
Article:
Smith, RD, Dziedzic, KS, Quicke, JG et al. (3 more authors) (2019) Identification and evaluation of self-reported physical activity instruments in adults with osteoarthritis: A systematic review. Arthritis Care and Research, 71 (2). pp. 237-251. ISSN 2151-464X

https://doi.org/10.1002/acr.23787

Title: Identification and evaluation of self-reported physical activity instruments in adults with osteoarthritis: A systematic review.

Running head: Measuring physical activity in osteoarthritis

Authors: Robert D. Smith, PhD[1];

Krysia S. Dziedzic, PhD[2];

Jonathan G. Quicke, PhD[2];

Melanie A. Holden, PhD[2];

Gretl A. McHugh, PhD, MSc Public Health, B.Nurs (Hons) [3];

Emma L. Healey, PhD[2].

Affiliations:

[1] School of Nursing, The University of Hong Kong, Hong Kong, China. [2]Research Institute for Primary Care and Health Sciences, Keele University, Keele, Staffordshire, United Kingdom, ST5 5BG. [3] School of Healthcare, University of Leeds, Leeds, LS2 9JT.

Funding: This work was funded by a Keele University ACORN studentship and supported by a National Institute for Health Research (NIHR) Programme Grant (RP-PG-0407-10386). This research was also funded by the Arthritis Research UK Centre in Primary Care grant (Grant Number 18139).

Corresponding Author: Robert Smith

Email: robsmith@hku.hk

Fax: +852 2872 6079

Abstract word count: 250

Manuscript word Count: 3776/3800

Tables/Figures: 6/6

Abstract

Objective: To identify and evaluate the measurement properties of self-report physical activity (PA) instruments suitable for those with osteoarthritis (OA).

Methods: A comprehensive two-stage systematic review using multiple electronic databases from inception until July 2018. Stage One sought to identify all self-reported PA instruments used in populations with joint pain attributable to OA in the foot, knee, hip or hand. Stage Two searched for and appraised studies investigating the measurement properties of the instruments identified. For both stages all articles were screened for study eligibility criteria, completed data extraction using the Qualitative Attributes and Measurement Properties of Physical Activity Questionnaires (QAPAQ) checklist, and conducted methodology quality assessments using a modified COnsensus-based Standards for the selection of health Measurement Instruments (COSMIN) checklist. Measurement properties for each physical activity instrument were evaluated and combined using narrative synthesis.

Results: Stage One identified 23 unique self-report PA instruments. Stage Two identified 53 studies that evaluated the measurement properties of 13 of the 23 instruments identified. Instrument reliability varied from inadequate to adequate (ICC=≥0.7). Instrument construct and criterion validity assessment demonstrated small to moderate correlations with direct measures of PA. Responsiveness was assessed in only 1 instrument and was unable to detect changes in comparison to accelerometers.

Conclusion: While many instruments were identified as potentially suitable for use in individuals with OA, none demonstrated adequate measurement properties across all domains of reliability, validity and responsiveness. Further high-quality assessment of

self-reported PA instruments is required before such measures can be recommended for use in OA research.

Significance and innovation:

- Physical activity (PA) is a recommended core treatment for osteoarthritis (OA) and is a commonly used outcome in clinical trials, therefore accurately measuring current PA levels and changes in PA in individuals with OA is vital.
- This systematic review updates and builds on a previous systematic review examining the measurement properties of PA instruments suitable for adults with OA, collecting evidence from 53 studies.
- This study highlights the need for high-quality assessment (following COSMIN guidelines) across all measurement properties of self-reported PA instruments before such measures can be recommended for use in OA research.

Introduction

Osteoarthritis (OA) is a clinical syndrome of joint pain with varying degrees of limitation in physical function and reduced quality of life and most commonly affects the knee, hip, hand and foot (1). Physical activity (PA), such as therapeutic strengthening exercises or aerobic exercise, can reduce joint pain symptoms and improve physical function. PA is recommended as a core treatment for people with OA in the foot, knee, hip or hand (2, 3). However, pain is an important predictor of physical inactivity (4) and less than half of adults with OA are meeting the current guideline of 150 minutes of moderate intensity PA per week (5, 6). Accurately measuring current PA levels and changes of PA in individuals with OA is important in research.

PA can be measured using direct methods such as accelerometry or indirect methods such as self-reported PA instruments (7). Self-reported PA instruments are a popular approach for measuring levels of PA in larger population studies (8). This is due to their ease of use, their ability to allow immediate access to information about an individual's PA, and the low cost involved in their administration to a large number of study participants (9). To accurately measure PA using self-report instruments, the appropriate instrument must be selected according to the demographics of the participants (10). An example are instruments developed specifically to measure PA for adults age 65 years and over (11).

Multi domain instruments such as the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) and the Knee injury and Osteoarthritis Outcome Score (KOOS), have been designed specifically for use in populations with OA. While these multi domain instruments do measure PA as a component or sub-scale score, they have been excluded from this review as their purpose is not to assess PA levels

explicitly in terms of frequency, duration and intensity, which are required to make comparisons to current PA guidelines.

To date there is still no consensus on which self-reported PA instrument is the most suitable for OA research. In 2011, Terwee et al evaluated the measurement properties of PA instruments in OA populations but focused solely on those with a diagnosis of knee or hip OA (12). This previous systematic review identified 9 studies, however none of these included the Physical Activity Scale for the Elderly (PASE) (13), an instrument that has more recently been used in OA research (14-16). Other systematic reviews that have evaluated the measurement properties of PA instruments for adults in non-joint pain populations restricted to adults aged between 18-65 years or adults aged 65 years or over (7, 8, 11). Therefore, there is a gap in the literature for a comprehensive, broader and updated systematic review that captures relevant information regarding the measurement of PA in those with OA, a group that are most commonly aged 45 years and over. Rather than just focusing on those with a diagnosis of OA, by including studies that have evaluated the measurement properties of relevant instruments in other populations (i.e. 1. those with joint pain attributable to OA in the foot, knee, hip or hand and 2. community dwelling adults in the same age bracket as those with OA), it will be possible to identify and evaluate the measurement properties of a range of instruments suitable for those with OA. To our knowledge, no instrument measuring PA levels has been specifically developed for populations with OA. Instruments developed for other populations, such as general adult or elderly adult populations, have been used in OA research. It is, therefore, important to understand how well these instruments reflect the construct of PA levels in OA populations by assessing the instruments' measurement properties as defined in the COSMIN taxonomy (17).

A two-stage systematic review was conducted and aimed to identify and evaluate the measurement properties of self-report physical activity (PA) instruments suitable for those with OA.

Patients and Methods

Stage One identified all self-report PA instruments used in published research involving populations aged 45 years and over with joint pain attributable to OA in the feet, knee, hips or hands. The age range and joint sites were selected following the National Institute for Health and Care Excellence guideline on the management of OA and the most commons peripheral joints affected by OA (1). Stage Two subsequently identified all the published evidence on the measurement properties of the instruments identified in Stage One. Both stages of the systematic review involved electronic database searching of MEDLINE, EMBASE and Web of Science from inception until 19th July 2018 combined with hand searching of reference lists from included articles. The primary reviewer (RS) screened all titles and the abstracts, full articles were independently double reviewed by the primary reviewer and at least one of the secondary reviewing team (MH, JQ, EH, GM, KD), with any disagreements resolved via consensus discussion between reviewers. Titles and abstracts were reviewed by the primary reviewer only due to time limitations of the secondary reviewers, to minimise risk of reviewer error, 10% of all titles and abstracts were independently reviewed with at least one of the secondary review team.

Stage One

Selection Criteria

The selection criteria for Stage One were quantitative research studies that focused on populations with joint pain attributable to OA in the foot, knee, hip or hand and measured self-reported PA (Table 1). Populations were included if other sites of pain were present alongside pain in the foot, knee, hip or hand. Due to cases where study sample include both OA and inflammatory arthritis populations, we only include those with more than 50% of the sample having OA or joint pain attributable to OA. Search terms for articles in Stage One were synthesised from previous joint pain and PA systematic review search strategies (18, 19). The full search strategy for Stage One is shown in appendix 1.

add Table 1 here

Data extraction

Data extraction for Stage One involved extracting the citation of the included studies and identifying the self-reported PA instrument used. Data extraction was conducted by two different reviewers independently (the primary reviewer and one of the secondary reviewers). As the aim of Stage One was simply to identify studies and instruments no further data extraction or quality assessment was conducted.

Stage Two

Selection Criteria

The selection criteria for Stage Two were studies that performed an evaluation of the at least one measurement property of the instruments identified in Stage One in

populations with joint pain attributable to OA, or community dwelling adults of a similar age (aged 45 years and over). For purposes of describing all instruments included in Stage Two, articles that described the instruments attributes (the settings, recall period, purpose) were also retrieved. The search strategy for Stage Two was constructed using a high sensitivity search term filter for identifying articles on measurement tool properties (20). This filter was combined with the name of the instrument identified in Stage One of this review. The full search strategy for Stage Two is shown in appendix 2.

Data extraction and quality assessment

In Stage Two, the Quality Assessment of Physical Activity Questionnaires checklist (QAPAQ) was used to extract data and conduct a preliminary quality assessment (21). The QAPAQ is a comprehensive checklist of all the measurement properties and qualitative attributes of self-report PA instruments and has been used in previous systematic reviews evaluating measurement properties of self-report PA (7, 11, 12). A comprehensive quality assessment of the articles identified in Stage Two was conducted using the COnsensus-based Standards for the selection of health Measurement Instruments (COSMIN) checklist (22). The COSMIN checklist has been used in previous systematic reviews that have assessed the quality of other self-reported instruments (23-26). To reduce reviewer burden within this systematic review, the COSMIN was modified by removing items on generalisability and interpretability already covered in the QAPAQ (21).

Following quality assessment, a previously used grading system was conducted to assign a quantitative score to the evidence of each instrument's measurement properties and the quality of that evidence (23-25). The grading system combined the

strength of evidence (using the COSMIN checklist) (Appendix 3) to a criteria for each measurement property (10) (Appendix 4), which was extracted using the QAPAQ (21). For the purposes of this systematic review construct validity was defined in terms convergent construct validity in which the self-reported instrument reflects PA measured objectively, such as accelerometers or heart rate monitoring. In criterion validity the gold standard measurement for PA in the review was considered as double-labelled water (DLW). Measurement error was not formally assessed as a COSMIN criterion as we could not identify a minimal important change reported for any of the instruments, measurement error has been reported when evaluated by studies.

Results

Stage One

From the search of the electronic databases and hand searching of reference lists of included studies, 20,292 articles were identified which reduced to 20,116 following removal of duplicates. Ninety-one studies comprising 23 unique self-reported PA instruments met the inclusion criteria and were included in the review. This is indicated by a PRISMA flowchart (Figure 1). Included studies focused on knee OA (n=52), knee and/or on hip OA (n=22), hip OA (n=8), general joint pain or multiple sites of OA (n=4) foot pain or foot OA (n=3) and knee pain (n=2) populations. Thirty-two of the studies were longitudinal cohort studies, 29 were randomized controlled trials, 18 were cross-sectional studies, 9 studies examined the measurement properties of instruments and 3 were systematic reviews. Seventeen studies were conducted in the United States (USA), 13 in Australia and the United Kingdom (UK), 12 in the Netherlands, 5 in

Canada and Germany, 4 in Switzerland and Denmark, 3 in Sweden, Brazil and Portugal, and Norway each and 1 in Greece, Spain, Japan, and Iran, two studies were multi-country studies across Europe.

add Figure 1 here

PA instruments identified

The self-reported instruments of PA (n=23) used in the included studies identified in Stage One are listed in appendix 3. The most common PA instruments used were the Physical Activity Scale for the Elderly (PASE) (used in 34 studies), and the International Physical Activity Questionnaire- Short Form (IPAQ-SF) (used in 17 studies). Nineteen of the instruments identified were multi-item self-reported PA questionnaires and 5 were single item PA instruments.

Stage Two

Within Stage Two of the systematic review, 3,661 articles were identified, with 54 meeting the inclusion criteria (Figure 2). Of those, nine (16%) evaluated the measurement properties of one or more of the identified PA instruments in adults with joint pain attributable to OA (knee =3; hip =3; combined hip and knee =3). Forty-five articles (84%) evaluated the measurement properties of the PA instruments in community dwelling adult populations aged 45 and over (adults aged 65 years and over = 20; aged 45-64 years = 25). The majority of studies were conducted in Australia (n=9), USA (n=8), the Netherlands (n=5), Japan (n=4) and China (n=4). Thirty-five studies evaluated construct validity, 36 evaluated reliability or measurement error, two studies examined content validity, two examined criterion validity, two evaluated

internal consistency and one evaluated responsiveness. A summary of the characteristics of the articles included in Stage Two have been included (Appendix 6).

add Figure 2 here

Of the 23 instruments identified in Stage One, 13 (56.5%) had a least one measurement property evaluated in either a population with joint pain attributable to OA or a community dwelling adult population aged 45 years and over. Table 2 describes the characteristics of these instruments.

add Table 2 here

Measurement properties of the PA instruments in populations with joint pain attributable to OA

There were no instruments identified in Stage One and evaluated in Stage Two which demonstrated full adequacy across all measurement property domains in populations with joint pain attributable to OA (Table 3). Criterion validity, internal consistency, content validity, structural validity and responsiveness were not assessed in any of the instruments. There was no evidence of any measurement properties for the Active Australia Survey (AAS), modified Baecke, Incidental And Planned Activity Questionnaire For Older People (IPEQ), Short Questionnaire To Assess Health Enhancing Physical Activity (SQUASH), Short Telephone Activity Recall Questionnaire (STAR) or Zutphen Physical Activity Questionnaire in populations with joint pain attributable to OA.

In terms of reliability, the only multi-item instruments with correlations or ICC above 0.7 in studies deemed to be of good-to-excellence methodological quality were the Beacke, Human Activity Profile (HAP), IPAQ-SF and PASE in populations with joint

pain attributable to OA (27-30). While the quality evidence rated as fair, all the single scale instruments (Activity Rating Scale (ARS), Tegner scale and University Of California, Los Angeles Activity (UCLAA) scale) demonstrated correlations above 0.7 in populations with joint pain attributable to OA for reliability (29). The measurement error of HAP, IPAQ-SF and PASE has been evaluated, while there is no minimally important change index to assess the adequacy of measurement error in these instruments. The proportion of error in IPAQ-SF and PASE were large compared to their maximal possible scoring range, while the HAP was small. Suggesting large measurement error in populations with joint pain attributable to OA in the IPAQ-SF and PASE (28, 30-33) (Table 3).

For construct validity in populations with joint pain attributable to OA, the Baecke, IPAQ-SF and PASE demonstrated only low to moderate correlations (0.06-0.49) with accelerometers (30-33) (Table 3).

Measurement properties of the PA instruments in community dwelling adult aged 45 and over

There were no instruments identified in Stage One and evaluated in Stage Two which demonstrated full adequacy across all measurement property domains in community dwelling adult aged 45 and over (Table 3). Structural validity was not assessed in any of the instruments (Table 4.).

In terms of reliability, the AAS displayed adequate reliability in one study (34) but inadequate reliability in two studies (35, 36). The modified Baecke demonstrated reliability in three studies above and below adequate reliability(37-39). The HAP, IPEQ and STAR demonstrated adequate reliability in three studies (40-42). The IPAQ-SF in 7 studies (43-50), and the PASE in 8 studies both demonstrated

reliability above and below adequate reliability (13, 51-56). Measurement error had been assessed in the PASE in one study; finding a relatively small standard error measurement (SEM) (3.3-8.5) to the maximal scoring range of the PASE (0-400) (56).

The PASE and modified Baecke were the only instruments to have criterion validity evaluated and this was in community dwelling older adults aged 45 and over. Both demonstrated a moderate correlation to DLW, in another study the PASE also demonstrated a non-significant correlation to DLW (51, 57, 58).

For construct validity, the AAS correlation with accelerometers was assessed in 5 studies and ranged from 0.39-0.61, all demonstrating some moderate correlations (34, 36, 43, 59, 60). The Modified Baecke demonstrated non-significance in a correlation with heart rate monitoring (37). The HAP showed moderate correlations to accelerometers in a single study(40). IPEQ showed a low correlation to accelerometers in a single study (61). The IPAQ-SF was evaluated for construct validity in 9 studies, correlations to accelerometers ranged from non-significant to moderate correlations(44, 46-49, 62-65). The PASE was evaluated for construct validity in 5 studies, correlations to accelerometers ranged from low to moderate correlations (51-53, 66, 67). The SQUASH demonstrated high agreement with heart monitoring in a single study (68). The STAR demonstrated low correlations with accelerometers in a single study (42). The Zutphen demonstrated moderate correlations with accelerometers (69).

The IPAQ-SF and PASE were evaluated for internal consistency, each in a single study. In both the IPAQ-SF and PASE internal consistency was deemed adequate. The AAS and IPAQ-SF were assessed for their content validity by cognitive

interviews about the understanding of the items in the instrument (50, 56). In both the AAS and IPAQ-SF terminology used in items were confusing or unclear to participants, making recall difficult (70, 71). Responsiveness was evaluated in the IPEQ and was evaluated to be less responsive to changes in PA levels compared to accelerometer (61).

add Table 3 here

add Table 4 here

Methodological quality of the included studies

For reliability, eight studies were evaluated as poor quality as a small sample size was used (n=<50) (29, 31-33, 38, 39, 46, 72), sample sizes below 50 are considered too small for evaluating measurement properties (10). Five studies that assessed reliability were evaluated as fair quality as their sample size was above 50, but they used a correlation rather than test for agreement (intra-class correlation) (35-37, 43, 54). Fourteen studies were evaluated as good quality with sample sizes larger than 50 but smaller than 100 (10), and seven studies were evaluated as excellent quality with sample sizes greater than 100 (10). One good quality study evaluated measurement error in a sample size <100 (56).

The two studies that evaluated criterion validity were evaluated as poor quality due to their sample size (57, 58). Of the studies evaluating construct validity: seven were evaluated as poor quality due to sample size (31-33, 38, 39, 46, 60, 67); three were evaluated as fair quality (45, 59, 68), as while the sample size was deemed appropriate, these studies used pedometers or heart monitors rather than accelerometers; twelve studies were evaluated as good quality with sample sizes

larger than 50 but small than 100 (27, 30, 34, 36, 42, 48, 49, 51, 52, 66, 69, 73); and 10 studies were evaluated as excellent quality with sample sizes greater than 100 (35, 40, 43, 47, 53, 61-64, 74), only one of the studies in this review used hypothesis testing to evaluate construct validity(49). Responsiveness was assessed in one study, which was evaluated as excellent quality due to a large sample size above 100 participants and a comparison with an accelerometer. Two studies of excellent quality assessed content validity using cognitive interviews (70, 71).

Discussion

Stage One of this systematic review identified 23 self-reported PA instruments that have been used previously in populations with joint pain attributable to OA. However, based on the findings from Stage Two of this systematic review, it is still not clear which instrument is most appropriate for use in those with OA. This is due to the lack of evidence of adequate measurement properties for all the instruments identified.

Reliability and internal consistency

In both populations, most self-report instruments demonstrated adequate test-retest reliability. Although methodological quality ranged from poor to excellent. This suggests that these self-report instruments are reliable in measuring levels of PA in test re-test evaluations. Two studies evaluated internal consistency, one in the IPAQ-SF and one in the PASE, both were of good methodological quality and indicated adequate consistency of all the items (Cronbach's alpha=≥0.70).

Criterion validity and construct validity

None of the instruments demonstrated strong correlations (above 0.70) with direct measures of PA, such as accelerometers or heart monitors, in those with joint pain attributable to OA or community dwelling older adults aged 45 years and over. Two studies evaluated criterion validity using the gold standard measurement of PA (DLW) (57, 58), but these only demonstrated low or not statistically significant correlations and were based on small samples below 50 participants. The implication of low to moderate criterion and construct validity of these instruments is that researchers cannot be certain the degree to which instruments reflect actual PA levels, particularly as there were no clear pattern in the self-report instruments regarding over-or-underestimating PA level compared to direct measures (75).

Content validity

Notably, only two studies evaluated content validity. Both were conducted on community dwelling adult populations aged 45 years and over and examined AAS and IPAQ-SF [15, 34]. These studies highlighted participant misinterpretation of both PA definitions and the questions used within these instruments. Gaining a clearer understanding of the difficulties demonstrated with interpreting definitions of PA and the questions contained within self-report PA instruments more generally would be useful.

Responsiveness

None of the studies examined the responsiveness of the instruments in those with joint pain attributable to OA, and only one study evaluated responsiveness (using the IPEQ) in community dwelling older adults aged 45 years and over. It is therefore unclear how sensitive the self-report PA instruments identified are to detecting changes in PA levels in populations with joint pain attributable to OA. This is a major limitation when

evaluating PA interventions aimed at increasing PA levels in these populations (76). None of the studies identified in this review evaluated formally addressed structural validity or cross-cultural validity in any of the instruments in any of our populations of interest.

The studies that evaluated measurement properties in populations with joint pain attributable to OA identified in this review were limited to only those in the knee and hip. None of the studies in Stage Two included those with joint pain in the foot or hand attributable to OA. This lack of evidence also limits comparisons of the measurement properties between different joints of pain attributable to OA.

Strengths and limitations

This systematic review used a comprehensive search strategy including multiple electronic databases, reference list screening from included studies. It is also original in its inclusion of studies of populations with joint pain attributable to OA and community dwelling older adults aged 45 and over. This study has used the gold standard tool for assessing study quality in outcome measures (22), as well as a previously published standardized form for extracting data on measurement properties of PA instruments (21).

Despite identifying many studies in Stage Two (n=54), it is difficult to determine to what degree the findings in community dwelling adults aged 45 years and over are generalisable to similar aged adults with OA or joint pain attributable to OA. The review focused on the most common sites of OA for the review in adults aged 45 and over, where the prevalence of OA is most common (1), the findings of this review may not be generalisable to younger people with post-traumatic OA.

Conclusion

This systematic review has demonstrated that there is limited evidence for the measurement properties of previously used self-report PA instruments in populations with joint pain attributable to OA. Further high methodological quality evaluation of additional measurement properties is required for commonly used instruments for this population. It is particularly recommended that such studies use larger sample sizes of at least 50, or ideally larger than 100 participants (10). Such studies will allow researchers to make appropriately informed decisions when selecting self-reported PA instruments in OA research. While the evidence that was identified demonstrated adequate test re-test reliability in a couple of instruments, overall the evidence on validity and responsiveness was lacking. Investigations into content validity may particularly help researchers to identify areas within self-reported PA instruments that may cause participants to misinterpret the questions and therefore report PA inaccurately. Evaluation of the responsiveness of PA instruments commonly used in randomized controlled trials focused on OA is highly recommended (76), especially if PA is the primary outcome. Future studies should also consider building the evidence base focused on reliability of PA instruments by examining correlations with direct measures of PA in OA populations.

Acknowledgements

We would like to thank Jo Jordan for her protocol feedback and systematic review methodological advice.

Author contributions

RS was the overall lead for the work for the systematic review and was involved at all stages of the paper. RS, ELH, GAM and KSD conceived and designed the review. All authors were involved in study searching, quality assessment and data extraction checking and editing drafts of the paper. All authors have approved the final version.

Role of the funding source

The lead author (RS) was funded by a Keele University ACORN studentship. The funder had no involvement in in the design, collection, analysis, writing or publishing of this paper. This work was also supported by a National Institute for Health Research (NIHR) Programme Grant (RP-PG-0407-10386). This research was also funded by the Arthritis Research UK Centre in Primary Care grant (Grant Number 18139). KD and ELH are part-funded by the National Institute for Health Research (NIHR) Collaborations for Leadership in Applied Health Research and Care West Midlands. KD is also part-funded by a Knowledge Mobilisation Research Fellowship (KMRF-2014-03-002) from the NIHR. MAH is supported by the National Institute for Health Research (NIHR) School for Primary Care Research. JQ is funded by a National Institute for Health Research (NIHR) Academic Clinical Lectureship in Physiotherapy, awarded as part of Professor Christian Mallen's NIHR Research Professorship (NIHR-RP-2014-026). The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the NIHR, HEE or the Department of Health and Social Care.

Conflict of interest

There is no conflict of interest for any of the authors.

References

1. National Institue for Health and Care Excellence. Osteoarthritis: care and management. Clinical guideline [CG177]. 2014.

2. McAlindon TE, Bannuru RR, Sullivan M, Arden N, Berenbaum F, Bierma-Zeinstra S, et al. OARSI guidelines for the non-surgical management of knee osteoarthritis. Osteoarthritis and Cartilage. 2014;22(3):363-88.

3. Fernandes L, Hagen KB, Bijlsma JW, Andreassen O, Christensen P, Conaghan PG, et al. EULAR recommendations for the non-pharmacological core management of hip and knee osteoarthritis. Annals of the Rheumatic Diseases. 2013;72(7):1125-35.

4. Plooij B, Scherder EJ, Eggermont LH. Physical inactivity in aging and dementia: a review of its relationship to pain. Journal of Clinical Nursing. 2012;21(21-22):3002-8.

5. Herbolsheimer F, Schaap LA, Edwards MH, Maggi S, Otero Á, Timmermans EJ, et al. Physical Activity Patterns Among Older Adults With and Without Knee Osteoarthritis in Six European Countries. Arthritis Care & Research. 2016;68(2):228-36.

6. Holden MA, Nicholls EE, Young J, Hay EM, Foster NE. Exercise and physical activity in older adults with knee pain: a mixed methods study. Rheumatology. 2015;54(3):413-23.

7. Van Poppel MN, Chinapaw MJ, Mokkink LB, Van Mechelen W, Terwee CB. Physical activity questionnaires for adults. Sports Medicine. 2010;40(7):565-600.

8. Helmerhorst HHJ, Brage S, Warren J, Besson H, Ekelund U. A systematic

review of reliability and objective criterion-related validity of physical activity questionnaires. International Journal of Behavioral Nutrition and Physical Activity. 2012;9(1):1.

9. Warren JM, Ekelund U, Besson H, Mezzani A, Geladas N, Vanhees L. Assessment of physical activity–a review of methodologies with reference to epidemiological research: a report of the exercise physiology section of the European Association of Cardiovascular Prevention and Rehabilitation. European Journal of Cardiovascular Prevention & Rehabilitation. 2010;17(2):127-39.

10. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. Journal of Clinical Epidemiology. 2007;60(1):34-42.

11. Forsén L, Loland NW, Vuillemin A, Chinapaw MJ, van Poppel MN, Mokkink LB, et al. Self-administered physical activity questionnaires for the elderly. Sports Medicine. 2010;40(7):601-23.

12. Terwee C, Bouwmeester W, van Elsland S, de Vet H, Dekker J. Instruments to assess physical activity in patients with osteoarthritis of the hip or knee: a systematic review of measurement properties. Osteoarthritis and Cartilage. 2011;19(6):620-33.

13. Washburn RA, McAuley E, Katula J, Mihalko SL, Boileau RA. The physical activity scale for the elderly (PASE): evidence for validity. Journal of Clinical Epidemiology. 1999;52(7):643-51.

14. Dunlop DD, Song J, Semanik PA, Sharma L, Chang RW. Physical activity levels and functional performance in the osteoarthritis initiative: a graded

relationship. Arthritis & Rheumatology. 2011;63(1):127-36.

15. Foster NE, Healey EL, Holden MA, Nicholls E, Whitehurst DG, Jowett S, et al. A multicentre, pragmatic, parallel group, randomised controlled trial to compare the clinical and cost-effectiveness of three physiotherapy-led exercise interventions for knee osteoarthritis in older adults: the BEEP trial protocol (ISRCTN: 93634563). BMC Musculoskeletal Disorders. 2014;15(1):254.

16. Fransen M, Su S, Harmer A, Blyth FM, Naganathan V, Sambrook P, et al. A longitudinal study of knee pain in older men: Concord Health and Ageing in Men Project. Age and Ageing. 2013;43(2):206-12.

17. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. Journal of Clinical Epidemiology. 2010;63(7):737-45.

18. Fransen M, McConnell S, Harmer AR, Van der Esch M, Simic M, Bennell KL. Exercise for osteoarthritis of the knee. Cochrane Database of Systematic Reviews. 2015(1).

19. Baker PRA, Francis DP, Soares J, Weightman AL, Foster C. Community wide interventions for increasing physical activity. Cochrane Database of Systematic Reviews. 2015(1).

20. Terwee CB, Jansma EP, Riphagen II, de Vet HC. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. Quality of Life Research. 2009;18(8):1115-23.

21. Terwee CB, Mokkink LB, van Poppel MN, Chinapaw MJ, van Mechelen W, de Vet HC. Qualitative attributes and measurement properties of physical activity questionnaires. Sports Medicine. 2010;40(7):525-37.

22. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Quality of Life Research. 2010;19(4):539-49.

23. Schellingerhout JM, Verhagen AP, Heymans MW, Koes BW, Henrica C, Terwee CB. Measurement properties of disease-specific questionnaires in patients with neck pain: a systematic review. Quality of Life Research. 2012;21(4):659-70.

24. Saether R, Helbostad JL, Riphagen II, Vik T. Clinical tools to assess balance in children and adults with cerebral palsy: a systematic review. Developmental Medicine & Child Neurology. 2013;55(11):988-99.

25. Weldam SW, Schuurmans MJ, Liu R, Lammers J-WJ. Evaluation of Quality of Life instruments for use in COPD care and research: a systematic review. International Journal of Nursing Studies. 2013;50(5):688-707.

26. Gilchrist F, Rodd H, Deery C, Marshman Z. Assessment of the quality of measures of child oral health-related quality of life. BMC Oral Health. 2014;14(1):1.

27. Merom D, Delbaere K, Cumming R, Voukelatos A, Rissel C, Van der Ploeg HP, et al. Incidental and planned exercise questionnaire for seniors: validity and responsiveness. Medicine & Science in Sports & Exercise. 2014;46(5):947-54.

28. Ono R, Hirata S, Yamada M, Nishiyama T, Kurosaka M, Tamura Y. Reliability

and validity of the Baecke physical activity questionnaire in adult women with hip disorders. BMC musculoskeletal disorders. 2007;8(1):1.

29. Bennell KL, Hinman RS, Crossley KM, Metcalf BR. Is the Human Activity Profile a useful measure in people with knee osteoarthritis? Journal of Rehabilitation Research and Development. 2004;41(4):621.

30. Naal FD, Impellizzeri FM, Leunig M. Which is the best activity rating scale for patients undergoing total joint arthroplasty? Clinical Orthopaedics and Related Research. 2009;467(4):958-65.

31. Casartelli NC, Bolszak S, Impellizzeri FM, Maffiuletti NA. Reproducibility and Validity of the Physical Activity Scale for the Elderly (PASE) Questionnaire in Patients After Total Hip Arthroplasty. Physical Therapy. 2015;95(1):86-94.

32. Blikman T, Stevens M, Bulstra SK, van den Akker-Scheek I, Reininga IH. Reliability and validity of the Dutch version of the International Physical Activity Questionnaire in patients after total hip arthroplasty or total knee arthroplasty. Journal of Orthopaedic & Sports Physical Therapy. 2013;43(9):650-9.

33. Bolszak S, Casartelli NC, Impellizzeri FM, Maffiuletti NA. Validity and reproducibility of the Physical Activity Scale for the Elderly (PASE) questionnaire for the measurement of the physical activity level in patients after total knee arthroplasty. BMC Musculoskeletal Disorders. 2014;15(1):1.

34. Svege I, Kolle E, Risberg MA. Reliability and validity of the Physical Activity Scale for the Elderly (PASE) in patients with hip osteoarthritis. BMC Musculoskeletal Disorders. 2012;13(1):1.

35. Hertogh EM, Monninkhof EM, Schouten EG, Peeters PH, Schuit AJ. Validity of the Modified Baecke Questionnaire: comparison with energy expenditure according to the doubly labeled water method. International Journal of Behavioral Nutrition and Physical Activity. 2008;5(1):1.

36. Schuit AJ, Schouten EG, Westerterp KR, Saris WH. Validity of the Physical Activity Scale for the Elderly (PASE): according to energy expenditure assessed by the doubly labeled water method. Journal of Clinical Epidemiology. 1997;50(5):541-6.

37. Colbert LH, Matthews CE, Havighurst TC, Kim K, Schoeller DA. Comparative validity of physical activity measures in older adults. Medicine and Science in Sports and Exercise. 2011;43(5):867.

38. Fjeldsoe BS, Winkler EA, Marshall AL, Eakin EG, Reeves MM. Active adults recall their physical activity differently to less active adults: test–retest reliability and validity of a physical activity survey. Health Promotion Journal of Australia. 2013;24(1):26-31.

39. Brown WJ, Burton NW, Marshall AL, Miller YD. Reliability and validity of a modified self-administered version of the Active Australia physical activity survey in a sample of mid-age women. Australian and New Zealand Journal of Public Health. 2008;32(6):535-41.

40. Pettee GK, McClain JJ, Lee CD, Swan PD, Alvar BA, Mitros MR, et al. Evaluation of physical activity measures used in middle-aged women. Medicine and Science in Sports and Exercise. 2009;41(7):1403-12.

41. Pols MA, Peeters P, Kemper HC, Collette H. Repeatability and relative validity

of two physical activity questionnaires in elderly women. Medicine and Science in Sports and Exercise. 1996;28(8):1020-5.

42. Pols MA, Peeters PH, Bueno-de-Mesquita HB, Ocke MC, Wentink CA, Kemper HC, et al. Validity and repeatability of a modified Baecke questionnaire on physical activity. International Journal of Epidemiology. 1995;24(2):381-8.

43. Voorrips LE, Ravelli AC, Petra C, Dongelmans A, Deurenberg P, van Staveren WA. A physical activity questionnaire for the elderly. Diet and Physical Activity as Determinants of Nutritional Status in Elderly Women. 1991:43.

44. de Carvalho Bastone A, de Souza Moreira B, Alvarenga Vieira R, Noce Kirkwood R, Domingues Dias JM, Corrêa Dias R. Validation of the human activity profile questionnaire as a measure of physical activity levels in older community-dwelling women. Journal of Aging & Physical Activity. 2014;22(3).

45. Delbaere K, Hauer K, Lord SR. Evaluation of the incidental and planned activity questionnaire for older people. British Journal of Sports Medicine. 2010;44(14):1029-34.

46. Matthews CE, Ainsworth BE, Hanby C, Pate RR, Addy C, Freedson PS, et al. Development and testing of a short physical activity recall questionnaire. Medicine and Science in Sports and Exercise. 2005;37(6):986-94.

47. Brown W, Trost S, Bauman A, Mummery K, Owen N. Test-retest reliability of four physical activity measures used in population surveys. Journal of Science and Medicine in Sport. 2004;7(2):205-15.

48. Booth ML, Ainsworth BE, Pratt M, Ekelund U, Yngve A, Sallis JF, et al.

International physical activity questionnaire: 12-country reliability and validity. Medicine and Science in Sports and Exercise. 2003;195(9131/03):3508-1381.

49. Deng H, Macfarlane D, Thomas G, Lao X, Jiang C, Cheng K, et al. Reliability and validity of the IPAQ-Chinese: the Guangzhou Biobank Cohort study. Medicine and Science in Sports and Exercise. 2008;40(2):303.

50. Mäder U, Martin BW, Schutz Y, Marti B. Validity of four short physical activity questionnaires in middle-aged persons. Medicine and Science in Sports and Exercise. 2006;38(7):1255-66.

51. Tomioka K, Iwamoto J, Saeki K, Okamoto N. Reliability and validity of the International Physical Activity Questionnaire (IPAQ) in elderly adults: the Fujiwara-kyo Study. Journal of Epidemiology. 2011;21(6):459-65.

52. Cerin E, Barnett A, Cheung M-c, Sit CH, Macfarlane DJ, Chan W-m. Reliability and validity of the IPAQ-L in a sample of Hong Kong urban older adults: does neighborhood of residence matter? Journal of Aging and Physical Activity. 2012;20(4):402-20.

53. Ryan DJ, Wullems JA, Stebbings GK, Morse CI, Stewart CE, Onambele-Pearson GL. Reliability and validity of the international physical activity questionnaire compared to calibrated accelerometer cut-off points in the quantification of sedentary behaviour and physical activity in older adults. PloS one. 2018;13(4):e0195712.

54. Ayvat E, Kilinc M, Kirdi N. The Turkish version of the physical activity scale for the elderly (PASE): Its cultural adaptation, validation, and reliability. Turkish Journal of Medical Sciences. 2017;47(3):908-15.

55. Dinger M, Oman F, Taylor E, Vesely S, Able J. Stability and convergent validity of the Physical Activity Scale for the Elderly (PASE). Journal of Sports Medicine and Physical Fitness. 2004;44(2):186.

56. Hagiwara A, Ito N, Sawai K, Kazuma K. Validity and reliability of the Physical Activity Scale for the Elderly (PASE) in Japanese elderly people. Geriatrics & Gerontology International. 2008;8(3):143-51.

57. Ngai SP, Cheung RT, Lam PL, Chiu JK, Fung EY. Validation and reliability of the Physical Activity Scale for the Elderly in Chinese population. Journal of Rehabilitation Medicine. 2012;44(5):462-5.

58. Vaughan K, Miller WC. Validity and reliability of the Chinese translation of the Physical Activity Scale for the Elderly (PASE). Disability and Rehabilitation. 2013;35(3):191-7.

59. Alqarni AM, Vennu V, Alshammari SA, Bindawas SM. Cross-cultural adaptation and validation of the Arabic version of the physical activity scale for the elderly among community-dwelling older adults in Saudi Arabia. Clinical Interventions in Aging. 2018;13:419-27.

60. Heesch KC, Hill RL, Van Uffelen JG, Brown WJ. Are Active Australia physical activity questions valid for older adults? Journal of Science and Medicine in Sport. 2011;14(3):233-7.

61. Freene N, Waddington G, Chesworth W, Davey R, Cochrane T. Validating two self-report physical activity measures in middle-aged adults completing a group exercise or home-based physical activity program. Journal of Science and Medicine in Sport. 2014;17(6):611-6.

62. Grimm EK, Swartz AM, Hart T, Miller NE, Strath SJ. Comparison of the IPAQ-Short Form and accelerometry predictions of physical activity in older adults. Journal of Aging and Physical Activity. 2012;20(1):64-79.

63. Dyrstad SM, Hansen BH, Holme IM, Anderssen SA. Comparison of self-reported versus accelerometer-measured physical activity. Medicine and Science in Sports and Exercise. 2014;46(1):99-106.

64. Steene-Johannessen J, Anderssen SA, van der Ploeg HP, Hendriksen IJ, Donnelly AE, Brage S, et al. Are Self-report Measures Able to Define Individuals as Physically Active or Inactive? Medicine and Science in Sports and Exercise. 2016;48(2):235-44.

65. Lipert A, Jegier A. Comparison of Different Physical Activity Measurement Methods in Adults Aged 45 to 64 Years Under Free-Living Conditions. Clinical Journal of Sport Medicine. 2017;27(4):400-8.

66. Harada ND, Chiu V, King AC, Stewart AL. An evaluation of three self-report physical activity instruments for older adults. Medicine and Science in Sports and Exercise. 2001;33(6):962-70.

67. Washburn R, Ficker J. Physical Activity Scale for the Elderly (PASE): the relationship with activity measured by a portable accelerometer. Journal of Sports Medicine and Physical Fitness. 1999;39(4):336.

68. de Hollander EL, Zwart L, de Vries SI, Wendel-Vos W. The SQUASH was a more valid tool than the OBiN for categorizing adults according to the Dutch physical activity and the combined guideline. Journal of Clinical Epidemiology. 2012;65(1):73-81.

69. Harris TJ, Owen CG, Victor CR, Adams R, Ekelund U, Cook DG. A comparison of questionnaire, accelerometer, and pedometer: measures in older people. Medicine and Science in Sports and Exercise. 2009;41(7):1392-402.

70. Heesch KC, Van Uffelen J, Brown WJ. How do older adults respond to active Australia physical activity questions? Lessons from cognitive interviews. Journal of Aging & Physical Activity. 2014;22(1).

71. Heesch KC, Van Uffelen JG, Hill RL, Brown WJ. What do IPAQ questions mean to older adults? Lessons from cognitive interviews. International Journal of Behavioral Nutrition and Physical Activity. 2010;7(1):1.

72. Bilek LD, Venema DM, Camp KL, Lyden ER, Meza JL. Evaluation of the human activity profile for use with persons with arthritis. Arthritis Care & Research. 2005;53(5):756-63.

73. Hurtig-Wennlöf A, Hagströmer M, Olsson L. The International Physical Activity Questionnaire modified for the elderly: aspects of validity and feasibility. Public Health Nutrition. 2010;13(11):1847-54.

74. Winkler E, Waters L, Eakin E, Fjeldsoe B, Owen N, Reeves M. Is measurement error altered by participation in a physical activity intervention? Medicine and Science in Sports and Exercise. 2013;45(5):1004-11.

75. Washburn RA, Smith KW, Jette AM, Janney CA. The Physical Activity Scale for the Elderly (PASE): development and evaluation. Journal of Clinical Epidemiology. 1993;46(2):153-62.

76. Prince SA, Adamo KB, Hamel ME, Hardt J, Gorber SC, Tremblay M. A

comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review. International Journal of Behavioral Nutrition and Physical Activity. 2008;5(1):1.

77. Quicke JG, Foster NE, Croft PR, Ogollah RO, Holden MA. Change in physical activity level and clinical outcomes in older adults with knee pain: a secondary analysis from a randomised controlled trial. BMC Musculoskeletal Disorders. 2018;19(1):59.

78. Armstrong T, Bauman AE, Davies J. Physical activity patterns of Australian adults: results of the 1999 National Physical Activity Survey: Australian Institute of Health and Welfare; 2000.

79. Baecke JA, Burema J, Frijters J. A short questionnaire for the measurement of habitual physical activity in epidemiological studies. The American Journal of Clinical Nutrition. 1982;36(5):936-42.

80. Moore DS, Ellis R, Allen PD, Monroe PA, Cherry KE, O'Neil CE, et al. Construct validation of physical activity surveys in culturally diverse older adults: a comparison of four commonly used questionnaires. Research Quarterly for Exercise and Sport. 2008;79(1):42-50.

Tables

Table 1: Selection criteria for articles in Stage One.

Inclusion	Exclusion
Age range that includes participants 45 years old or over(1).	
At least 50% of the study participants have OA or joint pain attributable to OA in the foot, knee, hip and hand(1).	Over 50% of the study participants with inflammatory arthritis.
Measurement instrument of PA using a reproducible self-reported questionnaire.	A measure of physical fitness rather than a measure of daily PA participation.
Self-reported PA used as a primary or secondary outcome measure.	Direct measures of PA. For example, accelerometers and calorimetry.
All research settings (hospital, primary care, community settings, etc.)	Not written in English.
All quantitative research methodologies (RCTS, cross-sectional, etc.)	Case study research design of a single subject.

Table 2: Characteristics of the PA instruments included in Stage Two.

Instrument and associated study	Construct	Setting	Recall Period	Purpose	Target population	Justification	Format	Interpretability	Ease of use
				Multi-item					
Active Australia Survey (AAS) (77)	Leisure time PA	leisure time activities at different intensities	7 Days	To assess knowledge of health benefits for PA in adult populations	Developed for adults aged 18-65, can be used internationally	Offers data on PA that can be implemented into self-report survey or interviewing	9 items, self-report on time spend during activities or frequency	Total score in time spent physically active during a week and time spent sedentary	Short time taken to complete

							of activities		
Baecke (78)	Habitual PA across three domains; work related, leisure time and sport	Activities in: occupation, sport and leisure time	Usual week	To assess habitual physical activities for epidemiological studies	Young adults	At the time of development, no appropriate instrument was available for use in epidemiological studies	16 items, Self-report questionnaire with closed answered questions	Scores are given in three indices; work, sport, leisure time. These scores are not interpretable outside of the Baecke	Small number of multiple choice questions
Modified Baecke (39)	Physical activities in household and leisure	Household activities and leisure	One year	Modified to better suite elderly population from the	Elderly adults, aged 65 years and over	Original Baecke not appropriate for elderly populations.	Interviewer administered, not self-report	Time spent PA in hours for one week. Scores can be compared to recommendatio	Interviewer required, takes 30 minutes

	sporting activities	sporting activities		original Baecke				ns on PA levels for health benefits	to complete.
Human Activity Profile (HAP) (28)	Energy expenditure or physical fitness	Daily activities	Same day	Originally developed as indicator of quality of life in pulmonary rehabilitation	Clinical and healthy populations	Previously developed instruments had floor and ceiling effects	94 items in a list, each one a daily activity	Scores give average levels of activity and maximal achievable activity	Closed answer questions, time taken to complete: 1-2 minutes
Incidental And Planned Activity Questionnaire For	Incidental and planned physical activities	Gym or home, activities in daily life	7 days or 3 months	Used in longitudinal epidemiology studies to	Frailer populations	Other instruments for adults aged 45 years and over have too	10 items, on planned or structured exercises	Scores are interpretable to time spent physically active	self-complete instrument, quick

Older People (IPEQ) (41)				assess levels of PA		many items for survey use	and activities in daily living		to complete
International Physical Activity Questionnaire (IPAQ-SF & IPAQ-LF) (44)	Energy expenditure in a week. There is a long version and short version	Long version includes; different settings Short version does not separate settings	Two versions; last week and usual week	Research to compare populations in levels of PA	Adults, 18-65 years old. Different languages available	A generic outcome measure of PA to be used in any adult population internationally	Short version: 4 items, Long version: 27 items. Closed questions, some with continuous scale answer	Scores given in energy expenditure per week, scores can be compared to recommendations on PA levels for health benefits	Short version requires minimal time and effort. Long version takes longer and requires recall in

									different aspects of PA
Physical Activity Scale For The Elderly (PASE) (79)	Time spent participating in PA	PA in various settings at work, home and leisure time	Leisure activities, occupational activities and household activities	Research to assess PA in elderly adults	Elderly adults, aged 65 years and over	None of the generic measures of PA are appropriate for elderly adults	32 items within the six different domains	Scores given as a total score, total score not interpretable in a meaningful way	Questions are easy to fill out with full instruction, short recall period, 32 items is a high number

Short Questionnaire To Assess Health Enhancing Physical Activity (SQUASH) (68)	Habitual activities	Leisure activities, travelling activities, household activities, activities at work	Normal week over past few months	A self-report measure with comparable scores to recommendations of levels of physical activities for health benefits	All adult populations	Required a measurement where scores were interpretable to quantify weekly PA levels	11 items asking questions on PA in different settings	Scores can be classified for recommended PA levels	Very short, simple to complete
Short Telephone Activity Recall Questionnaire (STAR) (42)	Classification of PA in moderate and vigorous levels of PA	All PA	Last 7 days	A telephone administered short instrument to classify individuals in	All adult populations	A need for a quick-to-complete measure of PA over the telephone	3 items, two versions available; open responses and	Responders can be classified into different levels of PA	Very quick to administer

				different levels of PA			closed responses		
Zutphen (69)	Daily physical activities	Leisure-time, walking, household activities, sporting activities and hobbies.	7 days, although some items differ	Used to assess levels of PA in a longitudinal study	Designed for a study in older male adults, but has been used in male and female adults since	Developed as an appropriate measure of PA over time for a longitudinal study	17 items, open and closed questions	Total score given as energy expenditure	Short with minimal requirements for completion
					Single item				

Activity Rating Scale (ARS) (29)	Physical activities	All physical activities	Past year	To assess level of PA in one item	Patient with knee disorder	No valid single item measure of PA	1 item: with a 5-point scale	Scoring range from 0-4	Only one item
Tegner Scale (29)	Physical activities	All physical activities	Past week	To assess level of PA in one item	Knee injury	No valid single item measure of PA	1 item: with a 10-level response	Each value on the scale identifies individuals at an interpretable level of PA	Only one item
University Of California, Los Angeles Activity Scale	Physical activities	All physical activities	Past week	To assess level of PA in one item	Joint replacement surgery	No valid single item measure of PA	1 item: with a 10-level response	Each value on the scale identifies individuals at an interpretable level of PA	Only one item

(UCLAA)

(29)

Table 3: Summary of each instruments measurement properties included in Stage Two.

Instrument and associated studies	Reliability	Measurement error	Criterion validity	Construct validity	Other measurement properties
Populations with joint pain attributable to OA					
Multi-item					
Active Australia Survey (AAS)	0	0	0	0	0
Baecke	ICC=0.87, good quality(27)	0	0	Convergent construct validity, correlation to	0

				accelerometer= 0.49, good quality(27)	
Modified Baecke	0	0	0	0	0
Human Activity Profile (HAP)	ICC= 0.95, 0.96, excellent quality(28). ICC=0.60, 0.83, poor quality(72)	SEM=3, excellent quality(28)	0	0	0
Incidental And Planned Activity Questionnaire For Older People (IPEQ)	0	0	0	0	0

International Physical Activity Questionnaire Short Form (IPAQ-SF)	ICC= 0.76, 0.87, Excellent quality(29). ICC=0.5, fair quality(31).	SEM=2487, SDC=1039, fair quality(31).	0	Convergent construct validity, correlation to accelerometer= 0.29, fair quality(31).	0
Physical Activity Scale For The Elderly (PASE)	ICC=0.77, poor quality(33). ICC=0.58, 0.77, poor quality(32). ICC=0.77, fair quality(30).	SEM= 23-35%, SDC= 63-97%, fair quality(30, 32). SEM= 31, SDC= 87, poor quality(33).	0	Convergent construct validity, correlation to accelerometer=0.3, poor quality(33). correlation to accelerometer=0.06, 0.45, poor quality(32). correlation to accelerometer=0.27, good quality(30)	0
Short Questionnaire	0	0	0	0	0

To Assess Health Enhancing Physical Activity (SQUASH)					
Short Telephone Activity Recall Questionnaire (STAR)	0	0	0	0	0
Zutphen	0	0	0	0	0
			Single item		

Activity Rating Scale (ARS)	Kappa=0.65, 0.88, fair quality(29)	0	0	0	No floor or ceiling effect, fair quality(29)
Tegner	Kappa=0.54, 0.84, fair quality(29)	0	0	0	No floor or ceiling effect, fair quality(29)
University Of California, Los Angeles Activity Scale (UCLAA)	Kappa=0.80, 0.86, fair quality(29)	0	0	0	No floor or ceiling effect, fair quality(29)

Community dwelling adults aged 45 and over					
		Multi-item			
Active Australia Survey (AAS)	Spearman's rank=0.58, 0.64, good quality(35). Spearman's rank=0.32, fair quality(36). Spearman's rank=0.76, fair quality(34)	0	0	Correlation to accelerometer=0.48, 0.52, good quality(35). Correlation to pedometers=0.42, good quality(59). Correlation to accelerometer=0.39, 0.49, good quality(36). Correlation to accelerometer=0.49, 0.56, good quality(60). Correlation to accelerometer=0.45, 0.61, good quality(34).	Wide range of limitations in items in terms of content validity, excellent quality(70)

Baecke	0	0	0	0	0
Modified Baecke	Spearman's rank=0.65, 0.89, fair quality(38). Correlation=0.73, 0.82, poor quality(37). Spearman's rank=0.86, poor quality(39).	0	Correlation with DLW, r=0.54, poor quality(57).	Correlation to heart rate monitoring= NS, poor quality(37). Correlation to PASE, good quality(80).	0
Human Activity Profile (HAP)	ICC=0.79, 0.94, good quality(40)	0	0	Correlation to accelerometer=0.52, 0.55, good quality(40)	0

Incidental And Planned Activity Questionnaire For Older People (IPEQ)	ICC=0.80, 0.84, good quality(41).	0	0	Correlation to accelerometer=0.17, excellent quality(61)	IPEQ responsiveness index=0.31, ActiGraph responsiveness index=0.65, excellent quality(61)
International Physical Activity Questionnaire Short Form (IPAQ-SF)	ICC=0.68, excellent quality(43). Spearman's rank=0.46-0.96, good quality(44). ICC=0.84, excellent quality(45). Spearman's	0	0	Correlation to accelerometer= 0.30-0.33, good quality(44). Correlation to accelerometer= NS, good quality(62). Correlation to accelerometer= 0.30-0.33, poor quality(46). Correlation to	Content validity showed that definitions were confusing and recall was difficult, good quality(71). Internal consistency, Cronbach alpha=0.70, good quality(50).

rank=0.54, poor quality(46).	accelerometer= 0.38-0.56, good quality(47).
ICC=0.5, 0.65, excellent quality(47).	Correlation to accelerometer= 0.39, good quality(48).
ICC=0.86, good quality(48).	Correlation to accelerometer= 0.33, excellent quality(63).
Spearman's rank= 0.26, good quality(49).	Correlation to accelerometer= 0.29, good quality(49).
ICC=0.99, good quality(50).	Correlation to accelerometer=NS, excellent quality (64). Sig difference to accelerometer, poor quality(65).

Physical Activity Scale For The Elderly (PASE)	ICC=0.60, good quality(51). ICC=0.60, good quality(52). ICC=0.65, good quality(53). ICC=0.75, good quality(66). ICC=0.68-0.84, good quality(13). ICC=0.81, fair quality(54). ICC=0.79, good quality(55). ICC=0.90-0.98, good quality(56).	SEM= 3.3–8.5, good quality(56).	Correlation with DLW=NS, good quality(51). Correlation with DLW=0.58, poor quality(58).	Correlation to accelerometer= 0.36, good quality(51). Correlation to accelerometer= 0.43, good quality(52). Spearman's rank correlation to accelerometer= 0.16, fair quality(53). Correlation to accelerometer=0.52, 0.59, good quality(66). Correlation to accelerometer= 0.49, poor quality(67).	Internal consistency, Cronbach alpha=0.71-0.75, good quality(56).

Short Questionnaire To Assess Health Enhancing Physical Activity (SQUASH)	0	0	0	Agreement with heart monitors= 97.6%, fair quality(68).	0
Short Telephone Activity Recall Questionnaire (STAR)	Kappa= 0.57-0.76, excellent quality(42).	0	0	Correlation to accelerometer= 0.14-0.15, good quality(42).	0
Zutphen	0	0	0	Correlation to accelerometer= 0.34, good quality(69).	0

Table 4: Grading of each instruments' measurement properties using COSMIN checklist and QAPAQ.

	Reliability and measurement error		Criterion validity		Construct validity using objective measure		Internal consistency		Content validity		Structural validity		Responsiveness	
	Joint pain	Older adults	Joint pain	Older adults	Joint pain	Older adults	Joint pain	Older adults	Joint pain	Older adults	Joint pain	Older adults	Joint pain	Older adults
Active Australia Survey (AAS)	0	++	0	0	0	---	0	0	0	---	0	0	0	0
Activity Rating Scale (ARS)*	+	0	0	0	0	0	0	0	0	0	0	0	0	0
Baecke	++	0	0	0	0	0	0	0	0	0	0	0	0	0
Modified Baecke	0	?	0	?	0	?	0	0	0	0	0	0	0	0
Human Activity Profile (HAP)	+++	0	0	0	0	--	0	0	0	0	0	0	0	0

International Physical Activity Questionnaire Short Form (IPAQ-SF)	+++	±	0	0	?	---	0	+	0	---	0	0	0	0
Incidental And Planned Activity Questionnaire For Older People (IPEQ)	0	++	0	0	0	---	0	0	0	0	0	0	0	+++
Physical Activity Scale For The Elderly (PASE)	++	±	0	--	--	---	0	+	0	0	0	0	0	0
Short Questionnaire To Assess Health Enhancing Physical Activity (SQUASH)	0	0	0	0	0	+	0	0	0	0	0	0	0	0

Short Telephone Activity Recall Questionnaire (STAR)	0	±	0	0	0	--	0	0	0	0	0	0	0	0
Tegner*	+	0	0	0	0	0	0	0	0	0	0	0	0	0
(University Of California, Los Angeles Activity Scale) UCLAA*	+	0	0	0	0	0	0	0	0	0	0	0	0	0
Zutphen	0	0	0	0	0	--	0	0	0	0	0	0	0	0

Key: '?' indicates unclear findings due to study quality; '±' indicates conflicting findings. * indicates single scale items. Strength of the evidence was given based on quality of articles assessed by the COSMIN [6]. 'Joint pain' refers to joint pain attributable to OA. Instruments were given a positive, negative or zero score for the corresponding measurement property based on criteria (10) (Appendix 3 & 4).

Figure 1: PRISMA flowchart of included articles from Stage One.

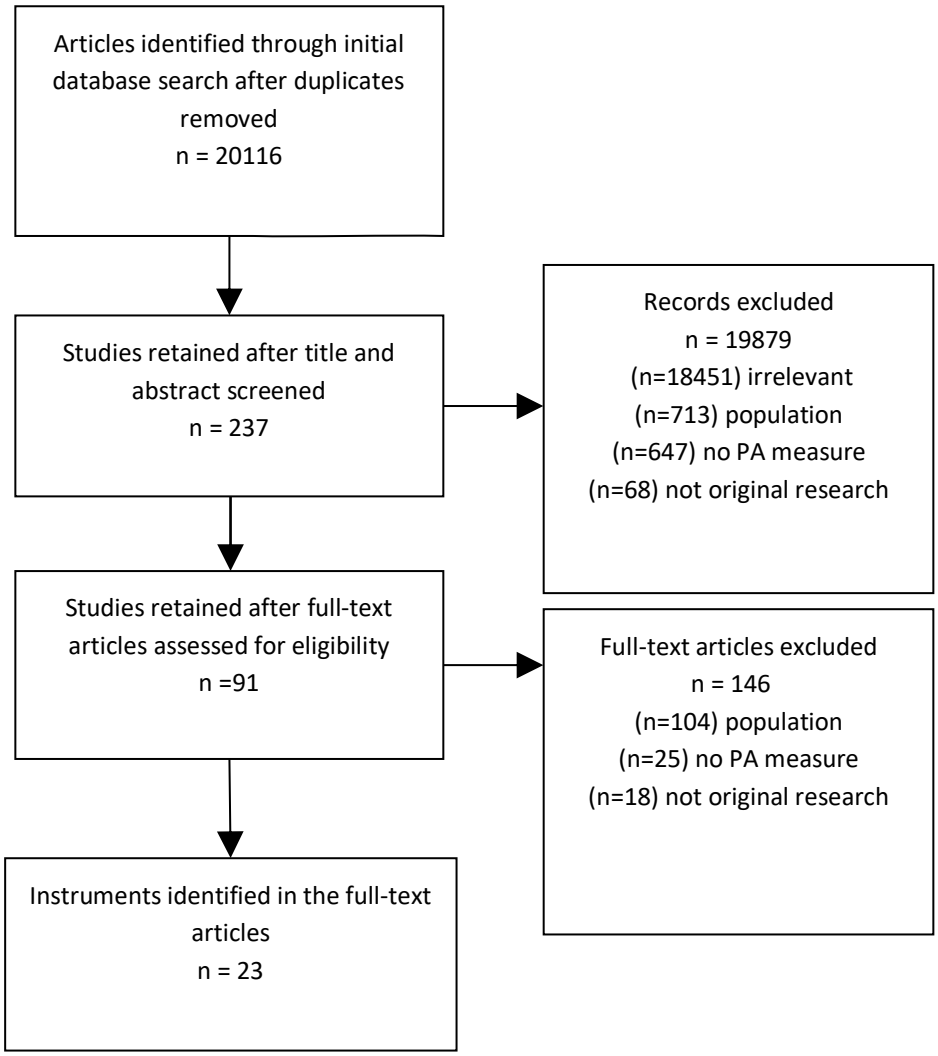

Figure 2: PRISMA Flowchart of included articles from Stage Two.