



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

Journal:	<i>BMJ Open</i>
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

SCHOLARONE™  
Manuscripts

# 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.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

**INTRODUCTION**

The Oxford Knee Score (OKS) is a widely used patient reported outcome measure (PROM), originally developed in 1998 to be used in clinical trials for assessing the patient-perceived outcomes of knee replacement surgery. In this form it has proven to be reliable, valid and responsive.(1, 2) The remit of 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.

1       **METHODS**

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

5  
6       Study procedures and assessments

7               This study took place at an orthopaedic centre between June 2011 and  
8       August 2012. Patients were eligible for inclusion if they were referred for knee  
9       problems, had a confirmed diagnosis of knee OA and were enrolled in the non-  
10      operative management pathway for their knee OA (as recommended by the  
11      National Institute of Clinical Excellence (NICE)(6)). Treatments for patients were  
12      tailored individually, taking into account patients' preferences and needs. As such,  
13      they represented standard practice in the NHS. All patients who met these criteria  
14      were sent an invitation letter containing information about the study, consent forms  
15      and baseline questionnaires. Patients who consented to participate in the study  
16      were asked to complete the OKS(2) the Intermittent and Constant Osteoarthritis  
17      Pain (ICOAP)(7) the Knee Injury and Osteoarthritis Score-Physical function Short  
18      form (KOOS-PS)(8) and SF12(9) patient-reported questionnaires.

19              The OKS is a 12 item questionnaire. It's item content was devised using  
20      patient interviews, which addresses pain and functional impairment in relation to  
21      their knee, in patients who are undergoing knee replacement surgery.(2) Likert  
22      responses are recommended to be scored from 0 to 4, which are summed to  
23      produce a summary score of 0 (worst) to 48 (best)(10). More recently, we  
24      presented evidence (in the context of joint replacement) that supported the original  
25      conceptual basis of the OKS using its composite summary scales, but which also

1  
2  
3 26 offered an option to perform additional analyses using pain and function  
4  
5 27 subscales.(11) The Pain Component Score (OKS-PCS) consists of items 2, 3, 7,  
6  
7 28 11 and 12 and the Functional Component Score (OKS-FCS) consists of items 1, 4,  
8  
9 29 5, 6, 8, 9 and 10. Subscale raw scores are standardized from 0 (worst) to 100  
10  
11 30 (best). Patients completed the OKS at baseline, 2 and 5 days (for test-retest  
12  
13 31 reliability) and at 3 months.

14  
15 32 We asked the patients to complete the KOOS-PS and ICOAP at baseline  
16  
17 33 and 3 month follow up. These scores were developed to measure pain and  
18  
19 34 functional disability related to knee OA, and are now a recommended outcome  
20  
21 35 measures by the Osteoarthritis Research Society International (OARSI).

22  
23 36 The KOOS-PS consists of 7 Likert-response items and was developed from  
24  
25 37 a longer version of the questionnaire (KOOS(12)) using Rasch analysis to measure  
26  
27 38 physical function in patients with various degrees of knee OA. It is scored as the  
28  
29 39 KOOS from 0 (best) to 4 (worst), with a summary raw score ranging from 0 to 28.  
30  
31 40 The score is converted to a true interval score that ranges from 0 (best) to 100  
32  
33 41 (worst). The ICOAP is an 11 item questionnaire whose items were informed from  
34  
35 42 focus groups with patients with hip or knee OA. It has two subscales that measure  
36  
37 43 the intermittent and constant pain with a standardized summary score ranging from  
38  
39 44 0 (best) to 100 (worst).

40  
41 45 Patients also completed the generic SF-12, a 12-item general health  
42  
43 46 measure with 8 items that have Likert-type response categories and 4 items with  
44  
45 47 dichotomous (yes/no) response categories. The SF-12 is scored as a Physical  
46  
47 48 Component Summary (PCS) and Mental Component Summary (MCS) ranging  
48  
49 49 from 0 (worst) to 100 (best).

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65  
66  
67  
68  
69  
70  
71  
72  
73  
74  
75  
76  
77  
78  
79  
80  
81  
82  
83  
84  
85  
86  
87  
88  
89  
90  
91  
92  
93  
94  
95  
96  
97  
98  
99  
100  
101  
102  
103  
104  
105  
106  
107  
108  
109  
110  
111  
112  
113  
114  
115  
116  
117  
118  
119  
120  
121  
122  
123  
124  
125  
126  
127  
128  
129  
130  
131  
132  
133  
134  
135  
136  
137  
138  
139  
140  
141  
142  
143  
144  
145  
146  
147  
148  
149  
150  
151  
152  
153  
154  
155  
156  
157  
158  
159  
160  
161  
162  
163  
164  
165  
166  
167  
168  
169  
170  
171  
172  
173  
174  
175  
176  
177  
178  
179  
180  
181  
182  
183  
184  
185  
186  
187  
188  
189  
190  
191  
192  
193  
194  
195  
196  
197  
198  
199  
200  
201  
202  
203  
204  
205  
206  
207  
208  
209  
210  
211  
212  
213  
214  
215  
216  
217  
218  
219  
220  
221  
222  
223  
224  
225  
226  
227  
228  
229  
230  
231  
232  
233  
234  
235  
236  
237  
238  
239  
240  
241  
242  
243  
244  
245  
246  
247  
248  
249  
250  
251  
252  
253  
254  
255  
256  
257  
258  
259  
260  
261  
262  
263  
264  
265  
266  
267  
268  
269  
270  
271  
272  
273  
274  
275  
276  
277  
278  
279  
280  
281  
282  
283  
284  
285  
286  
287  
288  
289  
290  
291  
292  
293  
294  
295  
296  
297  
298  
299  
300  
301  
302  
303  
304  
305  
306  
307  
308  
309  
310  
311  
312  
313  
314  
315  
316  
317  
318  
319  
320  
321  
322  
323  
324  
325  
326  
327  
328  
329  
330  
331  
332  
333  
334  
335  
336  
337  
338  
339  
340  
341  
342  
343  
344  
345  
346  
347  
348  
349  
350  
351  
352  
353  
354  
355  
356  
357  
358  
359  
360  
361  
362  
363  
364  
365  
366  
367  
368  
369  
370  
371  
372  
373  
374  
375  
376  
377  
378  
379  
380  
381  
382  
383  
384  
385  
386  
387  
388  
389  
390  
391  
392  
393  
394  
395  
396  
397  
398  
399  
400  
401  
402  
403  
404  
405  
406  
407  
408  
409  
410  
411  
412  
413  
414  
415  
416  
417  
418  
419  
420  
421  
422  
423  
424  
425  
426  
427  
428  
429  
430  
431  
432  
433  
434  
435  
436  
437  
438  
439  
440  
441  
442  
443  
444  
445  
446  
447  
448  
449  
450  
451  
452  
453  
454  
455  
456  
457  
458  
459  
460  
461  
462  
463  
464  
465  
466  
467  
468  
469  
470  
471  
472  
473  
474  
475  
476  
477  
478  
479  
480  
481  
482  
483  
484  
485  
486  
487  
488  
489  
490  
491  
492  
493  
494  
495  
496  
497  
498  
499  
500  
501  
502  
503  
504  
505  
506  
507  
508  
509  
510  
511  
512  
513  
514  
515  
516  
517  
518  
519  
520  
521  
522  
523  
524  
525  
526  
527  
528  
529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539  
540  
541  
542  
543  
544  
545  
546  
547  
548  
549  
550  
551  
552  
553  
554  
555  
556  
557  
558  
559  
560  
561  
562  
563  
564  
565  
566  
567  
568  
569  
570  
571  
572  
573  
574  
575  
576  
577  
578  
579  
580  
581  
582  
583  
584  
585  
586  
587  
588  
589  
590  
591  
592  
593  
594  
595  
596  
597  
598  
599  
600  
601  
602  
603  
604  
605  
606  
607  
608  
609  
610  
611  
612  
613  
614  
615  
616  
617  
618  
619  
620  
621  
622  
623  
624  
625  
626  
627  
628  
629  
630  
631  
632  
633  
634  
635  
636  
637  
638  
639  
640  
641  
642  
643  
644  
645  
646  
647  
648  
649  
650  
651  
652  
653  
654  
655  
656  
657  
658  
659  
660  
661  
662  
663  
664  
665  
666  
667  
668  
669  
670  
671  
672  
673  
674  
675  
676  
677  
678  
679  
680  
681  
682  
683  
684  
685  
686  
687  
688  
689  
690  
691  
692  
693  
694  
695  
696  
697  
698  
699  
700  
701  
702  
703  
704  
705  
706  
707  
708  
709  
710  
711  
712  
713  
714  
715  
716  
717  
718  
719  
720  
721  
722  
723  
724  
725  
726  
727  
728  
729  
730  
731  
732  
733  
734  
735  
736  
737  
738  
739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
750  
751  
752  
753  
754  
755  
756  
757  
758  
759  
760  
761  
762  
763  
764  
765  
766  
767  
768  
769  
770  
771  
772  
773  
774  
775  
776  
777  
778  
779  
780  
781  
782  
783  
784  
785  
786  
787  
788  
789  
790  
791  
792  
793  
794  
795  
796  
797  
798  
799  
800  
801  
802  
803  
804  
805  
806  
807  
808  
809  
810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863  
864  
865  
866  
867  
868  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917  
918  
919  
920  
921  
922  
923  
924  
925  
926  
927  
928  
929  
930  
931  
932  
933  
934  
935  
936  
937  
938  
939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
970  
971  
972  
973  
974  
975  
976  
977  
978  
979  
980  
981  
982  
983  
984  
985  
986  
987  
988  
989  
990  
991  
992  
993  
994  
995  
996  
997  
998  
999  
1000

50           Lastly, we asked the patients to complete a transition question in regards to

51           the change they experienced from the baseline measurement: “Compared to one

52           week before your clinic visit, please indicate how much your knee problem has

53           changed?” The question had three response options: “1. My knee has got better; 2.

54           My knee has stayed the same; 3. My knee has got worse”.

55           We supplemented patient reported outcome data with information on their

56           body mass index (BMI) and the degree of structural changes observed in the knee,

57           which was available from the patients’ medical records. An orthopaedic surgeon

58           (LDJ) performed Kellgren-Lawrence (K-L) grading using available knee OA

59           radiographs. The degree of structural changes in the knee was classified using (K-

60           L) grading.(13) In the absence of X-rays, we assessed intra-operative

61           documentation from previous knee arthroscopy or available MRIs to examine the

62           extent of cartilage loss and confirm the diagnosis of osteoarthritis.

63

64           Statistical methods

65           The recommended minimum sample sizes for validation studies (based on

66           optimal numbers for correlations) often range from 50 to 100.(14, 15) For

67           confirmatory factor analysis (CFA) the literature agrees with a minimum sample

68           size of about 100-150 or about 10 subjects per questionnaire item.(16, 17) These

69           sample sizes are required for data analyses and should be adjusted (i.e.

70           increased) for the risk of loss to follow up. In this study we stopped recruiting when

71           the dataset enabled us to perform CFA with at least 10 subjects per item.

72           We analysed the data using SPSS version 20 and LISREL V 8.80. Baseline

73           and 3 month follow up scores were generally non-normally distributed and change

74           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 than 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.<sup>(21)</sup> 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  $\chi^2$  ( $p > 0.05$ ), standardised root mean square residual (SRMR)  $> 0.08$ , comparative fit index (CFI)  $> 0.95$ , root mean square error of approximation (RMSEA):  $< 0.05$  close fit,  $< 0.08$  good fit,  $< 0.1$  satisfactory fit; RMSEA p test of close fit  $> 0.05$ .<sup>(23)</sup> Additionally, we used the Chi-square ( $\chi^2$ ) difference test and Parsimonious Normed Fit Index (PNFI) to compare the fit between the two models of the OKS and the ICOAP.<sup>(24)</sup> We calculated the  $\chi^2$  difference tests by looking at the difference of  $\chi^2$  of two models along with the difference in their degrees of freedom. We checked the  $\chi^2$  difference, with its degrees of freedom in the  $\chi^2$  distribution table. If this value is statistically significant, then the model with more degrees of freedom is favoured.

## 149 Responsiveness

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

159 We used a one sample t-test (2 tailed) to assess if the changes at 3 months  
160 for the OKS, its subscales (OKS-PCS and OKS-FCS), KOOS-PS and the ICOAP  
161 were significantly different from 0. We constructed a Cumulative Distribution  
162 Function (CDF) plot for the; (i) OKS, (ii) OKS-PCS and ICOAP, and (iii) OKS-FCS  
163 and KOOS-PS to examine the proportion of individual patients who experienced  
164 deterioration and improvement beyond the measurement error of the instrument at  
165 the individual level and to compare the proportion of change in pain and function  
166 detected by the different measures.

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

- 170 (i) the change scores in the OKS should correlate strongly with the change  
171 scores in the KOOS-PS and ICOAP,  
172 (ii) the change scores in the OKS-PCS should correlate more strongly with  
173 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

199 (MIC). The difference in the change score between the groups of patients who  
200 responded with “My knee has stayed the same” and “My knee has got better on  
201 the global item of change was taken as the minimal important difference (MID).  
202 Finally, the minimum change in the instrument that represents real change (beyond  
203 measurement error) was calculated using the Minimum Detectable Change  
204 ( $MDC_{90}$ ), which was obtained by multiplying the SEM with the z-value at the 90%  
205 level and the square root of two (to account for two measurement occasions).(27,  
206 28)

207

208

## 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  $\pm 4.4$  OKS points. SEM for the OKS-FCS was 6.2 with  $\pm 10.2$  90% CI for individual score and the SEM for the OKS-PCS was 6.9 with  $\pm 11.3$  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  $\pm 16.6$  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).<sup>(30)</sup> 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  $\pm 11.1$  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  $\rho$  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 than 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  $\chi^2$  in the two factor model was significant ( $\chi^2_{diff}>7.879$ , with  $df=1$ , at the  $\alpha=0.005$  level).

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

Factors	$\chi^2$ (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;  $\chi^2$ =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  $\chi^2$  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  $\chi^2$  test and both models had RMSEA values far above the lowest threshold of an acceptable fit. However, SRMR and CFI were acceptable for both scores. There was no significant reduction (at the 0.05 level) in  $\chi^2$  for the 2 factor model of the ICOAP ( $\chi^2_{diff}< 3.84$ , with  $df=1$ ).

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

	$\chi^2$ (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;  $\chi^2$ =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<sub>90</sub> 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.

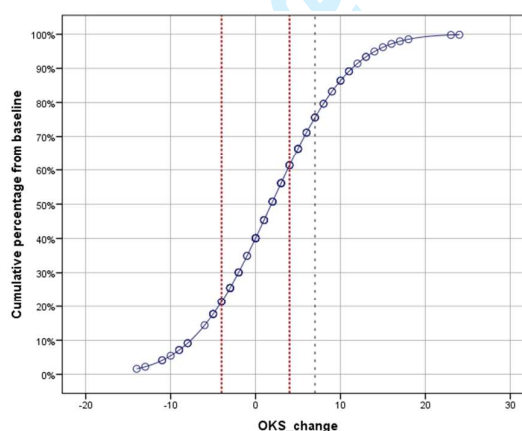


Figure 1. Cumulative percentage of patients experiencing the change on the OKS from baseline less or equal to the value on the x-axis. Red line marks the minimum detectable change beyond the measurement error of the score (MDC<sub>90</sub> of 4 points).

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 <sup>a</sup>	104	37.19 (25)	31.53 (25)	-5.66 (19)	<0.01	0.23
KOOS-PS <sup>a</sup>	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;<sup>a</sup> 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 <sub>90</sub>	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<sub>90</sub>=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).<sup>(31)</sup>

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  $\approx 7$  for the OKS,  $\approx 17$  for the OKS-PCS, and  $\approx 11$  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  $\approx 6$  for the OKS,  $\approx 14$  for the OKS-PCS, and  $\approx 10$  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  $MDC_{90}$  ( $\approx 4$  points for the OKS,  $\approx 11$  points for the OKS-PCS, and  $\approx 10$  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  $\chi^2$  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<sub>90</sub> 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).

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

**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](mailto:healthoutcomes@isis.ox.ac.uk)

**COMPETING INTEREST DECLARATION**

All authors have completed the Unified Competing Interest form at [www.icmje.org/coi\\_disclosure.pdf](http://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-arthritis. 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 Physicians; 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

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

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
<b>Title and abstract</b> <b>p.1 &amp; 2</b>	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
<b>Introduction</b>		
Background/rationale <b>p.3&amp;4</b>	2	Explain the scientific background and rationale for the investigation being reported
Objectives <b>p.4</b>	3	State specific objectives, including any prespecified hypotheses
<b>Methods</b>		
Study design <b>p.5</b>	4	Present key elements of study design early in the paper
Setting <b>p. 5&amp; 6</b>	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection
Participants <b>p. 5</b>	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 <b>n/a</b>	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable
Data sources/ measurement <b>p. 5&amp;6</b>	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 <b>p.12</b>	9	Describe any efforts to address potential sources of bias
Study size <b>p.7</b>	10	Explain how the study size was arrived at
Quantitative variables <b>p. 7-8</b>	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why
Statistical methods	12	<b>p.7-13</b> (a) Describe all statistical methods, including those used to control for confounding <b>n/a</b> (b) Describe any methods used to examine subgroups and interactions <b>p.8</b> (c) Explain how missing data were addressed <b>p. 14/15</b> (d) If applicable, explain how loss to follow-up was addressed <b>n/a</b> (e) Describe any sensitivity analyses
<b>Results</b>		
Participants	13*	<b>p.14</b> (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 <b>p.14</b> (b) Give reasons for non-participation at each stage <b>n/a</b> (c) Consider use of a flow diagram
Descriptive data	14*	<b>p.14</b> (a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders <b>p.14/15</b> (b) Indicate number of participants with missing data for each variable of interest <b>n/a</b> (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	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses
n/a		
<b>Discussion</b>		
Key results	18	Summarise key results with reference to study objectives
p.22/23		
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias
p.23		
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence
p.25/26		
Generalisability	21	Discuss the generalisability (external validity) of the study results
p.22/23, 25/26		
<b>Other information</b>		
Funding	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
p.28		

\*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:	<i>BMJ Open</i>
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

SCHOLARONE™  
Manuscripts

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  $\pm 6$  for the OKS ( $\pm 16$  for the OKS-PCS, and  $\pm 15$  for the OKS-FCS), Minimal important changes were  $\approx 7$  for the OKS ( $\approx 17$  for the OKS-PCS and  $\approx 11$  for the OKS-FCS) and minimal important differences were  $\approx 6$  for the OKS ( $\approx 14$  for the OKS-PCS and  $\approx 10$  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

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

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

subscale.(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 patients 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

1  
2  
3 51 week before your clinic visit, please indicate how much your knee problem has  
4  
5 52 changed?" The question had three response options: "1. My knee has got better; 2.  
6  
7 53 My knee has stayed the same; 3. My knee has got worse".  
8  
9

10 54 We supplemented patient reported outcome data with information on their  
11  
12 55 body mass index (BMI) and the degree of structural changes observed in the knee,  
13  
14 56 which was available from the patients' medical records. An orthopaedic surgeon  
15  
16 57 (LDJ) performed Kellgren-Lawrence (K-L) grading using available knee OA  
17  
18 58 radiographs. (13) The degree of structural changes in the knee was classified  
19  
20 59 using (K-L) grading. In the absence of X-rays, we assessed intra-operative  
21  
22 60 documentation from previous knee arthroscopy or available MRIs to examine the  
23  
24 61 extent of cartilage loss and confirm the diagnosis of osteoarthritis.  
25  
26  
27  
28  
29

### 30 Statistical methods

31  
32 64 The recommended minimum sample sizes for validation studies (based on  
33  
34 65 optimal numbers for correlations) often range from 50 to 100.(14, 15) For  
35  
36 66 confirmatory factor analysis (CFA) the literature agrees with a minimum sample  
37  
38 67 size of about 100-150 or about 10 subjects per questionnaire item.(16, 17) These  
39  
40 68 sample sizes are required for data analyses and should be adjusted (i.e.  
41  
42 69 increased) for the risk of loss to follow up. In this study we stopped recruiting when  
43  
44 70 the dataset enabled us to perform CFA with at least 10 subjects per item.  
45  
46

47 71 We analysed the data using SPSS version 20 and LISREL V 8.80. Baseline  
48  
49 72 and 3 month follow up scores were generally non-normally distributed and change  
50  
51 73 scores approximated to normal (except the ICOAP and the OKS-PCS). We used  
52  
53 74 non-parametric statistics, where appropriate. We did not use data imputation and  
54  
55 75 we excluded cases with missing data on analysis by analysis basis (unless  
56  
57  
58  
59  
60

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)

1  
2  
3 101 We examined the construct validity of the OKS summary scale and its  
4  
5 102 subscales by testing an *a priori* set of hypotheses about the expected relationships  
6  
7 103 between the instruments at baseline:  
8

9  
10 104 (i) the OKS and the physical component summary of the SF12 (PCS-12) are  
11  
12 105 measuring sufficiently similar constructs (SF-PCS measures self-reported physical  
13  
14 106 function and the OKS measures self-reported pain and physical functioning related  
15  
16 107 to the knee), so the correlation between these two instruments' scales should be  
17  
18 108 moderate and in the same direction,  
19

20  
21 109 (ii) the correlation between the OKS and the mental component summary of  
22  
23 110 the SF12 (MCS-12) should be weaker than the one between the PCS-12 and OKS  
24  
25 111 as these two scale constructs are not considered to be related to such an extent,  
26

27 112 (iii) the OKS and KOOS-PS are measuring a sufficiently similar construct  
28  
29 113 (the KOOS-PS measures self-reported knee function and the OKS measures self-  
30  
31 114 reported pain and physical functioning related to the knee) that the correlation  
32  
33 115 between these two measures should be strong and negative (as scores go in the  
34  
35 116 opposite direction),  
36

37  
38 117 (iv) the OKS and the ICOAP are measuring sufficiently similar constructs  
39  
40 118 (the ICOAP measures self-reported knee pain and the OKS measures self-  
41  
42 119 reported pain and physical functioning related to the knee) that the correlation  
43  
44 120 between these two measures should be strong and negative,  
45  
46

47 121 (v) the OKS-PCS should be correlated more with the ICOAP than with the  
48  
49 122 KOOS-PS and negatively, in each case (the OKS-PCS measures self-reported  
50  
51 123 knee pain as does the ICOAP),  
52  
53  
54  
55  
56  
57  
58  
59  
60



(vi) the OKS-FCS should be correlated more with the KOOS-PS than 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.<sup>(20)</sup> 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  $\chi^2$  ( $p>0.05$ ), standardised root mean square residual (SRMR) $>0.08$ , comparative fit index (CFI) $>0.95$ , root mean square error of approximation (RMSEA):  $<0.05$  close fit,  $<0.08$  good fit,  $<0.1$  satisfactory fit; RMSEA p test of close fit $>0.05$ .<sup>(23)</sup> Additionally, we used the Chi-square ( $\chi^2$ ) difference test and Parsimonious Normed Fit Index (PNFI) to compare the fit between the two models of the OKS and the ICOAP. <sup>(24)</sup> We calculated the  $\chi^2$  difference tests by looking at the difference of  $\chi^2$  of two models along with the difference in their degrees of freedom.

**Responsiveness**

The ability of a measure to detect meaningful clinical change (where it has occurred) over time is critical for the use and the application of a measure. <sup>(25)</sup>



1  
2  
3 149 This change might occur following an intervention, or just occur 'naturally' during a  
4  
5 150 period of observation. Generally, as with construct validity, responsiveness is  
6  
7 151 assessed by testing *a priori* hypotheses about the relationship of the changes in  
8  
9 152 one measure to the changes in another (validated) measure, or with reference to a  
10  
11 153 change in a gold standard (as with testing criterion validity). Responsiveness can  
12  
13 154 also be tested with reference to a transition item, where the responsiveness is  
14  
15 155 tested only in subjects who have reported that clinical change has occurred.  
16  
17

18  
19 156 We used a one sample t-test (2 tailed) to assess if the changes at 3 months  
20  
21 157 for the OKS, its subscales (OKS-PCS and OKS-FCS), KOOS-PS and the ICOAP  
22  
23 158 were significantly different from 0. We constructed a Cumulative Distribution  
24  
25 159 Function (CDF) plot for the; (i) OKS, (ii) OKS-PCS and ICOAP, and (iii) OKS-FCS  
26  
27 160 and KOOS-PS to examine the proportion of individual patients who experienced  
28  
29 161 deterioration and improvement beyond the measurement error of the instrument at  
30  
31 162 the individual level and to compare the proportion of change in pain and function  
32  
33 163 detected by the different measures.  
34  
35

36 164 As with construct validity, we tested the responsiveness by setting *a priori*  
37  
38 165 hypotheses about the direction and magnitude of changes of the validated  
39  
40 166 comparator instruments and the OKS:  
41

42 167 (i) the change scores in the OKS should correlate strongly with the change  
43  
44 168 scores in the KOOS-PS and ICOAP,  
45

46 169 (ii) the change scores in the OKS-PCS should correlate more strongly with  
47  
48 170 the change scores in the ICOAP than with the change scores in the KOOS-PS,  
49

50 171 (iii) change scores for the OKS-FCS should correlate more strongly with  
51  
52 172 change scores for the KOOS-PS than the change scores for the ICOAP.  
53

54 173 All correlations should be negative.  
55  
56  
57  
58  
59  
60

1  
2  
3 174  
4  
5 175       There was a concern about the amount of overall change that can be  
6  
7 176 experienced as a result of such a management pathway (which included a wide  
8  
9 177 range of individually tailored treatments administered to a heterogeneous sample),  
10  
11 178 so we additionally defined the construct of change using a patient rated item of  
12  
13 179 change. We then used the responses to this item to calculate anchor based values  
14  
15 180 of minimal important change and difference.  
16  
17  
18  
19 181

20  
21 182   **Interpretability**  
22

23 183       Interpretability is defined as the degree to which one can assign qualitative  
24  
25 184 meaning to a quantitative score.(20) In clinical trials, this issue can concern the  
26  
27 185 question of what is considered to be a 'good', 'bad' or 'indifferent' outcome (as  
28  
29 186 measured by a particular criterion or score) and what is considered to be a  
30  
31 187 clinically relevant change. The minimum amount of change that is discerned as  
32  
33 188 meaningful by patients is particularly important as it affects interpretation of study  
34  
35 189 results.  
36  
37

38  
39 190       We assessed the interpretability by relating the change in the PROMs  
40  
41 191 scores to the patient reported item of change (using an anchor based method) and  
42  
43 192 by relating the observed change in the score to its measurement error at the  
44  
45 193 individual level (using a distribution based method). Average change in the score  
46  
47 194 associated with the group of patients who responded with “My knee has got better”  
48  
49 195 on the transition item was taken as the anchor based minimal important change  
50  
51 196 (MIC). The difference in the change score between the groups of patients who  
52  
53 197 responded with “My knee has stayed the same” and “My knee has got better on  
54  
55 198 the global item of change was taken as the minimal important difference (MID).  
56  
57  
58  
59  
60

1  
2  
3 199 Finally, the minimum change in the instrument that represents real change (beyond  
4  
5 200 measurement error) was calculated using the Minimum Detectable Change  
6  
7 201 ( $MDC_{90}$ )(26, 27))  
8  
9  
10 202  
11  
12 203  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

204 **RESULTS**

206 **Sample characteristics.** 137 patients were recruited in the study. 21  
207 patients did not complete follow up questionnaires at 3 months, out of which 3  
208 patients were listed for a surgical procedure (2 osteotomies and 1 arthroplasty)  
209 before 3 month follow-up, 7 patients no longer wanted to participate in the study  
210 and 11 were lost to follow-up. 134 patients were included in the main baseline  
211 analysis of whom 67 (50 %) were male and 67 patients were female. The mean  
212 age of patients was 59 (SD 11). 70% of patients had information on Body Mass  
213 Index (BMI), out of whom 30% were classified as obese (BMI>30), 41% as  
214 overweight (BMI between 25 and 29.9), 29% as normal weight (BMI between 18.5  
215 and 24.9). No one was classified as underweight. All of the patients had a  
216 diagnosis of knee osteoarthritis. 2% of the patients had Kellgren-Lawrence (KL)  
217 grading of 0 (but evidence of cartilage loss on MRI scan), 8% had K-L of 1, 43%  
218 had K-L of 2, 16% had K-L of 3, 4% had K-L of 4. For 26% of cases, X-ray  
219 information was unavailable, of whom, 20% had their diagnosis confirmed on the  
220 basis of MRI, while 6% of patients did not have X-rays or MRIs accessible  
221 (however, these patients had the diagnosis of OA previously confirmed in the  
222 primary care setting, different trust, or in a private clinic). All patients underwent  
223 standard non-operative management of knee OA.(8)

224 116 (87%) out of 134 recruited patients returned the questionnaires at three  
225 month follow up. There was no difference in age or BMI between those patients  
226 who did not respond at three months versus those who did, but baseline OKS was  
227 different between these groups. The group that did not respond had scored, on  
228 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.

231

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).

235

236

	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

237

238

239

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 17 (when transformed to the 0-48 scoring system).

247

## 248 Reliability

249

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

251

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  $\pm 4.4$  OKS points. SEM for the OKS-FCS was 6.2 with  $\pm 10.2$  90% CI for individual score and the SEM for the OKS-PCS was 6.9 with  $\pm 11.3$  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  $\pm 15.9$  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  $\pm 11.1$  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  $\rho$  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 than with ICOAP. This evidence supports convergent and divergent validity of the OKS.

280

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)	/	/

286

287

**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  $\chi^2$  in the two factor model was significant ( $\chi^2_{diff} > 7.879$ , with  $df=1$ , at the  $\alpha=0.005$  level).

294

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

Factors	$\chi^2$ (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  $\chi^2$  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  $\chi^2$  test and both models had RMSEA values far above the lowest threshold of an acceptable fit. However, SRMR and CFI were acceptable for both scores. There was no significant reduction (at the 0.05 level) in  $\chi^2$  for the 2 factor model of the ICOAP ( $\chi^2_{diff} < 3.84$ , with  $df=1$ ).

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

	$\chi^2$ (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;  $\chi^2$ =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<sub>90</sub> 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.



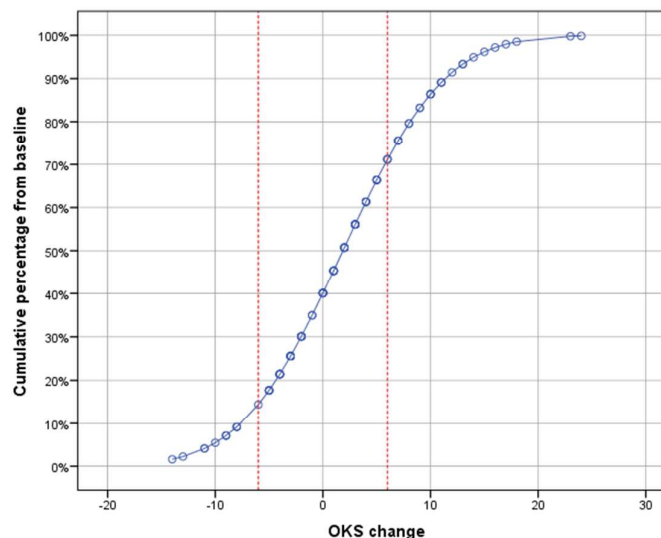


Figure 1. Cumulative percentage of patients experiencing the change on the OKS from baseline less or equal to the value on the x-axis. Red line marks the minimum detectable change beyond the measurement error of the score ( $MDC_{90}$  of 6 points).

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 <sup>a</sup>	104	37.19 (25)	31.53 (25)	-5.66 (19)	<0.01	0.23
KOOS-PS <sup>a</sup>	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; <sup>a</sup> 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 <sub>90</sub>	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<sub>90</sub>=minimum detectable change; MID=minimum important difference; MIC=minimum important change.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

**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  $\approx 7$  for the OKS,  $\approx 17$  for the OKS-PCS, and  $\approx 11$  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  $\approx 6$  for the OKS,  $\approx 14$  for the OKS-PCS, and  $\approx 10$  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  $MDC_{90}$  ( $\approx 6$  points for the OKS,  $\approx 16$  points for the OKS-PCS, and  $\approx 15$  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.<sup>(30)</sup> 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).<sup>(18)</sup> 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 <sup>(21)</sup>, 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<sub>90</sub> 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)

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

**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](mailto:healthoutcomes@isis.ox.ac.uk)

**COMPETING INTEREST DECLARATION**

All authors have completed the Unified Competing Interest form at [www.icmje.org/coi\\_disclosure.pdf](http://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.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

**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 Physicians; 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-arthritis. *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, Kraaijaat 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
<b>Title and abstract</b> <b>p.1 &amp; 2</b>	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
<b>Introduction</b>		
Background/rationale <b>p.3&amp;4</b>	2	Explain the scientific background and rationale for the investigation being reported
Objectives <b>p.4</b>	3	State specific objectives, including any prespecified hypotheses
<b>Methods</b>		
Study design <b>p.5</b>	4	Present key elements of study design early in the paper
Setting <b>p. 5&amp; 6</b>	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection
Participants <b>p. 5</b>	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 <b>n/a</b>	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable
Data sources/ measurement <b>p. 5&amp;6</b>	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 <b>p.12</b>	9	Describe any efforts to address potential sources of bias
Study size <b>p.7</b>	10	Explain how the study size was arrived at
Quantitative variables <b>p. 7-8</b>	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why
Statistical methods	12	<b>p.7-13</b> (a) Describe all statistical methods, including those used to control for confounding <b>n/a</b> (b) Describe any methods used to examine subgroups and interactions <b>p.8</b> (c) Explain how missing data were addressed <b>p. 14/15</b> (d) If applicable, explain how loss to follow-up was addressed <b>n/a</b> (e) Describe any sensitivity analyses
<b>Results</b>		
Participants	13*	<b>p.14</b> (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 <b>p.14</b> (b) Give reasons for non-participation at each stage <b>n/a</b> (c) Consider use of a flow diagram
Descriptive data	14*	<b>p.14</b> (a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders <b>p.14/15</b> (b) Indicate number of participants with missing data for each variable of interest <b>n/a</b> (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	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses
n/a		
<b>Discussion</b>		
Key results	18	Summarise key results with reference to study objectives
p.22/23		
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias
p.23		
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence
p.25/26		
Generalisability	21	Discuss the generalisability (external validity) of the study results
p.22/23, 25/26		
<b>Other information</b>		
Funding	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
p.28		

\*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<sub>90</sub>) was  $\pm 6$  for the OKS ( $\pm 16$  for the OKS-PCS, and  $\pm 15$  for the OKS-FCS), Minimal important changes were  $\approx 7$  for the OKS ( $\approx 17$  for the OKS-PCS and  $\approx 11$  for the OKS-FCS) and minimal important differences were  $\approx 6$  for the OKS ( $\approx 14$  for the OKS-PCS and  $\approx 10$  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.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

**INTRODUCTION**

The Oxford Knee Score (OKS) is a widely used patient reported outcome measure (PROM), originally developed in 1998 to be used in clinical trials for assessing the patient-perceived outcomes of knee replacement surgery. In this form it has proven to be reliable, valid and responsive.(1, 2) The remit of 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).

1       **METHODS**

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

5       Study procedures and assessments

6               This study took place at an orthopaedic centre between June 2011 and  
7       August 2012. Patients were eligible for inclusion if they were referred for knee  
8       problems, had a confirmed diagnosis of knee OA and were enrolled in the non-  
9       operative management pathway for their knee OA (as recommended by the  
10      National Institute of Clinical Excellence (NICE)(8)). Treatments for patients were  
11      tailored individually, taking into account patients' preferences and needs. As such,  
12      they represented standard practice in the NHS. All patients who met these criteria  
13      were sent an invitation letter containing information about the study, consent forms  
14      and baseline questionnaires. Patients who consented to participate in the study  
15      were asked to complete the OKS(2) the ~~Intermittent and Constant Osteoarthritis~~  
16      ~~Pain (ICOAP(6))~~, the ~~Knee Injury and Osteoarthritis Score-Physical function Short~~  
17      ~~form (KOOS-PS)(7))~~ and the SF12(9) patient-reported questionnaires.

18              The OKS is a 12 item questionnaire. It's item content was devised using  
19      patient interviews, which addresses pain and functional impairment in relation to  
20      their knee, in patients who are undergoing knee replacement surgery.(2) Likert  
21      responses are recommended to be scored from 0 to 4, which are summed to  
22      produce a summary score of 0 (worst) to 48 (best)(10). More recently, we  
23      presented evidence (in the context of joint replacement) that supported the original  
24      conceptual basis of the OKS using its composite summary scales, but which also  
25



26 offered an option to perform additional analyses using pain and function  
27 subscales.(11) The Pain Component Score (OKS-PCS) consists of items 2, 3, 7,  
28 11 and 12 and the Functional Component Score (OKS-FCS) consists of items 1, 4,  
29 5, 6, 8, 9 and 10. Subscale raw scores are standardized from 0 (worst) to 100  
30 (best). Patients completed the OKS at baseline, 2 and 5 days (for test-retest  
31 reliability) and at 3 months.

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

36 The KOOS-PS consists of 7 Likert-response items and was developed from  
37 a longer version of the questionnaire (KOOS(12)) using Rasch analysis to measure  
38 physical function in patients with various degrees of knee OA. It is scored as the  
39 KOOS from 0 (best) to 4 (worst), with a summary raw score ranging from 0 to 28.  
40 The score is converted to a true interval score that ranges from 0 (best) to 100  
41 (worst). The ICOAP is an 11 item questionnaire whose items were informed from  
42 focus groups with patients with hip or knee OA. It has two subscales that measure  
43 the intermittent and constant pain with a standardized summary score ranging from  
44 0 (best) to 100 (worst).

45 Patients also completed the generic SF-12, a 12-item general health  
46 measure with 8 items that have Likert-type response categories and 4 items with  
47 dichotomous (yes/no) response categories. The SF-12 is scored as a Physical  
48 Component Summary (PCS) and Mental Component Summary (MCS) ranging  
49 from 0 (worst) to 100 (best).

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65  
66  
67  
68  
69  
70  
71  
72  
73  
74  
75  
76  
77  
78  
79  
80  
81  
82  
83  
84  
85  
86  
87  
88  
89  
90  
91  
92  
93  
94  
95  
96  
97  
98  
99  
100  
101  
102  
103  
104  
105  
106  
107  
108  
109  
110  
111  
112  
113  
114  
115  
116  
117  
118  
119  
120  
121  
122  
123  
124  
125  
126  
127  
128  
129  
130  
131  
132  
133  
134  
135  
136  
137  
138  
139  
140  
141  
142  
143  
144  
145  
146  
147  
148  
149  
150  
151  
152  
153  
154  
155  
156  
157  
158  
159  
160  
161  
162  
163  
164  
165  
166  
167  
168  
169  
170  
171  
172  
173  
174  
175  
176  
177  
178  
179  
180  
181  
182  
183  
184  
185  
186  
187  
188  
189  
190  
191  
192  
193  
194  
195  
196  
197  
198  
199  
200  
201  
202  
203  
204  
205  
206  
207  
208  
209  
210  
211  
212  
213  
214  
215  
216  
217  
218  
219  
220  
221  
222  
223  
224  
225  
226  
227  
228  
229  
230  
231  
232  
233  
234  
235  
236  
237  
238  
239  
240  
241  
242  
243  
244  
245  
246  
247  
248  
249  
250  
251  
252  
253  
254  
255  
256  
257  
258  
259  
260  
261  
262  
263  
264  
265  
266  
267  
268  
269  
270  
271  
272  
273  
274  
275  
276  
277  
278  
279  
280  
281  
282  
283  
284  
285  
286  
287  
288  
289  
290  
291  
292  
293  
294  
295  
296  
297  
298  
299  
300  
301  
302  
303  
304  
305  
306  
307  
308  
309  
310  
311  
312  
313  
314  
315  
316  
317  
318  
319  
320  
321  
322  
323  
324  
325  
326  
327  
328  
329  
330  
331  
332  
333  
334  
335  
336  
337  
338  
339  
340  
341  
342  
343  
344  
345  
346  
347  
348  
349  
350  
351  
352  
353  
354  
355  
356  
357  
358  
359  
360  
361  
362  
363  
364  
365  
366  
367  
368  
369  
370  
371  
372  
373  
374  
375  
376  
377  
378  
379  
380  
381  
382  
383  
384  
385  
386  
387  
388  
389  
390  
391  
392  
393  
394  
395  
396  
397  
398  
399  
400  
401  
402  
403  
404  
405  
406  
407  
408  
409  
410  
411  
412  
413  
414  
415  
416  
417  
418  
419  
420  
421  
422  
423  
424  
425  
426  
427  
428  
429  
430  
431  
432  
433  
434  
435  
436  
437  
438  
439  
440  
441  
442  
443  
444  
445  
446  
447  
448  
449  
450  
451  
452  
453  
454  
455  
456  
457  
458  
459  
460  
461  
462  
463  
464  
465  
466  
467  
468  
469  
470  
471  
472  
473  
474  
475  
476  
477  
478  
479  
480  
481  
482  
483  
484  
485  
486  
487  
488  
489  
490  
491  
492  
493  
494  
495  
496  
497  
498  
499  
500  
501  
502  
503  
504  
505  
506  
507  
508  
509  
510  
511  
512  
513  
514  
515  
516  
517  
518  
519  
520  
521  
522  
523  
524  
525  
526  
527  
528  
529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539  
540  
541  
542  
543  
544  
545  
546  
547  
548  
549  
550  
551  
552  
553  
554  
555  
556  
557  
558  
559  
560  
561  
562  
563  
564  
565  
566  
567  
568  
569  
570  
571  
572  
573  
574  
575  
576  
577  
578  
579  
580  
581  
582  
583  
584  
585  
586  
587  
588  
589  
590  
591  
592  
593  
594  
595  
596  
597  
598  
599  
600  
601  
602  
603  
604  
605  
606  
607  
608  
609  
610  
611  
612  
613  
614  
615  
616  
617  
618  
619  
620  
621  
622  
623  
624  
625  
626  
627  
628  
629  
630  
631  
632  
633  
634  
635  
636  
637  
638  
639  
640  
641  
642  
643  
644  
645  
646  
647  
648  
649  
650  
651  
652  
653  
654  
655  
656  
657  
658  
659  
660  
661  
662  
663  
664  
665  
666  
667  
668  
669  
670  
671  
672  
673  
674  
675  
676  
677  
678  
679  
680  
681  
682  
683  
684  
685  
686  
687  
688  
689  
690  
691  
692  
693  
694  
695  
696  
697  
698  
699  
700  
701  
702  
703  
704  
705  
706  
707  
708  
709  
710  
711  
712  
713  
714  
715  
716  
717  
718  
719  
720  
721  
722  
723  
724  
725  
726  
727  
728  
729  
730  
731  
732  
733  
734  
735  
736  
737  
738  
739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
750  
751  
752  
753  
754  
755  
756  
757  
758  
759  
760  
761  
762  
763  
764  
765  
766  
767  
768  
769  
770  
771  
772  
773  
774  
775  
776  
777  
778  
779  
780  
781  
782  
783  
784  
785  
786  
787  
788  
789  
790  
791  
792  
793  
794  
795  
796  
797  
798  
799  
800  
801  
802  
803  
804  
805  
806  
807  
808  
809  
810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863  
864  
865  
866  
867  
868  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917  
918  
919  
920  
921  
922  
923  
924  
925  
926  
927  
928  
929  
930  
931  
932  
933  
934  
935  
936  
937  
938  
939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
970  
971  
972  
973  
974  
975  
976  
977  
978  
979  
980  
981  
982  
983  
984  
985  
986  
987  
988  
989  
990  
991  
992  
993  
994  
995  
996  
997  
998  
999  
1000  
1001  
1002  
1003  
1004  
1005  
1006  
1007  
1008  
1009  
1010  
1011  
1012  
1013  
1014  
1015  
1016  
1017  
1018  
1019  
1020  
1021  
1022  
1023  
1024  
1025  
1026  
1027  
1028  
1029  
1030  
1031  
1032  
1033  
1034  
1035  
1036  
1037  
1038  
1039  
1040  
1041  
1042  
1043  
1044  
1045  
1046  
1047  
1048  
1049  
1050  
1051  
1052  
1053  
1054  
1055  
1056  
1057  
1058  
1059  
1060  
1061  
1062  
1063  
1064  
1065  
1066  
1067  
1068  
1069  
1070  
1071  
1072  
1073  
1074  
1075  
1076  
1077  
1078  
1079  
1080  
1081  
1082  
1083  
1084  
1085  
1086  
1087  
1088  
1089  
1090  
1091  
1092  
1093  
1094  
1095  
1096  
1097  
1098  
1099  
1100  
1101  
1102  
1103  
1104  
1105  
1106  
1107  
1108  
1109  
1110  
1111  
1112  
1113  
1114  
1115  
1116  
1117  
1118  
1119  
1120  
1121  
1122  
1123  
1124  
1125  
1126  
1127  
1128  
1129  
1130  
1131  
1132  
1133  
1134  
1135  
1136  
1137  
1138  
1139  
1140  
1141  
1142  
1143  
1144  
1145  
1146  
1147  
1148  
1149  
1150  
1151  
1152  
1153  
1154  
1155  
1156  
1157  
1158  
1159  
1160  
1161  
1162  
1163  
1164  
1165  
1166  
1167  
1168  
1169  
1170  
1171  
1172  
1173  
1174  
1175  
1176  
1177  
1178  
1179  
1180  
1181  
1182  
1183  
1184  
1185  
1186  
1187  
1188  
1189  
1190  
1191  
1192  
1193  
1194  
1195  
1196  
1197  
1198  
1199  
1200  
1201  
1202  
1203  
1204  
1205  
1206  
1207  
1208  
1209  
1210  
1211  
1212  
1213  
1214  
1215  
1216  
1217  
1218  
1219  
1220  
1221  
1222  
1223  
1224  
1225  
1226  
1227  
1228  
1229  
1230  
1231  
1232  
1233  
1234  
1235  
1236  
1237  
1238  
1239  
1240  
1241  
1242  
1243  
1244  
1245  
1246  
1247  
1248  
1249  
1250  
1251  
1252  
1253  
1254  
1255  
1256  
1257  
1258  
1259  
1260  
1261  
1262  
1263  
1264  
1265  
1266  
1267  
1268  
1269  
1270  
1271  
1272  
1273  
1274  
1275  
1276  
1277  
1278  
1279  
1280  
1281  
1282  
1283  
1284  
1285  
1286  
1287  
1288  
1289  
1290  
1291  
1292  
1293  
1294  
1295  
1296  
1297  
1298  
1299  
1300  
1301  
1302  
1303  
1304  
1305  
1306  
1307  
1308  
1309  
1310  
1311  
1312  
1313  
1314  
1315  
1316  
1317  
1318  
1319  
1320  
1321  
1322  
1323  
1324  
1325  
1326  
1327  
1328  
1329  
1330  
1331  
1332  
1333  
1334  
1335  
1336  
1337  
1338  
1339  
1340  
1341  
1342  
1343  
1344  
1345  
1346  
1347  
1348  
1349  
1350  
1351  
1352  
1353  
1354  
1355  
1356  
1357  
1358  
1359  
1360  
1361  
1362  
1363  
1364  
1365  
1366  
1367  
1368  
1369  
1370  
1371  
1372  
1373  
1374  
1375  
1376  
1377  
1378  
1379  
1380  
1381  
1382  
1383  
1384  
1385  
1386  
1387  
1388  
1389  
1390  
1391  
1392  
1393  
1394  
1395  
1396  
1397  
1398  
1399  
1400  
1401  
1402  
1403  
1404  
1405  
1406  
1407  
1408  
1409  
1410  
1411  
1412  
1413  
1414  
1415  
1416  
1417  
1418  
1419  
1420  
1421  
1422  
1423  
1424  
1425  
1426  
1427  
1428  
1429  
1430  
1431  
1432  
1433  
1434  
1435  
1436  
1437  
1438  
1439  
1440  
1441  
1442  
1443  
1444  
1445  
1446  
1447  
1448  
1449  
1450  
1451  
1452  
1453  
1454  
1455  
1456  
1457  
1458  
1459  
1460  
1461  
1462  
1463  
1464  
1465  
1466  
1467  
1468  
1469  
1470  
1471  
1472  
1473  
1474  
1475  
1476  
1477  
1478  
1479  
1480  
1481  
1482  
1483  
1484  
1485  
1486  
1487  
1488  
1489  
1490  
1491  
1492  
1493  
1494  
1495  
1496  
1497  
1498  
1499  
1500  
1501  
1502  
1503  
1504  
1505  
1506  
1507  
1508  
1509  
1510  
1511  
1512  
1513  
1514  
1515  
1516  
1517  
1518  
1519  
1520  
1521  
1522  
1523  
1524  
1525  
1526  
1527  
1528  
1529  
1530  
1531  
1532  
1533  
1534  
1535  
1536  
1537  
1538  
1539  
1540  
1541  
1542  
1543  
1544  
1545  
1546  
1547  
1548  
1549  
1550  
1551  
1552  
1553  
1554  
1555  
1556  
1557  
1558  
1559  
1560  
1561  
1562  
1563  
1564  
1565  
1566  
1567  
1568  
1569  
1570  
1571  
1572  
1573  
1574  
1575  
1576  
1577  
1578  
1579  
1580  
1581  
1582  
1583  
1584  
1585  
1586  
1587  
1588  
1589  
1590  
1591  
1592  
1593  
1594  
1595  
1596  
1597  
1598  
1599  
1600  
1601  
1602  
1603  
1604  
1605  
1606  
1607  
1608  
1609  
1610  
1611  
1612  
1613  
1614  
1615  
1616  
1617  
1618  
1619  
1620  
1621  
1622  
1623  
1624  
1625  
1626  
1627  
1628  
1629  
1630  
1631  
1632  
1633  
1634  
1635  
1636  
1637  
1638  
1639  
1640  
1641  
1642  
1643  
1644  
1645  
1646  
1647  
1648  
1649  
1650  
1651  
1652  
1653  
1654  
1655  
1656  
1657  
1658  
1659  
1660  
1661  
1662  
1663  
1664  
1665  
1666  
1667  
1668  
1669  
1670  
1671  
1672  
1673  
1674  
1675  
1676  
1677  
1678  
1679  
1680  
1681  
1682  
1683  
1684  
1685  
1686  
1687  
1688  
1689  
1690  
1691  
1692  
1693  
1694  
1695  
1696  
1697  
1698  
1699  
1700  
1701  
1702  
1703  
1704  
1705  
1706  
1707  
1708  
1709  
1710  
1711  
1712  
1713  
1714  
1715  
1716  
1717  
1718  
1719  
1720  
1721  
1722  
1723  
1724  
1725  
1726  
1727  
1728  
1729  
1730  
1731  
1732  
1733  
1734  
1735  
1736  
1737  
1738  
1739  
1740  
1741  
1742  
1743  
1744  
1745  
1746  
1747  
1748  
1749  
1750  
1751  
1752  
1753  
1754  
1755  
1756  
1757  
1758  
1759  
1760  
1761  
1762  
1763  
1764  
1765  
1766  
1767  
1768  
1769  
1770  
1771  
1772  
1773  
1774  
1775  
1776  
1777  
1778  
1779  
1780  
1781  
1782  
1783  
1784  
1785  
1786  
1787  
1788  
1789  
1790  
1791  
1792  
1793  
1794  
1795  
1796  
1797  
1798  
1799  
1800  
1801  
1802  
1803  
1804  
1805  
1806  
1807  
1808  
1809  
1810  
1811  
1812  
1813  
1814  
1815  
1816  
1817  
1818  
1819  
1820  
1821  
1822  
1823  
1824  
1825  
1826  
1827  
1828  
1829  
1830  
1831  
1832  
1833  
1834  
1835  
1836  
1837  
1838  
1839  
1840  
1841  
1842  
1843  
1844  
1845  
1846  
1847  
1848  
1849  
1850  
1851  
1852  
1853  
1854  
1855  
1856  
1857  
1858  
1859  
1860  
1861  
1862  
1863  
1864  
1865  
1866  
1867  
1868  
1869  
1870  
1871  
1872  
1873  
1874  
1875  
1876  
1877  
1878  
1879  
1880  
1881  
1882  
1883  
1884  
1885  
1886  
1887  
1888  
1889  
1890  
1891  
1892  
1893  
1894  
1895  
1896  
1897  
1898  
1899  
1900  
1901  
1902  
1903  
1904  
1905  
1906  
1907  
1908  
1909  
1910  
1911  
1912  
1913  
1914  
1915  
1916  
1917  
1918  
1919  
1920  
1921  
1922  
1923  
1924  
1925  
1926  
1927  
1928  
1929  
1930  
1931  
1932  
1933  
1934  
1935  
1936  
1937  
1938  
1939  
1940  
1941  
1942  
1943  
1944  
1945  
1946  
1947  
1948  
1949  
1950  
1951  
1952  
1953  
1954  
1955  
1956  
1957  
1958  
1959  
1960  
1961  
1962  
1963  
1964  
1965  
1966  
1967  
1968  
1969  
1970  
1971  
1972  
1973  
1974  
1975  
1976  
1977  
1978  
1979  
1980  
1981  
1982  
1983  
1984  
1985  
1986  
1987  
1988  
1989  
1990  
1991  
1992  
1993  
1994  
1995  
1996  
1997  
1998  
1999  
2000  
2001  
2002  
2003  
2004  
2005  
2006  
2007  
2008  
2009  
2010  
2011  
2012  
2013  
2014  
2015  
2016  
2017  
2018  
2019  
2020  
2021  
2022  
2023  
2024  
2025  
2026  
2027  
2028  
2029  
2030  
2031  
2032  
2033  
2034  
2035  
2036  
2037  
2038  
2039  
2040  
2041  
2042  
2043  
2044  
2045  
2046  
2047  
2048  
2049  
2050  
2051  
2052  
2053  
2054  
2055  
2056  
2057  
2058  
2059  
2060  
2061  
2062  
2063  
2064  
2065  
2066  
2067  
2068  
2069  
2070  
2071  
2072  
2073  
2074  
2075  
2076  
2077  
2078  
2079  
2080  
2081  
2082  
2083  
2084  
2085  
2086  
2087  
2088  
2089  
2090  
2091  
2092  
2093  
2094  
2095  
2096  
2097  
2098  
2099  
2100  
2101  
2102  
2103  
2104  
2105  
2106  
2107  
2108  
2109  
2110  
2111  
2112  
2113  
2114  
2115  
2116  
2117  
2118  
2119  
2120  
2121  
2122  
2123  
2124  
2125  
2126  
2127  
2128  
2129  
2130  
2131  
2132  
2133  
2134  
2135  
2136  
2137  
2138  
2139  
2140  
2141  
2142  
2143  
2144  
2145  
2146  
2147  
2148  
2149  
2150  
2151  
2152  
2153  
2154  
2155  
2156  
2157  
2158  
2159  
2160  
2161  
2162  
2163  
2164  
2165  
2166  
2167  
2168  
2169  
2170  
2171  
2172  
2173  
2174  
2175  
2176  
2177  
2178  
2179  
2180  
2181  
2182  
2183  
2184  
2185  
2186  
2187  
2188  
2189  
2190  
2191  
2192  
2193  
2194  
2195  
2196  
2197  
2198  
2199  
2200  
2201  
2202  
2203  
2204  
2205  
2206  
2207  
2208  
2209  
2210  
2211  
221

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 than 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.<sup>(20)</sup> 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  $\chi^2$  ( $p > 0.05$ ), standardised root mean square residual (SRMR)  $> 0.08$ , comparative fit index (CFI)  $> 0.95$ , root mean square error of approximation (RMSEA):  $< 0.05$  close fit,  $< 0.08$  good fit,  $< 0.1$  satisfactory fit; RMSEA p test of close fit  $> 0.05$ .<sup>(23)</sup> Additionally, we used the Chi-square ( $\chi^2$ ) difference test and Parsimonious Normed Fit Index (PNFI) to compare the fit between the two models of the OKS and the ICOAP. <sup>(24)</sup> We calculated the  $\chi^2$  difference tests by looking at the difference of  $\chi^2$  of two models along with the difference in their degrees of freedom. ~~We checked the  $\chi^2$  difference, with its degrees of freedom in the  $\chi^2$  distribution table. If this value is statistically significant, then the model with more degrees of freedom is favoured.~~

## Responsiveness

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

159 We used a one sample t-test (2 tailed) to assess if the changes at 3 months  
160 for the OKS, its subscales (OKS-PCS and OKS-FCS), KOOS-PS and the ICOAP  
161 were significantly different from 0. We constructed a Cumulative Distribution  
162 Function (CDF) plot for the; (i) OKS, (ii) OKS-PCS and ICOAP, and (iii) OKS-FCS  
163 and KOOS-PS to examine the proportion of individual patients who experienced  
164 deterioration and improvement beyond the measurement error of the instrument at  
165 the individual level and to compare the proportion of change in pain and function  
166 detected by the different measures.

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

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

172 (ii) the change scores in the OKS-PCS should correlate more strongly with  
173 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.<sup>(20)</sup> 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



199 (MIC). The difference in the change score between the groups of patients who  
200 responded with “My knee has stayed the same” and “My knee has got better on  
201 the global item of change was taken as the minimal important difference (MID).  
202 Finally, the minimum change in the instrument that represents real change (beyond  
203 measurement error) was calculated using the Minimum Detectable Change  
204 ( $MDC_{90}$ ), which was obtained by multiplying the SEM with the z-value at the 90%  
205 level and the square root of two (to account for two measurement occasions). (26,  
206 27))

207

208



## 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  $\pm 4.4$  OKS points. SEM for the OKS-FCS was 6.2 with  $\pm 10.2$  90% CI for individual score and the SEM for the OKS-PCS was 6.9 with  $\pm 11.3$  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  $\pm 15.9$  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  $\pm 11.1$  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  $\rho$  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

283 poorly related to the SF12-MCS. The OKS-PCS correlated more with ICOAP than  
284 with KOOS-PS and the OKS-FCS correlated more with the KOOS-PS than with  
285 ICOAP. This evidence supports convergent and divergent validity of the OKS.

286

287 Table 2: Baseline Spearman's correlations between the scores. All correlations  
288 were significant at the 0.01 level (2-tailed). The number of cases with complete  
289 information that allowed the calculation of the correlation coefficients is in brackets  
290 for each correlation.

291

	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)	/	/

292

293

294 **Structural validity.** 122 pre-operative OKSs, 125 pre-operative ICOAP and  
295 113 pre-operative KOOS-PS were available for the CFA. Fit indices of one and two  
296 factor models for the OKS are presented in Table 3. Neither of the one and two  
297 factor models was rejected. Fit indices favoured the 2 factor model and the  
298 reduction in  $\chi^2$  in the two factor model was significant ( $\chi^2_{diff} > 7.879$ , with  $df=1$ , at  
299 the  $\alpha=0.005$  level).

300

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

Factors	$\chi^2$ (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

302

303

304

305

306

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.

307

308

309

CFA revealed that a one-factor KOOS-PS model was rejected by the  $\chi^2$  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  $\chi^2$  test and both models had RMSEA values far above the lowest threshold of an acceptable fit. However, SRMR and CFI were acceptable for both scores. There was no significant reduction (at the 0.05 level) in  $\chi^2$  for the 2 factor model of the ICOAP ( $\chi^2_{diff} < 3.84$ , with  $df=1$ ).

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

	$\chi^2$ (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;  $\chi^2$ =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<sub>90</sub> 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.

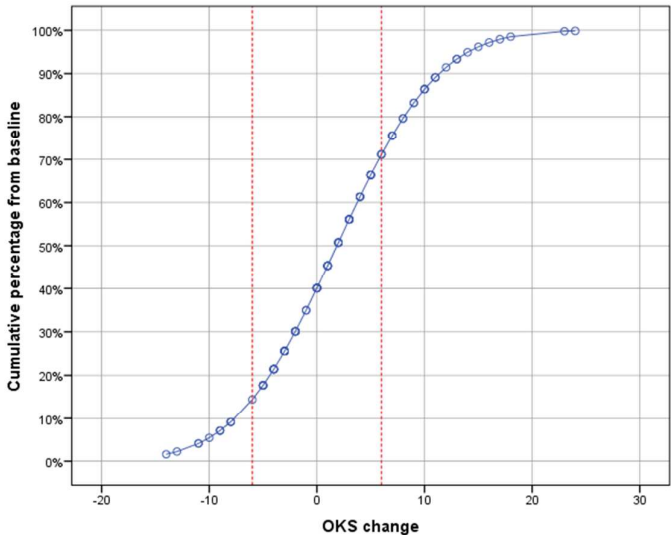


Figure 1. Cumulative percentage of patients experiencing the change on the OKS from baseline less or equal to the value on the x-axis. Red line marks the minimum detectable change beyond the measurement error of the score (MDC<sub>90</sub> of 6 points).

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 <sup>a</sup>	104	37.19 (25)	31.53 (25)	-5.66 (19)	<0.01	0.23
KOOS-PS <sup>a</sup>	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;<sup>a</sup> 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 <sub>90</sub>	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<sub>90</sub>=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  $\approx 7$  for the OKS,  $\approx 17$  for the OKS-PCS, and  $\approx 11$  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  $\approx 6$  for the OKS,  $\approx 14$  for the OKS-PCS, and  $\approx 10$  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  $MDC_{90}$  ( $\approx 6$  points for the OKS,  $\approx 16$  points for the OKS-PCS, and  $\approx 15$  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~~other 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<sub>90</sub> 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.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

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](mailto:healthoutcomes@isis.ox.ac.uk)

## COMPETING INTEREST DECLARATION

All authors have completed the Unified Competing Interest form at [www.icmje.org/coi\\_disclosure.pdf](http://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 Physicians; 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-arthritis. *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.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

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.