Osteoarthritis and Cartilage 24 (2016) 1317–1329

Osteoarthritis and Cartilage

Review

Knee Injury and Osteoarthritis Outcome Score (KOOS): systematic review and meta-analysis of measurement properties

N.J. Collins † ‡ *, C.A.C. Prinsen §, R. Christensen ||, E.M. Bartels ¶, C.B. Terwee §, E.M. Roos #

† School of Health and Rehabilitation Sciences, The University of Queensland, St Lucia, Queensland, Australia
‡ Department of Mechanical Engineering, Melbourne School of Engineering, The University of Melbourne, Parkville, Victoria, Australia
§ VU University Medical Center, Department of Epidemiology and Biostatistics, EMGO Institute for Health and Care Research, The Netherlands
|| Musculoskeletal Statistics Unit, The Parker Institute, Bispebjerg and Frederiksberg Hospital, Copenhagen, Denmark
¶ The Biochemistry and Physiology Laboratory, The Parker Institute, Bispebjerg and Frederiksberg Hospital, Copenhagen, Denmark
# Research Unit for Musculoskeletal Function and Physiotherapy, Institute of Sports Science and Clinical Biomechanics, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark

ARTICLE INFO

Article history:
Received 24 August 2015
Accepted 13 March 2016

Keywords:
Knee injury
Knee osteoarthritis
KOOS
Patient-reported outcome measures
Measurement properties
Meta-analysis

SUMMARY

Objective: To conduct a systematic review and meta-analysis to synthesize evidence regarding measurement properties of the Knee injury and Osteoarthritis Outcome Score (KOOS).
Design: A comprehensive literature search identified 37 eligible papers evaluating KOOS measurement properties in participants with knee injuries and/or osteoarthritis (OA). Methodological quality was evaluated using the COSMIN checklist. Where possible, meta-analysis of extracted data was conducted for all studies and stratified by age and knee condition; otherwise narrative synthesis was performed.
Results: KOOS has adequate internal consistency, test-retest reliability and construct validity in young and old adults with knee injuries and/or OA. The ADL subscale has better content validity for older patients and Sport/Rec for younger patients with knee injuries, while the Pain subscale is more relevant for painful knee conditions. The five-factor structure of the original KOOS is unclear. There is some evidence that the KOOS subscales demonstrate sufficient unidimensionality, but this requires confirmation. Although measurement error requires further evaluation, the minimal detectable change for KOOS subscales ranges from 14.3 to 19.6 for younger individuals, and ≥20 for older individuals. Evidence of responsiveness comes from larger effect sizes following surgical (especially total knee replacement) than non-surgical interventions.
Conclusions: KOOS demonstrates adequate content validity, internal consistency, test-retest reliability, construct validity and responsiveness for age- and condition-relevant subscales. Structural validity, cross-cultural validity and measurement error require further evaluation, as well as construct validity of KOOS Physical function Short form. Suggested order of subscales for different knee conditions can be applied in hierarchical testing of endpoints in clinical trials.
Systematic review registration: PROSPERO (CRD42011001603).

Introduction

Patient-reported outcome measures (PROMs) are used across medical disciplines to follow disease course or evaluate treatment outcomes. PROMs involve the patient's evaluation of any aspect of their health status, without interpretation of their response by another individual[1]. Study findings are more truthful if the PROM used has adequate measurement properties, and if the studies testing its measurement properties are of good or excellent

* Address correspondence and reprint requests to: N. Collins, School of Health and Rehabilitation Sciences, The University of Queensland, Brisbane QLD 4072, Australia.
E-mail addresses: n.collins1@uq.edu.au (N.J. Collins), c.prinsen@vumc.nl (C.A.C. Prinsen), robin.christensen@regionh.dk (R. Christensen), Else.Marie.Bartels@regionh.dk (E.M. Bartels), cb.terwee@vumc.nl (C.B. Terwee), eroos@health.sdu.dk (E.M. Roos).

http://dx.doi.org/10.1016/j.joca.2016.03.010
1063-4584/

methodological quality[2]. Specifically, a PROM should have content that is relevant for the construct of interest and the target population, measure intended dimensions, be stable on repeated measures, and be able to detect change in patients' perceived health status[3]. Since measurement properties can differ between different patient groups (e.g., patient characteristics, medical condition, intervention), PROMs need to be evaluated in different patient populations, ideally across multiple studies.

The Knee Injury and Osteoarthritis Outcome Score (KOOS) is a PROM intended for young, middle-aged and elderly adults with knee injury and/or knee osteoarthritis (OA), and can be used to monitor disease course and outcomes following surgical, pharmacological and other interventions[4]. KOOS holds five subscales: (1) Pain (9 items); (2) other Symptoms (7 items); (3) Activities of Daily Living (ADL, 17 items); (4) Sport and Recreation function (Sport/Rec, 5 items); and (5) knee-related Quality of Life (QoL, 4 items). Each subscale is scored separately from zero (extreme knee problems) to 100 (no knee problems). The KOOS Physical function Short form (KOOS-PS, 7 items) was later derived from the ADL and Sport/Rec subscales via Rasch analysis[5]. The clinical and research utility of KOOS is highlighted by large international patient datasets (>100,000 unique patient records) and frequent use in scientific publications. Importantly, KOOS has international accessibility, being free of charge and translated into >45 different language versions[6].

To provide clinicians and researchers with a single reference regarding KOOS measurement properties, we performed a systematic review and meta-analysis to evaluate the measurement properties of KOOS in people with knee injuries and/or OA.

Method

Review protocol

The protocol was developed according to the PRISMA statement[7] and prospectively registered (PROSPERO, CRD42011001603, 11 October 2011).

Literature search

A research librarian and researcher in arthritic diseases (E.M.B.) searched six bibliographic databases from 1998 (date of KOOS publication) to 16 January 2014, with no language restrictions (Medline via PubMed, EMBASE via OVID, CINAHL via EBSCO, Web of Science, Psycinfo via OVID, Cochrane Central Register of Controlled Trials). The following terms were searched as free text and key words (where applicable): (KOOS AND knee) OR (Knee Injury and Osteoarthritis Outcome Score). Reference lists of eligible papers and review papers were manually searched. The KOOS website[6] was reviewed to cross-check inclusion of all cross-cultural validation studies or papers published in non-English languages, and KOOS developers were consulted regarding known unpublished studies.

Eligibility criteria

Original full-text published studies were considered for inclusion, as well as PhD theses identified by the search strategy. Studies were eligible if: (1) the primary aim was to evaluate at least one measurement property (e.g., reliability, validity, responsiveness) or interpretability of KOOS/KOOS-PS; (2) they studied participants of any age suffering from any knee injury and/or knee OA; and (3) KOOS was patient-completed (e.g., paper, computer or touch-screen administration). Where appropriate, we included studies that utilized a research administrator to assist KOOS completion in populations with limited education (e.g., interviewer-administered), but excluded studies where medical practitioners administered KOOS in the clinic to reduce the risk of bias. No restrictions were placed on method of study recruitment, study venue, or KOOS language version. Studies were excluded if they used KOOS to assess participants for which KOOS was not designed (e.g., other lower limb conditions, asymptomatic cohorts), to evaluate treatment efficacy without assessing measurement properties, or to validate other measurement tools.

Article selection

Results of database searches were imported into Reference Manager 12 (Thomson Reuters, Philadelphia, USA). Two independent reviewers (N.J.C., E.M.R.) assessed titles, abstracts and full-text articles (where appropriate) for eligibility. Discrepancies were discussed to reach consensus, with unresolved cases taken to a third reviewer (C.B.T.).

Data extraction

One reviewer (N.J.C.) used a predefined spreadsheet to extract study characteristics and measurement property data on two separate occasions, blinded to the previous extraction[8]. There were minimal discrepancies between the two extractions, resolved by a second reviewer (R.C.). For studies published in languages other than English, we contacted corresponding authors to assist with translation of required data. When authors were unavailable, we utilized university volunteers who were native speakers.

Evaluation of measurement properties

We evaluated specific measurement properties as defined by the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) taxonomy[9,10]: content validity, structural validity and unidimensionality, internal consistency, test-retest reliability, measurement error, construct validity, cross-cultural validity and responsiveness (Table I). We also evaluated interpretability and feasibility. Criterion validity was not evaluated due to the lack of an established gold standard measure for domains captured by KOOS[13].

Data synthesis

For test-retest reliability, we calculated weighted mean intraclass correlation coefficients (ICCs) and 95% confidence intervals using a standard generic inverse variance random effects model[16]. ICC values were combined based on estimates derived from a Fisher transformation, $z = 0.5 \times \ln((1 + \text{ICC})/(1-\text{ICC}))$, which has an approximate variance, $(\text{Var}(z) = 1/(N\text{-}3))$, where N is the sample size. For each analysis, we qualitatively evaluated whether it was appropriate to pool data (eyeballing), and used quantitative methods (I^2) to evaluate the potential impact of heterogeneity on pooled estimates[17,18]. The likelihood of publication bias was addressed qualitatively, by contacting knee OA experts with an interest in psychometrics. We performed subgroup analyses for different age groups and knee conditions, comparing strata using meta-regression with study characteristics handled as independent class variables (SAS software, version 9.3).

For other measurement properties, we calculated weighted means (number of participants included per study) and 95% confidence intervals, where possible and appropriate (qualitative evaluation), as well as subgroup analyses for different age groups and knee conditions. Quantitative findings for each measurement

Table I
Quality criteria for evaluated measurement properties

Measurement property	Definition	Data management and interpretation
Reliability	Degree to which the KOOS is free from measurement error[9].	
Internal consistency	"Degree of interrelatedness among the items"[9].	Weighted mean Cronbach's alpha (95% CI). Considered adequate if pooled Cronbach's alpha $\geq$0.7 and <0.95[8], and subscale demonstrates unidimensionality[11].
Test-retest reliability	"Proportion of total variance in the measurements which is because of "true" differences among patients"[9].	Weighted mean ICC (95% CI) (random effects model*). Considered adequate if weighted mean ICC $\geq$0.7[8].
Measurement error	"The systematic and random error of a patient's score that is not attributed to true changes in the construct to be measured"[9]. Whether repeated KOOS administrations yield the same result when the patient's condition had not changed[12].	Weighted mean standard error of measurement (SEM) (95% CI). SDC was calculated for each study ($1.96 \times \sqrt{2} \times$ SEM), and weighted mean SDC (95% CI) calculated†. Measurement error considered adequate if pooled SDC smaller than minimal important change (MIC)[8].
Validity	"Degree to which a health-related PROM measures the construct(s) it purports to measure"[9].	
Content validity (including face validity)	Degree to which the content of the KOOS subscales "is an adequate reflection of the construct to be measured"[9]. This is the extent to which each KOOS subscale includes the most relevant and important aspects of the specific domain, as it relates to people with knee injury and knee OA[13].	Narrative synthesis. Considered to be adequate if the development paper reported clear descriptions of the measurement aim, target population, dimensions measured, and item selection process[8,13]. Target population (i.e., patients with knee injuries and/or knee OA) should have been involved in item selection, alongside other experts[8].
Construct validity (hypothesis testing)	Degree to which the KOOS scores were consistent with hypotheses based on the assumption that the KOOS "validly measures the construct to be measured"[9], being physical properties more so than mental or emotional properties.	Weighted mean correlation coefficients (95% CI). Hypothesised *a priori* that we would observe: (1) larger correlations between the six KOOS subscales and the Short Form-36 (SF-36) pain and function subscales, as they measure similar constructs (convergent validity); and (2) smaller correlations between the six KOOS subscales and the SF-36 mental health and emotional role subscales, as they measure dissimilar constructs (divergent validity). Considered adequate if correlations with instruments measuring similar constructs $\geq$0.5 OR at least 75% of the results in accordance with hypotheses AND correlations with related constructs are higher than with unrelated constructs[8].
Structural validity and unidimensionality	How well the KOOS items "are an adequate reflection of the dimensionality of the construct to be measured"[9].	Narrative synthesis of studies that evaluated the structural validity of all 42 original KOOS items, as well as studies that evaluated the unidimensionality of individual subscales. Considered adequate if factors explained at least 50% of the variance[8], or if item–scale correlation $\geq$ 0.3[14].
Responsiveness	"The ability of a health-related PROM to detect change over time in the construct to be measured"[9].	Weighted mean effect sizes and SRM were calculated (where possible) across different timeframes for: (1) non-surgical interventions (e.g., pharmacology, exercise and education, physical therapy); (2) surgical interventions generally targeted towards younger individuals with knee injury or early-stage knee OA (e.g., anterior cruciate ligament reconstruction, autologous chondrocyte implantation); and (3) total knee replacement. We hypothesised that‡: (1) larger effect sizes would be observed following surgical than non-surgical interventions; (2) largest effect sizes would be observed following total knee replacement; and (3) effect sizes would increase with increasing time following total knee replacement.
Interpretability	Degree to which qualitative or clinical meaning can be assigned to the KOOS subscale scores or change in scores[9].	
Percentage of missing items		Narrative synthesis
Description of how missing items were handled		Narrative synthesis
Floor and ceiling effects	Percentage of respondents with the lowest possible score (floor effects) and highest possible score (ceiling effects).	Narrative synthesis Considered adequate if less than 15% of respondents score the lowest (floor effects) or highest (ceiling effects) possible score[8].
Average scores	Reported baseline values for each KOOS subscale.	Weighted mean and standard deviation for overall cohort and relevant subgroups (fixed effects methods).
Minimal important change (MIC)	Smallest change in score that patients consider important.	Narrative synthesis
Feasibility		
Time to complete	Reported time taken for participants to complete the KOOS.	Narrative synthesis

CI, confidence interval. ICC, intra-class correlation coefficient.

* ICC values combined using a Fisher transformation, $z = 0.5 \times \ln((1 + \text{ICC})/(1-\text{ICC}))$ which has an approximate variance, $\text{Var}(z) = 1/(N\text{-}3)$, where N is the sample size.

† SDC data from individual studies not pooled due to differences in formulas and a lack of reporting of methods used to calculate SDC.

‡ COSMIN specifies that responsiveness evaluation should involve *a priori* hypothesis testing of correlations of change in two simultaneously administered measures over time[15]. Calculation of effect sizes (score change divided by baseline variation of a single instrument) without *a priori* hypotheses is considered to be inadequate as it only reflects the magnitude of the change score, rather than the validity of the change score[15]. However, the majority of included papers that evaluated responsiveness did so prior to widespread uptake of COSMIN guidelines. Thus, we chose to specify *a priori* hypotheses regarding the expected magnitude of effect sizes observed between different interventions.

property were summarized in tables, and interpreted based on predefined quality criteria[8] (Table I).

Quality assessment

Methodological quality of included studies was assessed using the COSMIN checklist with four-point rating scale ('excellent', 'good', 'fair', 'poor')[10]. Papers published in non-English languages were excluded from quality assessment due to limited resources for obtaining full translations. Two reviewers (N.J.C., C.A.C.P.) independently rated each study, blinded to authors, affiliations and publishing journal. Discrepancies were discussed to reach consensus, and taken to a third reviewer (C.B.T.) for resolution as required. Overall ratings of methodological quality were based on COSMIN guidelines.

When we repeated meta-analyses stratified by quality rating, pooled outcomes were generally not different between COSMIN ratings. Thus, irrespective of COSMIN rating, all studies were included in meta-analyses (including non-English papers).

Best evidence synthesis

We synthesized findings for measurement properties and methodological quality (COSMIN rating) by applying levels of evidence (defined in Table II)[19].

Results

From 1,254 unique records, we included 37 studies evaluating 19 different language versions of KOOS (Fig. 1). Subgroup analyses were performed separately based on age (young, old), and on knee condition (knee OA, ACL, focal cartilage lesions, mixed). Study characteristics, including subgroup categories, are summarized in Table III (additional details: Supplementary Tables S1 and S2). Four non-English papers were included[21,23,31,42].

Content validity

We found evidence of content validity for the five KOOS subscales from four studies[4,22,43,54]. No studies evaluated content validity of the KOOS-PS items as a stand-alone PROM. The original KOOS publication reported clear descriptions of the measurement aim, target population, dimensions measured, and item selection process[4]. Patients with knee injuries and other experts (physical therapists, orthopaedic surgeons) were involved in domain and item selection, and nominated short- and long-term symptoms and functional difficulties resulting from ACL or meniscus injury[4]. The relevance of these factors to patients with post-traumatic knee OA was evaluated in 75 people 20 years after meniscal surgery. Symptoms most frequently rated highly were incorporated in KOOS. The Western Ontario and MacMaster Universities (WOMAC) Osteoarthritis Index 3.0, which has established content validity[55], was included in KOOS to ensure content validity for older people with OA[4].

In subsequent studies, van Meer *et al.*[54] evaluated the relevance of KOOS items to young adults with ACL insufficiency, orthopaedic surgeons, sports physicians and physical therapists. Engelhart *et al.*[43] interviewed young adults who were candidates for or had undergone articular cartilage repair for focal knee cartilage lesions. All Sport/Rec (5/5) and QoL (4/4) items were considered relevant by $\geq$ 75% of participants (Supplementary Table S3). The ACL cohort rated 5/9 Pain items, 6/7 Symptoms items, and 5/17 ADL items as relevant[54]. Those with focal cartilage lesions rated 9/9 Pain items, 7/7 Symptoms items and 15/17 ADL items as relevant[43].

Roos *et al.*[22] reported that >90% of patients scheduled for primary TKR ($n = 105$) rated improvement in Pain, Symptoms, ADL and QoL as extremely or very important when deciding to have surgery[22]. 51% rated improvement in Sport/Rec function to be extremely/very important in surgical decision-making.

Structural validity and unidimensionality

Two studies applied factor analysis to all 42 KOOS items to confirm the five-subscale structure of the original KOOS. De Groot *et al.*[27] evaluated 262 older adults with knee OA ± previous ACL reconstruction, whereas Almangoush *et al.*[47] evaluated 129 young adults with ACL and/or meniscal injuries. Both studies scored a COSMIN rating of 'poor' for the quality of structural validity assessment; hence, particular methodological features are likely to influence outcomes. Rather than using confirmatory factor analysis to evaluate whether items fit a predetermined structure (five subscales), both studies used exploratory factor analysis (principal component analysis), which is recommended for use during PROM development or if confirmatory factor analysis does not support the hypothesized structure[15]. The sample size of Almangoush *et al.*[47] was markedly below COSMIN-recommended minimum of five participants per item (i.e., minimum sample size of 210), although De Groot *et al.*[27] included 6.2 participants per item and was rated as 'good' using COSMIN criteria. Thus, while both studies found that all 42 items loaded on one factor, firm conclusions confirming the KOOS five-factor structure cannot be made from available data.

Two studies applied principal component analysis to individual subscales to evaluate within-scale dimensionality. Hoogeboom *et al.*[44] evaluated ADL in 284 older adults with knee OA. All 17 ADL items loaded on a single major factor, explaining 54.2% of the variance. Roos *et al.*[20] performed five separate factor analyses to evaluate unidimensionality of each KOOS subscale separately in 142 young adults with ACL, meniscal or tibiofemoral joint cartilage injury. The ADL, Sport/Rec and QoL subscales appeared unidimensional (percentage variance explained by the principal factor ranging from 55 to 73%). Although two separate factor analyses for Pain and Symptoms found that items within each subscale loaded on two major factors, with the principal factor explaining 44% and 41% of the variance, respectively, the four items that loaded higher on the second factor had loads of at least 0.42 with the principal

Table II
Criteria for recommendations based on best evidence synthesis

Level of evidence	Definition
Strong evidence:	Evidence from pooled data, or consistent findings in multiple studies (qualitative synthesis), including at least two studies of 'good' methodological quality; OR evidence from one study of 'excellent' methodological quality.
Moderate evidence:	Evidence from pooled data, or consistent findings in multiple studies (qualitative synthesis), including at least two studies of 'fair' methodological quality; OR evidence from one study of 'good' methodological quality.
Limited evidence:	Evidence from one study of fair methodological quality.
Conflicting evidence:	Conflicting findings, or population-specific properties.
Unknown evidence:	Only studies of poor methodological quality; OR no evidence available.

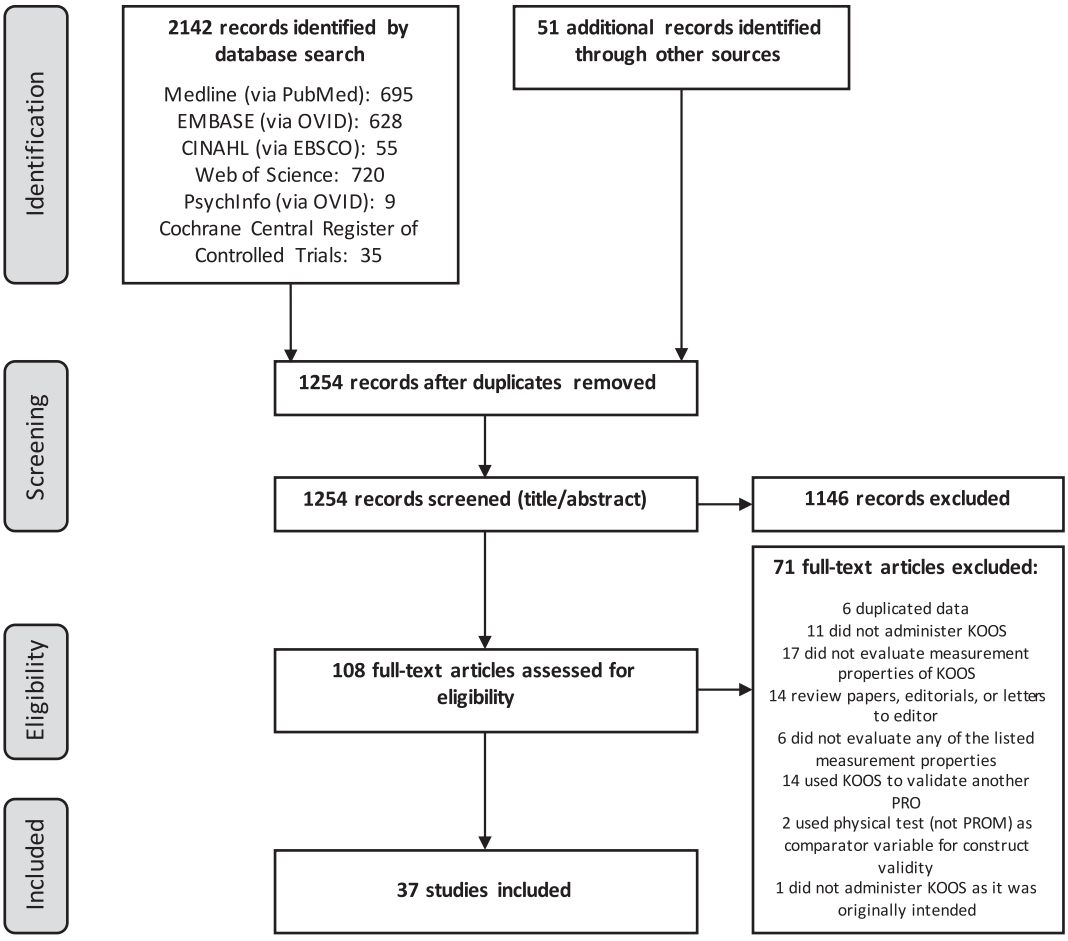

Fig. 1. PRISMA flow diagram.

factor in each subscale. This suggests that the items are sufficiently related to the principal factor to allow Pain and Symptoms to each be scored as one subscale.

Comins *et al.*[26] performed Rasch analysis on each subscale in 200 young adults who had undergone ACL reconstruction. Although rated as 'good' for structural validity methodology, the stability of results may have been influenced by the sample size, which was smaller than recommended for Rasch analyses ($n > 250$)[56]. Thus, although the Sport/Rec and QoL subscales fit the Rasch model, but not Pain, Symptoms or ADL, the sample size limits the ability to draw conclusions from these findings.

Four studies used item–scale correlation to evaluate dimensionality of individual subscales (Supplementary Table S4). Xie *et al.*[24] evaluated Singapore English and Singapore Chinese versions of KOOS in older people with OA. All Sport/Rec and QoL items had correlations ≥0.3 with their respective subscale score, as did all ADL items in the English version. Suboptimal correlations were found for Pain items 1 and 2, Symptoms items 2 and 6, and ADL items 14 and 15. In young people with ACL injuries, all Pain, ADL and Sport/Rec items had correlations ≥0.3 with their respective subscale score[29]. All QoL items, except item 2, had correlations ≥0.3 with the subscale score. Symptoms items 1, 4, 5 and 6 had correlations <0.3 with the subscale score. In contrast, in young adults with ACL reconstruction, all items in all subscales had item–scale correlations ≥0.3[40]. In patients with ACL, meniscus or combined injuries, all items in all subscales had item–scale correlations ≥0.3[45].

The seven KOOS-PS items were derived using Rasch analysis of data from older adults with knee OA, and formed a unidimensional structure[5]. The quality of structural validity evaluation used in this study was rated as 'good'. Franchignoni *et al.*[49] found that, in patients with knee OA, KOOS-PS had a unidimensional structure when evaluated using principal component analysis (COSMIN rating of 'fair' for quality of structural validity assessment). The authors also performed Rasch analysis, finding that 5/7 KOOS-PS items fit the Rasch model (item 1 showed overfit [infit MnSq 0.64], and item 5 showed borderline infit [MnSq = 1.24]). Harris *et al.*[51] performed confirmatory factor analysis on KOOS-PS (quality of structural validity evaluation rated as 'good'). The authors reported mixed results. Although the one-factor model was rejected by findings of two tests (chi-square test; root mean square error of approximation [RMSEA]), the comparative fit index and standardised root mean residuals suggested good fit to the one-factor model. Thus, the majority of evidence suggests that KOOS-PS has a unidimensional structure.

Internal consistency

Pooled values for Cronbach's alpha from 25 KOOS and six KOOS-PS unique cohorts were between 0.70 and 0.95 for all subscales (Table IV; Supplementary Table S5). Findings did not change when internal consistency data were split by age or knee condition.

Test-retest reliability and measurement error

Pooled data from 26 unique cohorts (Supplementary Table S6) revealed that all subscales demonstrated adequate test-retest

Table III
Characteristics and subgroup categories (condition, intervention, age) of included studies

Source	Language version	Condition	Intervention	*N* (females)	Age (mean ± SD) [category]	Content validity	Structural validity/ unidimen-sionality	Internal consistency	Test-retest reliability	Measurement error	Hypothesis testing	Cross-cultural translation/ validity	Responsive-ness	Interpretability/ feasibility
Roos 1998[4]	English	ACL	Surgery	21 (12)	32 ± NR [young]	✓			✓		✓		✓	
Roos 1998[20]	Swedish	Mixed	Surgery	142 (53)	39.6 ± NR [young]		✓	✓	✓		✓	✓	✓	✓
Kessler 2003[21]	German	OA	Surgery	a. 90 (58) b. 30 (NR)	a. 67.3 ± NR [old] b. NR†			✓	✓		✓	✓	✓	✓
Roos 2003[22]	Swedish	OA	Surgery	105 (66)	71.3 ± NR [old]	✓			✓		✓		✓	✓
Seo 2006[23]	Korean	Mixed	Surgery	108 (19)	31.5 ± NR [young]			✓	✓		✓	✓	✓	
Xie 2006[24]	a. Singapore English b. Singapore Chinese	OA	Surgery	a. 127 (97) b. 131 (116)	a. 65.3 ± 7.9 [old] b. 67.8 ± 7.1 [old]		✓	✓	✓		✓	✓		✓
Paker 2007[25]	Turkish	OA	NR	50 (43)	62.8 ± 9.5 [old]			✓	✓		✓	✓		✓
Comins 2008[26]	Danish	ACL	Surgery	200 (70)	29.4 ± NR [young]		✓	✓						
de Groot 2008[27]	Dutch	a. ACL + OA† b. OA c. OA d. OA e. OA	a. Surgery b. Surgery c. None d. Surgery e. Surgery	a. 36 (8) b. 62 (20) c. 47 (24) d. 63 (32) e. 54 (42)	a. 36 ± NR† b. 56 ± NR† c. 65 ± NR [old] d. 61 ± NR [old] e. 71 ± NR [old]		✓	✓	✓	✓	✓	✓		✓
Ornetti 2008[28]	French	OA	a. None b. Surgery	a. 37 (27) b. 30 (22)	a. 70 ± 10 [old] b. 71 ± 10 [old]			✓	✓	✓	✓	✓	✓	✓
Perrucio 2008[5,*]	Mixed	OA	Mixed	2145 (1532)	65 ± NR [old]		✓	✓						✓
Salavati 2008[29]	Persian	Mixed	None	147 (16)	31.4 ± 9 [young]		✓	✓	✓		✓	✓		✓
Bekkers, 2009[30]	Dutch	Focal cartilage lesion	Surgery	a. 40 (12) b. 36 (NR)	a. 35 ± 12 [young] b. NR†			✓	✓	✓	✓		✓	✓
Chaipinyo 2009[31]	Thai	OA	NR	a. 25 (18) b. 48 (37)	a. 63.4 ± 7 [old] b. 66.2 ± 7.3 [old]			✓	✓		✓	✓		
Davis 2009[32,*]	English	OA	Surgery	248 (156)	64.5 ± 10.3 [old]			✓			✓		✓	
Goncalves 2009[33]	Portuguese	OA	Non-pharmacological, non-surgical	a. 223 (156) b. 65 (46) c. 84 (57) d. 74 (53)	a. 66.5 ± 10.8 [old] b. 66.8 ± 10.8 [old] c. 67.6 ± 10.2 [old] d. 64.9 ± 11.2 [old]			✓	✓		✓	✓	✓	✓
Ornetti 2009[34,*]	French	OA	a. None b. Pharmacological	a. 49 (35) b. 38 (NR)	a. 72 ± 9 [old] b. NR†				✓		✓		✓	✓
Davis 2010[35]	English	OA	Surgery	44 (33)	64.2 ± 11.5 [old]								✓	
Goncalves 2010[36,*]	Portuguese	OA	Non-pharmacological, non-surgical	a. 85 (63) b. 51 (36) c. 34 (27)	a. 65.7 ± 6.9 [old] b. 66 ± 6.6 [old] c. 65.3 ± 7.3 [old]			✓	✓		✓		✓	✓
Gudbergsen 2011[37]	Danish	OA	Non-pharmacological, non-surgical	20 (20)	66.5 ± 7 [old]				✓					✓
Nakamura 2011[38]	Japanese	OA	NR	58 (51)	68.9 ± 9.4 [old]			✓	✓		✓	✓		✓
Ruyssen-Witrand 2011[39,*]	French	OA	a. None b. None c. Surgery	a. 30 (20) b. 128 (93) c. 60 (41)	a. 69.3 ± 10.9 [old] b. 70.9 ± 10.5 [old] c. 71 ± 10.3 [old]			✓	✓		✓		✓	
Salavati 2011[40]	Persian	ACL	Surgery	57 (18)	25.6 ± 3.4 [young]		✓	✓	✓	✓	✓			✓
Bond 2012[41,*]	NR	OA	Pharmacological	156 (107)	61.2 ± 9.2 [old]						✓		✓	
Briem 2012[42]	Icelandic	Mixed	Mixed	145 (93)	55.6 ± 18.5†			✓	✓	✓	✓	✓	✓	
Engelhart 2012[43]	NR	Focal cartilage lesion	Surgery	54 (22)	34.1 ± NR [young]	✓		✓	✓	✓	✓		✓	
Hoogeboom 2012[44]	Dutch	OA	None	284 (NR)	NR†		✓				✓			✓
Monticone 2012[45]	Italian	Mixed	NR	224 (90)	48.2 ± 21.2†		✓	✓	✓		✓	✓		✓
Vaquero 2014[46]	Spanish	Focal cartilage lesion	Surgery	20 (6)	41.3 ± 14 [young]			✓	✓		✓	✓	✓	✓

Almangoush 2013[47]	Arabic	Mixed	None	129 (30)	30.8 ± NR [young]	✓		✓		✓		✓				✓	
Ebert 2013[48]	English	Focal cartilage lesion	Surgery	104 (42)	37.9 ± 11.6 [young]								✓				
Franchignoni 2013[49]	Italian	OA	None	200 (147)	69.4 ± 9.5 [old]	✓		✓									
Gul 2013[50,*]	Turkish	OA	NR	80 (71)	58.9 ± 8.7[†]			✓		✓		✓					
Harris 2013[51,*]	English	OA	Mixed	134 (67)	59 ± 11[†]			✓		✓				✓			
Monticone 2013[52]	Italian	OA	Surgery	148 (94)	71 ± 6.5 [old]	✓				✓				✓		✓	
Paradowski 2013[53]	Polish	ACL	Surgery	72 (20)	29.8 ± 9.2 [young]			✓		✓		✓		✓		✓	
van Meer 2013[54]	Dutch	ACL	a. None	a. 50 (20)	a. 27 ± NR [young]	✓		✓		✓		✓		✓		✓	
			b. None	b. 100 (25)	b. 26 ± NR [young]												
			c. Surgery	c. 50 (12)	c. 28 ± NR [young]												

NR, not reported.
[*] Study evaluated KOOS-PS.
[†] Excluded from relevant subgroup analysis. 'Mixed' refers to non-homogenous cohorts with respect to language version, condition or intervention (see Table S1 for additional details).

reliability (ICC range 0.85–0.9, $I^2 \geq 68.9\%$) (Table IV). Pooled data confirmed adequate reliability for age (0.83–0.9) and condition (0.83–0.95) subgroups.

For the five KOOS subscales, the pooled smallest detectable change (SDC) for individuals ranged from 15.7 (ADL) to 25.1 (Sport/Rec) (nine studies, Table IV; Supplementary Table S7). The SDC was greater for older adults and those with knee OA than for younger and ACL cohorts. For KOOS-PS, the SDC was 18.6. Differences in patient groups prevented comparison of SDC data[27] with MIC data[52]. For KOOS-PS, one study reported that the SDC (16) was higher than the MIC (12) in older individuals with knee OA[51].

Construct validity

Most studies that evaluated construct validity conducted hypothesis testing of KOOS subscales against SF-36 subscales. Our hypotheses regarding convergent and divergent validity (Table I) were supported when all data were pooled, and when split by age group and knee condition (Table V; Supplementary Table S8).

Cross-cultural validity

We could not evaluate cross-cultural validity, as included studies translated the KOOS but did not undertake formal cross-cultural validation processes (e.g., confirmatory factor analysis, differential item function).

Responsiveness

Consistent with our hypotheses (Table I), we generally observed larger effect sizes and standardized response means (SRM) with surgical than non-surgical interventions (Fig. 2; Supplementary Table S9). We also observed largest effect sizes following TKR. However, we did not see a trend for increasing effect size or SRM with increasing time post-TKR, with effect sizes remaining stable at 3, 6 and 12 months.

Interpretability and feasibility

Average completion time for KOOS ranged from 7 to 12 min[37,45,47], and was not evaluated for KOOS-PS (Supplementary Table S10). Manual data entry and scoring KOOS took 10 min[21]. Most studies reported missing items (range: 0–5%). In an older TKR cohort, 74% of all Sport/Rec items were considered to be "not applicable"[22]. Multiple studies identified floor and ceiling effects for Sport/Rec (Supplementary Table S10). Pain, Symptoms and ADL demonstrated no floor effects, although some ceiling effects were observed pre- and post-ACL reconstruction[27,53,54]. For QoL, some studies demonstrated floor effects pre-TKR[27] and ceiling effects post-TKR[22]. One study found no floor or ceiling effects for KOOS-PS[34]. Table IV presents pooled mean scores for each subscale (individual study scores in Supplementary Table S11).

Two studies calculated the MIC using anchor-based methods. In 148 participants who underwent post-TKR rehabilitation, the MIC for Pain was 16.7, Symptoms 10.7, ADL 18.4, Sport/Rec 12.5 and QoL 15.6[52]. In 116 people with knee OA undergoing non-surgical management, the MIC for KOOS-PS was 12[51].

COSMIN quality ratings of included papers

Thirty-four articles underwent methodological quality rating (Supplementary Table S12). The majority of studies used classical test theory, with three articles utilizing item response theory. For internal consistency, most studies scored 'good' ($k = 13$) or 'fair' ($k = 7$), generally due to inadequate sample size or description of

Table IV
Pooled subscale values, Cronbach's alpha, ICC, SEM and SDC (overall, and split by age and condition subgroups)

	Pain mean (SD)	Symptoms mean (SD)	ADL mean (SD)	Sport/recreation mean (SD)	QoL mean (SD)	KOOS-PS mean (SD)
Pooled values (baseline scores)						
Overall	54.7 (19.2)$^{31;2882}$	55.1 (18.6)$^{31;2882}$	59.1 (20.1)$^{31;2882}$	28.5 (27.3)$^{30;2598}$	35.5 (19.8)$^{30;2598}$	49.9 (15.4)$^{7;880}$
Age: Young	58.8 (20.7)$^{10;786}$	55.8 (18.7)$^{10;786}$	64.3 (21.2)$^{10;786}$	35.6 (25.4)$^{10;786}$	32.5 (19.1)$^{10;786}$	–
Age: Old	50.3 (17.1)$^{16;1345}$	52 (18)$^{16;1345}$	53.1 (17.8)$^{16;1345}$	21.3 (24.4)$^{16;1345}$	34.7 (18.9)$^{16;1345}$	51.8 (14.5)$^{5;666}$
Condition: OA	50.7 (17.9)$^{18;1691}$	52.9 (18.4)$^{18;1691}$	54.2 (18.6)$^{18;1691}$	21.9 (24.9)$^{17;1407}$	35.1 (19.2)$^{17;1407}$	49.9 (15.4)$^{7;880}$
Condition: FCL	62.8 (19.8)$^{4;218}$	64.5 (20.2)$^{4;218}$	71.5 (20)$^{4;218}$	31.3 (26.9)$^{4;218}$	32.3 (21.4)$^{4;218}$	–
Condition: ACL	78 (13.3)$^{3;150}$	69.7 (13.8)$^{3;150}$	80.4 (11.9)$^{3;150}$	49.9 (21.6)$^{3;150}$	43.5 (16.7)$^{3;150}$	–
Condition: Mixed	55.3 (21.8)$^{5;787}$	53.4 (19.2)$^{5;787}$	60.6 (23.7)$^{5;787}$	33.5 (26.5)$^{5;787}$	34.2 (20.3)$^{5;787}$	–
	Pain mean (95% CI)	Symptoms mean (95% CI)	ADL mean (95% CI)	Sport/recreation mean (95% CI)	QoL mean (95% CI)	KOOS-PS mean (95% CI)
Pooled Cronbach's α						
Overall	0.86 (0.86–0.86)$^{24;2241}$	0.72 (0.71–0.72)$^{24;2241}$	0.93 (0.92–0.93)$^{25;2441}$	0.89 (0.89–0.89)$^{25;2441}$	0.79 (0.79–0.79)$^{24;2241}$	0.89 (0.89–0.89)$^{6;2778}$
Age: Young	0.86 (0.86–0.86)$^{10;959}$	0.68 (0.67–0.69)$^{10;959}$	0.92 (0.92–0.92)$^{10;959}$	0.86 (0.86–0.87)$^{10;959}$	0.76 (0.76–0.77)$^{10;959}$	–
Age: Old	0.83 (0.82–0.83)$^{10;815}$	0.73 (0.73–0.74)$^{10;815}$	0.92 (0.92–0.92)$^{11;1015}$	0.9 (0.89–0.9)$^{11;1015}$	0.8 (0.79–0.8)$^{10;815}$	0.89 (0.89–0.89)$^{5;2698}$
Condition: OA	0.84 (0.83–0.84)$^{11;877}$	0.74 (0.73–0.74)$^{11;877}$	0.92 (0.92–0.92)$^{12;1077}$	0.9 (0.9–0.9)$^{12;1077}$	0.8 (0.8–0.8)$^{11;877}$	0.86 (0.86–0.86)$^{5;741}$
Condition: FCL	0.9 (0.9–0.9)$^{3;114}$	0.77 (0.76–0.79)$^{3;114}$	0.96 (0.96–0.96)$^{3;114}$	0.9 (0.9–0.91)$^{3;114}$	0.82 (0.81–0.84)$^{3;114}$	–
Condition: ACL	0.85 (0.84–0.85)$^{3;329}$	0.72 (0.71–0.74)$^{3;329}$	0.93 (0.92–0.93)$^{3;329}$	0.85 (0.84–0.85)$^{3;329}$	0.79 (0.78–0.8)$^{3;329}$	–
Condition: Mixed	0.88 (0.88–0.89)$^{6;885}$	0.68 (0.67–0.7)$^{6;885}$	0.93 (0.93–0.94)$^{6;885}$	0.89 (0.89–0.9)$^{6;885}$	0.77 (0.77–0.78)$^{6;885}$	0.9 (–)$^{1;2037}$
Pooled ICC						
Overall	0.89 (0.87–0.91)$^{26;1616}$	0.87 (0.84–0.89)$^{26;1616}$	0.9 (0.87–0.92)$^{26;1616}$	0.85 (0.8–0.89)$^{26;1616}$	0.87 (0.83–0.9)$^{26;1616}$	0.86 (0.81–0.89)$^{4;207}$
Age: Young	0.9 (0.86–0.93)$^{11;717}$	0.88 (0.83–0.91)$^{11;717}$	0.9 (0.84–0.93)$^{11;717}$	0.86 (0.77–0.91)$^{11;717}$	0.89 (0.84–0.93)$^{11;717}$	–
Age: Old	0.9 (0.86–0.93)$^{10;529}$	0.88 (0.83–0.92)$^{10;529}$	0.89 (0.84–0.93)$^{10;529}$	0.83 (0.73–0.9)$^{10;529}$	0.84 (0.77–0.89)$^{10;529}$	0.86 (0.81–0.9)$^{3;127}$
Condition: OA	0.9 (0.85–0.92)$^{12;612}$	0.87 (0.83–0.91)$^{12;612}$	0.89 (0.84–0.93)$^{12;612}$	0.83 (0.73–0.89)$^{12;612}$	0.84 (0.78–0.88)$^{12;612}$	0.86 (0.81–0.89)$^{4;207}$
Condition: FCL	0.9 (0.8–0.95)$^{3;93}$	0.91 (0.82–0.96)$^{3;93}$	0.93 (0.82–0.97)$^{3;93}$	0.92 (0.79–0.97)$^{3;93}$	0.95 (0.9–0.98)$^{3;93}$	–
Condition: ACL	0.89 (0.8–0.94)$^{4;192}$	0.87 (0.78–0.93)$^{4;192}$	0.87 (0.73–0.94)$^{4;192}$	0.84 (0.66–0.93)$^{4;192}$	0.88 (0.78–0.93)$^{4;192}$	–
Condition: Mixed	0.91 (0.86–0.94)$^{6;684}$	0.86 (0.79–0.91)$^{6;684}$	0.91 (0.84–0.95)$^{6;684}$	0.85 (0.73–0.92)$^{6;684}$	0.86 (0.79–0.91)$^{6;684}$	–
Pooled SEM						
Overall	6 (5.8–6.2)$^{9;525}$	5.2 (4.9–5.4)$^{9;525}$	5.7 (5.4–6)$^{9;525}$	9.1 (8.6–9.6)$^{9;525}$	6.8 (6.6–7)$^{9;525}$	6.7(–)$^{1;112}$
Age: Young	5.2 (4.9–5.4)$^{5;362}$	5.9 (5.6–6.2)$^{5;362}$	5.2 (4.9–5.4)$^{5;362}$	7.1 (6.7–7.4)$^{5;\ 362}$	6.4 (6.2–6.6)$^{5;\ 362}$	–
Age: Old	10.1 (–)$^{1;47}$	7.2 (–)$^{1;47}$	11.7 (–)$^{1;47}$	24.6 (–)$^{1;47}$	10.8 (–)$^{1;47}$	–
Condition: OA	9.5 (9.4–9.6)$^{2;100}$	7.6 (7.5–7.7)$^{2;100}$	8.6 (8–9.2)$^{2;100}$	17.7 (16.4–19)$^{2;100}$	9 (8.7–9.3)$^{2;100}$	6.7(–)$^{1;112}$
Condition: FCL	9.5 (–)$^{1;54}$	11.1 (–)$^{1;54}$	10.7 (–)$^{1;54}$	10.8 (–)$^{1;54}$	7.4 (–)$^{1;54}$	–
Condition: ACL	4.6 (4.3–4.9)$^{3;179}$	5.6 (5.3–6)$^{3;179}$	4.7 (4.4–5)$^{3;179}$	7 (6.4–7.6)$^{3;179}$	5.9 (5.6–6.2)$^{3;179}$	–
Condition: Mixed	3.9 (3.8–4)$^{2;157}$	3.8 (3.7–3.9)$^{2;157}$	3.3 (3.3–3.4)$^{2;157}$	5.4 (5.3–5.5)$^{2;157}$	6.1 (5.9–6.3)$^{2;157}$	–
Pooled SDC						
Overall	16.6 (16–17.2)$^{9;525}$	17.4 (16.7–18)$^{9;525}$	15.7 (15–16.4)$^{9;525}$	25.1 (23.7–26.5)$^{9;525}$	18.8 (18.3–19.3)$^{9;525}$	18.6(–)$^{1;112}$
Age: Young	14.3 (13.6–14.9)$^{5;362}$	16.4 (15.5–17.2)$^{5;362}$	14.3 (13.5–15.1)$^{5;362}$	19.6 (18.7–20.6)$^{5;362}$	17.8 (17.3–18.2)$^{5;362}$	–
Age: Old	28 (–)$^{1;47}$	20 (–)$^{1;47}$	32.4 (–)$^{1;47}$	68.2 (–)$^{1;47}$	29.9 (–)$^{1;47}$	–
Condition: OA	26.4 (26–26.7)$^{2;100}$	21.2 (20.9–21.4)$^{2;100}$	23.8 (22.1–25.4)$^{2;100}$	49.1 (45.5–52.7)$^{2;100}$	24.9 (24–25.9)$^{2;100}$	18.6(–)$^{1;112}$
Condition: FCL	26.3 (–)$^{1;54}$	30.8 (–)$^{1;54}$	29.7 (–)$^{1;54}$	29.9 (–)$^{1;54}$	20.5 (–)$^{1;54}$	–
Condition: ACL	12.7 (12–13.4)$^{3;179}$	15.6 (14.6–16.5)$^{3;179}$	12.9 (12.1–13.7)$^{3;179}$	19.2 (17.6–20.9)$^{3;179}$	16.3 (15.4–17.2)$^{3;179}$	–
Condition: Mixed	10.8 (10.5–11)$^{2;157}$	10.7 (10.4–10.9)$^{2;157}$	9.2 (9–9.4)$^{2;157}$	14.9 (14.5–15.2)$^{2;157}$	16.9 (16.3–17.5)$^{2;157}$	–

Cronbach's alpha was used as a measure of internal consistency, ICC as a measure of test-retest reliability, and SEM and SDC as measures of measurement error. Numbers in superscript represent number of cohorts; total number of participants. FCL: focal cartilage lesion; ACL: anterior cruciate ligament injury; SD: standard deviation; SEM: standard error of measurement. (–) 95% CI not applicable (only 1 study in subgroup).

missing items. Most studies scored 'fair' for test-retest reliability ($k = 16$) and measurement error ($k = 7$), largely due to dissimilar test conditions or inadequate time interval between testing occasions (<1 week or >4 weeks between administrations). Construct validity studies scored mostly 'fair' ($k = 13$) or 'good' ($k = 9$), typically for not adequately describing the comparator instrument. For cross-cultural translation, 12/13 papers scored 'fair' for not reporting independent translators or describing pre-test sample characteristics. There was considerable variability in the quality of studies evaluating structural validity and unidimensionality. Only one study scored 'excellent', which evaluated unidimensionality of the individual subscales. Only 2/19 responsiveness studies utilized COSMIN-recommended hypothesis testing for expected correlations with a comparator instrument ('good', $k = 1$; 'fair', $k = 1$). The remaining 17 studies were rated 'poor' ($k = 12$) or 'fair' ($k = 5$) due to lack of comparator instrument or formulation of *a priori* hypotheses.

When meta-analyses were rerun for each COSMIN rating, there was no change in outcome for test-retest reliability (across studies rated as 'fair' or 'poor') or construct validity (across studies rated as 'good', 'fair' or 'poor') (Supplementary Tables S13 and S14). The Symptoms subscale did not demonstrate adequate internal consistency when data from studies rated as 'poor' were pooled ($\alpha = 0.67$).

Best evidence synthesis

Table VI presents the best evidence synthesis for each measurement property.

Discussion

Findings suggest that KOOS can be used to evaluate young and old patients, and those with different knee conditions, in terms of internal consistency, test-retest reliability and construct validity. As intended, ADL has better content validity for older patients, Sport/Rec has better content validity for younger patients, and Pain is more relevant for painful conditions. We are unable to conclude regarding the five-factor structure of the KOOS, due to limitations with sample size and methodology. Subscales demonstrate

Table V
Summary of construct validity data. Values represent weighted mean correlation coefficients (95% confidence intervals)

		SF-36 bodily Pain	SF-36 physical function	SF-36 role-emotional	SF-36 mental health
Overall	Pain	0.62 (0.61–0.62)$^{23;2064}$	0.55 (0.55–0.56)$^{22;2024}$	0.22 (0.22–0.23)$^{21;1924}$	0.27 (0.27–0.28)$^{21;1924}$
	Symptoms	0.47 (0.47–0.48)$^{22;2024}$	0.42 (0.41–0.42)$^{23;2064}$	0.18 (0.17–0.19)$^{21;1924}$	0.22 (0.22–0.23)$^{21;1924}$
	ADL	0.61 (0.6–0.61)$^{23;2104}$	0.67 (0.66–0.67)$^{24;2144}$	0.32 (0.31–0.32)$^{21;1924}$	0.32 (0.31–0.32)$^{21;1924}$
	Sport/Recreation	0.47 (0.46–0.48)$^{22;2024}$	0.54 (0.53–0.55)$^{22;2024}$	0.22 (0.22–0.23)$^{21;1924}$	0.18 (0.18–0.19)$^{21;1924}$
	QoL	0.53 (0.52–0.53)$^{22;2024}$	0.52 (0.51–0.52)$^{22;2024}$	0.27 (0.27–0.28)$^{21;1924}$	0.28 (0.27–0.29)$^{21;1924}$
	KOOS-PS	0.69 (–)$^{1;85}$	0.58 (–)$^{1;85}$	0.27 (–)$^{1;85}$	0.04 (–)$^{1;85}$
Age: Young	Pain	0.68 (0.68–0.69)$^{11;886}$	0.57 (0.57–0.58)$^{10;846}$	0.22 (0.21–0.22)$^{9;746}$	0.31 (0.3–0.33)$^{9;746}$
	Symptoms	0.43 (0.42–0.44)$^{10;846}$	0.41 (0.4–0.43)$^{11;886}$	0.15 (0.14–0.16)$^{9;746}$	0.22 (0.2–0.23)$^{9;746}$
	ADL	0.64 (0.64–0.65)$^{10;846}$	0.67 (0.67–0.68)$^{11;886}$	0.28 (0.27–0.29)$^{9;746}$	0.3 (0.29–0.31)$^{9;746}$
	Sport/Recreation	0.53 (0.52–0.54)$^{10;846}$	0.61 (0.6–0.61)$^{10;846}$	0.27 (0.26–0.28)$^{9;746}$	0.23 (0.22–0.24)$^{9;746}$
	QoL	0.5 (0.5–0.5)$^{10;846}$	0.5 (0.5–0.51)$^{10;846}$	0.25 (0.24–0.26)$^{9;746}$	0.25 (0.23–0.26)$^{9;746}$
	KOOS-PS	–	–	–	–
Age: Old	Pain	0.53 (0.52–0.54)$^{9;856}$	0.48 (0.47–0.48)$^{9;856}$	0.19 (0.18–0.2)$^{9;856}$	0.21 (0.2–0.22)$^{9;856}$
	Symptoms	0.5 (0.49–0.5)$^{9;856}$	0.38 (0.37–0.39)$^{9;856}$	0.16 (0.15–0.16)$^{9;856}$	0.22 (0.21–0.23)$^{9;856}$
	ADL	0.54 (0.52–0.55)$^{9;856}$	0.63 (0.63–0.64)$^{9;856}$	0.33 (0.32–0.34)$^{9;856}$	0.3 (0.29–0.31)$^{9;856}$
	Sport/Recreation	0.39 (0.38–0.41)$^{9;856}$	0.45 (0.43–0.46)$^{9;856}$	0.17 (0.16–0.18)$^{9;856}$	0.1 (0.1–0.11)$^{9;856}$
	QoL	0.54 (0.54–0.55)$^{9;856}$	0.5 (0.49–0.51)$^{9;856}$	0.26 (0.25–0.27)$^{9;856}$	0.28 (0.27–0.28)$^{9;856}$
	KOOS-PS	0.69 (–)$^{1;85}$	0.58 (–)$^{1;85}$	0.27 (–)$^{1;85}$	0.04 (–)$^{1;85}$
Condition: OA	Pain	0.54 (0.53–0.55)$^{10;918}$	0.49 (0.48–0.5)$^{10;918}$	0.2 (0.19–0.21)$^{10;918}$	0.21 (0.2–0.22)$^{10;918}$
	Symptoms	0.5 (0.5–0.51)$^{10;918}$	0.4 (0.39–0.4)$^{10;918}$	0.17 (0.17–0.18)$^{10;918}$	0.23 (0.22–0.24)$^{10;918}$
	ADL	0.56 (0.55–0.57)$^{11;998}$	0.65 (0.64–0.66)$^{11;998}$	0.34 (0.33–0.35)$^{10;918}$	0.3 (0.29–0.3)$^{10;918}$
	Sport/Recreation	0.39 (0.38–0.41)$^{10;918}$	0.46 (0.44–0.47)$^{10;918}$	0.17 (0.16–0.18)$^{10;918}$	0.1 (0.09–0.11)$^{10;918}$
	QoL	0.55 (0.54–0.56)$^{10;918}$	0.51 (0.5–0.52)$^{10;918}$	0.27 (0.26–0.28)$^{10;918}$	0.27 (0.27–0.28)$^{10;918}$
	KOOS-PS	0.69 (–)$^{1;85}$	0.58 (–)$^{1;85}$	0.27 (–)$^{1;85}$	0.04 (–)$^{1;85}$
Condition: FCL	Pain	0.72 (0.71–0.73)$^{3;113}$	0.72 (0.72–0.73)$^{2;73}$	0.39 (0.38–0.39)$^{2;73}$	0.53 (0.53–0.53)$^{2;73}$
	Symptoms	0.63 (0.62–0.65)$^{2;73}$	0.59 (0.58–0.6)$^{3;113}$	0.29 (0.28–0.3)$^{2;73}$	0.4 (0.38–0.42)$^{2;73}$
	ADL	0.74 (0.72–0.75)$^{2;73}$	0.67 (0.65–0.69)$^{3;113}$	0.43 (0.41–0.44)$^{2;73}$	0.61 (0.61–0.61)$^{2;73}$
	Sport/Recreation	0.7 (0.69–0.7)$^{2;73}$	0.63 (0.62–0.65)$^{2;73}$	0.36 (0.35–0.37)$^{2;73}$	0.35 (0.34–0.37)$^{2;73}$
	QoL	0.57 (0.54–0.6)$^{2;73}$	0.62 (0.61–0.64)$^{2;73}$	0.31 (0.3–0.32)$^{2;73}$	0.36 (0.32–0.4)$^{2;73}$
	KOOS-PS	–	–	–	–
Condition: ACL	Pain	0.66 (0.65–0.67)$^{4;250}$	0.57 (0.56–0.58)$^{4;250}$	0.26 (0.23–0.28)$^{3;150}$	0.28 (0.25–0.31)$^{3;150}$
	Symptoms	0.42 (0.4–0.43)$^{4;250}$	0.4 (0.39–0.41)$^{4;250}$	0.19 (0.16–0.22)$^{3;150}$	0.24 (0.21–0.27)$^{3;150}$
	ADL	0.57 (0.56–0.58)$^{4;250}$	0.62 (0.61–0.63)$^{4;250}$	0.26 (0.23–0.28)$^{3;150}$	0.24 (0.22–0.27)$^{3;150}$
	Sport/Recreation	0.49 (0.48–0.5)$^{4;250}$	0.65 (0.64–0.65)$^{4;250}$	0.29 (0.27–0.31)$^{3;150}$	0.24 (0.21–0.26)$^{3;150}$
	QoL	0.4 (0.39–0.42)$^{4;250}$	0.48 (0.47–0.5)$^{4;250}$	0.24 (0.22–0.27)$^{3;150}$	0.3 (0.27–0.32)$^{3;150}$
	KOOS-PS	–	–	–	–
Condition: Mixed	Pain	0.69 (0.68–0.69)$^{5;747}$	0.6 (0.59–0.61)$^{5;747}$	0.22 (0.22–0.23)$^{5;747}$	0.31 (0.31–0.32)$^{5;747}$
	Symptoms	0.44 (0.43–0.45)$^{5;747}$	0.42 (0.41–0.43)$^{5;747}$	0.18 (0.17–0.18)$^{5;747}$	0.19 (0.18–0.21)$^{5;747}$
	ADL	0.67 (0.66–0.67)$^{5;747}$	0.7 (0.7–0.71)$^{5;747}$	0.29 (0.29–0.3)$^{5;747}$	0.33 (0.31–0.34)$^{5;747}$
	Sport/Recreation	0.54 (0.53–0.55)$^{5;747}$	0.6 (0.6–0.61)$^{5;747}$	0.26 (0.26–0.27)$^{5;747}$	0.26 (0.25–0.27)$^{5;747}$
	QoL	0.55 (0.54–0.55)$^{5;747}$	0.52 (0.52–0.53)$^{5;747}$	0.28 (0.27–0.29)$^{5;747}$	0.27 (0.25–0.28)$^{5;747}$
	KOOS-PS	–	–	–	–

Numbers in superscript represent number of cohorts; total number of participants. (–) 95% CI not applicable (only 1 study in subgroup).
SF-36 subscales are scored from 0 (worst health) to 100 (best health). KOOS is scored 0 (worst) to 100 (best), whereas KOOS-PS was scored 0 (best) to 100 (worst), providing negative correlations with SF-36 to reflect positive relationships (i.e., best health on SF-36 and best score on KOOS-PS). Negative values have been removed from table for consistent presentation of results.

sufficient unidimensionality to be scored individually, which should be confirmed using recommended methods. Although measurement error requires further evaluation, change in a young patient's score between 14.3 and 19.6 represents a true change, while for older patients a change of at least 20 is required for all subscales. Evidence of responsiveness comes from larger effect sizes following surgical than non-surgical interventions, especially TKR, although effect size magnitudes did not change over time post-TKR.

Recommendations for future studies of KOOS measurement properties

Although current guidelines state that evaluation of test-retest reliability is required for all new language versions of PROMs, our data suggest otherwise. Consistent findings of test-retest reliability across groups that have used different language versions of KOOS suggest that test-retest reliability should not be a priority when developing new language versions. Rather, resources should be directed towards evaluation of other measurement properties, such as content validity, internal consistency and structural validity. A limitation of included studies was that cross-cultural validation was not performed during language translation. This should be evaluated for existing and new language versions.

Because the SDC has clinical utility for interpreting change scores in individual patients, future studies should prioritize evaluation of KOOS measurement error in different age and condition subgroups. Notably, SDC can be calculated from existing ICC data, negating the need for new data collection. While KOOS is sufficiently sensitive to detect changes in surgical and non-surgical interventions, our findings highlight the need for further responsiveness evaluation using COSMIN-recommended methods. Studies evaluating structural validity, following COSMIN recommendations for 'excellent' or 'good' methodological quality, are required to confirm KOOS's five-factor structure, although our findings suggest that overlap between subscales will be identified (e.g., Pain, ADL).

KOOS-PS requires further evaluation regarding its measurement properties, particularly to confirm construct validity findings across multiple cohorts. However, findings that a 12-item short form evaluating physical function may demonstrate more optimal measurement properties[49] highlights the need for further research on short form structure.

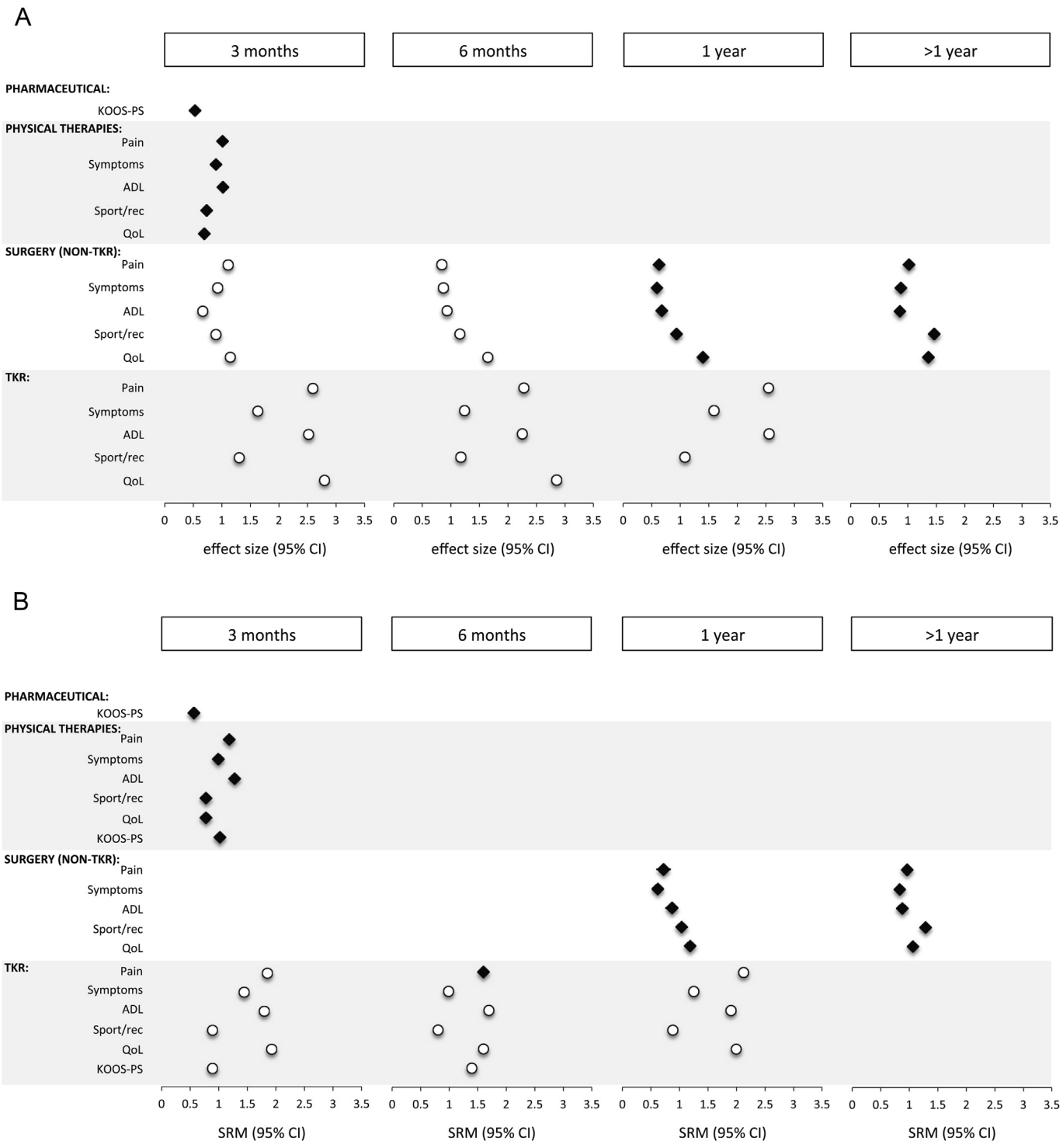

Fig. 2. KOOS responsiveness. Black diamonds represent weighted mean effect sizes (A) or SRM (B) with 95% CIs. White circles represent effect sizes or SRM from individual study data.

Recommendations regarding hierarchical testing of KOOS endpoints in clinical trials

When using KOOS subscales as endpoints in clinical trials, it is recommended that multiplicity be addressed by hierarchical testing of endpoints, defined *a priori* and reported in the study protocol[57]. Based on our findings, we suggest the following hierarchical orders. In patients with ACL injury, QoL and Sport/Rec have the highest content validity, demonstrate the greatest room for improvement, and show largest effect sizes following surgical ACL reconstruction. We recommend that these are tested first and second in clinical trials of ACL patients. For patients with knee OA undergoing TKR, we recommend that Pain, ADL and QoL are tested first, second and third, as they have the greatest relevance and room for improvement, and show largest effect sizes post-TKR.

Table VI
Summary of findings for each measurement property (adapted from[19])

Source	Content validity*	Structural validity	Unidimensionality†	Internal consistency‡	Test-retest reliability	Measurement error	Construct validity	Responsiveness
Pain	±	?	±	+++	++	?	++	+
Symptoms	++	?	±	+++	++	?	++	+
ADL	±	?	±	+++	++	?	++	+
Sport/recreation	±	?	+++	+++	++	?	+++	+
QoL	++	?	+++	+++	++	?	+++	+
KOOS-PS	?	?	±	+++	++	?	+	?

+++ or — Strong evidence. ++ or – Moderate evidence. + or - Limited evidence. ± Conflicting evidence or population-specific properties. ? Unknown evidence. See Table II for definitions.

* Methodological quality not evaluated using COSMIN checklist, and not considered in applying recommendations.

† Recommendations based on studies that utilized principal component analysis or Rasch analysis.

‡ Internal consistency summary based on Cronbach's alpha, and preliminary evidence that KOOS subscales are likely to be sufficiently unidimensional.

Strengths and limitations

This is one of few systematic reviews using meta-analyses to pool data on measurement properties. Although we did not restrict inclusion by publication language, we didn't have adequate resources to conduct full translation of four non-English papers. Although we could extract measurement property data, inability to perform COSMIN ratings means that the methodological quality of these studies is unknown.

Conclusion

KOOS users can be confident of content validity, internal consistency, test-retest reliability, construct validity and responsiveness for age- and condition-relevant subscales. Structural validity, cross-cultural validity and measurement error require further evaluation, and construct validity of KOOS-PS. Guidelines provided can assist researchers and clinicians to make decisions on using KOOS in patients with knee conditions, and in hierarchical testing of endpoints for clinical trials.

Author contributions

EMR conceived the study. EMR, NJC, RC, EMB and CBT contributed to the study design. EMB, NJC and EMR performed the literature search. NJC and EMR screened articles for eligibility. NJC and CACP performed quality ratings. NJC extracted data. NJC and RC performed statistical analyses. NJC and EMR wrote the manuscript. CACP, RC, EMB and CBT edited the manuscript. All authors read and approved the final manuscript. Dr Collins (n.collins1@uq.edu.au) takes responsibility for the integrity of the work as a whole.

Competing interests

Professor Roos is the developer of the KOOS. The KOOS is freely available, with no licencing required for academic or commercial use. Dr Terwee is one of the developers of the COSMIN checklist. Other authors declare no conflicts of interest.

Funding sources

NJC was supported by a National Health and Medical Research Council (Australia) Research Training (Post-Doctoral) Fellowship (628918), and is currently supported by a UQ Postdoctoral Research Fellowship. RC and EMB are supported by unrestricted grants from the Oak Foundation. EMR is supported by the Swedish Medical Research Council. Funding bodies had no role in the study design, writing of the manuscript or in the decision to submit the manuscript for publication.

Acknowledgements

The authors wish to thank the researchers and students who assisted with extracting information from non-English papers.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.joca.2016.03.010.

References

1. Patrick DL, Burke LB, Powers JH, Scott JA, Rock EP, Dawisha S, *et al.* Patient-reported outcomes to support medical product labeling claims: FDA perspective. Value Health 2007;10(Suppl 2):S125–37.
2. Tugwell P, Boers M, D'Agostino MA, Beaton D, Boonen A, Bingham 3rd CO, *et al.* Updating the OMERACT filter: implications of filter 2.0 to select outcome instruments through assessment of "truth": content, face, and construct validity. J Rheumatol 2014;41:1000–4.
3. Boers M, Kirwan JR, Wells G, Beaton D, Gossec L, d'Agostino MA, *et al.* Developing core outcome measurement sets for clinical trials: OMERACT filter 2.0. J Clin Epidemiol 2014;67:745–53.
4. Roos EM, Roos HP, Lohmander LS, Ekdahl C, Beynnon BD. Knee Injury and Osteoarthritis Outcome Score (KOOS)—development of a self-administered outcome measure. J Orthop Sports Phys Ther 1998;28:88–96.
5. Perruccio AV, Stefan Lohmander L, Canizares M, Tennant A, Hawker GA, Conaghan PG, *et al.* The development of a short measure of physical function for knee OA KOOS-Physical Function Shortform (KOOS-PS) – an OARSI/OMERACT initiative. Osteoarthritis Cartilage 2008;16:542–50.
6. Roos EM. http://www.koos.nu. vol.s 2015.
7. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, *et al.* The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. J Clin Epidemiol 2009;62:e1–e34.
8. 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. J Clin Epidemiol 2007;60:34–42.
9. 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. J Clin Epidemiol 2010;63:737–45.
10. Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res 2012;21:651–7.
11. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, *et al*. COSMIN Checklist Manual. VU University Medical Centre; 2012.
12. Collins NJ, Roos EM. Patient-reported outcomes for total hip and knee arthroplasty: commonly used instruments and attributes of a "good" measure. Clin Geriatric Med 2012;28:367–94.
13. Reeve BB, Wyrwich KW, Wu AW, Velikova G, Terwee CB, Snyder CF, *et al*. ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. Qual Life Res 2013;22:1889–905.
14. Nunnally JC, Bernstein IH. Psychometric Theory. 3rd edn. New York: McGraw-Hill; 1994.
15. de Vet HC, Terwee CB, Mokkink LB, Knol DL. Measurement in Medicine: A Practical Guide. Cambridge: Cambridge University Press; 2011.
16. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986;7:177–88.
17. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003;327:557–60.
18. Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, *et al*. GRADE guidelines: 7. Rating the quality of evidence–inconsistency. J Clin Epidemiol 2011;64:1294–302.
19. Schellingerhout JM, Verhagen AP, Heymans MW, Koes BW, de Vet HC, Terwee CB. Measurement properties of disease-specific questionnaires in patients with neck pain: a systematic review. Qual Life Res 2012;21:659–70.
20. Roos EM, Roos HP, Ekdahl C, Lohmander LS. Knee injury and osteoarthritis outcome score (KOOS)–validation of a Swedish version. Scand J Med Sci Sports 1998;8:439–48.
21. Kessler S, Lang S, Puhl W, Stove J. The Knee Injury and Osteoarthritis Outcome Score–a multifunctional questionnaire to measure outcome in knee arthroplasty. Z Orthop Ihre Grenzgeb 2003;141:277–82.
22. Roos EM, Toksvig-Larsen S. Knee injury and Osteoarthritis Outcome Score (KOOS) – validation and comparison to the WOMAC in total knee replacement. Health Qual Life Outcomes 2003;1:17.
23. Seo SS, Chung KC, Kim YB. Assessment of validity, reliability and responsiveness of Korean Knee Injury and Osteoarthritis Outcome Score (KOOS) for the knee injury. J Korean Orthop Assoc 2006;41:441–53.
24. Xie F, Li SC, Roos EM, Fong KY, Lo NN, Yeo SJ, *et al*. Cross-cultural adaptation and validation of Singapore English and Chinese versions of the Knee injury and Osteoarthritis Outcome Score (KOOS) in Asians with knee osteoarthritis in Singapore. Osteoarthritis Cartilage 2006;14:1098–103.
25. Paker N, Bugdayci D, Sabirli F, Ozel S, Ersoy S. Knee injury and osteoarthritis outcome score: reliability and validation of the Turkish version. Turkiye Klinikleri J Med Sci 2007;27:350–6.
26. Comins J, Brodersen J, Krogsgaard M, Beyer N. Rasch analysis of the Knee injury and Osteoarthritis Outcome Score (KOOS): a statistical re-evaluation. Scand J Med Sci Sports 2008;18:336–45.
27. de Groot IB, Favejee MM, Reijman M, Verhaar JA, Terwee CB. The Dutch version of the knee injury and osteoarthritis outcome score: a validation study. Health Qual Life Outcomes 2008;6:16.
28. Ornetti P, Parratte S, Gossec L, Tavernier C, Argenson JN, Roos EM, *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;16:423–8.
29. Salavati M, Mazaheri M, Negahban H, Sohani SM, Ebrahimian MR, Ebrahimi I, *et al*. Validation of a Persian-version of Knee injury and Osteoarthritis Outcome Score (KOOS) in Iranians with knee injuries. Osteoarthritis Cartilage 2008;16:1178–82.
30. Bekkers JE, de Windt TS, Raijmakers NJ, Dhert WJ, Saris DB. Validation of the Knee Injury and Osteoarthritis Outcome Score (KOOS) for the treatment of focal cartilage lesions. Osteoarthritis Cartilage 2009;17:1434–9.
31. Chaipinyo K. Test-retest reliability and construct validity of Thai version of Knee Osteoarthritis Outcome Score (KOOS). Thai J Phys Ther 2009;31:67–76.
32. Davis AM, Perruccio AV, Canizares M, Hawker GA, Roos EM, Maillefert JF, *et al*. Comparative, validity and responsiveness of the HOOS-PS and KOOS-PS to the WOMAC physical function subscale in total joint replacement for osteoarthritis. Osteoarthritis Cartilage 2009;17:843–7.
33. Goncalves RS, Cabri J, Pinheiro JP, Ferreira PL. Cross-cultural adaptation and validation of the Portuguese version of the Knee injury and Osteoarthritis Outcome Score (KOOS). Osteoarthritis Cartilage 2009;17:1156–62.
34. Ornetti P, Perruccio AV, Roos EM, Lohmander LS, Davis AM, Maillefert JF. Psychometric properties of the French translation of the reduced KOOS and HOOS (KOOS-PS and HOOS-PS). Osteoarthritis Cartilage 2009;17:1604–8.
35. Davis AM, Lohmander LS, Wong R, Venkataramanan V, Hawker GA. Evaluating the responsiveness of the ICOAP following hip or knee replacement. Osteoarthritis Cartilage 2010;18:1043–5.
36. Goncalves RS, Cabri J, Pinheiro JP, Ferreira PL, Gil J. Reliability, validity and responsiveness of the Portuguese version of the Knee injury and Osteoarthritis Outcome Score–Physical Function Short-form (KOOS-PS). Osteoarthritis Cartilage 2010;18:372–6.
37. Gudbergsen H, Bartels EM, Krusager P, Waehrens EE, Christensen R, Danneskiold-Samsoe B, *et al*. Test-retest of computerized health status questionnaires frequently used in the monitoring of knee osteoarthritis: a randomized crossover trial. BMC Musculoskelet Disord 2011;12:190.
38. Nakamura N, Takeuchi R, Sawaguchi T, Ishikawa H, Saito T, Goldhahn S. Cross-cultural adaptation and validation of the Japanese Knee Injury and Osteoarthritis Outcome Score (KOOS). J Orthop Sci 2011;16:516–23.
39. Ruyssen-Witrand A, Fernandez-Lopez CJ, Gossec L, Anract P, Courpied JP, Dougados M. Psychometric properties of the OARSI/OMERACT osteoarthritis pain and functional impairment scales: ICOAP, KOOS-PS and HOOS-PS. Clin Exp Rheumatol 2011;29:231–7.
40. Salavati M, Akhbari B, Mohammadi F, Mazaheri M, Khorrami M. Knee injury and Osteoarthritis Outcome Score (KOOS); reliability and validity in competitive athletes after anterior cruciate ligament reconstruction. Osteoarthritis Cartilage 2011;19:406–10.
41. Bond M, Davis A, Lohmander S, Hawker G. Responsiveness of the OARSI-OMERACT osteoarthritis pain and function measures. Osteoarthritis Cartilage 2012;20:541–7.
42. Briem K. Reliability, validity and responsiveness of the Icelandic version of the knee injury and osteoarthritis outcome score (KOOS). Laeknabladid 2012;98:403–7.

43. Engelhart L, Nelson L, Lewis S, Mordin M, Demuro-Mercon C, Uddin S, *et al*. Validation of the Knee Injury and Osteoarthritis Outcome Score subscales for patients with articular cartilage lesions of the knee. Am J Sports Med 2012;40:2264–72.
44. Hoogeboom TJ, de Bie RA, den Broeder AA, van den Ende CH. The Dutch lower extremity functional scale was highly reliable, valid and responsive in individuals with hip/knee osteoarthritis: a validation study. BMC Musculoskelet Disord 2012;13:117.
45. Monticone M, Ferrante S, Salvaderi S, Rocca B, Totti V, Foti C, *et al*. Development of the Italian version of the knee injury and osteoarthritis outcome score for patients with knee injuries: cross-cultural adaptation, dimensionality, reliability, and validity. Osteoarthritis Cartilage 2012;20:330–5.
46. Vaquero J, Longo UG, Forriol F, Martinelli N, Vethencourt R, Denaro V. Reliability, validity and responsiveness of the Spanish version of the Knee Injury and Osteoarthritis Outcome Score (KOOS) in patients with chondral lesion of the knee. Knee Surg Sports Traumatol Arthrosc 2014;22:104–8.
47. Almangoush A, Herrington L, Attia I, Jones R, Aldawoudy A, Abdul Aziz A, *et al*. Cross-cultural adaptation, reliability, internal consistency and validation of the Arabic version of the Knee injury and Osteoarthritis Outcome Score (KOOS) for Egyptian people with knee injuries. Osteoarthritis Cartilage 2013;21:1855–64.
48. Ebert JR, Smith A, Wood DJ, Ackland TR. A comparison of the responsiveness of 4 commonly used patient-reported outcome instruments at 5 years after matrix-induced autologous chondrocyte implantation. Am J Sports Med 2013;41:2791–9.
49. Franchignoni F, Salaffi F, Giordano A, Carotti M, Ciapetti A, Ottonello M. Rasch analysis of the 22 knee injury and osteoarthritis outcome score-physical function items in Italian patients with knee osteoarthritis. Arch Phys Med Rehabil 2013;94:480–7.
50. Gul ED, Yilmaz O, Bodur H. Reliability and validity of the Turkish version of the knee injury and osteoarthritis outcome score-physical function short-form (KOOS-PS). J Back Musculoskelet Rehabil 2013;26:461–6.
51. Harris KK, Dawson J, Jones LD, Beard DJ, Price AJ. Extending the use of PROMs in the NHS—using the Oxford Knee Score in patients undergoing non-operative management for knee osteoarthritis: a validation study. BMJ Open 2013;3:e003365.
52. Monticone M, Ferrante S, Salvaderi S, Motta L, Cerri C. Responsiveness and minimal important changes for the Knee Injury and Osteoarthritis Outcome Score in subjects undergoing rehabilitation after total knee arthroplasty. Am J Phys Med Rehabil 2013;92:864–70.
53. Paradowski PT, Witonski D, Keska R, Roos EM. Cross-cultural translation and measurement properties of the polish version of the Knee injury and Osteoarthritis Outcome Score (KOOS) following anterior cruciate ligament reconstruction. Health Qual Life Outcomes 2013;11:107.
54. van Meer BL, Meuffels DE, Vissers MM, Bierma-Zeinstra SM, Verhaar JA, Terwee CB, *et al*. Knee injury and Osteoarthritis Outcome Score or International Knee Documentation Committee Subjective Knee form: which questionnaire is most useful to monitor patients with an anterior cruciate ligament rupture in the short term? Arthroscopy 2013;29:701–15.
55. Collins NJ, Misra D, Felson DT, Crossley KM, Roos EM. 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(Suppl 11):S208–28.
56. Chen WH, Lenderking W, Jin Y, Wyrwich KW, Gelhorn H, Revicki DA. Is Rasch model analysis applicable in small sample size pilot studies for assessing item characteristics? An example using PROMIS pain behavior item bank data. Qual Life Res 2014;23:485–93.
57. Roos EM, Engelhart L, Ranstam J, Anderson AF, Irrgang J, Marx RG, *et al*. ICRS recommendation document: patient-reported outcome instruments for use in patients with articular cartilage defects. Cartilage 2011;2:122–36.