EXTENDING THE USE OF PROMS IN THE NHS: USING THE OXFORD KNEE SCORE TO MONITOR THE PROGRESSION OF KNEE OSTEOARTHRITIS. A VALIDATION STUDY

Journal:	*BMJ Open*
Manuscript ID:	bmjopen-2013-003365
Article Type:	Research
Date Submitted by the Author:	07-Jun-2013
Complete List of Authors:	Harris, Kristina; University of Oxford, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences Dawson, Jill; University of Oxford, Department of Public Health Jones, Luke; University of Oxford, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences Beard, David; University of Oxford, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences Price, Andrew; University of Oxford, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences
<b>Primary Subject Heading</b>:	Health services research
Secondary Subject Heading:	Research methods
Keywords:	Knee < ORTHOPAEDIC & TRAUMA SURGERY, Musculoskeletal disorders < ORTHOPAEDIC & TRAUMA SURGERY, STATISTICS & RESEARCH METHODS, HEALTH SERVICES ADMINISTRATION & MANAGEMENT

EXTENDING THE USE OF PROMS IN THE NHS: USING THE OXFORD KNEE SCORE TO MONITOR THE PROGRESSION OF KNEE OSTEOARTHRITIS. A VALIDATION STUDY

ABSTRACT

Objectives To assess the validity of the OKS for use in patients undergoing non-operative management for their knee OA within the 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, ICOAP, KOOS-PS, at baseline and three month follow up, transition item of change at three months.

Results The OKS summary scale and its pain and functional component subscales demonstrated good test-retest reliability (ICC 0.93, 0.91, 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 the OKS with other validated measures (KOOS-PS, ICOAP, SF12), which provided evidence of construct validity and responsiveness of the score and its subscales. Confirmatory Factor Analysis confirmed the structural validity of the OKS. However, there was a lack of satisfactory evidence of structural validity for the ICOAP and KOOS. Minimal important changes, minimal important differences and the precision of the change score were calculated for the OKS, its subscales, the ICOAP and the KOOS-PS.

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

Article summary

Article focus

This study examines the measurement properties of the OKS, an instrument designed for patients undergoing knee replacement surgery, when used in patients undergoing non-operative treatment for knee OA.

Key messages

The OKS summary scale, and its pain and functional component subscales, were found to have acceptable evidence of measurement properties.

The findings of this study support the use of the OKS with groups of patients (in research/audit) and for individuals (in clinical practice) who are undergoing non-operative treatment for knee OA, in addition to patients undergoing total knee replacement.

Strengths and limitations

Despite the reliability, construct validity and responsiveness of the OKS and its subscales have been proven to be satisfactory when used in patients undergoing non-operative management of knee OA, there might be a need to further verify its content validity in this extended context.

INTRODUCTION

The Oxford Knee Score (OKS) is a widely used patient reported outcome measure

(PROM), originally developed in1998 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 the OKS was extended in 2009, when it was adopted by the 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 post operation, in order to monitor and benchmark the performance of health providers.

The increasing popularity of the 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 the OKS in clinical practice as a means of standardizing clinical assessment, monitoring individual's self-reported health state across the spectrum of 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 the 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 (i.e., 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 the 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 the OKS remains as reliable, valid and responsive when used with patients who are at earlier stages of their disease management.

The aim of our study was to assess the measurement properties of the OKS when used with patients who are undergoing non operative management for knee OA, by examining its reliability, validity, responsiveness and interpretability when applied in this context.

METHODS

We obtained ethical approval for a prospective cohort study from a local ethics committee (11/SC/005). Informed consent was obtained from all participants in the study.

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)(6)). 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 Intermittent and Constant Osteoarthritis Pain (ICOAP)(7) the Knee Injury and Osteoarthritis Score-Physical function Short form (KOOS-PS)(8) and SF12(9) patient-reported questionnaires.

The OKS is a 12 item questionnaire. It's item content was devised using patient interviews, which addresses 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 the 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 standardized from 0 (worst) to 100 (best). Patients completed the OKS at baseline, 2 and 5 days (for test-retest reliability) and at 3 months.

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

The 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 the 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). The 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 standardized 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. The 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 regards 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. The degree of structural changes in the knee was classified using (K-L) grading.(13) In the absence of X-rays, we assessed intra-operative documentation from previous knee arthroscopy or available MRIs to examine the extent of cartilage loss and confirm the diagnosis of osteoarthritis.

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 subjects per questionnaire item.(16, 17) These sample sizes are required for data analyses and should be adjusted (i.e. 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 subjects per item.

We analysed the data using SPSS version 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 the ICOAP and the OKS-PCS). We used non-parametric statistics, where appropriate. We did not use data imputation and we excluded cases with missing data on analysis by analysis basis (unless mentioned otherwise). We examined the following measurement properties of the 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 inter-correlation of all items) and free from measurement error. We used Cronbach's alpha to assess the internal consistency of the OKS summary scale and its subscales. Alpha 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 the 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 the OKS on the individual level, we calculated the precision of individual scores at 90% CI level by multiplying the standard error of measurement (SEM) by the 2-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.(21)

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:

(i) the OKS and the physical component summary of the SF12 (PCS-12) are measuring sufficiently similar constructs (SF-PCS measures self-reported physical function and the 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,

(ii) the correlation between the OKS and the mental component summary of the SF12 (MCS-12) should be weaker than the one between the PCS-12 and OKS as these two scale constructs are not considered to be related to such an extent,

(iii) the OKS and KOOS-PS are measuring a sufficiently similar construct (the KOOS-PS measures self-reported knee function and the 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),

(iv) the OKS and the ICOAP are measuring sufficiently similar constructs (the ICOAP measures self-reported knee pain and the OKS measures self-reported pain and physical functioning related to the knee) that the correlation between these two measures should be strong and negative,

(v) the OKS-PCS should be correlated more with the ICOAP than with the KOOS-PS and negatively, in each case (the OKS-PCS measures self-reported knee pain as does the ICOAP),

(vi) the OKS-FCS should be correlated more with the KOOS-PS that the ICOAP and negatively (the OKS-FCS measures self-reported knee function, as does the KOOS-PS).

We classified correlations (r) as: r=0 to 0.29 as none/weak; r= 0.3 to 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 (i.e. latent variable) that is supposed to be measured.(21) For instance, if a measure is unidimensional (i.e. 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 the OKS by conducting Confirmatory Factor Analysis (CFA) that tested the fit of the one and two factor models of the OKS to the data, using LISREL V8.80 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, <0.08good fit, <0.1 satisfactory fit; RMSEA p test of close fit>0.05.(23) Additionally, we used the Chi-square (χ2) difference test and Parsimonious Normed Fit Index (PNFI) to compare the fit between the two models of the OKS and the ICOAP.(24) We calculated the χ2 difference tests by looking at the difference of χ2 of two models along with the difference in their degrees of freedom. We checked the χ2 difference, with its degrees of freedom in the χ2 distribution table. If this value is statistically significant, then the model with more degrees of freedom is favoured.

Responsiveness

The ability of a measure to detect meaningful clinical change (where it has occurred) over time is critical for the use and the 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 subjects 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 the OKS, its subscales (OKS-PCS and OKS-FCS), KOOS-PS and the ICOAP were significantly different from 0. We constructed a Cumulative Distribution Function (CDF) plot for the; (i) OKS, (ii) OKS-PCS and ICOAP, and (iii) 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 the OKS:

(i) the change scores in the OKS should correlate strongly with the change scores in the KOOS-PS and ICOAP,

(ii) the change scores in the OKS-PCS should correlate more strongly with the change scores in the ICOAP than with the change scores in the KOOS-PS,

(iii) change scores for the OKS-FCS should correlate more strongly with change scores for the KOOS-PS than the change scores for the ICOAP.

All correlations should be negative.

There was a concern about the amount of overall change that can 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 and difference.

Interpretability

Interpretability is defined as the degree to which one can assign qualitative meaning to a quantitative score.(26) 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 interpretation of study results.

We assessed the interpretability by relating the change in the PROMs 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). 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 minimal important change

(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$_{90}$), which was obtained by multiplying the SEM with the z-value at the 90% level and the square root of two (to account for two measurement occasions).(27, 28)

RESULTS

Sample characteristics. 137 patients were recruited in the study. 21 patients did not complete follow up questionnaires at 3 months, out of which 3 patients were listed for a surgical procedure (2 osteotomies and 1 arthroplasty) before 3 month follow-up, 7 patients no longer wanted to participate in the study and 11 were lost to follow-up. 134 patients were included in the main baseline analysis of whom 67 (50 %) were male and 67 patients were female. The mean age of patients was 59 (SD 11), which is about 10 years less than the average age of the developmental sample of the OKS. 70% of patients had information on Body Mass Index (BMI), out of whom 30% were classified as obese (BMI>30), 41% as overweight (BMI between 25 and 29.9), 29% as normal weight (BMI between 18.5 and 24.9). No one was classified as underweight. All of the patients had a diagnosis of knee osteoarthritis. 2% of the patients had Kellgren-Lawrence (KL) grading of 0 (but evidence of cartilage loss on MRI scan), 8% had K-L of 1, 43% had K-L of 2, 16% had K-L of 3, 4% had K-L of 4. For 26% of cases, X-ray information was unavailable, of whom, 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 the primary care setting, different trust, or in a private clinic). All patients underwent standard non-operative management of knee OA.(29)

116 (87%) out of 134 recruited patients returned the questionnaires at three month follow up. There was no difference in age or BMI between those patients who did not respond at three 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 the OKS than responders at three months (Independent samples t-test, p<0.05). A summary of the baseline scores is presented in Table 1.

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

	N		Mean (SD)	Median	Percentiles	
	Valid	Missing			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 (26)	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

Reliability

Cronbach's alpha for the 12-item OKS was 0.94, 0.88 for the OKS-FCS and 0.90 for the OKS-PCS. For the ICOAP and KOOS-PS, the Cronbach's alpha was 0.97 and 0.94 respectively. The alpha 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, 0.91-0.95) for the summary OKS, 0.91 (95% CI, 0.88-0.94) for the OKS-PCS and 0.92 (95% CI, 0.90-0.95) for the OKS-FCS.

The standard error of measurement (SEM) for the summary OKS was 2.65 and the confidence in individual single score at 90% was ±4.4 OKS points. SEM for the OKS-FCS was 6.2 with ±10.2 90% CI for individual score and the SEM for the OKS-PCS was 6.9 with ±11.3 points as 90% CI for individual score (noting that the OKS-PCS and the OKS-FCS are presented on a different scale than the OKS).

The SEM for the ICOAP was 10.1 with ±16.6 points as 90% CI for individual score. We calculated the SEM for the ICOAP by using the test-retest reliability that was reported in the developmental study (0.85).(30) For the 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. The SEM for the 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 the OKS with comparator instruments. Spearman's ρ between the baseline OKS, KOOS-PS, ICOAP, SF12-MCS and SF-12-PCS are shown in Table 2. The OKS correlated strongly with the KOOS-PS and ICOAP. The correlation between the SF12-PCS and the OKS was slightly higher than expected. As expected, the OKS was most poorly related to the SF12-MCS. The OKS-PCS correlated more with ICOAP than with KOOS-PS and the OKS-FCS correlated more with the KOOS-PS that with ICOAP. This evidence supports convergent and divergent validity of the OKS.

Table 2: Baseline Spearman's correlations between the scores. 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.

	OKS	OKS-PCS	OKS-FCS
ICOAP	-.879 (115)	-.884 (117)	-.792 (121)
KOOS-PS	-.849 (106)	-.779 (107)	-.867 (111)
PCS-12	.648 (121)	/	/
MCS-12	.370 (121)	/	/

Structural validity. 122 pre-operative OKSs, 125 pre-operative ICOAP and 113 pre-operative KOOS-PS were available for the CFA. Fit indices of one and two factor models for the OKS are presented in Table 3. Neither of the one and two factor models was rejected. Fit indices favoured the 2 factor model and the reduction in χ^2 in the two factor model was significant (χ2diff>7.879, with df=1, at the a=0.005 level).

Table 3. Fit indices of one and two-factor model of the OKS.

Factors	χ2 (p value)	df	RMSEA	90% CI RMSEA	RMSEA p test	CFI	SRMR	PNFI
1	71.32 (p=0.06)	54	0.052	0.00-0.08	0.44	0.99	0.043	0.80
2	56.64 (p=0.34)	53	0.024	0.0-0.06	0.83	1	0.039	0.79

Note. F=number of factors; 2 =chi-square; df=degrees of freedom; RMSEA=root mean square of approximation; CI=confidence intervals; p-value for test of close fit (RMSEA<.05); SRMR=standardized root mean square residual; CFI-comparative t index; PNFI=parsimonious normed fit index.

CFA revealed that a one-factor KOOS-PS model was rejected by the χ2 test and its RMSEA was above the highest acceptable threshold of an acceptable fit (0.1) (Table 4). The SRMR was acceptable and CFI was on the threshold of a good fit. Both one and two factor ICOAP models were rejected by the χ2 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 χ2 for the 2 factor model of the ICOAP (χ2diff< 3.84, with df=1).

Table 4. Fit indices of one and two-factor model of the ICOAP and KOOS-PS.

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

Note. F=number of factors; 2 =chi-square; df=degrees of freedom; RMSEA=root mean square of approximation; CI=confidence intervals; p-value for test of close fit (RMSEA<.05); SRMR=standardized root mean square residual; CFI-comparative t index; PNFI=parsimonious normed fit index.

Responsiveness

Figure 1 shows the CDF plot for the OKS. The plot demonstrates that, based on the OKS summary score, approximately 20% of patients in the study experienced deterioration in health state, at three month follow up, that was beyond the MDC_{90} of 4 points, approximately 40% of patients experienced improvement and 40% of patients did not experience change beyond this value. Also, about 25% of the patients experienced improvement that was beyond the MIC of 7 points on the OKS.

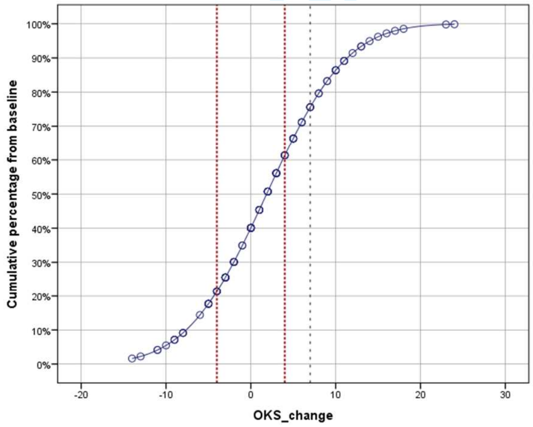

Figure 1. Cumulative percentage of patients experiencing the change on the OKS from baseline less or equal to the value on the x-axis. Red line marks the minimum detectable change beyond the measurement error of the score (MDC_{90} of 4 points).

Table 5 shows the mean baseline, three month follow-up change scores, and p values for the significance of 3 month change and ES for the 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 the OKS-FCS.

Table 5: Significance of change in OKS, its subscales (OKS-PCS and OKS-FCS), ICOAP and KOOS-PS scores at three 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[a]	104	37.19 (25)	31.53 (25)	-5.66 (19)	<0.01	0.23
KOOS-PS[a]	92	39.42 (18)	34.88 (20)	-4.5 (14)	<0.01	0.25

Note. N=number of complete cases available for calculation of 3 month follow up; SD=standard deviation; ES=effect size;[a] The ICOAP and the KOOS-PS represent severity of the disease in the opposite direction from the OKS and its subscales.

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

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

	ICOAP	KOOS-PS
OKS	-.674 (96)	-.617 (87)
OKS-PCS	-.669 (99)	-.551 (88)
OKS-FCS	-.598 (100)	-.622 (90)

Note. 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.

Interpretability

Tables 7 and 8 present the percentage of responses for different response categories, effect sizes 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 the 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. Here, the OKS and OKS-PCS mean differences were close to (and generally just above) scale MDC/MID values and thus likely beyond measurement error, while the OKS-FCS mean differences were just less than the subscale’s MDC/MID values. All OKS scales’ mean differences were greater than the scales’ relevant SEM values.

Table 9 presents the summary of interpretability indices.

Table 7: Number (N) and percentage of responses for different response categories with effect sizes (ES), mean score changes by response category and ANOVA tests for linear trend for the 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	.7	.1	-.2
	P-value for linear trend	<.001	<.001	<.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	.8	.2	-.1
	P-value for linear trend	<.001	<.001	<.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	.5	.1	-.3
	P-value for linear trend	<.001	<.001	<.001

Table 8: Number (N) and percentage of responses for different response categories with effect sizes (ES), mean score changes by response category and ANOVA tests for linear trend for the ICOAP and the 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	-.6	-.3	.1
	P-value for linear trend	<.003	<.003	<.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	-.8	-.3	.1
	P-value for linear trend	<.001	<.001	<.001

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

	Distribution based	Anchor based	
	MDC_{90}	MID	MIC
OKS	±4	6.4	7.1
OKS-PCS	±11	14.3	17.3
OKS-FCS	±10	9.5	10.6
ICOAP	±17	7.8	13.4
KOOS-PS	±11	7.8	12.0

Note. MDC_{90}=minimum detectable change; MID=minimum important difference; MIC=minimum important change.

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 are undergoing non-operative treatment for knee OA. The OKS summary scale and its subscales were validated against the KOOS-PS, the ICOAP (measures developed for use in patients with knee OA) and the SF-12 by testing logical *a priori* hypotheses regarding the construct validity and responsiveness of the OKS and its subscales in comparison to these other (validated) measures. Thus, CFA demonstrated excellent fit and confirmed the structural validity of the OKS and both subscales. Furthermore, assessment of test-retest reliability demonstrated that the OKS and its subscales could all be used both at group and individual levels (clinical practice).(31)

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 the OKS, ≈17 for the OKS-PCS, and ≈11 for the OKS-FCS can be used in cohort studies to assess if the change in the OKS (from baseline) is clinically relevant. Anchor based MID of ≈6 for the OKS, ≈14 for the OKS-PCS, and ≈10 for the 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 the MDC_{90} (≈4 points for the OKS, ≈11 points for the OKS-PCS, and ≈10 points for the 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 the OKS is used in a different population of patients (i.e. patients undergoing knee replacement surgery).

Limitations

Even though the reliability, construct validity and responsiveness of the OKS and its subscales have been proven to be 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.(32) The items for the 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 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).(33) On the other hand it may be assumed to be appropriate if the context is considered to be ‘similar enough’.(20) Another argument is that it is unrealistic to have a new/different measure (and a new study conducted to design and test one) for every possible sub category 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, but as a minimum should check that the measurement properties are still otherwise maintained. Any further examination of the content validity of the OKS in this extended context would necessitate a new study (based on qualitative interviews) being undertaken.

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

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

The 11-item ICOAP had a Cronbach's alpha of 0.97 (compared to the alpha of the OKS-PCS of 0.9) and the alpha was 0.94 for the KOOS-PS (compared to the alpha of 0.87 for the OKS-FCS). A high alpha value can mean that some of the items on a scale are redundant and this seems to be more of a concern for the ICOAP and KOOS-PS than for the OKS subscales. Furthermore, the reliability and precision of the score was better for the OKS and its subscales than for the KOOS-PS and ICOAP, which makes it more suitable to be used in clinical practice.

There was evidence to support both one and two factor models of the OKS, but no acceptable evidence of structural validity was found for the KOOS-PS or the ICOAP. The KOOS-PS and the one and two-factor ICOAP models were rejected by the $\chi2$ 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 the ICOAP and KOOS-PS. It seems that these measures performed less well than the OKS in this regard. First, due to the fact that the ICOAP has low precision at the individual level (the MDC_{90} is 4 points larger than the 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 the ICOAP or the 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 the OKS and its subscales). This could indicate problems with the sensitivity of these scores to change. Third, whilst 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 the PROM), this lack of symmetry seems to be more pronounced for the KOOS-PS and ICOAP.

Implications for clinicians and policymakers

In this study, we obtained evidence that supports the use of the 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, the OKS has demonstrated evidence of validity, reliability, and responsiveness in measuring the health state 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.

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 the 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.(34) Most importantly, more work is required to understand how results of the 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 the OKS might have on the quality of care that patients receive (i.e. the effect on the quality of service and influence on patients' clinical outcomes)(35).

ACKNOWLEDGMENTS

A copy of the OHS and OKS questionnaires and permission to use this measure can be acquired from Isis Innovation Ltd, the technology transfer company of the University of Oxford via website: http://www.isis-innovation.com/outcomes/index.html or email: healthoutcomes@isis.ox.ac.uk

COMPETING INTEREST DECLARATION

All authors have completed the Unified Competing Interest form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare that KKH, LDJ, AJP, DJB have no financial interests that may be relevant to the submitted work. JD is one of the original inventors of the OHS and OKS. She has received consultancy payments, via Isis Innovation, in relation to work involving both questionnaires.

CONTRIBUTIONS

Conception and design: JD, DJB, AJP

Acquisition of data: KKH, LDJ

Analysis and interpretation of data: KKH, JD, DJB, AJP

Drafting of the article and revision it critically for important intellectual content: KKH, JD, LDJ, DJB, AJP

Final approval of the article: KKH, JD, LDJ, DJB, AJP

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.

ETHICS APPROVAL

This study obtained ethics approval from the Oxfordshire Research Ethics Committee B (11/SC/005). Informed consent was obtained from all participants in the study.

ROLE OF THE FUNDING SOURCE

The research was supported under the general programme of research undertaken by the Nuffield Department of Orthopaedics, Rheumatology & Musculoskeletal Sciences as a NIHR Biomedical Research Unit.

DATA SHARING STATEMENT

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

REFERENCES

1. Dawson JF, R. Churchman, D. Verjee-Lorenz, A. Clayson, D. Oxford Knee Score (OKS) User Manual. 2010.

2. Dawson J, Fitzpatrick R, Murray D, Carr A. Questionnaire on the perceptions of patients about total knee replacement. Journal of Bone and Joint Surgery British Volume. 1998 Jan;80(1):63-9. PubMed PMID: 9460955. Epub 1998/02/14. eng.

3. Department of Health. Guidance of the Routine Collection of Patient Reported Outcome Measures (PROMs). In: Department of Health, editor. London2008.

4. Devlin NJ, Appleby J. GettinG the most out of proms. King's 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(9):1125-35.

6. Conaghan PG, Dickson J, Grant RL. Guidelines: Care and management of osteoarthritis in adults: summary of NICE guidance. BMJ: British Medical Journal. 2008;336(7642):502.

7. Hawker GA, Davis AM, French MR, Cibere J, Jordan JM, March L, et al. Development and preliminary psychometric testing of a new OA pain measure--an OARSI/OMERACT initiative. Osteoarthritis and Cartilage. 2008 Apr;16(4):409-14. PubMed PMID: 18381179. Epub 2008/04/03. eng.

8. 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 and Cartilage. 2008 May;16(5):542-50. PubMed PMID: 18294869. Epub 2008/02/26. eng.

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. Medical Care. 1996 Mar;34(3):220-33. PubMed PMID: 8628042. Epub 1996/03/01. eng.

10. Murray D, Fitzpatrick R, Rogers K, Pandit H, Beard D, Carr A, et al. The use of the Oxford hip and knee scores. The Journal of bone and joint surgery British volume. 2007;89(8):1010.

11. Harris K, Dawson J, Doll H, Field RE, Murray DW, Fitzpatrick R, et al. Can pain and function be distinguished in the Oxford Knee Score in a meaningful way? An exploratory and confirmatory factor analysis. Quality of Life Research. 2013:1-8.

12. Roos EM, Roos HP, Lohmander LS, Ekdahl C, Beynnon BD. Knee Injury and Osteoarthritis Outcome Score (KOOS)--development of a self-administered outcome measure. Journal of Orthopaedic and Sports Physical Therapy. 1998 Aug;28(2):88-96. PubMed PMID: 9699158. Epub 1998/08/12. eng.

13. Kellgren J, Lawrence J. Radiological assessment of osteo-arthrosis. Ann Rheum Dis. 1957;16(4):494-502.

14. De Vet HCW, Terwee CB, Mokkink LB, Knol DL. Measurement in Medicine a Practical Guide Cambridge: Cambridge University Press; 2011. Available from: http://public.eblib.com/EBLPublic/PublicView.do?ptiID=802925.

15. Kline P. An easy guide to factor analysis. 1993.

16. Ding L, Velicer WF, Harlow LL. Effects of estimation methods, number of indicators per factor, and improper solutions on structural equation modeling fit indices. Structural Equation Modeling: A Multidisciplinary Journal. 1995 1995/01/01;2(2):119-43.

17. Bentler PM, Chou CP. Practical issues in structural modeling. Sociological Methods & Research. 1987;16(1):78-117.

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(2):420.

20. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use: Oxford University Press, USA; 2008.

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

22. Association AER, Association AP, Education NCoMi, Educational JCoSf, Testing P. Standards for educational and psychological testing: Amer Educational Research Assn; 1999.

23. Browne MW, Cudeck R. Alternative ways of assessing model fit. Testing structural equation models. 1993;154:136–62.

24. Schumacker RE, Lomax RG. A beginner's guide to structural equation modeling. Mahwah, N.J.: Lawrence Erlbaum Associates; 2004.

25. Fitzpatrick R, Davey C, Buxton M, Jones D. Evaluating patient-based outcome measures for use in clinical trials. 1998 Contract No.: 14.

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

27. Beckerman H, Roebroeck M, Lankhorst G, Becher J, Bezemer P, Verbeek A. Smallest real difference, a link between reproducibility and responsiveness. Quality of Life Research. 2001;10(7):571-8.

28. De Vet HC, Terwee CB, Ostelo RW, Beckerman H, Knol DL, Bouter LM. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. Health and Quality of Life Outcomes. 2006;4(1):54.

29. National Institute for Health and Clinical Excellence (NICE). Osteoarthritis. National clinical guideline for care and management in adults. London: Royal College of Physicans; 2008.

30. Hawker G, Davis A, French M, Cibere J, Jordan J, March L, et al. Development and preliminary psychometric testing of a new OA pain measure–an OARSI/OMERACT initiative. Osteoarthritis and Cartilage. 2008;16(4):409-14.

31. Charter RA, Feldt LS. Confidence intervals for true scores: Is there a correct approach? Journal of Psychoeducational Assessment. 2001;19(4):350-64.

32. Rothman M, Burke L, Erickson P, Leidy NK, Patrick DL, Petrie CD. 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 in Health. 2009;12(8):1075-83.

33. Nunnally, JC. Psychometric theory. New York McGraw-Hill. 1994.

34. Dawson J, Doll H, Fitzpatrick R, Jenkinson C, Carr AJ. The routine use of patient reported outcome measures in healthcare settings. BMJ (Clinical research ed). 2010;340:c186.

35. Snyder CF, Aaronson NK, Choucair AK, Elliott TE, Greenhalgh J, Halyard MY, et al. Implementing patient-reported outcomes assessment in clinical practice: a review of the options and considerations. Quality of Life Research. 2012;21(8):1305-14.

STROBE Statement—Checklist of items that should be included in reports of ***cohort studies***

	Item No	Recommendation
Title and abstract **p.1 & 2**	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found
Introduction		
Background/rationale **p.3&4**	2	Explain the scientific background and rationale for the investigation being reported
Objectives **p.4**	3	State specific objectives, including any prespecified hypotheses
Methods		
Study design **p.5**	4	Present key elements of study design early in the paper
Setting **p. 5& 6**	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection
Participants **p. 5**	6	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up
		(*b*) For matched studies, give matching criteria and number of exposed and unexposed
Variables **n/a**	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable
Data sources/ measurement **p. 5&6**	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group
Bias **p.12**	9	Describe any efforts to address potential sources of bias
Study size **p.7**	10	Explain how the study size was arrived at
Quantitative variables **p. 7-8**	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why
Statistical methods	12	**p.7-13** (*a*) Describe all statistical methods, including those used to control for confounding
		n/a (*b*) Describe any methods used to examine subgroups and interactions
		p.8 (*c*) Explain how missing data were addressed
		p. 14/15 (*d*) If applicable, explain how loss to follow-up was addressed
		n/a (*e*) Describe any sensitivity analyses
Results		
Participants	13*	**p.14** (a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed
		p.14 (b) Give reasons for non-participation at each stage
		n/a (c) Consider use of a flow diagram
Descriptive data	14*	**p.14** (a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders
		p.14/15 (b) Indicate number of participants with missing data for each variable of interest
		n/a (c) Summarise follow-up time (eg, average and total amount)
Outcome data	15*	Report numbers of outcome events or summary measures over time
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and

n/a		their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included
		(*b*) Report category boundaries when continuous variables were categorized
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period
Other analyses **n/a**	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses
Discussion		
Key results **p.22/23**	18	Summarise key results with reference to study objectives
Limitations **p.23**	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias
Interpretation **p.25/26**	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence
Generalisability **p.22/23, 25/26**	21	Discuss the generalisability (external validity) of the study results
Other information		
Funding **p.28**	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based

*Give information separately for exposed and unexposed groups.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at http://www.strobe-statement.org.

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

Journal:	*BMJ Open*
Manuscript ID:	bmjopen-2013-003365.R1
Article Type:	Research
Date Submitted by the Author:	19-Jul-2013
Complete List of Authors:	Harris, Kristina; University of Oxford, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences Dawson, Jill; University of Oxford, Department of Public Health Jones, Luke; University of Oxford, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences Beard, David; University of Oxford, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences Price, Andrew; University of Oxford, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences
<b>Primary Subject Heading</b>:	Health services research
Secondary Subject Heading:	Research methods
Keywords:	Knee < ORTHOPAEDIC & TRAUMA SURGERY, Musculoskeletal disorders < ORTHOPAEDIC & TRAUMA SURGERY, STATISTICS & RESEARCH METHODS, HEALTH SERVICES ADMINISTRATION & MANAGEMENT

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

ABSTRACT

Objectives To assess the validity of the OKS for use in patients undergoing non-operative management for their knee OA within the 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, ICOAP, KOOS-PS, at baseline and three month follow up, transition item of change at three months.

Results The OKS summary scale and its pain and functional component subscales demonstrated good test-retest reliability (ICC 0.93, 0.91, 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 the OKS with other validated measures (KOOS-PS, ICOAP, SF12), which provided evidence of its construct validity and responsiveness.. Confirmatory Factor Analysis confirmed the structural validity of the OKS.

However, there was a lack of satisfactory evidence of structural validity for the ICOAP and KOOS. The minimum detectable change (MDC_{90}) was ±6 for the OKS (±16 for the OKS-PCS, and ±15 for the OKS-FCS), Minimal important changes were ≈ 7 for the OKS (≈17 for the OKS-PCS and ≈11 for the OKS-FCS) and minimal important differences were ≈6 for the OKS (≈14 for the OKS-PCS and ≈10 for the OKS-FCS). These values were also calculated for the ICOAP and the KOOS-PS.

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

Article focus

- The 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 the OKS in clinical practice, across the spectrum of OA disease.
- The aim of this study was to assess the measurement properties of the OKS when used with (individuals and groups of) patients who are undergoing non operative management for knee OA and compare it with most commonly used measures in this population of patients.

Key Messages

- The OKS, and its pain and functional component subscales, have acceptable evidence of its measurement properties when used in patients (individual and groups) undergoing non-operative treatment for knee OA.
- The OKS performed better than the ICOAP and the KOOS-PS (widely used outcome measures for knee OA) on several counts.

Strength 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.

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 post operation, in order to monitor and benchmark the performance of health providers.

The increasing popularity of the 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 the OKS in clinical practice as a means of standardizing clinical assessment, monitoring individual's self-reported health state across the spectrum of 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 the 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 (i.e., 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 the 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 the 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 the 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

We obtained ethical approval for a prospective cohort study from a local ethics committee (11/SC/005). Informed consent was obtained from all participants in the study.

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 SF12(9) patient-reported questionnaires.

The OKS is a 12 item questionnaire. It’s item content was devised using patient interviews, which addresses 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 the 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 standardized from 0 (worst) to 100 (best). Patients completed the OKS at baseline, 2 and 5 days (for test-retest reliability) and at 3 months.

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

The 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 the 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). The 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 standardized 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. The 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 regards 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 intra-operative documentation from previous knee arthroscopy or available MRIs to examine the extent of cartilage loss and confirm the diagnosis of osteoarthritis.

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 subjects per questionnaire item.(16, 17) These sample sizes are required for data analyses and should be adjusted (i.e. 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 subjects per item.

We analysed the data using SPSS version 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 the ICOAP and the OKS-PCS). We used non-parametric statistics, where appropriate. We did not use data imputation and we excluded cases with missing data on analysis by analysis basis (unless mentioned otherwise). We examined the following measurement properties of the 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 inter-correlation of all items) and free from measurement error. We used Cronbach’s alpha to assess the internal consistency of the OKS summary scale and its subscales. Alpha 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 the 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 the OKS on the individual level, we calculated the precision of individual scores at 90% CI level by multiplying the standard error of measurement (SEM) by the 2-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:

(i) the OKS and the physical component summary of the SF12 (PCS-12) are measuring sufficiently similar constructs (SF-PCS measures self-reported physical function and the 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,

(ii) the correlation between the OKS and the mental component summary of the SF12 (MCS-12) should be weaker than the one between the PCS-12 and OKS as these two scale constructs are not considered to be related to such an extent,

(iii) the OKS and KOOS-PS are measuring a sufficiently similar construct (the KOOS-PS measures self-reported knee function and the 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),

(iv) the OKS and the ICOAP are measuring sufficiently similar constructs (the ICOAP measures self-reported knee pain and the OKS measures self-reported pain and physical functioning related to the knee) that the correlation between these two measures should be strong and negative,

(v) the OKS-PCS should be correlated more with the ICOAP than with the KOOS-PS and negatively, in each case (the OKS-PCS measures self-reported knee pain as does the ICOAP),

(vi) the OKS-FCS should be correlated more with the KOOS-PS that the ICOAP and negatively (the OKS-FCS measures self-reported knee function, as does the KOOS-PS).

We classified correlations (r) as: r=0 to 0.29 as none/weak; r= 0.3 to 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 (i.e. latent variable) that is supposed to be measured.(20) For instance, if a measure is unidimensional (i.e. 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 the OKS by conducting Confirmatory Factor Analysis (CFA) that tested the fit of the one and two factor models of the OKS to the data, using LISREL V8.80 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, <0.08good fit, <0.1 satisfactory fit; RMSEA p test of close fit>0.05.(23) Additionally, we used the Chi-square (χ2) difference test and Parsimonious Normed Fit Index (PNFI) to compare the fit between the two models of the OKS and the ICOAP. (24) We calculated the χ2 difference tests by looking at the difference of χ2 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 the 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 subjects 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 the OKS, its subscales (OKS-PCS and OKS-FCS), KOOS-PS and the ICOAP were significantly different from 0. We constructed a Cumulative Distribution Function (CDF) plot for the; (i) OKS, (ii) OKS-PCS and ICOAP, and (iii) 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 the OKS:

(i) the change scores in the OKS should correlate strongly with the change scores in the KOOS-PS and ICOAP,

(ii) the change scores in the OKS-PCS should correlate more strongly with the change scores in the ICOAP than with the change scores in the KOOS-PS,

(iii) change scores for the OKS-FCS should correlate more strongly with change scores for the KOOS-PS than the change scores for the ICOAP.

All correlations should be negative.

There was a concern about the amount of overall change that can 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 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 interpretation of study results.

We assessed the interpretability by relating the change in the PROMs 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). 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 minimal important change (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. 137 patients were recruited in the study. 21 patients did not complete follow up questionnaires at 3 months, out of which 3 patients were listed for a surgical procedure (2 osteotomies and 1 arthroplasty) before 3 month follow-up, 7 patients no longer wanted to participate in the study and 11 were lost to follow-up. 134 patients were included in the main baseline analysis of whom 67 (50 %) were male and 67 patients were female. The mean age of patients was 59 (SD 11). 70% of patients had information on Body Mass Index (BMI), out of whom 30% were classified as obese (BMI>30), 41% as overweight (BMI between 25 and 29.9), 29% as normal weight (BMI between 18.5 and 24.9). No one was classified as underweight. All of the patients had a diagnosis of knee osteoarthritis. 2% of the patients had Kellgren-Lawrence (KL) grading of 0 (but evidence of cartilage loss on MRI scan), 8% had K-L of 1, 43% had K-L of 2, 16% had K-L of 3, 4% had K-L of 4. For 26% of cases, X-ray information was unavailable, of whom, 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 the primary care setting, different trust, or in a private clinic). All patients underwent standard non-operative management of knee OA.(8)

116 (87%) out of 134 recruited patients returned the questionnaires at three month follow up. There was no difference in age or BMI between those patients who did not respond at three 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 the OKS than responders at three months

(Independent samples t-test, p<0.05). A summary of the baseline scores is presented in Table 1.

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

	N		Mean (SD)	Median	Percentiles	
	Valid	Missing			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

For comparison, in the developmental study of the 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 of in the developmental study sample of 17 (when transformed to the 0-48 scoring system).

Reliability

Cronbach's alpha for the 12-item OKS was 0.94, 0.88 for the OKS-FCS and 0.90 for the OKS-PCS. For the ICOAP and KOOS-PS, the Cronbach's alpha was

0.97 and 0.94 respectively. The alpha 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, 0.91-0.95) for the summary OKS, 0.91 (95% CI, 0.88-0.94) for the OKS-PCS and 0.92 (95% CI, 0.90-0.95) for the OKS-FCS.

The standard error of measurement (SEM) for the summary OKS was 2.65 and the confidence in individual single score at 90% was ±4.4 OKS points. SEM for the OKS-FCS was 6.2 with ±10.2 90% CI for individual score and the SEM for the OKS-PCS was 6.9 with ±11.3 points as 90% CI for individual score (noting that the OKS-PCS and the OKS-FCS are presented on a different scale than the OKS). The SEM for the ICOAP was 9.68 with ±15.9 points as 90% CI for individual score. We calculated the SEM for the ICOAP by using the test-retest reliability that was reported in the developmental study (0.85)(6). For the 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) The SEM for the 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 the OKS with comparator instruments. Spearman's ρ between the baseline OKS, KOOS-PS, ICOAP, SF12-MCS and SF-12-PCS are shown in Table 2. The OKS correlated strongly with the KOOS-PS and ICOAP. The correlation between the SF12-PCS and the OKS was slightly higher than expected. As expected, the OKS was most poorly related to the SF12-MCS. The OKS-PCS correlated more with ICOAP than with KOOS-PS and the OKS-FCS correlated more with the KOOS-PS that with ICOAP. This evidence supports convergent and divergent validity of the OKS.

Table 2: Baseline Spearman's correlations between the scores. 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.

	OKS	OKS-PCS	OKS-FCS
ICOAP	-.879 (115)	-.884 (117)	-.792 (121)
KOOS-PS	-.849 (106)	-.779 (107)	-.867 (111)
PCS-12	.648 (121)	/	/
MCS-12	.370 (121)	/	/

Structural validity. 122 pre-operative OKSs, 125 pre-operative ICOAP and 113 pre-operative KOOS-PS were available for the CFA. Fit indices of one and two factor models for the OKS are presented in Table 3. Neither of the one and two factor models was rejected. Fit indices favoured the 2 factor model and the reduction in χ^2 in the two factor model was significant (χ2diff>7.879, with df=1, at the a=0.005 level).

Table 3. Fit indices of one and two-factor model of the OKS.

Factors	χ2 (p value)	df	RMSEA	90% CI RMSEA	RMSEA p test	CFI	SRMR	PNFI
1	71.32 (p=0.06)	54	0.052	0.00-0.08	0.44	0.99	0.043	0.80
2	56.64 (p=0.34)	53	0.024	0.0-0.06	0.83	1	0.039	0.79

Note. F=number of factors; 2 =chi-square; df=degrees of freedom; RMSEA=root mean square of approximation; CI=confidence intervals; p-value for test of close fit (RMSEA<.05); SRMR=standardized root mean square residual; CFI-comparative t index; PNFI=parsimonious normed fit index.

CFA revealed that a one-factor KOOS-PS model was rejected by the χ2 test and its RMSEA was above the highest acceptable threshold of an acceptable fit (0.1) (Table 4). The SRMR was acceptable and CFI was on the threshold of a good fit.

Both one and two factor ICOAP models were rejected by the χ2 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 χ2 for the 2 factor model of the ICOAP (χ2diff< 3.84, with df=1).

Table 4. Fit indices of one and two-factor model of the ICOAP and KOOS-PS.

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

Note. F=number of factors; 2 =chi-square; df=degrees of freedom; RMSEA=root mean square of approximation; CI=confidence intervals; p-value for test of close fit (RMSEA<.05); SRMR=standardized root mean square residual; CFI-comparative t index; PNFI=parsimonious normed fit index.

Responsiveness

Figure 1 shows the CDF plot for the OKS. The plot demonstrates that, based on the OKS summary score, approximately 15% of patients in the study experienced deterioration in health state, at three month follow up, that 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 the patients experienced improvement that was beyond the MIC of 7 points on the OKS.

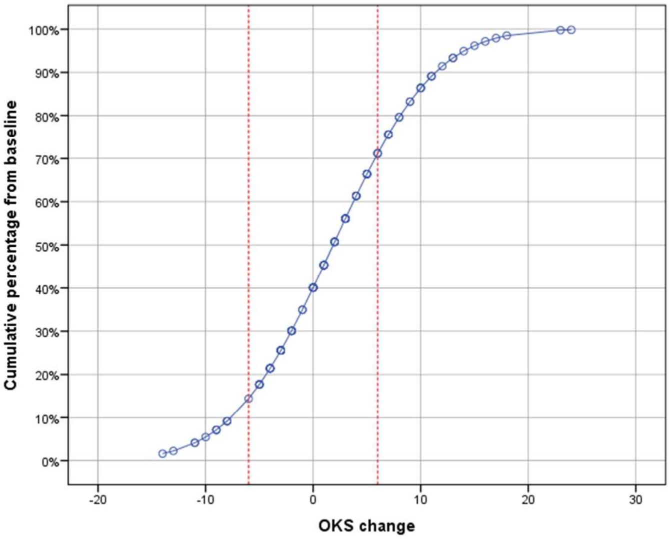

Figure 1. Cumulative percentage of patients experiencing the change on the OKS from baseline less or equal to the value on the x-axis. Red line marks the minimum detectable change beyond the measurement error of the score (MDC$_{90}$ of 6 points).

Table 5 shows the mean baseline, three month follow-up change scores, and p values for the significance of 3 month change and ES for the 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 the OKS-FCS.

Table 5: Significance of change in OKS, its subscales (OKS-PCS and OKS-FCS), ICOAP and KOOS-PS scores at three 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[a]	104	37.19 (25)	31.53 (25)	-5.66 (19)	<0.01	0.23
KOOS-PS[a]	92	39.42 (18)	34.88 (20)	-4.5 (14)	<0.01	0.25

Note. N=number of complete cases available for calculation of 3 month follow up; SD=standard deviation; ES=effect size;[a] The ICOAP and the KOOS-PS represent severity of the disease in the opposite direction from the OKS and its subscales.

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

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

	ICOAP	KOOS-PS
OKS	-.674 (96)	-.617 (87)
OKS-PCS	-.669 (99)	-.551 (88)
OKS-FCS	-.598 (100)	-.622 (90)

Note. 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.

Interpretability

Tables 7 and 8 present the percentage of responses for different response categories, effect sizes 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 the 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 interpretability indices.

Table 7: Number (N) and percentage of responses for different response categories with effect sizes (ES), mean score changes by response category and ANOVA tests for linear trend for the mean score across the three response categories for the 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	.7	.1	-.2
	P-value for linear trend	<.001	<.001	<.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	.8	.2	-.1
	P-value for linear trend	<.001	<.001	<.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	.5	.1	-.3
	P-value for linear trend	<.001	<.001	<.001

Table 8: Number (N) and percentage of responses for different response categories with effect sizes (ES), mean score changes by response category and ANOVA tests for linear trend for the mean score across the three response categories for the ICOAP and the 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	-.6	-.3	.1
	P-value for linear trend	<.003	<.003	<.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	-.8	-.3	.1
	P-value for linear trend	<.001	<.001	<.001

Table 9: Anchor based and distribution based MIC/MID values for the 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

Note. MDC_{90}=minimum detectable change; MID=minimum important difference; MIC=minimum important change.

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 are undergoing non-operative treatment for knee OA. The OKS summary scale and its subscales were validated against the KOOS-PS, the ICOAP (measures developed for use in patients with knee OA) and the SF-12 by testing logical *a priori* hypotheses regarding the construct validity and responsiveness of the OKS and its subscales in comparison to these other (validated) measures. Thus, CFA demonstrated excellent fit and confirmed the structural validity of the OKS and both subscales. Furthermore, assessment of test-retest reliability demonstrated that the OKS and its subscales could all be used both at 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 the OKS, ≈17 for the OKS-PCS, and ≈11 for the OKS-FCS can be used in cohort studies to assess if the change in the OKS (from baseline) is clinically relevant. Anchor based MID of ≈6 for the OKS, ≈14 for the OKS-PCS, and ≈10 for the 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 the MDC_{90} (≈6 points for the OKS, ≈16 points for the OKS-PCS, and ≈15 points for the 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 the OKS is used in a different population of patients (i.e. patients undergoing knee replacement surgery).

Limitations

Even though the reliability, construct validity and responsiveness of the OKS and its subscales have been proven to be 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 the 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 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 sub category 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 the 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, 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 done with awareness of its caveats. However, regardless of the shortcomings of the transition item, the same was used in the comparative analysis of interpretability between the OKS, its subscales, the KOOS-PS and the ICOAP and in terms of drawing conclusions about the comparative performance between the scores, this is not such a source of concern.

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

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

The 11-item ICOAP had a Cronbach's alpha of 0.97 (compared to the alpha of the OKS-PCS of 0.9) and the alpha was 0.94 for the KOOS-PS (compared to the alpha of 0.87 for the OKS-FCS). A high alpha value can mean that some of the items on a scale are redundant and this seems to be more of a concern for the ICOAP and KOOS-PS than for the OKS subscales. Furthermore, the reliability and precision of the score was better for the OKS and its subscales than for the KOOS-PS and ICOAP, which makes it more suitable to be used in clinical practice.

There was evidence to support both one and two factor models of the OKS, but no acceptable evidence of structural validity was found for the KOOS-PS or the ICOAP. The KOOS-PS and the one and two-factor ICOAP models were rejected by the $\chi 2$ 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 the ICOAP and KOOS-PS. It seems that these measures performed less well than the OKS in this regard. First, due to the fact that the ICOAP has low precision at the individual level (the MDC_{90} is almost 10 points larger than the 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 the ICOAP or the 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 the OKS and its subscales). This could indicate problems with the sensitivity of these scores to change. Third, whilst 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 the PROM), this lack of symmetry seems to be more pronounced for the KOOS-PS and ICOAP.

Implications for clinicians and policymakers

In this study, we obtained evidence that supports the use of the 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, the OKS has demonstrated evidence of validity, reliability, and responsiveness in measuring the health state 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 generalizable to the primary care setting. Studies have shown no significant differences in the 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 the 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 the 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 the OKS might have on the quality of care that patients receive (i.e. the effect on the quality of service and influence on patients' clinical outcomes). (34)

ACKNOWLEDGMENTS

A copy of the OHS and OKS questionnaires and permission to use this measure can be acquired from Isis Innovation Ltd, the technology transfer company of the University of Oxford via website: http://www.isis-innovation.com/outcomes/index.html or email: healthoutcomes@isis.ox.ac.uk

COMPETING INTEREST DECLARATION

All authors have completed the Unified Competing Interest form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare that KKH, LDJ, AJP, DJB have no financial interests that may be relevant to the submitted work. JD is one of the original inventors of the OHS and OKS. She has received consultancy payments, via Isis Innovation, in relation to work involving both questionnaires.

CONTRIBUTIONS

Conception and design: JD, DJB, AJP

Acquisition of data: KKH, LDJ

Analysis and interpretation of data: KKH, JD, DJB, AJP

Drafting of the article and revision it critically for important intellectual content: KKH, JD, LDJ, DJB, AJP

Final approval of the article: KKH, JD, LDJ, DJB, AJP

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.

ETHICS APPROVAL

This study obtained ethics approval from the Oxfordshire Research Ethics Committee B (11/SC/005). Informed consent was obtained from all participants in the study.

ROLE OF THE FUNDING SOURCE

The research was supported under the general programme of research undertaken by the Nuffield Department of Orthopaedics, Rheumatology & Musculoskeletal Sciences as a NIHR Biomedical Research Unit.

DATA SHARING STATEMENT

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

REFERENCES

1. Dawson J, Fitzpatrick M, Churchman D, Verjee-Lorenz A, Claysonm D. User Manual for the Oxford Knee Score (OKS). 2010.

2. Dawson J, Fitzpatrick R, Murray D, Carr A. Questionnaire on the perceptions of patients about total knee replacement. Journal of Bone and Joint Surgery British Volume. 1998 Jan;80(1):63-9. PubMed PMID: 9460955. Epub 1998/02/14. eng.

3. Department of Health. Guidance of the Routine Collection of Patient Reported Outcome Measures (PROMs). In: Department of Health, editor. London2008.

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. Quality of Life Research. 2008;17(9):1125-35.

6. Hawker GA, Davis AM, French MR, Cibere J, Jordan JM, March L, et al. Development and preliminary psychometric testing of a new OA pain measure--an OARSI/OMERACT initiative. Osteoarthritis and Cartilage. 2008 Apr;16(4):409-14. PubMed PMID: 18381179. Epub 2008/04/03. eng.

7. 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 and Cartilage. 2008 May;16(5):542-50. PubMed PMID: 18294869. Epub 2008/02/26. eng.

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. Medical Care. 1996 Mar;34(3):220-33. PubMed PMID: 8628042. Epub 1996/03/01. eng.

10. Murray DW, Fitzpatrick R, Rogers K, Pandit H, Beard DJ, Carr AJ, et al. The use of the Oxford hip and knee scores. Journal of Bone and Joint Surgery British Volume. 2007 August 1, 2007;89-B(8):1010-4.

11. Harris K, Dawson J, Doll H, Field R, Murray D, Fitzpatrick R, et al. Can pain and function be distinguished in the Oxford Knee Score in a meaningful way? An exploratory and confirmatory factor analysis. Quality of Life Research. 2013 2013/03/23:1-8. English.

12. Roos EM, Roos HP, Lohmander LS, Ekdahl C, Beynnon BD. Knee Injury and Osteoarthritis Outcome Score (KOOS)--development of a self-administered outcome measure. Journal of Orthopaedic and Sports Physical Therapy. 1998 Aug;28(2):88-96. PubMed PMID: 9699158. Epub 1998/08/12. eng.

13. Kellgren J, Lawrence J. Radiological assessment of osteo-arthrosis. Annals of the Rheumatic Diseases. 1957;16(4):494-502.

14. De Vet HCW, Terwee CB, Mokkink LB, Knol DL. Measurement in Medicine a Practical Guide Cambridge: Cambridge University Press; 2011. Available from: 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. Sociological Methods & Research. 1987;16(1):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. Structural Equation Modeling: A Multidisciplinary Journal. 1995 1995/01/01;2(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(2):420-8.

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

21. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use: Oxford University Press, USA; 2008.

22. AERA, APA, NCME. Standards for educational and psychological testing. Washington: American Educational Research Association; 1999. p. 194.

23. Browne MW, Cudeck R. Alternative ways of assessing model fit. Testing structural equation models. 1993;154:136–62.

24. Schumacker RE, Lomax RG. A beginner's guide to structural equation modeling. Mahwah, N.J.: Lawrence Erlbaum Associates; 2004.

25. Fitzpatrick R, Davey C, Buxton M, Jones D. Evaluating patient-based outcome measures for use in clinical trials. Health Technology Assessment 1998;2(14):1-74.

26. Beckerman H, Roebroeck M, Lankhorst G, Becher J, Bezemer P, Verbeek A. Smallest real difference, a link between reproducibility and responsiveness. Quality of Life Research. 2001;10(7):571-8.

27. De Vet HC, Terwee CB, Ostelo RW, Beckerman H, Knol DL, Bouter LM. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. Health and Quality of Life Outcomes. 2006;4(1):54.

28. Ornetti P, Perruccio A, Roos E, Lohmander L, Davis A, Maillefert J. Psychometric properties of the French translation of the reduced KOOS and HOOS (KOOS-PS and HOOS-PS). Osteoarthritis and Cartilage. 2009;17(12):1604-8.

29. Charter RA, Feldt LS. Confidence intervals for true scores: Is there a correct approach? Journal of Psychoeducational Assessment. 2001;19(4):350-64.

30. Rothman M, Burke L, Erickson P, Leidy NK, Patrick DL, Petrie CD. 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 in Health. 2009;12(8):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(6):771-6.

32. Hopman-Rock M, De Bock GH, Bijlsma JW, Springer MP, Hofman A, Kraaimaat FW. The pattern of health care utilization of elderly people with arthritic pain in the hip or knee. International Journal for Quality in Health Care. 1997;9(2):129-37.

33. Dawson J, Doll H, Fitzpatrick R, Jenkinson C, Carr AJ. The routine use of patient reported outcome measures in healthcare settings. BMJ (Clinical research ed). 2010;340:c186.

34. Snyder CF, Aaronson NK, Choucair AK, Elliott TE, Greenhalgh J, Halyard MY, et al. Implementing patient-reported outcomes assessment in clinical practice: a review of the options and considerations. Quality of Life Research. 2012;21(8):1305-14.

STROBE Statement—Checklist of items that should be included in reports of ***cohort studies***

	Item No	**Recommendation**
Title and abstract **p.1 & 2**	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found
Introduction		
Background/rationale **p.3&4**	2	Explain the scientific background and rationale for the investigation being reported
Objectives **p.4**	3	State specific objectives, including any prespecified hypotheses
Methods		
Study design **p.5**	4	Present key elements of study design early in the paper
Setting **p. 5& 6**	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection
Participants **p. 5**	6	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up
		(*b*) For matched studies, give matching criteria and number of exposed and unexposed
Variables **n/a**	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable
Data sources/ measurement **p. 5&6**	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group
Bias **p.12**	9	Describe any efforts to address potential sources of bias
Study size **p.7**	10	Explain how the study size was arrived at
Quantitative variables **p. 7-8**	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why
Statistical methods	12	**p.7-13** (*a*) Describe all statistical methods, including those used to control for confounding
		n/a (*b*) Describe any methods used to examine subgroups and interactions
		p.8 (*c*) Explain how missing data were addressed
		p. 14/15 (*d*) If applicable, explain how loss to follow-up was addressed
		n/a (*e*) Describe any sensitivity analyses
Results		
Participants	13*	**p.14** (a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed
		p.14 (b) Give reasons for non-participation at each stage
		n/a (c) Consider use of a flow diagram
Descriptive data	14*	**p.14** (a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders
		p.14/15 (b) Indicate number of participants with missing data for each variable of interest
		n/a (c) Summarise follow-up time (eg, average and total amount)
Outcome data	15*	Report numbers of outcome events or summary measures over time
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and

n/a		their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included
		(*b*) Report category boundaries when continuous variables were categorized
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period
Other analyses **n/a**	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses
Discussion		
Key results **p.22/23**	18	Summarise key results with reference to study objectives
Limitations **p.23**	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias
Interpretation **p.25/26**	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence
Generalisability **p.22/23, 25/26**	21	Discuss the generalisability (external validity) of the study results
Other information		
Funding **p.28**	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based

*Give information separately for exposed and unexposed groups.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at http://www.strobe-statement.org.

EXTENDING THE USE OF PROMS IN THE NHS: USING THE OXFORD KNEE SCORE ~~TO MONITOR THE PROGRESSION OF KNEE OSTEOARTHRITIS~~IN PATIENTS UNDERGOING NON-OPERATIVE MANAGEMENT FOR KNEE OSTEOARTHRITIS. A VALIDATION STUDY

ABSTRACT

Objectives To assess the validity of the OKS for use in patients undergoing non-operative management for their knee OA within the 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, ICOAP, KOOS-PS, at baseline and three month follow up, transition item of change at three months.

Results The OKS summary scale and its pain and functional component subscales demonstrated good test-retest reliability (ICC 0.93, 0.91, 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 the OKS with other validated measures (KOOS-PS, ICOAP, SF12), which provided evidence of its construct validity and responsiveness. ~~of the score and its subscales~~. Confirmatory Factor Analysis confirmed the structural validity of the OKS. However, there was a lack of satisfactory evidence of structural validity for the ICOAP and KOOS. The minimum detectable change (MDC_{90}) was ±6 for the OKS (±16 for the OKS-PCS, and ±15 for the OKS-FCS), Minimal important changes were ≈ 7 for the OKS (≈17 for the OKS–PCS and ≈11 for the OKS–FCS) and minimal important differences were ≈6 for the OKS (≈14 for the OKS–PCS and ≈10 for the OKS–FCS)~~-~~.~~and the , minimal important differences and the precision of the change score were also calculated for the OKS, its subscales,~~ These ~~These~~ values were also calculated for the ICOAP and the KOOS-PS.

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

Article focus

- The 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 the OKS in clinical practice, across the spectrum of OA disease.
- The aim of this study was to assess the measurement properties of the OKS when used with (individuals and groups of) patients who are undergoing non operative management for knee OA and compare it with most commonly used measures in this population of patients.

Key Messages

- The OKS, and its pain and functional component subscales, have acceptable evidence of its measurement properties when used in patients (individual and groups) undergoing non-operative treatment for knee OA.
- The OKS performed better than the ICOAP and the KOOS-PS (widely used outcome measures for knee OA) on several counts.

Strength 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 in1998 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 the OKS was extended in 2009, when it was adopted by the 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 post operation, in order to monitor and benchmark the performance of health providers.

The increasing popularity of the 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 the OKS in clinical practice as a means of standardizing clinical assessment, monitoring individual’s self-reported health state across the spectrum of 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 the 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 (i.e., 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 the 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 the 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 the 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

We obtained ethical approval for a prospective cohort study from a local ethics committee (11/SC/005). Informed consent was obtained from all participants in the study.

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 ~~Intermittent and Constant Osteoarthritis Pain (~~ICOAP(6)~~)~~, the ~~Knee Injury and Osteoarthritis Score-Physical function Short form (~~KOOS-PS~~)~~(7), and the SF12(9) patient-reported questionnaires.

The OKS is a 12 item questionnaire. It's item content was devised using patient interviews, which addresses 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 the 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 standardized from 0 (worst) to 100 (best). Patients completed the OKS at baseline, 2 and 5 days (for test-retest reliability) and at 3 months.

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

The 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 the 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). The 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 standardized 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. The 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 regards 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 intra-operative documentation from previous knee arthroscopy or available MRIs to examine the extent of cartilage loss and confirm the diagnosis of osteoarthritis.

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 subjects per questionnaire item.(16, 17) These sample sizes are required for data analyses and should be adjusted (i.e. 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 subjects per item.

We analysed the data using SPSS version 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 the ICOAP and the OKS-PCS). We used non-parametric statistics, where appropriate. We did not use data imputation and we excluded cases with missing data on analysis by analysis basis (unless mentioned otherwise). We examined the following measurement properties of the 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 inter-correlation of all items) and free from measurement error. We used Cronbach's alpha to assess the internal consistency of the OKS summary scale and its subscales. Alpha 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 the 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 the OKS on the individual level, we calculated the precision of individual scores at 90% CI level by multiplying the standard error of measurement (SEM) by the 2-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:

(i) the OKS and the physical component summary of the SF12 (PCS-12) are measuring sufficiently similar constructs (SF-PCS measures self-reported physical function and the 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,

(ii) the correlation between the OKS and the mental component summary of the SF12 (MCS-12) should be weaker than the one between the PCS-12 and OKS as these two scale constructs are not considered to be related to such an extent,

(iii) the OKS and KOOS-PS are measuring a sufficiently similar construct (the KOOS-PS measures self-reported knee function and the 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),

(iv) the OKS and the ICOAP are measuring sufficiently similar constructs (the ICOAP measures self-reported knee pain and the OKS measures self-reported pain and physical functioning related to the knee) that the correlation between these two measures should be strong and negative,

(v) the OKS-PCS should be correlated more with the ICOAP than with the KOOS-PS and negatively, in each case (the OKS-PCS measures self-reported knee pain as does the ICOAP),

(vi) the OKS-FCS should be correlated more with the KOOS-PS that the ICOAP and negatively (the OKS-FCS measures self-reported knee function, as does the KOOS-PS).

We classified correlations (r) as: r=0 to 0.29 as none/weak; r= 0.3 to 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 (i.e. latent variable) that is supposed to be measured.(20) For instance, if a measure is unidimensional (i.e. 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 the OKS by conducting Confirmatory Factor Analysis (CFA) that tested the fit of the one and two factor models of the OKS to the data, using LISREL V8.80 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, <0.08good fit, <0.1 satisfactory fit; RMSEA p test of close fit>0.05.(23) Additionally, we used the Chi-square (χ2) difference test and Parsimonious Normed Fit Index (PNFI) to compare the fit between the two models of the OKS and the ICOAP. (24) We calculated the χ2 difference tests by looking at the difference of χ2 of two models along with the difference in their degrees of freedom. ~~We checked the χ2 difference, with its degrees of freedom in the χ2 distribution table. If this value is statistically significant, then the model with more degrees of freedom is favoured.~~

Responsiveness

The ability of a measure to detect meaningful clinical change (where it has occurred) over time is critical for the use and the 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 subjects 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 the OKS, its subscales (OKS-PCS and OKS-FCS), KOOS-PS and the ICOAP were significantly different from 0. We constructed a Cumulative Distribution Function (CDF) plot for the; (i) OKS, (ii) OKS-PCS and ICOAP, and (iii) 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 the OKS:

(i) the change scores in the OKS should correlate strongly with the change scores in the KOOS-PS and ICOAP,

(ii) the change scores in the OKS-PCS should correlate more strongly with the change scores in the ICOAP than with the change scores in the KOOS-PS,

(iii) change scores for the OKS-FCS should correlate more strongly with change scores for the KOOS-PS than the change scores for the ICOAP.

All correlations should be negative.

There was a concern about the amount of overall change that can 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 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 interpretation of study results.

We assessed the interpretability by relating the change in the PROMs 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). 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 minimal important change

(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})~~, which was obtained by multiplying the SEM with the z-value at the 90% level and the square root of two (to account for two measurement occasions).(~~(26, 27))

RESULTS

Sample characteristics. 137 patients were recruited in the study. 21 patients did not complete follow up questionnaires at 3 months, out of which 3 patients were listed for a surgical procedure (2 osteotomies and 1 arthroplasty) before 3 month follow-up, 7 patients no longer wanted to participate in the study and 11 were lost to follow-up. 134 patients were included in the main baseline analysis of whom 67 (50 %) were male and 67 patients were female. The mean age of patients was 59 (SD 11)~~, which is about 10 years less than the average age of the developmental sample of the OKS~~. 70% of patients had information on Body Mass Index (BMI), out of whom 30% were classified as obese (BMI>30), 41% as overweight (BMI between 25 and 29.9), 29% as normal weight (BMI between 18.5 and 24.9). No one was classified as underweight. All of the patients had a diagnosis of knee osteoarthritis. 2% of the patients had Kellgren-Lawrence (KL) grading of 0 (but evidence of cartilage loss on MRI scan), 8% had K-L of 1, 43% had K-L of 2, 16% had K-L of 3, 4% had K-L of 4. For 26% of cases, X-ray information was unavailable, of whom, 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 the primary care setting, different trust, or in a private clinic). All patients underwent standard non-operative management of knee OA.(8)

116 (87%) out of 134 recruited patients returned the questionnaires at three month follow up. There was no difference in age or BMI between those patients who did not respond at three 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 the OKS than responders at three months (Independent samples t-test, p<0.05). A summary of the baseline scores is presented in Table 1.

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

	N		Mean (SD)	Median	Percentiles	
	Valid	Missing			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

For comparison, in the developmental study of the 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 of in the developmental study sample of 17 (when transformed to the 0-48 scoring system).

Reliability

Cronbach's alpha for the 12-item OKS was 0.94, 0.88 for the OKS-FCS and 0.90 for the OKS-PCS. For the ICOAP and KOOS-PS, the Cronbach's alpha was

0.97 and 0.94 respectively. The alpha 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, 0.91-0.95) for the summary OKS, 0.91 (95% CI, 0.88-0.94) for the OKS-PCS and 0.92 (95% CI, 0.90-0.95) for the OKS-FCS.

The standard error of measurement (SEM) for the summary OKS was 2.65 and the confidence in individual single score at 90% was ±4.4 OKS points. SEM for the OKS-FCS was 6.2 with ±10.2 90% CI for individual score and the SEM for the OKS-PCS was 6.9 with ±11.3 points as 90% CI for individual score (noting that the OKS-PCS and the OKS-FCS are presented on a different scale than the OKS). The SEM for the ICOAP was 9.68 with ±15.9 points as 90% CI for individual score. We calculated the SEM for the ICOAP by using the test-retest reliability that was reported in the developmental study (0.85)(6). For the 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) The SEM for the 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 the OKS with comparator instruments. Spearman's ρ between the baseline OKS, KOOS-PS, ICOAP, SF12-MCS and SF-12-PCS are shown in Table 2. The OKS correlated strongly with the KOOS-PS and ICOAP. The correlation between the SF12-PCS and the OKS was slightly higher than expected. As expected, the OKS was most poorly related to the SF12-MCS. The OKS-PCS correlated more with ICOAP than with KOOS-PS and the OKS-FCS correlated more with the KOOS-PS that with ICOAP. This evidence supports convergent and divergent validity of the OKS.

Table 2: Baseline Spearman's correlations between the scores. 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.

	OKS	OKS-PCS	OKS-FCS
ICOAP	-.879 (115)	-.884 (117)	-.792 (121)
KOOS-PS	-.849 (106)	-.779 (107)	-.867 (111)
PCS-12	.648 (121)	/	/
MCS-12	.370 (121)	/	/

Structural validity. 122 pre-operative OKSs, 125 pre-operative ICOAP and 113 pre-operative KOOS-PS were available for the CFA. Fit indices of one and two factor models for the OKS are presented in Table 3. Neither of the one and two factor models was rejected. Fit indices favoured the 2 factor model and the reduction in χ^2 in the two factor model was significant (χ2diff>7.879, with df=1, at the a=0.005 level).

Table 3. Fit indices of one and two-factor model of the OKS.

Factors	χ2 (p value)	df	RMSEA	90% CI RMSEA	RMSEA p test	CFI	SRMR	PNFI
1	71.32 (p=0.06)	54	0.052	0.00-0.08	0.44	0.99	0.043	0.80
2	56.64 (p=0.34)	53	0.024	0.0-0.06	0.83	1	0.039	0.79

Note. F=number of factors; 2 =chi-square; df=degrees of freedom; RMSEA=root mean square of approximation; CI=confidence intervals; p-value for test of close fit (RMSEA<.05); SRMR=standardized root mean square residual; CFI-comparative t index; PNFI=parsimonious normed fit index.

CFA revealed that a one-factor KOOS-PS model was rejected by the χ2 test and its RMSEA was above the highest acceptable threshold of an acceptable fit (0.1) (Table 4). The SRMR was acceptable and CFI was on the threshold of a good fit.

Both one and two factor ICOAP models were rejected by the χ2 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 χ2 for the 2 factor model of the ICOAP (χ2diff< 3.84, with df=1).

Table 4. Fit indices of one and two-factor model of the ICOAP and KOOS-PS.

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

Note. F=number of factors; 2 =chi-square; df=degrees of freedom; RMSEA=root mean square of approximation; CI=confidence intervals; p-value for test of close fit (RMSEA<.05); SRMR=standardized root mean square residual; CFI-comparative fit index; PNFI=parsimonious normed fit index.

Responsiveness

Figure 1 shows the CDF plot for the OKS. The plot demonstrates that, based on the OKS summary score, approximately 15% of patients in the study experienced deterioration in health state, at three month follow up, that 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 the patients experienced improvement that was beyond the MIC of 7 points on the OKS.

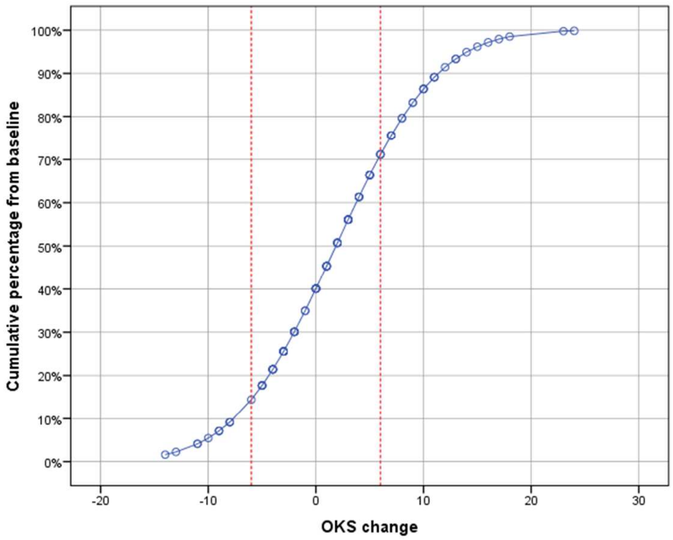

Figure 1. Cumulative percentage of patients experiencing the change on the OKS from baseline less or equal to the value on the x-axis. Red line marks the minimum detectable change beyond the measurement error of the score (MDC$_{90}$ of 6 points).

Table 5 shows the mean baseline, three month follow-up change scores, and p values for the significance of 3 month change and ES for the 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 the OKS-FCS.

Table 5: Significance of change in OKS, its subscales (OKS-PCS and OKS-FCS), ICOAP and KOOS-PS scores at three 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[a]	104	37.19 (25)	31.53 (25)	-5.66 (19)	<0.01	0.23
KOOS-PS[a]	92	39.42 (18)	34.88 (20)	-4.5 (14)	<0.01	0.25

Note. N=number of complete cases available for calculation of 3 month follow up; SD=standard deviation; ES=effect size;[a] The ICOAP and the KOOS-PS represent severity of the disease in the opposite direction from the OKS and its subscales.

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

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

	ICOAP	KOOS-PS
OKS	-.674 (96)	-.617 (87)
OKS-PCS	-.669 (99)	-.551 (88)
OKS-FCS	-.598 (100)	-.622 (90)

Note. 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.

Interpretability

Tables 7 and 8 present the percentage of responses for different response categories, effect sizes 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 the 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. ~~Here, the OKS and OKS-PCS mean differences were close to (and generally just above) scale MDC/MID values and thus likely beyond measurement error, while the OKS-FCS mean differences were just less than the subscale’s MDC/MID values. All OKS scales’ mean differences were greater than the scales’ relevant SEM values.~~

Table 9 presents the summary of interpretability indices.

Table 7: Number (N) and percentage of responses for different response categories with effect sizes (ES), mean score changes by response category and ANOVA tests for linear trend for the mean score across the three response categories for the 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	.7	.1	-.2
	P-value for linear trend	<.001	<.001	<.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	.8	.2	-.1
	P-value for linear trend	<.001	<.001	<.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	.5	.1	-.3
	P-value for linear trend	<.001	<.001	<.001

Table 8: Number (N) and percentage of responses for different response categories with effect sizes (ES), mean score changes by response category and ANOVA tests for linear trend for the mean score across the three response categories for the ICOAP and the 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	-.6	-.3	.1
	P-value for linear trend	<.003	<.003	<.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	-.8	-.3	.1
	P-value for linear trend	<.001	<.001	<.001

Table 9: Anchor based and distribution based MIC/MID values for the 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

Note. MDC_{90}=minimum detectable change; MID=minimum important difference; MIC=minimum important change.

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 are undergoing non-operative treatment for knee OA. The OKS summary scale and its subscales were validated against the KOOS-PS, the ICOAP (measures developed for use in patients with knee OA) and the SF-12 by testing logical *a priori* hypotheses regarding the construct validity and responsiveness of the OKS and its subscales in comparison to these other (validated) measures. Thus, CFA demonstrated excellent fit and confirmed the structural validity of the OKS and both subscales. Furthermore, assessment of test-retest reliability demonstrated that the OKS and its subscales could all be used both at 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 the OKS, ≈17 for the OKS-PCS, and ≈11 for the OKS-FCS can be used in cohort studies to assess if the change in the OKS (from baseline) is clinically relevant. Anchor based MID of ≈6 for the OKS, ≈14 for the OKS-PCS, and ≈10 for the 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 the MDC_{90} (≈6 points for the OKS, ≈16 points for the OKS-PCS, and ≈15 points for the 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 the OKS is used in a different population of patients (i.e. patients undergoing knee replacement surgery).

Limitations

Even though the reliability, construct validity and responsiveness of the OKS and its subscales have been proven to be 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 the 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 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) ~~On the other hand it may be assumed to be appropriate if the context is considered to be ‘similar enough’.(20)~~ A counter~~nother~~ argument is that it is unrealistic to have a new/different measure (and a new study conducted to design and test one) for every possible sub category 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 the 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, 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 done with awareness of its caveats. However, regardless of the shortcomings of the- transition item, the same was used in the comparative analysis of interpretability between the OKS, its subscales, the KOOS-PS and the ICOAP and in terms of drawing conclusions about the comparative performance between the scores, this is not such a source of concern.

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

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

The 11-item ICOAP had a Cronbach's alpha of 0.97 (compared to the alpha of the OKS-PCS of 0.9) and the alpha was 0.94 for the KOOS-PS (compared to the alpha of 0.87 for the OKS-FCS). A high alpha value can mean that some of the items on a scale are redundant and this seems to be more of a concern for the ICOAP and KOOS-PS than for the OKS subscales. Furthermore, the reliability and precision of the score was better for the OKS and its subscales than for the KOOS-PS and ICOAP, which makes it more suitable to be used in clinical practice.

There was evidence to support both one and two factor models of the OKS, but no acceptable evidence of structural validity was found for the KOOS-PS or the ICOAP. The KOOS-PS and the one and two-factor ICOAP models were rejected by the $\chi 2$ 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 the ICOAP and KOOS-PS. It seems that these measures performed less well than the OKS in this regard. First, due to the fact that the ICOAP has low precision at the individual level (the MDC_{90} is almost 10 points larger than the 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 the ICOAP or the 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 the OKS and its subscales). This could indicate problems with the sensitivity of these scores to change. Third, whilst 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 the PROM), this lack of symmetry seems to be more pronounced for the KOOS-PS and ICOAP.

Implications for clinicians and policymakers

In this study, we obtained evidence that supports the use of the 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, the OKS has demonstrated evidence of validity, reliability, and responsiveness in measuring the health state 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 generalizable to the primary care setting. Studies have shown no significant differences in the 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 the 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 the 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 the OKS might have on the quality of care that patients receive (i.e. the effect on the quality of service and influence on patients' clinical outcomes). (34)

ACKNOWLEDGMENTS

A copy of the OHS and OKS questionnaires and permission to use this measure can be acquired from Isis Innovation Ltd, the technology transfer company of the University of Oxford via website: http://www.isis-innovation.com/outcomes/index.html or email: healthoutcomes@isis.ox.ac.uk

COMPETING INTEREST DECLARATION

All authors have completed the Unified Competing Interest form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare that KKH, LDJ, AJP, DJB have no financial interests that may be relevant to the submitted work. JD is one of the original inventors of the OHS and OKS. She has received consultancy payments, via Isis Innovation, in relation to work involving both questionnaires.

CONTRIBUTIONS

Conception and design: JD, DJB, AJP

Acquisition of data: KKH, LDJ

Analysis and interpretation of data: KKH, JD, DJB, AJP

Drafting of the article and revision it critically for important intellectual content: KKH, JD, LDJ, DJB, AJP

Final approval of the article: KKH, JD, LDJ, DJB, AJP

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.

ETHICS APPROVAL

This study obtained ethics approval from the Oxfordshire Research Ethics Committee B (11/SC/005). Informed consent was obtained from all participants in the study.

ROLE OF THE FUNDING SOURCE

The research was supported under the general programme of research undertaken by the Nuffield Department of Orthopaedics, Rheumatology & Musculoskeletal Sciences as a NIHR Biomedical Research Unit.

DATA SHARING STATEMENT

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

REFERENCES

1. Dawson J, Fitzpatrick M, Churchman D, Verjee-Lorenz A, Claysonm D. User Manual for the Oxford Knee Score (OKS). 2010.

2. Dawson J, Fitzpatrick R, Murray D, Carr A. Questionnaire on the perceptions of patients about total knee replacement. Journal of Bone and Joint Surgery British Volume. 1998 Jan;80(1):63-9. PubMed PMID: 9460955. Epub 1998/02/14. eng.

3. Department of Health. Guidance of the Routine Collection of Patient Reported Outcome Measures (PROMs). In: Department of Health, editor. London2008.

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. Quality of Life Research. 2008;17(9):1125-35.

6. Hawker GA, Davis AM, French MR, Cibere J, Jordan JM, March L, et al. Development and preliminary psychometric testing of a new OA pain measure--an OARSI/OMERACT initiative. Osteoarthritis and Cartilage. 2008 Apr;16(4):409-14. PubMed PMID: 18381179. Epub 2008/04/03. eng.

7. 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 and Cartilage. 2008 May;16(5):542-50. PubMed PMID: 18294869. Epub 2008/02/26. eng.

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. Medical Care. 1996 Mar;34(3):220-33. PubMed PMID: 8628042. Epub 1996/03/01. eng.

10. Murray DW, Fitzpatrick R, Rogers K, Pandit H, Beard DJ, Carr AJ, et al. The use of the Oxford hip and knee scores. Journal of Bone and Joint Surgery British Volume. 2007 August 1, 2007;89-B(8):1010-4.

11. Harris K, Dawson J, Doll H, Field R, Murray D, Fitzpatrick R, et al. Can pain and function be distinguished in the Oxford Knee Score in a meaningful way? An exploratory and confirmatory factor analysis. Quality of Life Research. 2013 2013/03/23:1-8. English.

12. Roos EM, Roos HP, Lohmander LS, Ekdahl C, Beynnon BD. Knee Injury and Osteoarthritis Outcome Score (KOOS)--development of a self-administered outcome measure. Journal of Orthopaedic and Sports Physical Therapy. 1998 Aug;28(2):88-96. PubMed PMID: 9699158. Epub 1998/08/12. eng.

13. Kellgren J, Lawrence J. Radiological assessment of osteo-arthrosis. Annals of the Rheumatic Diseases. 1957;16(4):494-502.

14. De Vet HCW, Terwee CB, Mokkink LB, Knol DL. Measurement in Medicine a Practical Guide Cambridge: Cambridge University Press; 2011. Available from: 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. Sociological Methods & Research. 1987;16(1):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. Structural Equation Modeling: A Multidisciplinary Journal. 1995 1995/01/01;2(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(2):420-8.

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

21. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use: Oxford University Press, USA; 2008.

22. AERA, APA, NCME. Standards for educational and psychological testing. Washington: American Educational Research Association; 1999. p. 194.

23. Browne MW, Cudeck R. Alternative ways of assessing model fit. Testing structural equation models. 1993;154:136–62.

24. Schumacker RE, Lomax RG. A beginner's guide to structural equation modeling. Mahwah, N.J.: Lawrence Erlbaum Associates; 2004.

25. Fitzpatrick R, Davey C, Buxton M, Jones D. Evaluating patient-based outcome measures for use in clinical trials. Health Technology Assessment 1998;2(14):1-74.

26. Beckerman H, Roebroeck M, Lankhorst G, Becher J, Bezemer P, Verbeek A. Smallest real difference, a link between reproducibility and responsiveness. Quality of Life Research. 2001;10(7):571-8.

27. De Vet HC, Terwee CB, Ostelo RW, Beckerman H, Knol DL, Bouter LM. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. Health and Quality of Life Outcomes. 2006;4(1):54.

28. Ornetti P, Perruccio A, Roos E, Lohmander L, Davis A, Maillefert J. Psychometric properties of the French translation of the reduced KOOS and HOOS (KOOS-PS and HOOS-PS). Osteoarthritis and Cartilage. 2009;17(12):1604-8.

29. Charter RA, Feldt LS. Confidence intervals for true scores: Is there a correct approach? Journal of Psychoeducational Assessment. 2001;19(4):350-64.

30. Rothman M, Burke L, Erickson P, Leidy NK, Patrick DL, Petrie CD. 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 in Health. 2009;12(8):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(6):771-6.

32. Hopman-Rock M, De Bock GH, Bijlsma JW, Springer MP, Hofman A, Kraaimaat FW. The pattern of health care utilization of elderly people with arthritic pain in the hip or knee. International Journal for Quality in Health Care. 1997;9(2):129-37.

33. Dawson J, Doll H, Fitzpatrick R, Jenkinson C, Carr AJ. The routine use of patient reported outcome measures in healthcare settings. BMJ (Clinical research ed). 2010;340:c186.

34. Snyder CF, Aaronson NK, Choucair AK, Elliott TE, Greenhalgh J, Halyard MY, et al. Implementing patient-reported outcomes assessment in clinical practice: a review of the options and considerations. Quality of Life Research. 2012;21(8):1305-14.