Osteoarthritis and Cartilage 27 (2019) 754–761

Osteoarthritis and Cartilage

A 12-item short form of the Hip disability and Osteoarthritis Outcome Score (HOOS-12): tests of reliability, validity and responsiveness

B. Gandek † ‡ *, E.M. Roos §, P.D. Franklin †, J.E. Ware Jr. † ‡

† University of Massachusetts Medical School, Worcester, MA, USA
‡ John Ware Research Group, Watertown, MA, USA
§ Department of Sports Science and Clinical Biomechanics, University of Southern Denmark, Odense, Denmark

ARTICLE INFO

Article history:
Received 17 April 2018
Accepted 18 September 2018

Keywords:
HOOS
Hip
Osteoarthritis
Patient-reported outcome measures
Psychometrics

SUMMARY

Objective: To evaluate reliability, validity and responsiveness of HOOS-12, a 12-item short form of the 40-item Hip disability and Osteoarthritis Outcome Score (HOOS). HOOS-12 provides Pain, Function and Quality of Life (QOL) scale scores and a summary hip impact score.

Design: Data from 1,273 FORCE-TJR hip osteoarthritis (OA) patients who completed HOOS before and six and 12 months after total hip replacement (THR) were analyzed. HOOS-12 includes a pain frequency item and three items measuring pain during increasingly difficult (sitting/lying, walking, stairs) activities; function items about standing, rising from sitting, getting in/out of a car, and walking on an uneven surface; and the 4-item HOOS QOL scale. Percent computable scale scores, floor and ceiling effects, internal consistency reliability, validity (scale correlations, tests of known groups validity using one-way analysis of variance (ANOVA)), and responsiveness (effect sizes (ES), standardized response means (SRM)) were compared for HOOS-12, full-length HOOS, HOOS-PS and HOOS, JR.

Results: Internal consistency reliability was above 0.70 for all HOOS-12 scales and above 0.90 for the HOOS-12 Summary score. Validity and responsiveness of HOOS-12 Pain, Function and QOL scales were satisfactory and reached similar conclusions as comparable full-length HOOS scales. The HOOS-12 Summary score was highly responsive in discriminating between groups who differed in global ratings of post-THR change in physical capabilities and had high ES and SRM standardized response means.

Conclusions: HOOS-12 was a reliable and valid alternative to HOOS in THR patients with moderate to severe OA and provided three domain-specific and summary hip impact scores with substantially reduced respondent burden.

Introduction

Patient-reported outcome measures (PROMs) play an important role in understanding the patient's perspective on the impact of hip osteoarthritis (OA) and other hip disorders and their treatment[1]. One of the most widely-used hip-specific PROMs is the 40-item Hip disability and Osteoarthritis Outcome Score (HOOS)[2], which was developed to measure the impact of hip disability in patients with hip osteoarthritis[3] and non-arthritis hip disorders[4]. HOOS has the advantage of providing domain-specific measures of Pain, Symptoms, Function (Activities of Daily Living (ADL) and Sport/Recreation), and Quality of Llife (QOL), but its respondent burden has been viewed as too great for routine use in registries or clinical care. Efforts to construct brief versions of the HOOS that included items from multiple domains resulted in measures that provided only one summary score and lost the specificity of changes in pain, function and QOL[5]. However, in clinical care, treatment can vary for persistent pain as opposed to functional limitations, making the global assessment less informative. For registries and clinical care, a brief but comprehensive hip-specific PROM that is reliable, valid and responsive and allows for construction of domain-specific scores as well as an overall hip impact score would be optimal.

HOOS-12 is a new 12-item short form that provides domain-specific scores for pain, function, and hip-specific QOL, while also representing content across domains sufficiently to enable construction of a summary measure of overall hip impact. It contains 12 items selected from the HOOS, including four Pain items, four

* Address correspondence and reprint requests to: B. Gandek, University of Massachusetts Medical School, Department of Orthopedics and Physical Rehabilitation, 55 Lake Avenue North, Worcester, MA, 01655, USA.
E-mail addresses: barbara.gandek@umassmed.edu (B. Gandek), eroos@health.sdu.dk (E.M. Roos), patricia.franklin@umassmed.edu (P.D. Franklin), john.ware@jwrginc.com (J.E. Ware).

https://doi.org/10.1016/j.joca.2018.09.017
1063-4584/

Function items, and four QOL items (Fig. 1). HOOS-12 items were selected based on item content; coverage of a wide measurement range and high item information in item response theory (IRT) models; computerized adaptive test (CAT) simulations to identify items that best matched patients' levels of pain and function; scale-level internal consistency reliability, validity and responsiveness; and qualitative feedback from translation developers, clinicians and patients. The HOOS-12 item selection process is described in detail in a separate paper, along with item selection for a companion measure, a 12-item short form of the Knee injury and Osteoarthritis Outcome Score (KOOS-12)[6]. This paper evaluates the reliability, validity and responsiveness of HOOS-12 and compares HOOS-12 psychometric properties to those of the original (full-length) HOOS and its derivative forms. A companion paper examines the psychometric properties of KOOS-12[7].

Methods

Study design and participants

Data came from the Function and Outcomes Research for Comparative Effectiveness in Total Joint Replacement (FORCE-TJR) research cohort, which includes more than 30,000 patients of 200 diverse surgeons throughout the U.S[8]. FORCE-TJR surveys were completed by patients pre-TJR and six and 12 months post-TJR, at their surgeon's office or at home, either as paper-pencil surveys or on the Internet. Data from a random sample of $n = 1{,}281$ hip OA patients who had a total hip replacement (THR) between 2011 and 2014 (Item Selection sample) was used to select items for HOOS-12[6]. An independent random sample of $n = 1{,}273$ hip OA patients who had a THR between 2011 and 2014 (Cross-Validation sample)

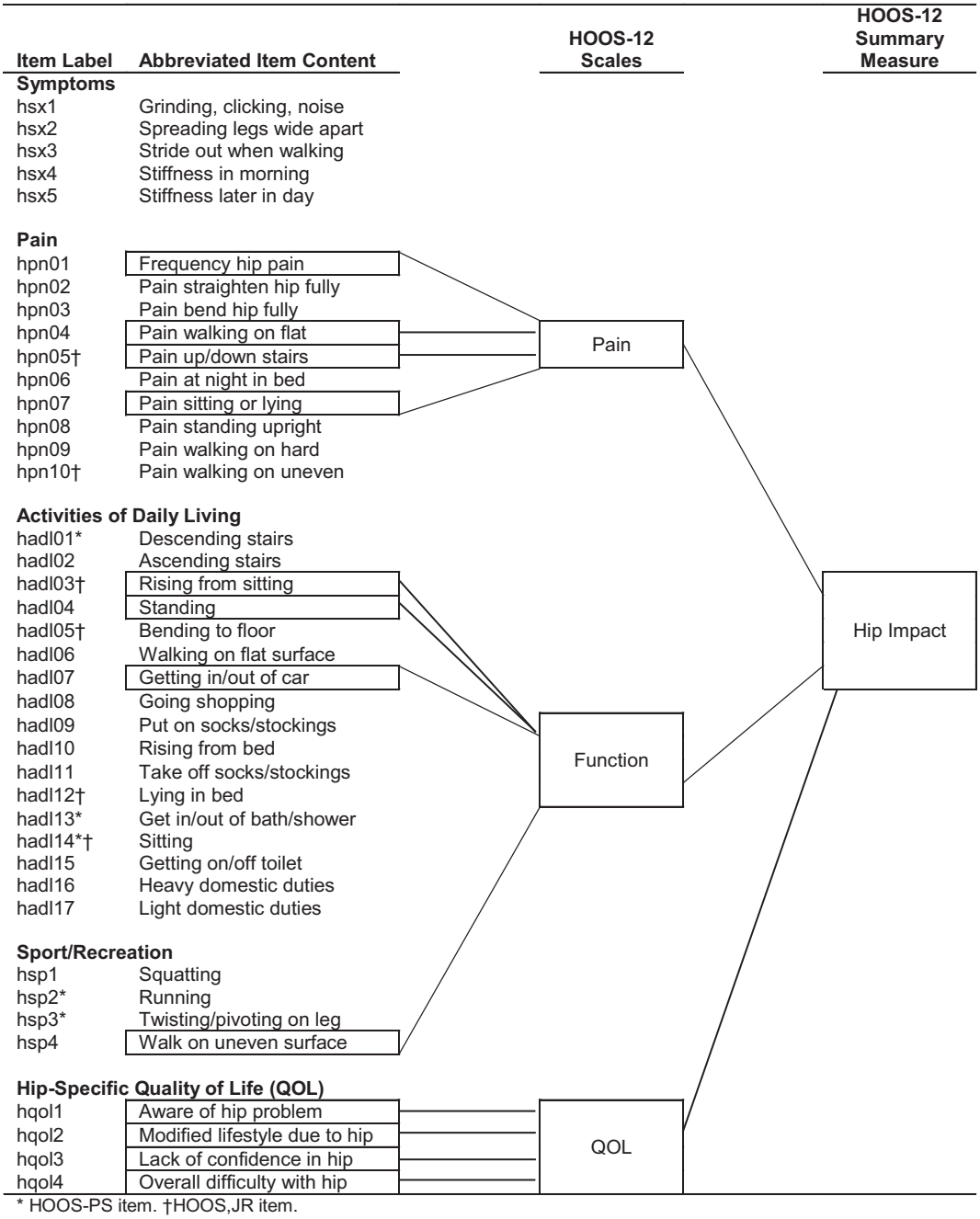

Fig. 1. HOOS-12 measurement model.

was analyzed in this paper, to independently evaluate the psychometric properties of HOOS-12. FORCE-TJR and this study were approved by the University of Massachusetts Medical School Institutional Review Board.

The Cross-Validation sample had a mean age of 64.8 (SD = 9.7, range 30–95); 60.3% were female. 92.2% were White non-Hispanic, 4.9% Black non-Hispanic, 1.4% Hispanic, and 1.5% of other race and ethnicity. Overall, 18.2% were a high school graduate or had less education, 26.2% had some education post-high school, 23.6% were college graduates, 28.0% had some post-college education, 1.9% had other education and 2.0% were missing education. Sociodemographic characteristics were similar to those of the Item Selection sample.

Measures

HOOS-12 is scored as three domain-specific scales measuring Pain (number of items $k = 4$), Function ($k = 4$) and hip-specific Quality of Life ($k = 4$) (Fig. 1), using the method of summated ratings[9], in which item responses in a scale are simply summed. Scale scores calculated using this method and using more complex IRT-based scoring correlated 0.98 and had similar known groups validity (see Methods). Therefore, the summated ratings method was adopted for scoring HOOS-12. A person-specific value is imputed for missing item data within a scale, if $\geq$ 50% of items in the scale are answered. To facilitate interpretation, scores are transformed so 0 is the worst and 100 is the best measured score. This is in line with the original HOOS scales, which also are scored using the method of summated ratings and a similar imputation method for missing item data, to produce Pain ($k = 10$), Symptoms ($k = 5$), Function in ADL ($k = 17$), Function in Sport/Recreation ($k = 4$), and QOL ($k = 4$) scale scores with possible ranges from 0 to 100[10].

HOOS-12 also produces a summary hip impact score (HOOS-12 Summary score), which is calculated as the average of the HOOS-12 Pain, Function and QOL scale scores. Advantages of a single summary score include addressing the issue of analyzing multiple outcome measures, relevant for example in randomized controlled trials. A summary score that can be disaggregated into its components (Pain, Function, QOL) provides the best of both worlds. It reduces the need for multiple comparisons, while enabling the interpretation of specific outcomes as needed, for example in patient-clinician communication and in systematic reviews and meta-analysis. To evaluate methods for constructing a summary measure from the three HOOS-12 scales, a principal components analysis of their inter-correlations was conducted on the Item Selection sample, to see if loadings for each scale were equivalent or if scale scores needed to be standardized and weighted prior to calculating a summary score. Loadings were equivalent and substantial (0.944–0.957) across the HOOS-12 scales, indicating that each scale contributed equally to measuring the underlying construct of hip impact. Correlations between a summary score calculated as the simple average of the three scale scores vs a summary score calculated using weighted scale scores were substantial ($r = 0.998$–0.999) at pre-THR and six and 12 months post-THR. Therefore, the HOOS-12 Summary score was calculated using the simpler method of averaging the three HOOS-12 scale scores. The HOOS-12 Summary score also ranges from 0 to 100, where 0 is the worst possible and 100 is the best possible score. A summary score is not calculated if any of the three scale scores is missing.

Two other HOOS derivative measures constructed by others were scored from the original HOOS for comparative purposes (Fig. 1). HOOS-PS is a 5-item measure of physical function with three ADL and two Sport/Recreation items and was developed in an OARSI/OMERACT initiative[11]. HOOS, JR contains two Pain and four ADL items and is scored to provide an overall measure of hip health[5]; it does not provide separate Pain and ADL scores. HOOS, JR has been accepted by the U.S. Center for Medicare & Medicaid Services for the PROM component of its Comprehensive Care for Joint Replacement model[12]. HOOS-PS and HOOS, JR were scored following their developers' methods, which require that all items in a scale must be answered to calculate a score[5,13]; the worst and best possible scores are 0 and 100, respectively.

To test construct validity, HOOS-12 scales were evaluated in relation to the SF-36 Health Survey, a general measure of physical and mental functioning and well-being[14]. The SF-36 is scored as eight scales, including measures of physical function ($k = 10$), bodily pain ($k = 2$) and mental health ($k = 5$), and summary Physical (PCS) and Mental (MCS) Component Scores[15]. SF-36 (Version 2.0) scales were scored so that 50 was the mean and 10 was the standard deviation in the U.S. general population[16].

Statistical analysis

Analyses focused on comparisons of the psychometric properties of HOOS-12 to those of the full-length HOOS scales from which the HOOS-12 was derived. In addition, psychometric properties of HOOS-PS and HOOS, JR were evaluated to provide information about their performance relative to HOOS-12.

The percent of respondents for whom scale scores could be calculated pre-THR and six and 12 months post-THR was examined. Internal consistency reliability of all scales was evaluated using Cronbach's coefficient alpha, which is based on the number of items in a scale and the mean inter-item correlation[17]. An alpha of 0.70 or higher is recommended for group-level comparisons, while a minimum reliability of 0.90–0.95 is considered acceptable when a measure is used with individual patients[18,19]. Reliability of the HOOS-12 Summary score was calculated based on Cronbach's alpha for the three HOOS-12 scales, their component weights and their covariances, following methods similar to those used to calculate reliability for SF-36 PCS and MCS[15]. In addition, because responsiveness of a measure is constrained if a high percentage of patients have the lowest or highest possible scores, floor and ceiling effects were evaluated and considered present if more than 15% of respondents had the lowest or highest possible scores, respectively[20]. Because many patients could be considered as "disease-free" after a successful THR, some ceiling effects were expected post-THR, as in previous studies[21].

Validity was evaluated using several approaches. Construct validity was evaluated by estimating Pearson product–moment correlations between HOOS-12 scales and full-length HOOS and general SF-36 measures, to determine if HOOS-12 scales correlated higher with measures of the same construct (convergent validity) than with measures of different constructs (discriminant validity). Correlations <0.30, 0.30–0.69, and ≥0.70 (equivalent to shared variances of <10%, 10% to <50%, and ≥50%) were considered as low, moderate and high, respectively. Because the HOOS-12 Pain and Function scales are subsets of the full-length HOOS Pain and full-length HOOS ADL and Sport/Recreation scales, high correlations between HOOS-12 and corresponding HOOS scales were expected. Moderate to high correlations were expected between HOOS-12 scales and SF-36 physical health measures and low correlations between HOOS-12 scales and SF-36 mental health measures.

Tests of known groups validity were used to compare the responsiveness of HOOS-12 and other HOOS measures, in terms of their ability to differentiate between groups who varied in ratings of change at 6 months post-THR, using one-way analysis of variance (ANOVA). Patients responded to a global item about their change in capability to do everyday physical activities at 6 months compared to pre-THR (lot more, more, same, less/lot less capable). In each ANOVA, the change score for a scale was the dependent variable

and the self-evaluated change in capability was the independent variable. Each ANOVA F-statistic indicated how strongly a scale discriminated between groups and thus provided information about that scale's statistical efficiency. To facilitate comparisons across scales, results were summarized using relative validity (RV) statistics (ratio of the F-statistic for each scale divided by the F-statistic for the full-length HOOS scale in domain-specific comparisons and divided by the HOOS-12 Summary score in comparisons across all scales), as in previous analyses[22]. In each set of comparisons, the denominator scale had RV = 1.0; 95% confidence intervals were derived using empirical bootstrap[23]. Within the pain and function domains, the null hypothesis of equal validity between HOOS-12 and original HOOS scales was tested. In addition, because the HOOS-12 Summary score contains multiple indicators of joint impact, it was hypothesized to be the most valid of all measures. HOOS, JR also provides a summary score, but its Pain items did not cover as wide a measurement range as HOOS-12 plus HOOS, JR does not include QOL items.

As a measurement property, responsiveness is best interpreted in relation to another measure captured simultaneously[24] using an approach such as the anchor-based method described above [25,26]. In addition, the responsiveness of all hip-specific scales and summary measures also was compared using the standardized effect size (ES; observed change score (post minus pre-THR) divided by the standard deviation of the pre-THR score)[27] and the standardized response mean (SRM; observed change score divided by the standard deviation of the change score)[28].

All analyses were performed using Stata Statistical Software: Release 11 (StataCorp LP, College Station, TX). Two-tailed tests were used to determine significant ($P < 0.05$) differences.

Results

Scale scores could be calculated for 98.5–99.7%, 98.7–99.6% and 97.4–98.8% of patients for the HOOS-12 Pain, Function and QOL scales, respectively (Table I). The percentage computable for HOOS-12 Pain and Function scales were similar to those for the full-length HOOS Pain and ADL scales and slightly higher than the percentage computable for the HOOS Sport/Recreation scale. The HOOS-12 Summary score could be calculated for 96.5–97.6% of patients.

Scores for HOOS-PS could only be calculated for 87.3–91.8% of patients, because the HOOS-PS scoring algorithm requires that all five items need to be answered to compute a score; in addition, missing data rates were high for the HOOS-PS running item in particular. HOOS, JR scores could be computed for only 94.1–95.5% of patients, due to the requirement that all items must be answered to calculate a score.

Internal consistency reliability of all three HOOS-12 scales was above 0.70 at all time points (Table I). While Cronbach's alpha for HOOS-12 Pain and Function was lower than alpha for corresponding full-length HOOS scales, the average inter-item correlations did not differ greatly between corresponding HOOS-12 and HOOS scales, indicating that reliability differences were primarily due to differences in scale length[29]. Reliability of HOOS, JR was 0.86 at all three time points, while reliability of the HOOS-12 Summary score exceeded 0.90 across time points.

Floor effects (percent with the lowest (worst) possible score) for all measures were very low (<2%) pre- and post-THR, with the exception of the HOOS Sport/Recreation and HOOS QOL scales pre-THR (Table II). Ceiling effects (percent with the highest (best) possible score) were negligible pre-THR. Post-THR, there were notable ceiling effects for almost all measures. For Pain, post-THR ceiling effects were 49–50% at six and 12 months for HOOS-12 Pain compared to 40–44% for the full-length HOOS Pain scale. For Function, post-THR ceiling effects were 33–41% for HOOS-12 Function compared to 16–27% for HOOS ADL, 15–21% for HOOS Sport/Recreation, and 16–23% for HOOS-PS. The HOOS-12 Summary score had the lowest percentage at the ceiling post-THR, at 12–14%. In comparison, the percentage at the ceiling post-THR was 27–36% for HOOS, JR, indicating that the HOOS-12 Summary score was better able to distinguish between patients at higher levels of hip health than the other summary measure.

Tests of construct validity supported convergent and discriminant validity of HOOS-12 at pre-THR and both post-THR time points and are presented for pre-THR data (Table III). The correlation of HOOS-12 Pain and HOOS Pain was very high ($r = 0.94$), indicating that all reliable variance in the full-length HOOS Pain scale was captured by the HOOS-12 Pain scale. The HOOS-12 Function scale also had high correlations with HOOS ADL ($r = 0.92$) and Sport/Recreation ($r = 0.75$). HOOS-12 Pain, Function and QOL scales had similar patterns of moderate correlations with SF-36 scales primarily measuring physical health (Bodily Pain, Physical Functioning, Physical Component Summary) and, in support of discriminant validity, relatively low correlations with SF-36 scales primarily measuring mental health (Mental Health, Mental Component Summary).

Tests of known groups validity indicated that all hip-specific measures were able to detect differences between groups differing in self-reported evaluation of change (lot more, more, same, or less) in capability to do everyday physical activities at 6 months post-THR (Table IV). Within the Pain domain, the HOOS-12 scale (RV = 1.29, 95% CI (1.09, 1.56)) was significantly ($P < 0.05$) more responsive to group differences than the full-length HOOS Pain scale (RV = 1.0). In comparisons across Function scales, the HOOS-12 Function scale had similar RV as HOOS ADL, HOOS Sport/Recreation and HOOS-PS, as hypothesized. The HOOS-12 Summary score was significantly ($P < 0.05$) more responsive than HOOS, JR, as

Table I
Percent computable scales, internal consistency reliability and average inter-item correlations for hip-specific measures

	k	% Computable			Cronbach's alpha			Inter-Item Correlation		
		Pre	6 m	12 m	Pre	6 m	12 m	Pre	6 m	12 m
HOOS-12 Pain	4	98.7	98.5	99.7	0.77	0.77	0.78	0.46	0.46	0.47
HOOS Pain	10	98.7	98.5	99.7	0.91	0.91	0.92	0.50	0.50	0.53
HOOS-12 Function	4	98.7	99.6	99.1	0.83	0.81	0.84	0.55	0.52	0.57
HOOS ADL	17	98.7	99.6	99.1	0.95	0.94	0.95	0.53	0.48	0.53
HOOS Sport/Recreation	4	98.2	96.6	98.3	0.83	0.87	0.87	0.55	0.63	0.63
HOOS-PS	5	91.8	87.3	87.8	0.80	0.77	0.77	0.44	0.40	0.40
HOOS/HOOS-12 QOL	4	98.4	97.4	98.8	0.80	0.79	0.82	0.50	0.48	0.53
HOOS-12 Summary	12	96.9	96.5	97.6	0.92	0.91	0.92	–	–	–
HOOS, JR	6	94.2	95.5	94.1	0.86	0.86	0.86	0.51	0.51	0.51

N = 1,273 (pre-THR), 909 (6 month post-THR), 757 (12 month post-THR).
k, Number of items; % Computable, Percent of respondents for whom scale score could be computed at pre-THR, 6 month and 12 months post-THR; ADL, Activities of Daily Living; QOL, Quality of Life.

Table II
Floor and ceiling effects for hip-specific measures

	k	% at Floor			% at Ceiling		
		Pre-THR	6 month	12 month	Pre-THR	6 month	12 month
HOOS-12 Pain	4	1.4	0.0	0.0	0.2	48.9	50.2
HOOS Pain	10	1.1	0.0	0.0	0.2	40.0	44.3
HOOS-12 Function	4	1.1	0.0	0.0	0.2	33.3	41.3
HOOS ADL	17	0.2	0.0	0.0	0.2	16.4	27.2
HOOS Sport/Recreation	4	14.8	0.8	0.2	0.0	14.8	20.6
HOOS-PS	5	0.7	0.0	0.0	0.0	15.6	22.6
HOOS/HOOS-12 QOL	4	11.6	0.0	0.0	0.0	17.4	20.1
HOOS-12 Summary	12	0.5	0.0	0.0	0.0	11.6	13.9
HOOS, JR	6	0.7	0.0	0.0	0.0	27.2	36.3

N = 438 patients with all scale scores at all three time points.
k, Number of items; % Floor, % with lowest possible score; % Ceiling, Percent with highest possible score; ADL, Activities of Daily Living; QOL, Quality of Life. All measures scored so 0 = worst possible and 100 = best possible score.

hypothesized. Comparison of mean change scores indicated that the HOOS-12 Summary score detected about a half SD greater improvement on average than HOOS, JR for those who rated themselves as most improved.

Effect sizes (ES) at six and 12 months post-THR were somewhat higher for the HOOS-12 Pain scale than the full-length HOOS Pain scale, while standardized response means (SRM) were similar (Table V). ES and SRM were similar for the HOOS-12 Function and HOOS ADL scales, while the HOOS Sport/Recreation scale had higher ES but lower SRM than other Function measures. HOOS-PS had the lowest ES of all measures. The ES for the QOL scale (2.77 and 3.04 at six and 12 months) were similar to those for the HOOS-12 Pain scale (2.81–3.07). ES for the HOOS-12 Summary score (2.90–3.16) was somewhat higher than the ES for HOOS, JR (2.34–2.56). The SRM also was higher for the HOOS-12 Summary (2.31–2.64) than HOOS, JR (2.01–2.21)).

Discussion

The objective of this study, which was achieved, was to evaluate the HOOS-12, a short form HOOS survey that had 70% lower respondent burden than the full-length HOOS, while allowing for the construction of domain-specific scales in addition to a comprehensive summary score. Construction of the HOOS-12 benefitted from use of modern psychometric methods to aid in item selection and from feedback by patients, clinicians, and researchers who developed translations of the HOOS and its companion measure, the KOOS. Results from this study indicated that a relatively short 12-item hip-specific survey can be constructed from HOOS items in a manner that: (1) substantially reduces respondent burden (12-item surveys can be completed by most patients in 2 min or less); (2) allows for scoring a profile of separate Pain, Function, and QOL scales with satisfactory reliability, which reach similar conclusions in tests of validity and responsiveness as full-length HOOS scales; and (3) achieves the advantages of a summary score with satisfactory validity and that is likely to improve responsiveness to change after THR.

The HOOS-12 Summary score provides an aggregate measure of hip impact across the Pain, Function and QOL domains. Summary measures have the advantage of providing results for one or two endpoints rather than multiple endpoints. Perhaps the best known PRO summary measures are the SF-36 Physical (PCS) and Mental (MCS) Component Summary scores[15], which reduce the number of statistical comparisons from eight scales to two summary measures when analyzing the SF-36. Similarly, the HOOS-12 Summary score reduces the number of endpoints from three scales to one aggregate measure, which can be used as the primary outcome and complemented by the three scale scores as secondary outcomes for more specific clinical interpretation. The HOOS-12 Summary score was the only summary measure with internal consistency reliability at or above 0.90, which is the minimum level often recommended when using a measure with individual patients. It also had the lowest (12–14%) ceiling effects and was the stronger summary measure in detecting differences between groups differing in self-

Table III
Descriptive Statistics, Reliability and Inter-scale correlations Among Hip-Specific and SF-36 Measures, pre-THR

	k	Mean	SD	Reliability	Pain		Function			QOL
					HOOS-12	HOOS	HOOS-12	HOOS ADL	HOOS Sport	HOOS QOL
Pain										
HOOS-12 Pain	4	40.0	16.7	0.77						
HOOS Pain	10	43.5	17.4	0.91	0.94					
Function										
HOOS-12 Function	4	42.5	19.1	0.83	0.76	0.82				
HOOS ADL	17	46.5	18.7	0.95	0.79	0.84	0.92			
HOOS Sport/Recreation	4	22.4	18.5	0.83	0.64	0.69	0.75	0.73		
HOOS/HOOS-12 QOL	4	25.4	17.7	0.80	0.61	0.64	0.65	0.66	0.67	
SF-36 Generic										
Bodily Pain	2	33.1	7.1	0.80	0.66	0.68	0.67	0.70	0.57	0.60
Physical Functioning	10	30.3	9.9	0.90	0.54	0.56	0.59	0.64	0.58	0.54
Mental Health	5	49.0	10.4	0.85	0.30	0.32	0.33	0.35	0.28	0.34
PCS	35	31.8	8.3	0.91	0.56	0.57	0.60	0.63	0.58	0.56
MCS	35	51.5	11.9	0.93	0.31	0.34	0.34	0.37	0.27	0.34

N = 1,212.
k, number of items. ADL, Activities of Daily Living; Sport, Sport/Recreation; QOL, Quality of Life; PCS, Physical Component Summary; MCS, Mental Component Summary. Reliability is internal consistency reliability (Cronbach's alpha), see text.
SE for all correlations = 0.029. All measures scored so 0 = worst possible and 100 = best possible score, except for SF-36 measures (US general population mean = 50, SD = 10; lower score = poorer health).

Table IV
Mean change scores (SD) and known-groups validity tests by self-evaluated change in physical activity at 6 months

	k	Mean (SD) Change Score by Change in Capability in Everyday Physical Activities*				F	RVWithinDomain (95% CI)	RV Across Domains (95% CI)
		Lot More ($n = 466$)	More ($n = 126$)	Same ($n = 46$)	Less ($n = 24$)			
HOOS-12 Pain	4	51.1 (17.7)	40.8 (23.5)	31.2 (24.9)	19.8 (29.8)	34.78	1.29 (1.09, 1.56)	0.69 (0.52, 0.86)
HOOS Pain	10	48.7 (17.3)	39.9 (21.3)	31.5 (23.1)	23.8 (24.9)	26.96	1.00	0.54 (0.40,0.66)
HOOS-12 Function	4	48.4 (19.2)	35.9 (20.4)	29.7 (20.8)	18.2 (23.0)	36.38	1.03 (0.84, 1.27)	0.72 (0.57, 0.90)
HOOS ADL	17	44.6 (17.7)	33.1 (19.3)	26.1 (20.4)	19.0 (19.5)	35.38	1.00	0.70 (0.54, 0.88)
HOOS Sport/Recreation	4	53.1 (23.7)	31.8 (25.6)	30.7 (25.7)	6.5 (31.6)	53.80	1.52 (1.07, 2.27)	1.07 (0.79, 1.42)
HOOS-PS	5	35.9 (16.6)	25.8 (17.9)	20.8 (16.4)	10.9 (16.4)	33.43	0.94 (0.74, 1.22)	0.66 (0.52, 0.86)
HOOS/HOOS-12 QOL	4	54.3 (21.7)	38.6 (23.8)	33.1 (21.9)	15.9 (25.3)	43.09	–	0.86 (0.69, 1.04)
HOOS-12 Summary	12	51.2 (16.9)	38.4 (20.0)	31.3 (20.2)	18.0 (22.4)	50.33	–	1.00
HOOS, JR	6	40.0 (16.3)	29.6 (17.4)	22.2 (17.3)	15.9 (16.4)	38.44	–	0.76 (0.60, 0.91)

k, Number of items; F, ANOVA F-statistic; RV, relative validity; CI, confidence interval; ADL, Activities of Daily Living; QOL, Quality of Life. All measures scored so 0 = worst possible and 100 = best possible score. All F-statistics $P < 0.001$.

* Item text (response options): Thinking about your everyday physical activities today (such as walking, climbing stairs, carrying groceries, or participating in sports); Compared to before your joint surgery, are you more or less capable now in your everyday physical activities because of your joint surgery? (A lot more capable now, somewhat more capable now, about the same, somewhat less capable now, a lot less capable now; fourth and fifth responses combined in ANOVA).

reported evaluation of change in capability to do everyday physical activities post-THR. However, because it combines scores from three scales, there are many different ways to achieve any particular HOOS-12 Summary score. Thus, summary scores should be interpreted in relation to HOOS-12 scale scores.

HOOS, JR also provides an overall measure of hip health. However, HOOS, JR does not allow for computation of domain scores and also had higher ceiling effects (27–36%) at six and 12 months post-THR than the HOOS-12 Summary score (12–14%). HOOS, JR also did not discriminate as well as the HOOS-12 Summary score between post-THR groups differing in change in capability to do everyday physical activities, and in particular did not detect as much change as HOOS-12 for groups who said they were a “lot more” capable post-THR. The superior performance of the HOOS-12 Summary score is likely related to selection of pain items for HOOS-12 that have a wider measurement range than those included in HOOS, JR[6], along with the inclusion of QOL items in HOOS-12, which provide a more comprehensive assessment of the functional and emotional impact of hip disorders. This is in line with recommendations from OARSI/OMERACT regarding outcome domains for inclusion in OA studies[30–32]. Finally, HOOS, JR also could not be scored for approximately 5% of patients, due to the requirement that all six items must be answered to score this measure.

The 4-item HOOS-12 Pain scale had better performance than the 10-item HOOS Pain scale in responding to differences in global change ratings post-THR, although the shorter scale had higher ceiling effects than the full–length scale post-THR. The 4-item scale is parsimonious, containing two items that span the least painful (sitting) and more painful (stairs) activities plus the item that was best at discriminating among respondents (walking on a flat surface) in IRT analyses[6], while also giving more proportional weight to pain frequency than the full-length HOOS scale. By strategic selection of items for inclusion, and by excluding items that were of limited usefulness in estimating the full-length HOOS Pain score in CAT simulations[6], the HOOS-12 Pain scale appears to be a more efficient measure than its full-length counterpart.

The HOOS-12 Function scale was reliable, valid and responsive, but did not perform as well as other HOOS function scales in some tests. In particular, the HOOS-12 Function scale had notably higher ceiling effects than other HOOS function measures. The most difficult activities such as running were not selected for HOOS-12 because they had relatively high rates of missing data and are not

Table V
Effect sizes and standardized response means for hip-specific measures

	k	Mean Score (SD)				Effect Size		Standardized Response Mean	
		6 month post-THR ($n = 676$)		12 month post-THR ($n = 547$)					
		Pre-THR Score	Change Score	Pre-THR Score	Change Score	6 Month	12 Month	6 Month	12 Month
HOOS-12 Pain	4	40.8 (16.6)	46.6 (21.5)	40.8 (15.9)	48.9 (19.6)	2.81	3.07	2.17	2.50
HOOS Pain	10	44.7 (17.5)	44.9 (20.0)	45.0 (17.0)	46.4 (18.7)	2.56	2.73	2.25	2.48
HOOS-12 Function	4	43.5 (19.3)	43.4 (21.3)	44.2 (19.1)	44.7 (20.6)	2.25	2.34	2.04	2.17
HOOS ADL	17	47.7 (18.7)	40.1 (19.8)	48.7 (18.8)	41.7 (19.4)	2.15	2.22	2.03	2.15
HOOS Sport/Recreation	4	23.3 (18.9)	45.6 (27.4)	23.7 (19.1)	50.6 (25.8)	2.41	2.65	1.66	1.96
HOOS-PS	5	52.7 (16.3)	31.9 (18.2)	53.3 (16.4)	33.8 (17.5)	1.96	2.06	1.76	1.94
HOOS/HOOS-12 QOL	4	26.2 (17.5)	48.3 (24.4)	26.3 (17.3)	52.7 (23.0)	2.77	3.04	1.98	2.29
HOOS-12 Summary	12	36.8 (15.9)	46.1 (20.0)	37.1 (15.4)	48.7 (18.5)	2.90	3.16	2.31	2.64
HOOS, JR	6	48.3 (15.3)	35.8 (17.9)	49.1 (14.7)	37.6 (17.0)	2.34	2.56	2.01	2.21

ADL, Activities of Daily Living; QOL, Quality of Life. All measures scored so 0 = worst possible and 100 = best possible score.

done or are viewed as unimportant by many THR patients[33]. HOOS-PS includes the running item and had a wider measurement range than the HOOS-12 Function scale, with an item threshold range of 3.7SD for HOOS-PS compared to a range of 2.8SD for items in the HOOS-12 Function scale in IRT analyses[6]. However HOOS-PS could not be scored for 8–12% of THR patients in this study due to missing item data, and had lower ES and SRM than the HOOS-12 Function scale. The HOOS Sport/Recreation scale also includes the most difficult HOOS item (running), thereby also extending the measurement range beyond that of the HOOS-12 Function scale. For studies of younger and more active patients, administering and scoring the HOOS Sport/Recreation scale along with the HOOS-12 is recommended, particularly because this only adds three more HOOS items to a questionnaire. In the future, consideration might be given to extending the range of hip-specific physical function measures by adding high performance response options to capture when it is “easy” or “very easy” to do common activities, rather than asking if respondents had no difficulty in performing intense activities that many respondents do not attempt to do. Tests of general physical function measures have shown that this approach can increase the measurement range[34].

Further research is needed into the performance of HOOS-12 in other THR samples, in patient populations other than hip OA patients having THR, and in countries outside the US. In addition, the performance of HOOS-12 should be evaluated in relation to other hip-specific measures, such as the Oxford Hip Score[35] and Harris Hip Score[36]. This study also could not evaluate test-retest reliability of the HOOS-12, although intraclass correlation coefficients for full-length HOOS scales have ranged from 0.75 to 0.97 in previous test-retest studies[37–39] and it is likely that HOOS-12 scales also would have satisfactory test-retest reliability. It also is notable that patients who rated their capability as “less” (compared to preoperatively) still had positive change scores on all measures, on average. This may reflect possible placebo effects from surgery or the impact of comorbid conditions on overall physical capability despite hip-specific improvement, and warrants further study. Finally, analyses reported here were based on HOOS-12 items embedded within the full-length HOOS; HOOS-PS and HOOS, JR items also were embedded within the HOOS. Additional studies should be conducted to confirm that HOOS-12 psychometric properties are similar when the short form is administered by itself.

The full-length HOOS may still offer some advantages in THR populations, including when used in research, in prediction analyses including clinical symptoms, and when THR is performed in patients with high physical activity levels that are better captured by the HOOS Sport/Recreation scale. However, HOOS-12 is a promising alternative to the full-length HOOS and HOOS derivatives and uniquely, it allows for estimation of domain-specific scores of pain, function and QOL plus an overall hip impact summary score, with reduced respondent burden compared to the full-length HOOS. While HOOS-12 domain-specific scales are important for clinical interpretation and systematic reviews of OA treatment, the HOOS-12 Summary score demonstrated potential to serve as an aggregate outcome measure for use in clinical trials, registries and quality initiatives.

Authors' contributions

All authors contributed to study conception and design, analysis and interpretation of the data, and drafting the article or revising it critically for important intellectual content. Dr. Gandek assembled the data and performed the data analysis. All authors read and approved the final manuscript. Dr. Gandek takes responsibility for the integrity of the work as a whole.

Competing interests

Professor Roos is developer of the HOOS, which is freely available with no licensing required for academic or commercial use. Other authors report no conflicts of interest.

Role of the funding source

This research was supported by AHRQ grant R03 HS024632 (Gandek PI) and a FORCE-TJR program project award (P50 HS018910, Franklin PI) to the Department of Orthopedics and Physical Rehabilitation at the University of Massachusetts Medical School. The funding sources did not play any role in the study design, collection, analysis or interpretation of data, in the writing of the manuscript, or in the decision to submit the manuscript for publication. The opinions expressed in this document are those of the authors and do not reflect the official position of AHRQ or the U.S. Department of Health and Human Services.

Distribution

HOOS-12 is available free of charge from www.koos.nu. This site also includes a guide to the different HOOS versions. No licensing or permission to use HOOS-12, HOOS or other questionnaires available from www.koos.nu is required.

Acknowledgements

The authors thank Jakob Bjorner, MD, PhD for psychometric consultation; Nina Deng, EdD for developing the relative validity bootstrapping software; Celeste Lemay RN, MPH, Wenyun Yang, MS and Hua Zheng, PhD for data and computer support; and the researchers, clinicians and patients who provided feedback on KOOS and HOOS item content and translations.

References

1. Nilsdotter A, Bremander A. Measures of hip function and symptoms. Arthritis Care Res 2011;63:S200–7.
2. Gagnier JJ, Huang H, Mullins M, Marinac-Dabic D, Ghambaryan A, Eloff B, *et al*. Measurement properties of patient-reported outcome measures used in patients undergoing total hip arthroplasty: a systematic review. JBJS Rev 2018;6:e2.
3. Nilsdotter AK, Lohmander LS, Klässbo M, Roos EM. Hip disability and osteoarthritis outcome score (HOOS)–validity and responsiveness in total hip replacement. BMC Musculoskel Disord 2003;4:10.
4. Kemp JL, Collins NJ, Roos EM, Crossley KM. Psychometric properties of patient-reported outcome measures for hip arthroscopic surgery. Am J Sports Med 2013;41:2065–73.
5. Lyman S, Lee YY, Franklin PD, Li W, Mayman DJ, Padgett DE. Validation of the HOOS, JR: a short-form hip replacement survey. Clin Orthop Relat Res 2016;474:1472–82.
6. Gandek B, Roos EM, Franklin PD, Ware JE,Jr. Item selection for 12-item short forms of the knee injury and osteoarthritis outcome score (KOOS-12) and hip disability and osteoarthritis outcome score (HOOS-12). 2019;27:746–753. https://doi.org/10.1016/j.joca.2018.11.011.
7. Gandek B, Roos EM, Franklin PD, Ware JE,Jr. A 12-item short form of the Knee injury and Osteoarthritis Outcome Score (KOOS-12): tests of reliability, validity and responsiveness. 2019;27:762–770. https://doi.org/10.1016/j.joca.2019.01.011.
8. Franklin PD, Allison JJ, Ayers DC. Beyond joint implant registries: a patient-centered research consortium for comparative effectiveness in total joint replacement. J Am Med Assoc 2012;308:1217–8.

9. Likert R. A technique for the measurement of attitudes. Arch Psychol 1932;140:5–55.
10. HOOS Scoring 2013. www.koos.nu.
11. Davis AM, Perruccio AV, Canizares M, Tennant A, Hawker GA, Conaghan PG, *et al*. The development of a short measure of physical function for hip OA HOOS-Physical Function Shortform (HOOS-PS): an OARSI/OMERACT initiative. Osteoarthritis Cartilage 2008;16:551–9.
12. Federal Register. Medicare program; Comprehensive Care for Joint Replacement Payment Model for Acute Care Hospitals Furnishing Lower Extremity Joint Replacement Services. Fed Regist 2015;80:73273–554.
13. HOOS-PS User's Guide 2016. Updated March 2016, www.koos.nu.
14. Ware Jr JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992;30:473–83.
15. Ware Jr JE, Kosinski M, Bayliss MS, McHorney CA, Rogers WH, Raczek A. Comparison of methods for the scoring and statistical analysis of SF-36 health profile and summary measures: summary of results from the Medical Outcomes Study. Med Care 1995;33:AS264–79.
16. Ware Jr JE. SF-36 health survey update. Spine 2000;25:3130–9.
17. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16:297–334.
18. Scientific Advisory Committee of the Medical Outcomes Trust. Assessing health status and quality-of-life instruments: attributes and review criteria. Qual Life Res 2002;11:193–205.
19. Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, *et al*. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Med Care 2007;45:S22–31.
20. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, *et al*. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007;60:34–42.
21. Paulsen A, Pedersen AB, Overgaard S, Roos EM. Feasibility of 4 patient-reported outcome measures in a registry setting. Acta Orthop 2012;83:321–7.
22. McHorney CA, Ware Jr JE, Raczek AE. The MOS 36-item short-form health survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. Med Care 1993;31:247–63.
23. Deng N, Allison JJ, Fang HJ, Ash AS, Ware Jr JE. Using the bootstrap to establish statistical significance for relative validity comparisons among patient-reported outcome measures. Health Qual Life Outcome 2013;11:89.
24. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, *et al*. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. BMC Med Res Methodol 2010;10:22.
25. Ware Jr JE, Keller SD. Interpreting general health measures. In: Spilker B, Ed. Quality of Life and Pharmacoeconomics in Clinical Trials. 2nd edn. Philadelphia, PA: Lippincott-Raven Publishers; 1996:445–60.
26. Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR. Clinical Significance Consensus Meeting Group. Methods to explain the clinical significance of health status measures. Mayo Clin Proc 2002;77:371–83.
27. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Med Care 1989;27:S178–89.
28. Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. Med Care 1990;28:632–42.
29. Nunnally JC, Bernstein IH. Psychometric Theory. New York: Mc-Graw Hill; 1994.
30. Pham T, Van Der Heijde D, Lassere M, Altman RD, Anderson JJ, Bellamy N, *et al*. Outcome variables for osteoarthritis clinical trials: the OMERACT-OARSI set of responder criteria. J Rheumatol 2003;30:1648–54.
31. Gossec L, Paternotte S, Bingham 3rd CO, Clegg DO, Coste P, Conaghan PG, *et al*. OARSI/OMERACT initiative to define states of severity and indication for joint replacement in hip and knee osteoarthritis. J Rheumatol 2011;38:1765–9.
32. Singh JA, Dowsey MM, Dohm M, Goodman SM, Leong AL, Scholte Voshaar MMJH, *et al*. Achieving consensus on total joint replacement trial outcome reporting using the OMERACT filter: endorsement of the final core domain set for total hip and total knee replacement trials for endstage arthritis. J Rheumatol 2017;44:1723–6.
33. Wiering B, de Boer D, Delnoij D. Asking what matters: the relevance and use of patient-reported outcome measures that were developed without patient involvement. Health Expect 2017;20:1330–41.
34. Liegl G, Gandek B, Fischer HF, Bjorner JB, Ware Jr JE, Rose M, *et al*. Varying the item format improved the range of measurement in patient-reported outcome measures assessing physical function. Arthritis Res Ther 2017 Mar 21;19(1):66.
35. Dawson J, Fitzpatrick R, Carr A, Murray D. Questionnaire on the perceptions of patients about total hip replacement. J.Bone Joint Surg Br 1996;78:185–90.
36. Harris WH. Traumatic arthritis of the hip after dislocation and acetabular fractures: treatment by mold arthroplasty. An end-result study using a new method of result evaluation. J Bone Joint Surg Am 1969;51:737–55.
37. Klässbo M, Larsson E, Mannevik E. Hip disability and osteoarthritis outcome score. An extension of the western ontario and McMaster universities osteoarthritis index. Scand JRheumatol 2003;32:46–51.
38. de Groot IB, Reijman M, Terwee CB, Bierma-Zeinstra SM, Favejee M, Roos EM, *et al*. Validation of the Dutch version of the hip disability and osteoarthritis outcome score. Osteoarthritis Cartilage 2007;15:104–9.
39. Blasimann A, Dauphinee SW, Staal JB. Translation, cross-cultural adaptation, and psychometric properties of the German version of the hip disability and osteoarthritis outcome score. J Orthop Sports PhysTher 2014;44:989–97.