compared to one year ago</u>, how would you rate your health in general <u>now</u>?

Much better now than one year ago	Somewhat better now than one year ago	About the same as one year ago	Somewhat worse now than one year ago	Much worse now than one year ago
☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

[3] Reprinted with permission from QualityMetric, www.qualitymetric.com

3 The following questions are about activities you might do during a typical day. Does your health now limit you in these activities? If so, how much?

		Yes, limited a lot	Yes, limited a little	No, not limited at all
a	Vigorous activities, such as running, lifting heavy objects, participating in strenuous sports	☐ 1	☐ 2	☐ 3
b	Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf	☐ 1	☐ 2	☐ 3
c	Lifting or carrying groceries	☐ 1	☐ 2	☐ 3
d	Climbing several flights of stairs	☐ 1	☐ 2	☐ 3
e	Climbing one flight of stairs	☐ 1	☐ 2	☐ 3
f	Bending, kneeling, or stooping	☐ 1	☐ 2	☐ 3
g	Walking more than a kilometre	☐ 1	☐ 2	☐ 3
h	Walking several hundred metres	☐ 1	☐ 2	☐ 3
i	Walking one hundred metres	☐ 1	☐ 2	☐ 3
j	Bathing or dressing yourself	☐ 1	☐ 2	☐ 3

4. During the past 4 weeks, how much of the time have you had any of the following problems with your work or other regular daily activities as a result of your physical health?

	All of the time	Most of the time	Some of the time	A little of the time	None of the time
a Cut down on the amount of time you spent on work or other activities	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
b Accomplished less than you would like	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
c Were limited in the kind of work or other activities	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
d Had difficulty performing the work or other activities (for example, it took extra effort)	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

5. During the past 4 weeks, how much of the time have you had any of the following problems with your work or other regular daily activities as a result of any emotional problems (such as feeling depressed or anxious)?

	All of the time	Most of the time	Some of the time	A little of the time	None of the time
a Cut down on the amount of time you spent on work or other activities	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
b Accomplished less than you would like	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
c Did work or other activities less carefully than usual	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

6. During the past 4 weeks, to what extent has your physical health or emotional problems interfered with your normal social activities with family, friends, neighbours, or groups?

Not at all	Slightly	Moderately	Quite a bit	Extremely
☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

7. How much bodily pain have you had during the past 4 weeks?

None	Very mild	Mild	Moderate	Severe	Very severe
☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 6

8. During the past 4 weeks, how much did pain interfere with your normal work (including both work outside the home and housework)?

Not at all	A little bit	Moderately	Quite a bit	Extremely
☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

9. These questions are about how you feel and how things have been with you during the past 4 weeks. For each question, please give the one answer that comes closest to the way you have been feeling. How much of the time during the past 4 weeks…

		All of the time	Most of the time	Some of the time	A little of the time	None of the time
		▼	▼	▼	▼	▼
a	Did you feel full of life?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
b	Have you been very nervous?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
c	Have you felt so down in the dumps that nothing could cheer you up?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
d	Have you felt calm and peaceful?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
e	Did you have a lot of energy?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
f	Have you felt downhearted and depressed?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
g	Did you feel worn out?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
h	Have you been happy?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
i	Did you feel tired?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

10. During the past 4 weeks, how much of the time has your physical health or emotional problems interfered with your social activities (like visiting with friends, relatives, etc.)?

All of the time	Most of the time	Some of the time	A little of the time	None of the time
▼	▼	▼	▼	▼
☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

11. How TRUE or FALSE is <u>each</u> of the following statements for you?

	Definitely true	Mostly true	Don't know	Mostly false	Definitely false
a I seem to get sick a little easier than other people	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
b I am as healthy as anybody I know	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
c I expect my health to get worse	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
d My health is excellent	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

Thank you for completing these questions!

(IQOLA SF-36v2 Standard, Australia (English))

Appendix D PFOOS Study – general and demographic questions

This survey asks for your view about your knee. This information will help us keep track of how you feel about your knee and how well you are able to do your usual activities. If you are unsure about how to answer a question, please give the best answer you can.

Q1 **Date of Birth:** **/ /** **(dd/mm/yyyy)**

Q2 **Gender:** **Male** **Female**

Q3 **Height:** **_____cm**

Q4 **Weight:** **_____ kg**

Q5 **Do you have knee pain in both knees?** (Please circle) **Yes** **No**

Q6 **In which knee do you experience the worse pain?** (Please circle) **Left** **Right**

Q7 **Is your knee pain aggravated by activities such as:** (Please circle)

Squatting	**Yes**	**No**
Walking up stairs	**Yes**	**No**
Walking down stairs	**Yes**	**No**
Rising from sitting	**Yes**	**No**
Running	**Yes**	**No**
Other _____	**Yes**	**No**

Q8 **How long have you experienced knee pain?** **_____months**

Q9 **How long have you experienced your current symptoms? _____months**

Q10 **Do you know what caused your knee pain?** (Please circle)

Yes	**No**
Traumatic onset (known injury/cause)	Insidious onset (no known cause)

Q11 **How would you rate your level of knee pain?** (Please place a tick in the box below) **(Pain severity where "0" = no pain and "10" = worst imaginable pain)**

0	1	2	3	4	5	6	7	8	9	10
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
No Pain										**Worst Imaginable Pain**

Q12 **Have you had knee surgery?** (Please circle) **Yes** **No**

Q13 **If Yes, what type of knee surgery have you had?_____**

Q14 **Overall, how would you rate your knee pain now?** (Please circle)

No Problem **Mild** **Moderate** **Somewhat Severe** **Severe**

Q15 **In the past week, how many hours of physical activity or sport (>30 mins) have you participated in? _____**

Q16 **Are you happy to be contacted in the future for further research?** (Please circle)

Yes **No**

Q17 **You will be asked to complete the three questionnaires in paper form and then also in a web-based form on a computer:**

Do you have easy access to a computer? (Please circle)

Yes **No**

Do you have a preference for a paper or web-based form? (Please circle)

Yes **No**

Appendix E DEMMI Study: The de Morton Mobility Index (DEMMI) [4]

	0	1	2
Bed			
1. Bridge	☐ unable	☐ able	
2. Roll onto side	☐ unable	☐ able	
3. Lying to sitting	☐ unable	☐ min assist ☐ supervision	☐ independent
Chair			
4. Sit unsupported in chair	☐ unable	☐ 10 sec	
5. Sit to stand from chair	☐ unable	☐ min assist ☐ supervision	☐ independent
6. Sit to stand without using arms	☐ unable	☐ able	
Static balance (no gait aid)			
7. Stand unsupported	☐ unable	☐ 10 sec	
8. Stand feet together	☐ unable	☐ 10 sec	
9. Stand on toes	☐ unable	☐ 10 sec	
10. Tandem stand with eyes closed	☐ unable	☐ 10 sec	
Walking			
11. Walking distance +/- gait aid *Gait aid (circle): nil/frame/stick/other*	☐ unable ☐ 5m	☐ 10m ☐ 20m	☐ 50m
12. Walking independence	☐ unable ☐ min assist ☐ supervision	☐ independent with gait aid	☐ independent without gait aid
Dynamic balance (no gait aid)			
13. Pick up pen from floor	☐ unable	☐ able	
14. Walks 4 steps backwards	☐ unable	☐ able	
15. Jump	☐ unable	☐ able	
COLUMN TOTAL SCORE:			

RAW SCORE TOTAL
(sum of column total scores) /19

DEMMI SCORE
(MDC_{90} = 9 points; MCID = 10 points) /100

easiest

sit unsupported

bridge

stand unsupported

sit to stand

roll

lie to sit

standing feet together

pick up pen from floor

walks backwards

walking distance

sit to stand no arms

walking independence

jump

stand on toes

tandem stand eyes closed

hardest

Raw-DEMMI Score Conversion Table

Raw Score	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
DEMMI score	0	8	15	20	24	27	30	33	36	39	41	44	48	53	57	62	67	74	85	100

Comments:
Signature:_____ Date:_____

[4] Reprinted from *Health and Quality of Life Outcomes, Volume 6,* de Morton NA et al., "*The de Morton Mobility Index (DEMMI): an essential health index for an ageing world"***,** 2008 with permission.

ITEM INSTRUCTIONS

Bed

Person is lying supine and is asked to bend their knees and lift their bottom clear of the bed.
Person is lying supine and is asked to roll onto one side without external assistance.
Person is lying supine and is asked to sit up over the edge of the bed.

Chair

Person is asked to maintain sitting balance for 10 seconds while seated on the chair, without
holding arm rests, slumping or swaying. Knees and feet are placed together and feet can be
resting on the floor.
Person is asked to rise from sitting to standing using the arm rests of the chair.
Person is asked to stand with their arms crossed over their chest.

Static Balance

The person is asked if they can stand for 10 seconds without
external support.
The person is asked if, for 10 seconds, they can stand with their feet together.
The person is asked if they can stand on their toes for 10 seconds.
The person is asked to place the heel of one foot directly in front of the other with their eyes
closed for 10 seconds.

Walking

Persons will be asked to walk with their current gait aid to where they can without a rest. Testing ceases if the person stops to rest. The person uses the gait aid that is currently most appropriate
for them. If either of two gait aids could be used, the aid that provides the person with the highest
level of independence should be used. Testing ceases once the person reaches 50 meters.
Independence is assessed over the person's maximum walking distance up to 50m (from item 11).

Dynamic Balance

A pen is placed 5 cm in front of the person's feet in standing. The person is asked if they can
pick the pen up off the floor.
Walks backwards 4 steps. Person remains steady throughout.
Person can jump. Both feet clear the ground. Person remains steady throughout.

Definitions

Minimal assistance = "hands on" physical but minimal assistance, primarily to guide movement. Supervision = another person monitors the activity without providing hands on assistance. May include verbal prompting.
Independent = the presence of another person is not considered necessary for safe mobility.

PROTOCOL FOR ADMINISTRATION OF THE DEMMI

Testing should be performed at the person's bedside.

1. Testing should be performed when the person has adequate medication eg. at least half an hour after pain or Parkinson's Disease medication.
2. The test should be administered in the sequence described in sections A-E: bed transfers, chair transfers, static balance, walking and dynamic balance.
3. Each item should be explained and, if necessary, demonstrated to the person.
4. Tick items to indicate item success or failure. Reasons for not testing items should be recorded.
5. Items should not be tested if either the test administrator or the person performing the test are reluctant to attempt the item.
6. Persons should be scored based on their first attempt.
7. If an item is not appropriate given a person's medical condition, the item should not be tested and the reason recorded.
8. Persons can be encouraged but feedback should not be provided regarding performance.
9. Three equipment items are required: chair with 45cm seat height with arm rests, a hospital bed or plinth and a pen.
10. The person administering the test manipulates person medical equipment during testing (eg. portable oxygen, drips, drains etc) unless the person requires minimal assistance to perform the test and then a 2nd person will be required to assist with medical equipment.
11. For persons that require a rest after each item due to shortness of breath, a 10 minute rest should be provided half way through testing i.e. after completing the chair transfers section.
12. For person's who have low level mobility and require a hoist to transfer in/out of bed or chair, the chair section can be administered before the bed section for these persons.
13. **Bed transfers:** the bed height should be appropriate for the individual person. A standardised hospital bed or plinth should be used for testing. The person cannot use an external device such as the monkey bar, bed rail, edge of bed or a bed pole. Additional pillows may be provided for persons who are unable to lie flat in supine.
14. **Chair transfers:** A standardised chair height of 45cm is required. Use a firm chair with arms.
15. **Balance:** Shoes cannot be worn for balance testing. The person cannot use external support to successfully complete any balance items. For sitting balance, neither the arm rests or the back of the chair can be used for external support. Standing balance tests should be performed with the person positioned between an elevated bed on one side and the test administrator on the other side. If a person displays unsteadiness or significant sway during testing, testing of that item should cease.
16. **Walking**: Appropriate shoes can be worn for walking tests. The same shoes must be worn for repeat testing.
17. **Scoring**: Using the conversion table provided, the raw score total must be converted to a DEMMI SCORE

The development of the DEMMI has been supported by a post graduate scholarship from the National Health and Medical Research Council of Australia (Dora Lush Postgraduate Scholarship, Grant no. 280632), funded by the HCF Health and Medical Research Foundation and also supported by The Northern Clinical Research Centre, Northern Health.

The DEMMI should be cited as: de Morton NA, Davidson M, Keating JL. **The de Morton Mobility Index (DEMMI): an essential health index for an ageing world.** *Health and Qualitv of Life Outcomes*