BMJ open

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

Kristina K Harris,[1] Jill Dawson,[2] Luke D Jones,[1] David J Beard,[1] Andrew J Price[1]

To cite: Harris KK, Dawson J, Jones LD, *et al*. 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. doi:10.1136/bmjopen-2013-003365

▸ Prepublication history for this paper is available online. To view these files please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2013-003365).

Received 7 June 2013
Revised 19 July 2013
Accepted 22 July 2013

[1]Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Botnar Research Centre Nuffield Orthopaedic Centre, Oxford, UK
[2]Department of Population Health, University of Oxford, Oxford, UK
Correspondence to
Dr Kristina K Harris; kristina.harris@ndorms.ox.ac.uk

ABSTRACT

Objectives: To assess the validity of the Oxford Knee Score (OKS) for use in patients undergoing non-operative management for their knee osteoarthritis (OA) within the National Health Service (NHS).

Design: Observational cohort study.

Setting: Single orthopaedic centre in England.

Participants: 134 patients undergoing non-operative management for knee OA.

Main outcome measures: OKS, the Intermittent and Constant Osteoarthritis Pain (ICOAP), the Knee Injury and Osteoarthritis Score-Physical Function Short Form (KOOS-PS), at baseline and 3-month follow-up, transition item of change at 3 months.

Results: The OKS summary scale and its pain and functional component subscales demonstrated good test–retest reliability (intraclass correlation coefficient 0.93, 0.91 and 0.92, respectively) and measurement precision which, allows its use with groups of patients with knee OA (research/audit) and with individuals (clinical practice). The results in this study were consistent with a priori set hypotheses about the relationship of OKS with other validated measures (KOOS-PS, ICOAP and short form 12 (SF-12)), which provided evidence of its construct validity and responsiveness. Confirmatory factor analysis confirmed the structural validity of OKS. However, there was a lack of satisfactory evidence of structural validity for ICOAP and KOOS. The minimum detectable change (MDC_{90}) was ±6 for OKS (±16 for the Pain Component Score (OKS-PCS) and ±15 for the Functional Component Score (OKS-FCS)). Minimal important changes were ≈7 for OKS (≈17 for OKS-PCS and ≈11 for OKS-FCS) and minimal important differences were ≈6 for OKS (≈14 for OKS-PCS and ≈10 for OKS-FCS). These values were also calculated for ICOAP and KOOS-PS.

Conclusions: The OKS summary scale, together with its pain and functional component subscales, has excellent measurement properties when used with patients with knee OA undergoing non-operative treatment and is superior to ICOAP and KOOS-PS for this purpose. This evidence provides support for the validity of the use of OKS when used across the spectrum of knee OA disease severity, both in research and clinical practice.

ARTICLE SUMMARY

Article focus

- The Oxford Knee Score (OKS) is a widely used patient-reported outcome measure that was originally developed to measure the outcomes of knee replacement surgery.
- There is a growing interest to use OKS in clinical practice, across the spectrum of osteoarthritis (OA) disease.
- The aim of this study was to assess the measurement properties of OKS when used with (individuals and groups of) patients who are undergoing non-operative management for knee OA and to compare it with the most commonly used measures in this population of patients.

Key messages

- OKS, as well as its pain and functional component subscales, has acceptable evidence of its measurement properties when used in patients (individual and groups) undergoing non-operative treatment for knee OA.
- OKS performed better than the Intermittent and Constant Osteoarthritis Pain (ICOAP) and the Knee Injury and Osteoarthritis Score-Physical Function Short Form (KOOS-PS, widely used outcome measures for knee OA) on several counts.

Strengths and limitations of this study

- This study has conducted a comprehensive examination of scores' measurement properties.
- There might be a need to additionally re-evaluate evidence on some of the measurement properties presented here (such as interpretability or content validity), using different methods.
- The impact of the routine use of such scores in clinical practice should also be evaluated.

INTRODUCTION

The Oxford Knee Score (OKS) is a widely used patient-reported outcome measure (PROM), originally developed in 1998, to be used in clinical trials for assessing the patient-perceived outcomes of knee replacement surgery. In this form, it has proven to be reliable, valid and responsive.[1 2] The remit of OKS was extended in 2009 when it was adopted by the National Health Service (NHS) PROMs programme in England and Wales as a primary outcome measure for knee replacement surgery.[3] Thus, OKS data are now collected on all patients undergoing knee replacement surgery preoperatively and at 6 months postoperation, in order to monitor and benchmark the performance of health providers.

The increasing popularity of OKS has also resulted in its being used for different populations and contexts from that for which it was originally developed. In particular, there has been a growing interest in using OKS in clinical practice as a means of standardising clinical assessment, monitoring the individual's self-reported health state across the spectrum of osteoarthritis (OA) disease and using the scores as an aid to clinical decision-making. Extending the potential uses of PROMs in this manner has generally been highlighted as an opportunity to achieve maximum benefit from these measures, although the challenges of the application of such systems have also been recognised.[4 5]

Using OKS as a single score across the patient pathway to aid diagnosis, monitor progression, assist in shared decision-making and measure the outcome of intervention offers great potential for continuity of care and understanding for patients. However, robust evidence is required of the score's overall validity (ie, the consistency of its measurement properties, such as reliability), when applied in these proposed new contexts. Generally, a measure is valid when applied to populations and contexts similar to the context in which the instrument was originally developed and tested, but measurement properties may change when the measure is applied in other contexts. The fact that OKS was developed and tested to be used in the knee OA context (albeit end stage) is justification for considering its application in people with knee OA 'in general', but evidence has not been presented demonstrating that OKS remains as reliable (both on an individual and a group level), valid and responsive when used with patients who are at earlier stages of their disease management.

The principal aim of our study was to assess the measurement properties of OKS when used with (individuals and groups of) patients who are undergoing non-operative management for knee OA, by examining its reliability, validity, responsiveness and interpretability when applied in this context. Furthermore, we examined some of the measurement properties of the two most commonly used measures in this population: the Intermittent and Constant Osteoarthritis Pain (ICOAP)[6] and the Knee Injury and Osteoarthritis Score-Physical Function Short Form (KOOS-PS).[7]

METHODS

Study procedures and assessments

This study took place at an orthopaedic centre between June 2011 and August 2012. Patients were eligible for inclusion if they were referred for knee problems, had a confirmed diagnosis of knee OA and were enrolled in the non-operative management pathway for their knee OA (as recommended by the National Institute of Clinical Excellence (NICE)[8]). Treatments for patients were tailored individually, taking into account patients' preferences and needs. As such, they represented standard practice in the NHS. All patients who met these criteria were sent an invitation letter containing information about the study, consent forms and baseline questionnaires. Patients who consented to participate in the study were asked to complete the OKS,[2] the ICOAP,[6] the KOOS-PS[7] and the short form 12 (SF-12)[9] patient-reported questionnaires.

OKS is a 12-item questionnaire. Its item content was devised using patient interviews, which address pain and functional impairment in relation to their knee, in patients who are undergoing knee replacement surgery.[2] Likert responses are recommended to be scored from 0 to 4, which are summed to produce a summary score of 0 (worst) to 48 (best).[10] More recently, we presented evidence (in the context of joint replacement) that supported the original conceptual basis of OKS using its composite summary scales, but which also offered an option to perform additional analyses using pain and function subscales.[11] The Pain Component Score (OKS-PCS) consists of items 2, 3, 7, 11 and 12 and the Functional Component Score (OKS-FCS) consists of items 1, 4, 5, 6, 8, 9 and 10. Subscale raw scores are standardised from 0 (worst) to 100 (best). Patients completed OKS at baseline, 2 and 5 days (for test–retest reliability) and at 3 months.

We asked the patients to complete KOOS-PS and ICOAP at baseline and at 3-month follow-up. These scores were developed to measure pain and functional disabilities related to knee OA and are now recommended outcome measures by the Osteoarthritis Research Society International (OARSI).

KOOS-PS consists of 7 Likert-response items and was developed from a longer version of the questionnaire (KOOS[12]) using Rasch analysis to measure physical function in patients with various degrees of knee OA. It is scored as KOOS from 0 (best) to 4 (worst), with a summary raw score ranging from 0 to 28. The score is converted to a true interval score that ranges from 0 (best) to 100 (worst). ICOAP is an 11-item questionnaire whose items were informed from focus groups with patents with hip or knee OA. It has two subscales that measure the intermittent and constant pain with a standardised summary score ranging from 0 (best) to 100 (worst).

Patients also completed the generic SF-12, a 12-item general health measure with 8 items that have Likert-type response categories and 4 items with

dichotomous (yes/no) response categories. SF-12 is scored as a physical component summary (PCS) and mental component summary (MCS) ranging from 0 (worst) to 100 (best).

Lastly, we asked the patients to complete a transition question in regard to the change they experienced from the baseline measurement: "Compared to one week before your clinic visit, please indicate how much your knee problem has changed?" The question had three response options: "1. My knee has got better; 2. My knee has stayed the same; 3. My knee has got worse."

We supplemented patient reported outcome data with information on their body mass index (BMI) and the degree of structural changes observed in the knee, which was available from the patients' medical records. An orthopaedic surgeon (LDJ) performed Kellgren-Lawrence (K-L) grading using available knee OA radiographs.[13] The degree of structural changes in the knee was classified using K-L grading. In the absence of X-rays, we assessed intraoperative documentation from previous knee arthroscopy or available MRIs to examine the extent of cartilage loss and confirm the diagnosis of OA.

Statistical methods

The recommended minimum sample sizes for validation studies (based on optimal numbers for correlations) often range from 50 to 100.[14 15] For confirmatory factor analysis (CFA), the literature agrees with a minimum sample size of about 100–150 or about 10 participants per questionnaire item.[16 17] These sample sizes are required for data analyses and should be adjusted (ie, increased) for the risk of loss to follow-up. In this study, we stopped recruiting when the dataset enabled us to perform CFA with at least 10 participants per item.

We analysed the data using SPSS V.20 and LISREL V.8.80. Baseline and 3-month follow-up scores were generally non-normally distributed and change scores approximated to normal (except ICOAP and OKS-PCS). We used non-parametric statistics, where appropriate. We did not use data imputation and excluded cases with missing data on analysis-by-analysis basis (unless mentioned otherwise). We examined the following measurement properties of OKS.

Reliability

Reliability is an estimation of the consistency and stability of a measure. It includes analysis of the extent to which a measure is internally consistent (measured by the intercorrelation of all items) and free from measurement error. We used Cronbach's α to assess the internal consistency of the OKS summary scale and its subscales. The α values of at least 0.7 are recommended in order to demonstrate internal consistency.[18] We calculated an intraclass correlation coefficient ($ICC_{2,1}$)[19] to assess the test–retest reliability of OKS and its subscales. Minimum ICC values of 0.7 are normally considered acceptable,[18] although higher values are required for the use of the score applied at an individual level. To inform the potential use of OKS on the individual level, we calculated the precision of individual scores at 90% CI level by multiplying the standard error (SE) of measurement (SEM) by the two-tailed z value at 90%.

Construct validity

The validity of a measure is concerned with whether a measure actually measures what it purports to measure.[20 21] The definition of validity has recently been further refined as: "The degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of a test."[22] Construct validity of a measure is supported by the accumulation of evidence obtained by testing hypotheses about the relationship that the measure exhibits with other (validated) measures.[20]

We examined the construct validity of the OKS summary scale and its subscales by testing an a priori set of hypotheses about the expected relationships between the instruments at baseline

1. OKS and PCS of SF-12 (PCS-12) are measuring sufficiently similar constructs (SF-PCS measures self-reported physical function and OKS measures self-reported pain and physical functioning related to the knee), so the correlation between these two instruments' scales should be moderate and in the same direction.
2. The correlation between OKS and MCS of SF-12 (MCS-12) should be weaker than the one between PCS-12 and OKS as these two scale constructs are not considered to be related to such an extent.
3. OKS and KOOS-PS are measuring a sufficiently similar construct (KOOS-PS measures self-reported knee function and OKS measures self-reported pain and physical functioning related to the knee) that the correlation between these two measures should be strong and negative (as scores go in the opposite direction).
4. OKS and ICOAP are measuring sufficiently similar constructs (ICOAP measures self-reported knee pain and OKS measures self-reported pain and physical functioning related to the knee) that the correlation between these two measures should be strong and negative.
5. OKS-PCS should be correlated more with ICOAP than with KOOS-PS and negatively, in each case (OKS-PCS measures self-reported knee pain as does ICOAP).
6. OKS-FCS should be correlated more with KOOS-PS than with the ICOAP. In each case the correlations should be negative (OKS-FCS measures self-reported knee function, as does KOOS-PS).

We classified correlations (r) as: r=0–0.29 as none/weak, r=0.3–0.69 as moderate and r>0.7 as strong.

Structural validity is one particular aspect of construct validity; it examines the extent to which the dimensionality of a measure corresponds to the construct (ie, latent variable) that is supposed to be measured.[20] For instance, if a measure is unidimensional (ie, it is

supposed to measure one construct, such as pain), all of its items will measure the same underlying construct. We examined the structural validity of OKS by conducting CFA that tested the fit of the one-factor and two-factor models of OKS to the data, using LISREL V.8.80 software. In line with the standard CFA testing guidelines, we considered the following indices as satisfactory: a non-significant χ^2 ($p>0.05$), standardised root mean square residual (SRMR) >0.08, comparative fit index (CFI) >0.95, root mean square error of approximation (RMSEA): <0.05 close fit, <0.08 good fit, <0.1 satisfactory fit; RMSEA p test of close fit >0.05.[23] Additionally, we used the χ^2 difference test and Parsimonious Normed Fit Index (PNFI) to compare the fit between the two models of OKS and ICOAP.[24] We calculated the χ^2 difference tests by looking at the difference of χ^2 of two models along with the difference in their degrees of freedom.

Responsiveness

The ability of a measure to detect meaningful clinical change (where it has occurred) over time is critical for the use and application of a measure.[25] This change might occur following an intervention, or just occur 'naturally' during a period of observation. Generally, as with construct validity, responsiveness is assessed by testing a priori hypotheses about the relationship of the changes in one measure to the changes in another (validated) measure, or with reference to a change in a gold standard (as with testing criterion validity). Responsiveness can also be tested with reference to a transition item, where the responsiveness is tested only in participants who have reported that clinical change has occurred.

We used a one-sample t test (2 tailed) to assess if the changes at 3 months for OKS, its subscales (OKS-PCS and OKS-FCS), KOOS-PS and ICOAP were significantly different from 0. We constructed a cumulative distribution function (CDF) plot for the following: (1) OKS, (2) OKS-PCS and ICOAP and (3) OKS-FCS and KOOS-PS to examine the proportion of individual patients who experienced deterioration and improvement beyond the measurement error of the instrument at the individual level and to compare the proportion of change in pain and function detected by the different measures.

As with construct validity, we tested the responsiveness by setting a priori hypotheses about the direction and magnitude of changes of the validated comparator instruments and OKS

1. The change scores in OKS should correlate strongly with the change scores in KOOS-PS and ICOAP.
2. The change scores in OKS-PCS should correlate more strongly with the change scores in ICOAP than with the change scores in KOOS-PS.
3. The change scores in OKS-FCS should correlate more strongly with the change scores in KOOS-PS than with the change scores in ICOAP.

All correlations should be negative.

There was concern about the amount of overall change that could be experienced as a result of such a management pathway (which included a wide range of individually tailored treatments administered to a heterogeneous sample), so we additionally defined the construct of change using a patient-rated item of change. We then used the responses to this item to calculate anchor-based values of minimal important change (MIC) and difference.

Interpretability

Interpretability is defined as the degree to which one can assign qualitative meaning to a quantitative score.[20] In clinical trials, this issue can concern the question of what is considered to be a 'good', 'bad' or 'indifferent' outcome (as measured by a particular criterion or score) and what is considered to be a clinically relevant change. The minimum amount of change that is discerned as meaningful by patients is particularly important as it affects the interpretation of the study results.

We assessed the interpretability by relating the change in the PROM scores to the patient-reported item of change (using an anchor-based method) and by relating the observed change in the score to its measurement error at the individual level (using a distribution-based method). The average change in the score associated with the group of patients who responded with "My knee has got better" on the transition item was taken as the anchor-based MIC. The difference in the change score between the groups of patients who responded with "My knee has stayed the same" and "My knee has got better" on the global item of change was taken as the minimal important difference (MID). Finally, the minimum change in the instrument that represents real change (beyond measurement error) was calculated using the minimum detectable change (MDC_{90}).[26 27]

RESULTS

Sample characteristics

A total of 137 patients were recruited in the study. Twenty-one patients did not complete the follow-up questionnaires at 3 months, of which 3 patients were listed for a surgical procedure (2 osteotomies and 1 arthroplasty) before the 3-month follow-up, 7 patients no longer wanted to participate in the study and 11 were lost to follow-up. In total, of the 134 patients included in the main baseline analysis, 67 (50%) were men and 67 were women. The mean age of the patients was 59 (SD 11). Seventy per cent of patients had information on BMI, of whom 30% were classified as obese (BMI>30), 41% as overweight (BMI between 25 and 29.9) and 29% as normal weight (BMI between 18.5 and 24.9). No one was classified as underweight. All the patients had a diagnosis of knee OA. Two per cent of the patients had K-L grading of 0 (but evidence of cartilage loss on an MRI scan), 8% had K-L of 1, 43% had

Table 1 Baseline scores for OKS, its subscales (OKS-PCS and OKS-FCS), ICOAP, KOOS-PS and SF-12 (PCS-12 and MCS-12)

	N				Percentiles	
	Valid	Missing	Mean (SD)	Median	25	75
OKS	121	13	29.3 (10)	30	22	37
OKS-PCS	123	11	57.4 (23)	57	43	75
OKS-FCS	137	7	66.5 (22)	70	50	85
ICOAP	124	10	37.8 (25)	31.8	16	57
KOOS-PS	112	22	40.5 (18)	38.6	32	49
PCS-12	130	4	36.7 (10)	35	29	45
MCS-12	130	4	51 (12)	56	43	60

ICOAP, Intermittent and Constant Osteoarthritis Pain; KOOS-PS, Knee Injury and Osteoarthritis Score-Physical Function Short Form; MCS-12, mental component summary of SF-12; OKS, Oxford Knee Score; OKS-FCS, Functional Component Score; OKS-PCS, Pain Component Score; PCS-12, physical component summary of SF-12; SF-12, short form 12.

K-L of 2, 16% had K-L of 3 and 4% had K-L of 4. For 26% of cases, X-ray information was unavailable, of which, 20% had their diagnosis confirmed on the basis of MRI, while 6% of patients did not have X-rays or MRIs accessible (however, these patients had the diagnosis of OA previously confirmed in a primary care setting, different trust or private clinic). All patients underwent standard non-operative management of knee OA.[8]

In total, 116 (87%) of 134 recruited patients returned the questionnaires at the 3-month follow-up. There was no difference in age or BMI between those patients who did not respond at 3 months versus those who did, but baseline OKS was different between these groups. The group that did not respond had scored, on average, 7.3 points lower (worse) on OKS than responders at 3 months (independent samples t test, $p<0.05$). A summary of the baseline scores is presented in table 1.

For comparison, in the developmental study of OKS, the median age of patients undergoing knee replacement was 73 and in this study the median age was 58 (mean 59).[2] There was also considerable difference in self-reported pain and functional disability between the patients in the two studies. The mean baseline OKS in this sample was 29, compared to the mean preoperative OKS in the developmental study sample of 17 (when transformed to the 0–48 scoring system).

Reliability

Cronbach's α for the 12-item OKS was 0.94, 0.88 for OKS-FCS and 0.90 for OKS-PCS. For ICOAP and KOOS-PS, Cronbach's α was 0.97 and 0.94, respectively. The α value did not change considerably if any of the items were sequentially removed from the total scores.

Test–retest reliability ICCs were 0.93 (95% CI 0.91 to 0.95) for the summary OKS, 0.91 (95% CI 0.88 to 0.94) for OKS-PCS and 0.92 (95% CI 0.90 to 0.95) for OKS-FCS.

SEM for the summary OKS was 2.65 and the confidence in individual single score at 90% was ±4.4 OKS points. SEM for OKS-FCS was 6.2 with ±10.2 90% CI for individual score and SEM for OKS-PCS was 6.9 with ±11.3 points as 90% CI for individual score (noting that OKS-PCS and OKS-FCS are presented on a different scale than OKS). SEM for ICOAP was 9.68 with ±15.9 points as 90% CI for individual score. We calculated SEM for ICOAP by using the test–retest reliability that was reported in the developmental study (0.85).[6] For KOOS-PS, this information for the English version of the questionnaire was not available, so we used the test–retest reliability value of 0.86 from the validation of the French version of the questionnaire.[28] SEM for KOOS-PS was 6.7 with ±11.1 points as 90% CI for individual score.

Construct validity

Construct validity (hypothesis testing)

All correlations were generally consistent with a priori hypotheses concerning the relationships of OKS with comparator instruments. Spearman's ρ between the baseline OKS, KOOS-PS, ICOAP, SF-12-MCS and SF-12-PCS is shown in table 2. OKS correlated strongly with KOOS-PS and ICOAP. The correlation between SF-12-PCS and OKS was slightly higher than expected. As expected, OKS was most poorly related to SF-12-MCS. OKS-PCS correlated more with ICOAP than with KOOS-PS and OKS-FCS correlated more with KOOS-PS than with ICOAP. This evidence supports convergent and divergent validity of OKS.

Table 2 Baseline Spearman's correlations between the scores

	OKS	OKS-PCS	OKS-FCS
ICOAP	−0.879 (115)	−0.884 (117)	−0.792 (121)
KOOS-PS	−0.849 (106)	−0.779 (107)	−0.867 (111)
PCS-12	0.648 (121)		
MCS-12	0.370 (121)		

All correlations were significant at the 0.01 level (2 tailed). The number of cases with complete information that allowed the calculation of the correlation coefficients is in brackets for each correlation.
ICOAP, Intermittent and Constant Osteoarthritis Pain; KOOS-PS, Knee Injury and Osteoarthritis Score-Physical Function Short Form; MCS-12, mental component summary of SF-12; OKS, Oxford Knee Score; OKS-FCS, Functional Component Score; OKS-PCS, Pain Component Score; PCS-12, physical component summary of SF-12; SF-12, short form 12.

Table 3 Fit indices of one-factor and two-factor models of OKS

Factors	χ^2 (p Value)	Df	RMSEA	90% CI RMSEA	RMSEA p test	CFI	SRMR	PNFI
1	71.32 (0.06)	54	0.052	0.00 to 0.08	0.44	0.99	0.043	0.80
2	56.64 (0.34)	53	0.024	0.00 to 0.06	0.83	1	0.039	0.79

p Value for test of close fit (RMSEA <0.05).
CFI, comparative fit index; df, degrees of freedom; OKS, Oxford Knee Score; PNFI, Parsimonious Normed Fit Index; RMSEA, root mean square error of approximation; SRMR, standardised root mean square residual.

Structural validity

In total, 122 preoperative OKSs, 125 preoperative ICOAP and 113 preoperative KOOS-PS were available for CFA. Fit indices of one-factor and two-factor models for OKS are presented in table 3. Neither of the one-factor and two-factor models was rejected. Fit indices favoured the two-factor model and the reduction in χ^2 in the two-factor model was significant (χ^2 diff >7.879, with df=1, at a=0.005 level).

CFA revealed that a one-factor KOOS-PS model was rejected by the χ^2 test and its RMSEA was above the highest acceptable threshold of an acceptable fit (0.1; table 4). SRMR was acceptable and CFI was on the threshold of a good fit. The one-factor and two-factor ICOAP models were rejected by the χ^2 test and both models had RMSEA values far above the lowest threshold of an acceptable fit. However, SRMR and CFI were acceptable for both scores. There was no significant reduction (at the 0.05 level) in χ^2 for the two-factor model of ICOAP (χ^2 diff <3.84, with df=1).

Responsiveness

Figure 1 shows the CDF plot for OKS. The plot demonstrates that, based on the OKS summary score, approximately 15% of patients in the study experienced deterioration in health state, at 3-month follow-up, which was beyond the MDC_{90} of 6 points, approximately 30% of patients experienced improvement and 55% of patients did not experience change beyond this value. Also, slightly less than 30% of patients experienced improvement that was beyond the MIC of 7 points on OKS.

Table 5 shows the mean baseline, 3-month follow-up change scores, and p values for the significance of 3-month change and effect sizes (ESs) for OKS, OKS-PCS, OKS-FCS, KOOS-PS and ICAOP for the overall cohort. All mean changes were significant at the 0.01 level (2-tailed t test) except OKS-FCS.

The correlations between the changes in OKS and changes in KOOS-PS and ICOAP were somewhat less than anticipated (0.67 and 0.62, respectively). As hypothesised, the changes in OKS-PCS correlated more with the changes in ICOAP (also assessing knee pain) than with the changes in KOOS-PS, and the changes in OKS-FCS correlated more strongly with the changes in KOOS-PS (also assessing knee function) than with the changes in ICOAP (table 6).

Interpretability

Tables 7 and 8 present the percentage of responses for different response categories, ES and mean score changes by response category. We conducted independent sample t tests for the equality of means between the mean scores for groups of patients who responded 'better' and 'the same' on the transition item. Only OKS, OKS-PCS and OKS-FCS had registered significant differences between the means (2 tailed, p<0.05) of groups who responded that they were better/the same. Table 9 presents the summary of the interpretability indices.

DISCUSSION

The OKS summary scale and its pain and functional component subscales were each found to have acceptable evidence of their measurement properties to support their use with groups of patients (research/audit) and for individuals (clinical practice) who were undergoing non-operative treatment for knee OA. The OKS summary scale and its subscales were validated against KOOS-PS, ICOAP (measures developed for use in patients with knee OA) and SF-12 by testing logical a priori hypotheses regarding the construct validity and responsiveness of OKS and its subscales in comparison

Table 4 Fit indices of one-factor and two-factor models of ICOAP and KOOS-PS

	χ^2 (p Value)	Df	RMSEA	90% CI RMSEA	RMSEA p test	CFI	SRMR	PNFI
ICOAP (1F)	242.31 (0.00)	44	0.19	0.17 to 0.22	0.00	0.95	0.064	0.75
ICOAP (2F)	228.19 (0.00)	43	0.19	0.16 to 0.21	0.00	0.96	0.057	0.74
KOOS-PS (1F)	40.88 (0.00)	14	0.13	0.09 to 0.18	0.00	0.98	0.046	

p Value for test of close fit (RMSEA <0.05).
CFI, comparative fit index; df, degrees of freedom; F, number of factors; ICOAP, Intermittent and Constant Osteoarthritis Pain; KOOS-PS, Knee Injury and Osteoarthritis Score-Physical Function Short Form; PNFI, Parsimonious Normed Fit Index; RMSEA, root mean square error of approximation; SRMR, standardised root mean square residual.

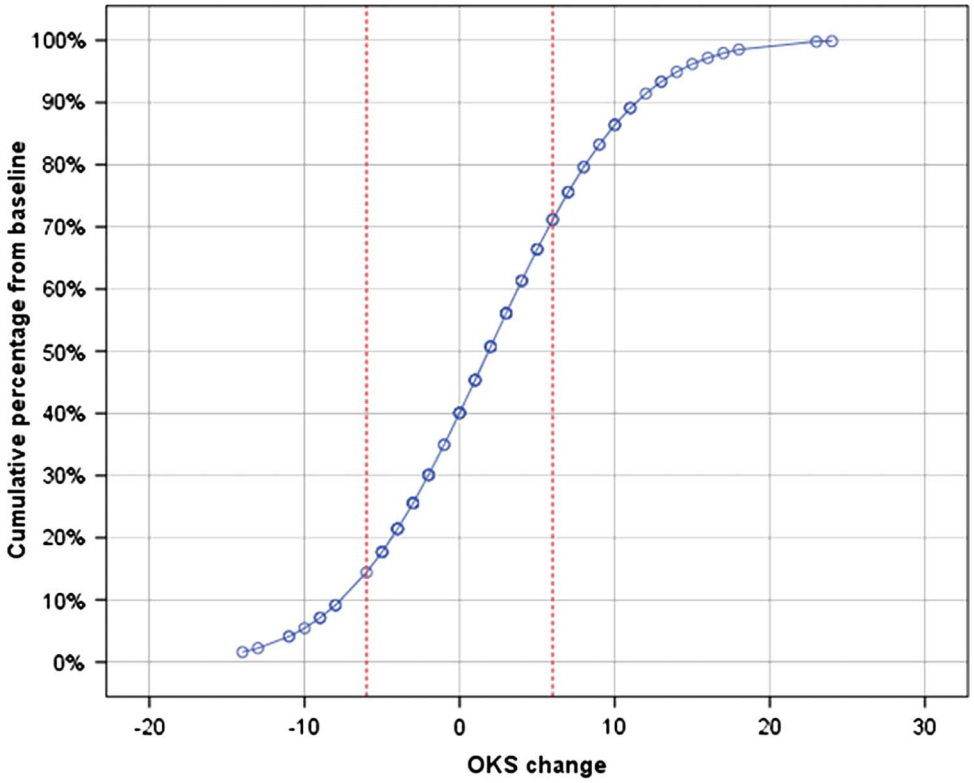

Figure 1 Cumulative percentage of patients experiencing the change on the Oxford Knee Score (OKS) from baseline less or equal to the value on the x axis. Red line marks the minimum detectable change (MDC_{90}) beyond the measurement error of the score (MDC_{90} of 6 points).

to these other (validated) measures. Thus, CFA demonstrated excellent fit and confirmed the structural validity of OKS and both subscales. Furthermore, assessment of the test–retest reliability demonstrated that OKS and its subscales could be used both at the group and individual levels (clinical practice).[29]

The OKS subscales can be used to specifically target the improvement or deterioration in pain or function, whether in research (as an endpoint or for sample size calculations) or in clinical practice. Anchor-based MIC of ≈7 for OKS, ≈17 for OKS-PCS and ≈11 for OKS-FCS can be used in cohort studies to assess if the change in OKS (from baseline) is clinically relevant. Anchor-based MID of ≈6 for OKS, ≈14 for OKS-PCS and ≈10 for OKS-FCS can be used in clinical trials to assess if the difference in change between two arms of treatment is clinically relevant. Finally, changes in individual patient scores beyond MDC_{90} (≈6 points for OKS, ≈16 points for OKS-PCS and ≈15 points for OKS-FCS) can be used as a benchmark of improvement or deterioration that is beyond the measurement error of the score. These values are likely to be different if OKS is used in a different population of patients (ie, patients undergoing knee replacement surgery).

Limitations

Even though the reliability, construct validity and responsiveness of OKS and its subscales have been proven to be

Table 5 Significance of change in OKS, its subscales (OKS-PCS and OKS-FCS), ICOAP and KOOS-PS scores at 3 months (one-sample t test)

	N	Baseline (SD)	3 Months (SD)	Change (SD)	p Value	ES
OKS	104	30.29 (10)	32.15 (11)	1.87 (7)	0.01	0.19
OKS-PCS	107	59.36 (22)	65.13 (24)	5.77 (17)	<0.01	0.26
OKS-FCS	108	67.22 (21)	68.66 (23)	1.44 (16)	0.4	0.07
ICOAP*	104	37.19 (25)	31.53 (25)	−5.66 (19)	<0.01	0.23
KOOS-PS*	92	39.42 (18)	34.88 (20)	−4.5 (14)	<0.01	0.25

*ICOAP and KOOS-PS represent severity of the disease in the opposite direction from OKS and its subscales.
ES, effect size; ICOAP, Intermittent and Constant Osteoarthritis Pain; KOOS-PS, Knee Injury and Osteoarthritis Score-Physical Function Short Form; N, number of complete cases available for calculation of 3-month follow-up; OKS, Oxford Knee Score; OKS-FCS, Functional Component Score; OKS-PCS, Pain Component Score.

Table 6 Spearman's correlations between the 3-month changes in OKS and its subscales (OKS-PCS and OKS-FCS), ICOAP and KOOS-PS

	ICOAP	KOOS-PS
OKS	−0.674 (96)	−0.617 (87)
OKS-PCS	−0.669 (99)	−0.551 (88)
OKS-FCS	−0.598 (100)	−0.622 (90)

All correlations are significant at the 0.01 level (2 tailed). The number of cases with complete information that allowed the calculation of the correlation coefficients is in brackets for each correlation.
ICOAP, Intermittent and Constant Osteoarthritis Pain; KOOS-PS, Knee Injury and Osteoarthritis Score-Physical Function Short Form; OKS, Oxford Knee Score; OKS-FCS, Functional Component Score; OKS-PCS, Pain Component Score.

satisfactory when used in patients undergoing non-operative management for their knee OA, there might be a need to further verify its content validity in this extended context.[30] The items for OKS were originally devised using a representative sample of patients with end-stage disease who were undergoing knee replacement surgery. It could be argued that the measure in its current form might not fully represent the concerns of this slightly different population of patients whose knee OA is generally at an earlier stage. If a measure is used in a different context or with a different type of patients than that which was used in its design/development, then the content validity may be suspect (in relation to the new/different usage).[18] A counterargument is that it is unrealistic to have a new/different measure (and a new study conducted to design and test one) for every possible subcategory of patient or type of treatment within all diseases or conditions. In such cases, a researcher should make a judgement about the best available/closest measure[21] but, as a minimum, should check that the measurement properties are still otherwise maintained. Any further examination of the content validity of OKS in this extended context would necessitate a new study (based on qualitative interviews) being undertaken.

One of the limitations concerns the use of the transition question with three response levels (better, the same, and worse). MIC/MID values depend on the number of response categories on the transition question. If, for instance, a response category 'a little better' was used instead of 'better', the final MIC value would have probably been smaller. Indeed, the methods of MIC/MID estimation have been a subject of debate within the scientific community and we would recommend that any application of the MIC/MID values presented in this paper is performed with awareness of its caveats. Regardless of this potential limitation of the transition item, the same method was used in the comparative analysis of interpretability between OKS, KOOS-PS, and ICOAP ensuring appropriate comparison between the scores.

Comparative performance of OKS and its subscales versus ICOAP and KOOS-PS in this study

Even though ICOAP and KOOS-PS are currently widely used as outcome measures for knee OA, OKS performed better in this study on several counts.

The 11-item ICOAP had Cronbach's α of 0.97 (compared to the α of OKS-PCS of 0.9) and 0.94 for KOOS-PS (compared to the α of 0.87 for OKS-FCS). A high α value can mean that some of the items on a scale are redundant and this seems to be more of a concern for ICOAP and KOOS-PS subscales than for the OKS subscale. Furthermore, the reliability and precision

Table 7 Number (N) and percentage of responses for different response categories with ESs, mean score changes by response category and ANOVA tests for linear trend for the mean score across the three response categories for OKS and its subscales (OKS-PCS and OKS-FCS)

	Better	Same	Worse
OKS			
N (% of responses)	30 (33)	26 (28)	36 (39)
Mean change (SD)	7.1 (8)	0.7 (6)	−1.88 (5)
ES	0.7	0.1	−0.2
p Value for linear trend	<0.001	<0.001	<0.001
OKS-PCS			
N (% of responses)	31 (33)	28 (30)	38 (35)
Mean change (SD)	17.27 (19)	2.93 (14)	−2.68 (11)
ES	0.8	0.2	−0.1
p Value for linear trend	<0.001	<0.001	<0.001
OKS-FCS			
N (% of responses)	28 (33)	26 (31)	30 (36)
Mean change (SD)	10.63 (14)	1.11 (16)	−6.35 (14)
ES	0.5	0.1	−0.3
p Value for linear trend	<0.001	<0.001	<0.001

ANOVA, analysis of variance; ESs, effect sizes; OKS, Oxford Knee Score; OKS-FCS, Functional Component Score; OKS-PCS, Pain Component Score.

Table 8 Number (N) and percentage of responses for different response categories with ESs, mean score changes by response category and ANOVA tests for linear trend for the mean score across the three response categories for ICOAP and KOOS-PS

	Better	Same	Worse
ICOAP			
N (% of responses)	32 (34)	27 (29)	35 (37)
Mean change (SD)	−13.42 (23)	−5.64 (17)	2.73 (16)
ES	−0.6	−0.3	0.1
p Value for linear trend	<0.003	<0.003	<0.003
KOOS-PS			
N (% of responses)	25 (31)	27 (33)	30 (37)
Mean change (SD)	−11.98 (15)	−4.22 (12)	1.61 (12)
ES	−0.8	−0.3	0.1
p Value for linear trend	<0.001	<0.001	<0.001

ANOVA, analysis of variance; ESs, effect sizes; ICOAP, Intermittent and Constant Osteoarthritis Pain; KOOS-PS, Knee Injury and Osteoarthritis Score-Physical Function Short Form.

of the score was better for OKS and its subscales than for KOOS-PS and ICOAP, which makes it more suitable to be used in clinical practice.

There was evidence to support the one-factor and two-factor models of OKS, but no acceptable evidence of structural validity was found for KOOS-PS or ICOAP. KOOS-PS and the one-factor and two-factor ICOAP models were rejected by the χ^2 test. Furthermore, RMSEAs were unacceptably high for both scales. The exploration of the sources of poor fit of these measures is beyond the scope of this study and future studies should investigate this problem further (perhaps also using exploratory factor analysis).

We have some concerns about the interpretability of ICOAP and KOOS-PS. It seems that these measures performed less well than OKS in this regard. First, owing to the fact that ICOAP has low precision at the individual level (MDC_{90} is almost 10 points larger than MIC), this makes it less suitable to interpret change scores in individual patients. Second, although around one-third of the patients in our sample reported being better following 3 months of non-operative management for knee OA, neither ICOAP nor KOOS-PS obtained statistically significant differences in the change score between the groups of patients who reported themselves to be better or the same (in contrast with OKS and its subscales). This could indicate problems with the sensitivity of these scores to change. Third, while there was some lack of symmetry between the mean change in the OKS score and its subscales in relation to the patient-rated item of change (patients who claim they had not experienced change on the global transition item actually experienced change as measured by PROM), this lack of symmetry seems to be more pronounced for KOOS-PS and ICOAP.

Table 9 Anchor-based and distribution-based MIC/MID values for OKS, its subscales, ICOAP and KOOS-PS

	Distribution based	Anchor based	
	MDC_{90}	MID	MIC
OKS	±6	6.4	7.1
OKS-PCS	±16	14.3	17.3
OKS-FCS	±15	9.5	10.6
ICOAP	±23	7.8	13.4
KOOS-PS	±16	7.8	12.0

ICOAP, Intermittent and Constant Osteoarthritis Pain; KOOS-PS, Knee Injury and Osteoarthritis Score-Physical Function Short Form; MDC_{90}, minimum detectable change; MIC, minimum important change; MID, minimum important difference; OKS, Oxford Knee Score; OKS-FCS, Functional Component Score; OKS-PCS, Pain Component Score.

Implications for clinicians and policymakers

In this study, we obtained evidence that supports the use of OKS and its pain and functional subscales in patients who are undergoing non-operative management for their knee. When used with patients in this context, OKS has demonstrated evidence of validity, reliability and responsiveness in measuring the state of health of individuals. The measure could be used in clinical practice to monitor disease progression in individual patients undergoing non-operative management for their knee OA or for hospital audit where the information from groups of patients is analysed to assess the effectiveness of current patient management pathways for treating OA in terms of health gain/deterioration.

Although this study was conducted on a sample of patients with knee OA presenting themselves in the secondary care setting, we consider that the findings presented here may be generalisable to the primary care setting. Studies have shown no significant differences in pain severity and function between the groups of patients with knee OA who get referred to secondary care and who do not.[31 32] Other factors, such as the chronicity of the disease, or complex interaction of psychological and social factors, are more associated with secondary care referral. However, further research, involving larger sample sizes is needed to confirm these findings.

The use of a single valid score across a patient pathway is a compelling goal when considering how to develop standardisation of patient care in the NHS. Our new evidence suggests extending the use of OKS in the patient pathway for managing knee OA may be possible. However, the practicalities and feasibility of widespread score administration need further exploration focusing on appropriate timing, frequency and method of score administration.[33] Most importantly, more work is required to understand how results of OKS, if adopted earlier in the pathway, should be interpreted to support patients in shared decision-making regarding treatment options and the influence that such routine use of OKS might have on the quality of care that patients receive (ie, the effect on the quality of service and influence on patients' clinical outcomes).[34]

Contributors JD, DB and AJP contributed to the conception and design. KKH and LDJ contributed to the acquisition of data. KKH, JD, DB and AJP were involved in the analysis and interpretation of the data. KKH, JD, LDJ, DB and AJP were responsible for drafting the article, revising it critically for important intellectual content and approving the final version of the article. All authors, external and internal, had full access to all of the data (including statistical reports and tables) in the study and can take responsibility for the integrity of the data and the accuracy of the data analysis.

Funding The research was supported by the Nuffield Orthopaedic Centre National Institute for Health Research (NIHR) Biomedical Research Unit into Musculoskeletal Disease and the Arthritis Research UK EOTC.

Competing interests JD is one of the original inventors of the Oxford Hip Score (OHS) and Oxford Knee Score (OKS). She has received consultancy payments, via Isis Innovation, in relation to work involving both questionnaires.

Ethics approval This study obtained ethics approval from the Oxfordshire Research Ethics Committee B (11/SC/005).

Patient consent Obtained.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement Anonymised data and statistical codes are available from the corresponding author.

REFERENCES

1. Dawson J, Fitzpatrick M, Churchman D, *et al*. User manual for the Oxford Knee Score (OKS). 2010.
2. Dawson J, Fitzpatrick R, Murray D, *et al*. Questionnaire on the perceptions of patients about total knee replacement. *J Bone Joint Surg Br* 1998;80:63–9.
3. Department of Health. *Guidance of the routine collection of Patient Reported Outcome Measures (PROMs)*. London: Department of Health, 2008.
4. Devlin NJ, Appleby J. *Getting the most out of PROMS*. The Kings Fund Office of health economics, 2010.
5. Valderas J, Alonso J. Patient reported outcome measures: a model-based classification system for research and clinical practice. *Qual Life Res* 2008;17:1125–35.
6. Hawker GA, Davis AM, French MR, *et al*. Development and preliminary psychometric testing of a new OA pain measure—an OARSI/OMERACT initiative. *Osteoarthritis Cartilage* 2008;16:409–14.
7. Perruccio AV, Stefan Lohmander L, Canizares M, *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.
8. National Institute for Health and Clinical Excellence (NICE). *Osteoarthritis. National clinical guideline for care and management in adults*. London: Royal College of Physicans, 2008.
9. Ware J Jr, Kosinski M, Keller SD. A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. *Med Care* 1996;34:220–33.
10. Murray DW, Fitzpatrick R, Rogers K, *et al*. The use of the Oxford hip and knee scores. *J Bone Joint Surg Br* 2007;89-B:1010–14.
11. Harris K, Dawson J, Doll H, *et al*. Can pain and function be distinguished in the Oxford Knee Score in a meaningful way? An exploratory and confirmatory factor analysis. *Qual Life Res* 2013:1–8. English.
12. Roos EM, Roos HP, Lohmander LS, *et al*. Knee injury and Osteoarthritis Outcome Score (KOOS)—development of a self-administered outcome measure. *J Orthop Sports Phys Ther* 1998;28:88–96.
13. Kellgren J, Lawrence J. Radiological assessment of osteo-arthrosis. *Ann Rheum Dis* 1957;16:494–502.
14. De Vet HCW, Terwee CB, Mokkink LB, *et al*. *Measurement in medicine: A practical guide*. Cambridge: Cambridge University Press, 2011. http://public.eblib.com/EBLPublic/PublicView.do?ptiID=802925
15. Kline P. An easy guide to factor analysis. 1993.
16. Bentler PM, Chou CP. Practical issues in structural modeling. *Socio Methods Res* 1987;16:78–117.
17. Ding L, Velicer WF, Harlow LL. Effects of estimation methods, number of indicators per factor, and improper solutions on structural equation modeling fit indices. *Struct Equ Modeling* 1995;2:119–43.
18. Nunnally JC, Bernstein IH. *Psychometric theory*. New York: McGraw-Hill, 1994.
19. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;86:420–8.
20. Mokkink LB, Terwee CB, Patrick 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.
21. Streiner DL, Norman GR. *Health measurement scales: a practical guide to their development and use*. Oxford University Press, 2008.
22. AERA, APA, NCME. *Standards for educational and psychological testing*. Washington: American Educational Research Association, 1999:194.
23. Browne MW, Cudeck R. Alternative ways of assessing model fit. *Sociological Methods & Research* 1992;21:230–58.
24. Schumacker RE, Lomax RG. *A beginner's guide to structural equation modeling*. Mahwah, NJ: Lawrence Erlbaum Associates, 2004.
25. Fitzpatrick R, Davey C, Buxton M, *et al*. Evaluating patient-based outcome measures for use in clinical trials. *Health Technol Assess* 1998;2:1–74.
26. Beckerman H, Roebroeck M, Lankhorst G, *et al*. Smallest real difference, a link between reproducibility and responsiveness. *Qual Life Res* 2001;10:571–8.
27. De Vet HC, Terwee CB, Ostelo RW, *et al*. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. *Health Qual Life Outcomes* 2006;4:54.
28. Ornetti P, Perruccio A, Roos E, *et al*. Psychometric properties of the French translation of the reduced KOOS and HOOS (KOOS-PS and HOOS-PS). *Osteoarthritis Cartilage* 2009;17:1604–8.
29. Charter RA, Feldt LS. Confidence intervals for true scores: is there a correct approach? *J Psychoeduc Assess* 2001;19:350–64.
30. Rothman M, Burke L, Erickson P, *et al*. Use of existing patient-reported outcome (PRO) instruments and their modification: the ISPOR good research practices for evaluating and documenting content validity for the use of existing instruments and their modification PRO Task Force Report. *Value Health* 2009;12:1075–83.
31. Mitchell H, Carr A, Scott D. The management of knee pain in primary care: factors associated with consulting the GP and referrals to secondary care. *Rheumatology* 2006;45:771–6.
32. Hopman-Rock M, De Bock GH, Bijlsma JW, *et al*. The pattern of health care utilization of elderly people with arthritic pain in the hip or knee. *Int J Qual Health Care* 1997;9:129–37.
33. Dawson J, Doll H, Fitzpatrick R, *et al*. The routine use of patient reported outcome measures in healthcare settings. *BMJ* 2010;340:c186.
34. Snyder CF, Aaronson NK, Choucair AK, *et al*. Implementing patient-reported outcomes assessment in clinical practice: a review of the options and considerations. *Qual Life Res* 2012;21:1305–14.