Osteoarthritis and Cartilage 27 (2019) 746–753

Osteoarthritis and Cartilage

OARSI OSTEOARTHRITIS RESEARCH SOCIETY INTERNATIONAL

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)

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 27 November 2018

Keywords:
KOOS
HOOS
Osteoarthritis
Patient reported outcome measures
Psychometrics
Item response theory

SUMMARY

Objective: To develop 12-item short forms (KOOS-12, HOOS-12) of the 42-item Knee injury and Osteoarthritis Outcome Score (KOOS) and 40-item Hip disability and Osteoarthritis Outcome Score (HOOS) that represent the full-length instruments sufficiently to provide joint-specific pain, function and quality of life (QOL) domain and summary joint impact scores. This paper describes KOOS-12 and HOOS-12 item selection. Subsequent papers will examine KOOS-12 and HOOS-12 reliability, validity and responsiveness.
Design: Items were selected based on qualitative information from patients, clinicians and KOOS/HOOS translators and analysis of data from 1,395 knee osteoarthritis (OA) and 1,281 hip OA patients from the FORCE-TJR cohort who completed KOOS or HOOS before and after total joint replacement (TJR). Item response theory models and computerized adaptive test (CAT) simulations were used to identify items that best measured patients' levels of pain and function pre- and post-TJR. KOOS-12/HOOS-12 items were selected based on content, coverage of a wide measurement range, high item information, item usage in CAT simulations, scale-level properties (reliability, validity, responsiveness), and qualitative information.
Results: KOOS-12 and HOOS-12 each included a pain frequency item and three items measuring pain during increasingly difficult activities (sitting/lying, walking, up/down stairs); function items about standing, rising from sitting, getting in/out of a car, and twisting/pivoting (KOOS-12) or walking on an uneven surface (HOOS-12); and the original 4-item QOL scale.
Conclusions: This study demonstrated the benefits of examining patient-reported outcome measures using modern psychometric methods, to create short forms with diverse content that provide domain-specific and summary joint impact scores.

Introduction

Joint-specific patient-reported outcome measures (PROMs) provide information directly from the patient about the joint-related impact of osteoarthritis (OA) and other joint disorders and their treatment effectiveness. PROM data collection is recommended in registries and clinical practice to monitor outcomes.[1] However, widely-used knee-specific and hip-specific PROMs have limitations. Instruments such as the 24-item Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC),[2] 42-item Knee injury and Osteoarthritis Outcome Score (KOOS)[3] and 40-item Hip disability and Osteoarthritis Outcome Score (HOOS)[4] have the advantage of providing scores for specific domains such as pain or function, but their respondent burden is often viewed as too great to allow for routine use in registries or clinical care. Shorter (5–7 item) forms have been developed from these instruments, including KOOS-PS[5] and HOOS-PS,[6] and KOOS, JR[7] and HOOS, JR,[8] but the first two forms only measure physical function and the latter forms only provide a summary score. Ideally for clinical use, a joint-specific PROM would be brief and provide a summary measure of overall knee or hip impact, but also allow for estimation of scales measuring

* 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.2018.11.011
1063-4584/

pain, function and quality of life (QOL), as recommended by OARSI/OMERACT and ICHOM initiatives.[9–12]

This is the first of three papers that describe development and initial evaluation of new 12-item KOOS and HOOS short forms: KOOS-12 and HOOS-12. The objective was to develop short forms that had 70–75% lower respondent burden than the full-length KOOS and HOOS, while representing item content sufficiently to construct domain-specific scales and a comprehensive joint impact score. This paper describes item selection for KOOS-12 and HOOS-12. Separate papers report the psychometric properties (reliability, validity and responsiveness) of KOOS-12[13] and HOOS-12[14] in knee and hip OA patients who had total joint replacement (TJR).

Methods

KOOS-12 and HOOS-12 measurement model

KOOS-12 and HOOS-12 were constructed to measure joint-specific pain, function, and QOL. All KOOS and HOOS items use attribution to the knee (KOOS) or hip (HOOS), such that patients are asked to focus on a specific joint in evaluating their pain or functional limitations rather than any other orthopedic or non-orthopedic co-morbid conditions. The full-length KOOS measures four knee-specific domains with five scales: Pain ($k = 9$ items), Symptoms ($k = 7$), Function (two scales, Activities of Daily Living (ADL, $k = 17$) and Sport/Recreation ($k = 5$)), and QOL ($k = 4$). Similarly, the full-length HOOS measures four hip-specific domains with five scales: Pain ($k = 10$), Symptoms ($k = 5$), Function (ADL ($k = 17$) and Sport/Recreation ($k = 4$)), and QOL ($k = 4$). KOOS-12 and HOOS-12 items were selected from the full-length instruments to construct Pain, Function, and QOL scales and to provide complementary content across domains in support of construction of a summary score. To represent the mixture of health states from end stage OA to successful TJR, four items were selected for each KOOS-12 and HOOS-12 Pain and Function scale. Pain and Function items that had content overlap (e.g., pain standing, difficulty standing) were evaluated across domains and selected at most for one scale (separately for KOOS-12 and HOOS-12), because including items of similar difficulty and similar content in both the Pain and Function scales would make measurement less efficient. Because both QOL scales contained only four items, the full-length QOL scale was included in KOOS-12 and in HOOS-12, which also allows for direct comparison of results across short and full-length instruments. Symptom items were not included in KOOS-12 and HOOS-12 because conceptually symptoms do not directly capture the impact of knee or hip disorders on QOL. In addition, the Symptom items are heterogeneous and any reduced set of Symptom items selected to supplement a short form should be specific to a diagnostic group.

Criteria for item selection

Items were selected for KOOS-12 and HOOS-12 Pain and Function scales based on qualitative (patient, clinician, translator) and quantitative information. Patient and clinician ratings of item importance came from studies used to develop the KOOS,[15] HOOS,[4] and WOMAC short forms[16–18] and to evaluate KOOS-PS/HOOS-PS.[19] Additional input about item importance was obtained from 18 knee/hip OA patients in Massachusetts and New York and TJR surgeons. In addition, feedback about any difficulty in translating KOOS or HOOS items and cross-cultural differences, obtained from 16 research groups in Europe, Asia and the Middle East, was considered. KOOS and HOOS translation literature also was reviewed.

Quantitative information used in item selection was obtained by examining scaling properties of the Pain and Function items using item response theory (IRT) modeling and computerized adaptive test (CAT) simulations. IRT models can be used to develop item banks; an item bank consists of a set of items that measure the same domain and parameters that describe the items' measurement properties.[20] Short fixed-length scales can be constructed by selecting a subset of items based on their measurement properties and empirical performance. In addition, item parameters are the foundation for CAT. Unlike fixed-length surveys in which everyone answers the same items, CATs administer only the most informative items to each respondent.[21] CAT simulations use existing data for all items in a bank to provide information about which items are most informative in measuring a domain with fewer items.

Study design and participants

Data for quantitative analyses 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.[22] FORCE-TJR surveys were completed by patients pre-TJR and 6 and 12 months post-TJR, at their surgeon's office or at home, via paper-pencil or the Internet. For KOOS-12 and HOOS-12 item selection, random samples of $n = 1,395$ total knee replacement (TKR) patients and $n = 1,281$ total hip replacement (THR) patients were selected from the FORCE-TJR database. All patients had knee or hip OA and obtained TJR surgery between 2011 and 2014. These samples included approximately 50% of all TKR and THR patients of non-white race and/or Hispanic ethnicity; the remaining 50% were reserved for KOOS-12 and HOOS-12 cross-validation analyses, reported in separate papers.[13,14] Sociodemographic characteristics of the Item Selection samples are in Table I.

TKR and THR samples exceeded generally accepted IRT standards ($N > 1,000$, representative sample, all responses for every item selected by some respondents).[23] Most patients had considerable impairment before TJR and improved substantially post-TJR. Therefore, many items were highly skewed, which meant that IRT analyses of only pre-TJR (or only post-TJR) data would have sparse data in some item by response cells. Therefore, datasets used in IRT modeling (Steps 1–6, below) were constructed by randomly selecting pre-TJR data for two thirds of respondents and 6 month or 12 month post-TJR data for the other third, which resulted in datasets with more normal distributions for each item.

FORCE-TJR and this study were approved by the University of Massachusetts Medical School Institutional Review Board.

Data analysis

Methods used to quantitatively evaluate the Pain and Function items are summarized with citations in the Supplementary Appendix and documented in detail elsewhere, including a 2003 series of papers to construct and evaluate a headache impact item bank and accompanying short form and CATs[24–26] and the PROMIS protocol for creating item banks and short forms.[27] These methods are described briefly below.

To select items for the KOOS-12 and HOOS-12, item descriptive statistics initially were examined, including missing data rates. IRT modeling assumptions then were evaluated. First, confirmatory factor analysis (CFA) models were used to verify that all items for each domain were sufficiently unidimensional to measure the same underlying construct (Step 1). Criteria used to evaluate model fit included the Comparative Fit Index (CFI), Tucker–Lewis Index (TLI), root-mean-square error of approximation (RMSEA), and the magnitude of standardized factor loadings (ideally ≥ 0.70).[28] CFA also was used to confirm that item pairs did not have any significant associations (local dependence) once the primary factor

Table I
Characteristics of item selection samples

	KOOS Sample (N = 1,395)	HOOS Sample (N = 1,281)
Age		
Mean (SD)	66.7 (8.9)	64.9 (9.8)
Range	37–92	32–99
Female (%)	70.0	61.8
Race/Ethnicity (%)		
White non-Hispanic	84.4	92.4
Black non-Hispanic	8.9	4.6
Hispanic	3.2	1.1
Other non-Hispanic	3.5	1.9
Education (%)		
High school graduate or less	26.8	20.1
Some post-high school	29.8	25.5
College graduate	18.4	21.4
Post-college education	20.9	20.0
Other/missing	4.1	4.0
N with 6 month, 12 month data	996, 859	899, 755

influencing a patient's response to an item (their overall level of pain or function) was controlled for (Step 2). Any item local dependence was accounted for in IRT modeling. The probability of each item response was plotted across score levels to make sure that each response was most likely to be selected over a unique interval of the response continuum (Step 3). In addition, differential item function (DIF) was examined to make sure that the likelihood of answering an item was the same for persons differing in age, gender or race/ethnicity who were at the same scale level (Step 4). If an item demonstrated DIF, it was not a good candidate for the short form.

Items in a bank then were calibrated using the two-parameter IRT generalized partial credit model (GPCM, Step 5). This model produces item slopes, which can vary across items and quantify how well each item discriminates between respondents, and item thresholds. Collectively, thresholds provide information about the range measured by each item's response categories. IRT models were evaluated to determine how informative each item was and where along the pain or function continuum each item was most informative (Step 6). IRT parameters also were used to conduct CAT simulations, to determine the five items that were best in estimating the total domain score (the score estimated by all items in the bank) for each respondent, at varying levels of pain or function (Step 7). Finally, regression analyses evaluated how well each pain or function item predicted the full KOOS or HOOS QOL scale pre- and post-TJR (Step 8).

A subset of items was selected for candidate short form scales, based on patient, clinician and translator feedback, item content, coverage of a wide measurement range and other IRT information, item usage in CAT simulations, and QOL regression results. Psychometric properties of candidate scales (internal consistency reliability, validity, responsiveness) were compared, to determine if any candidate scales had notably better psychometric performance, and a final set of items was selected for each scale. Reliability, validity and responsiveness of KOOS-12 and HOOS-12 are reported in separate papers.[13,14]

All QOL items were included in KOOS-12 and HOOS-12, so IRT models and CATs were not used to select QOL items. However, QOL item properties were examined.

Results

KOOS-12 Pain

The percentage of item-level missing data was low (1–3% per item pre-TKR, 1–4% post-TKR). A 1-factor CFA model supported unidimensionality overall, with factor loadings = 0.78–0.88 (Table II). This indicated that all nine pain items could be calibrated using one IRT model. There was no item local dependence, and no items showed DIF.

Table II presents information from the IRT model, including item slopes and thresholds on a mean = 0, SD = 1 metric. Each item had four thresholds ranging from b1 to b4, where b1 was the first threshold (e.g., between the responses “Extreme” and “Severe” for KOOS Pain Item 2 (labeled as *kpn2* for “KOOS” “Pain” “Item 2”)) and b4 was the fourth threshold (e.g., between “Mild” and “None” for *kpn2*). For ease of comparing the ranges, only the minimum (b1) and maximum (b4) thresholds are presented. Less difficult activities (activities that should be less painful because they are easier to do) had the lowest thresholds, while more difficult activities had higher thresholds. For example, sitting/lying (*kpn8*) had the lowest threshold of −2.30, while going up/down stairs (*kpn6*) had the highest threshold of 1.30. Slopes were low for *kpn2*, *kpn4* and *kpn7*, indicating these items did not discriminate as well between respondents as other items.

In CAT simulations, items *kpn5* (walking on flat) and *kpn9* (standing) were selected frequently by CAT for respondents at all pain levels (Table II). Items *kpn1* (pain frequency), *kpn2* (pain twisting/pivoting) and *kpn6* (up/down stairs) were selected frequently by CAT to estimate a pain score for respondents who had lower levels of pain, while items *kpn3* (straighten fully), *kpn7* (at night in bed), and *kpn8* (sitting/lying) were selected frequently for respondents who had higher pain levels. In regressions predicting KOOS QOL, knee pain frequency (*kpn1*) had the highest t-statistics (t = 6.17–7.13) of all pain and function items at pre-TKR and post-TKR (data not presented).

Items selected for the KOOS-12 Pain scale were pain frequency (*kpn1*) and pain severity in increasingly difficult activities (*kpn8* (sitting/lying), *kpn5* (walking on flat), *kpn6* (stairs)) (Table II). Pain frequency was the best predictor in KOOS QOL regressions and was selected frequently by CAT for respondents with less pain; pain frequency also is a different concept than pain during activities. Pain sitting/lying had the lowest IRT threshold and was selected frequently by CAT for respondents with the most pain. In contrast, pain going up/down stairs had the highest IRT threshold and was selected frequently by CAT for respondents with the least pain; this item also measures stair climbing with one item vs two stair items in the ADL scale. Reliability, validity and responsiveness of candidate 4-item scales (*kpn1, kpn6, kpn8,* plus *kpn3*, *kpn5* or *kpn9*) were similar (data not presented). Item *kpn5* was selected for KOOS-12 because walking on a flat surface was ranked as one of the most important activities by patients (along with stairs) and because this item best filled in the remaining gaps between short form item thresholds.

Table II
KOOS pain items: Results of confirmatory factor analysis, IRT analysis and CAT simulations

Item Label	Abbreviated content	Loading	IRT Slope	IRT Thresholds		Model Prob.	% of Times Item Used in CAT for Respondents at Different Pain Levels			
				Min	Max		Most Pain	More	Less	Least
kpn1*	Frequency knee pain	0.866	2.41	−0.78	0.87	0.447	0.2	4.9	18.0	19.5
kpn2†	Pain twisting/pivoting	0.825	1.73	−1.27	0.88	0.293	0.1	2.1	17.4	19.7
kpn3†	Pain straight fully	0.840	2.04	−1.73	0.36	0.104	19.8	19.0	3.9	0.0
kpn4	Pain bending fully	0.775	1.27	−1.32	0.76	0.022	0.1	0.0	0.1	9.7
kpn5*	Pain walking on flat	0.870	2.48	−1.96	0.47	0.001	19.8	19.9	19.8	10.5
kpn6*†	Pain up/down stairs	0.860	2.30	−1.30	1.30	0.032	5.2	18.8	19.9	20.0
kpn7	Pain at night in bed	0.852	1.73	−1.99	0.58	0.221	14.9	1.0	0.0	0.3
kpn8*	Pain sitting or lying	0.884	2.30	−2.30	0.52	0.516	19.9	14.5	1.1	0.5
kpn9†	Pain standing upright	0.880	2.78	−1.83	0.63	0.175	19.9	19.9	19.9	19.9

Item Selection Sample, N = 1,395.
kpn#, KOOS Pain item number; Loading, loading in one factor CFA; Model Prob., *P*-value for S-X2 fit statistic in IRT model. Item thresholds are on a scale where 0 is the mean and 1 is the SD; lower thresholds indicate activities that are easier to do and thus should be less painful. CAT results are the percent of the total number of times that the item was selected in a 5-item CAT for respondents with different levels of pain: Most pain, theta score < −1.0; More pain, theta score = −1.0 to 0; Less pain, theta score>0 to 1.0; Least pain, theta score>1.0; a lower theta score indicates that respondent has higher levels of knee pain. Each column sums to 100%. kpn1 responses:1 = Always, 2 = Daily,3 = Weekly, 4 = Monthly,5 = Never. kpn2-kpn9 responses:1 = Extreme,2 = Severe,3 = Moderate,4 = Mild,5 = None.
*KOOS-12 item. †KOOS, JR item.
CFI = 0.968, TLI = 0.957, RMSEA = 0.181 (90% CI 0.172, 0.189) in one-factor CFA.

KOOS-12 Function (ADL and Sport/Recreation)

Item-level missing data generally was low (1–3%), but post-TKR it was 6–8% for *ksp2* (running), *ksp3* (jumping), and *ksp5* (kneeling). Fit of a 1-factor CFA model was not optimal, but factor loadings were substantial (0.73–0.92). In addition, the correlation of ADL and Sport/Recreation factors was 0.76 in a 2-factor model. A bifactor model also supported unidimensionality (general factor OmegaH = 0.92, ADL and Sport/Recreation group factor OmegaH = 0.04, 0.02). Thus, CFA results overall indicated that all 22 items could be calibrated using one IRT model. Nineteen (out of 231) item pairs had item local dependence, all of which included items *ksp2* (running) or *ksp3* (jumping). To account for this in IRT modeling, the initial IRT model included all items except *ksp3*; IRT item parameters were fixed and a second IRT model was run to estimate parameters for *ksp3*. No items showed DIF. In the IRT model, item slopes generally were lower for Sport/Recreation items, but Sport/Recreation items had higher item difficulty thresholds, which raised the ceiling (Table III).

Translation developers raised concerns about international differences for some items. How people shop (*kadl08*) varies across

Table III
KOOS function items: Results of confirmatory factor analysis, IRT analysis and CAT simulations

Item Label	Abbreviated content	Loading	IRT Slope	IRT Thresholds		Model Prob.	% of Times Item Used in CAT for Respondents at Different Function Levels			
				Min	Max		Lowest	Lower	Higher	Highest
kadl01	Descending stairs	0.862	1.83	−1.42	1.34	0.184	0.0	0.0	0.0	0.0
kadl02	Ascending stairs	0.880	2.09	−1.44	1.23	0.848	0.0	0.0	2.8	7.2
kadl03*†‡	Rising from sitting	0.868	2.66	−1.56	1.05	0.824	1.5	1.5	10.7	19.7
kadl04*	Standing	0.878	2.84	−2.02	0.54	0.811	18.2	18.6	15.8	0.5
kadl05†‡	Bending to floor	0.824	2.02	−1.53	0.77	0.592	0.0	0.0	0.0	0.0
kadl06	Walking on flat surface	0.861	2.49	−2.04	0.49	0.323	1.8	0.2	0.2	0.0
kadl07*	Getting in/out of car	0.877	3.08	−1.78	1.11	0.160	18.2	19.7	18.8	19.7
kadl08	Going shopping	0.885	2.73	−1.59	0.81	0.150	9.9	19.6	19.7	9.2
kadl09‡	Put on socks/stockings	0.843	1.92	−1.80	0.64	0.970	0.0	0.0	0.0	0.0
kadl10‡	Rising from bed	0.871	2.76	−1.76	0.60	0.511	18.1	19.5	10.2	0.1
kadl11	Take off socks/stockings	0.851	2.02	−1.86	0.60	0.200	0.1	0.0	0.0	0.0
kadl12	Lying in bed	0.808	1.81	−1.93	0.52	0.660	0.0	0.0	0.0	0.0
kadl13	Get in/out bath/shower	0.809	1.79	−1.57	0.53	0.072	0.0	0.0	0.0	0.0
kadl14	Sitting	0.856	2.64	−2.36	0.25	0.812	10.0	0.5	0.0	0.0
kadl15	Getting on/off toilet	0.850	2.40	−1.94	0.65	0.880	1.6	0.0	0.0	0.0
kadl16	Heavy domestic duties	0.823	1.74	−0.99	1.56	0.137	0.0	0.0	1.3	16.2
kadl17	Light domestic duties	0.888	3.16	−2.11	0.55	0.152	19.6	19.7	19.7	19.8
ksp1‡	Squatting	0.776	1.04	0.01	1.90	0.119	0.2	0.1	0.1	0.2
ksp2	Running	0.915	0.98	0.50	2.05	0.207	0.2	0.1	0.1	0.3
ksp3	Jumping	0.907	0.98	0.53	1.82	0.790	0.2	0.1	0.2	6.7
ksp4*‡	Twisting/pivoting	0.795	1.19	−0.24	1.45	0.008	0.2	0.1	0.1	0.2
ksp5‡	Kneeling	0.730	0.91	0.07	2.52	0.009	0.2	0.1	0.1	0.1

Item Selection Sample, N = 1,395.
kadl#, KOOS ADL item number; ksp#, KOOS Sport/Recreation item number; Loading, loading in one factor CFA; Model Prob., *P*-value for S-X2 fit statistic in IRT model. Item thresholds are on a scale where 0 is the mean and 1 is the SD; lower thresholds indicate activities that are easier to do. CAT results are the percent of the total number of times that the item was selected in a 5-item CAT for respondents with different levels of function: Lowest function, theta score < −1.0; Lower function, theta = −1.0 to 0; Higher function, theta>0 to 1.0; Highest function, theta>1.0; a lower theta score indicates lower levels of function. Each column sums to 100%. All responses: 1 = Extreme, 2 = Severe, 3 = Moderate, 4 = Mild, 5 = None.
*KOOS-12 item. †KOOS, JR item. ‡KOOS-PS item.
CFI = 0.944, TLI = 0.938, RMSEA = 0.149 (90% CI 0.146, 0.152) in one-factor CFA.
CFI = 0.978, TLI = 0.973, RMSEA = 0.099(0.096, 0.103) in bifactor CFA.

countries; for example, driving to a large supermarket is different than walking to a small store. How people sleep and thus the difficulty of rising from a bed (*kadl10*) also differs, as do toileting practices (*kadl15*).[29,30] In addition, men in some cultures do not do light domestic activities such as cooking and dusting (*kadl17*), particularly older men. Therefore, these items were not included in KOOS-12.

After items were excluded due to notable missing data, overlap with activities in the KOOS-12 Pain scale, cross-cultural differences, or never being selected by CAT, the candidate items for the Function scale were: *kadl03, kadl04, kadl07, kadl11, kadl16, ksp1,* and *ksp4*. Items selected for the KOOS-12 Function scale were *kadl03* (rising from sitting), *kadl04* (standing), *kadl07* (getting in/out of a car) and *ksp4* (twisting/pivoting on knee) (Table III). Standing had one of the lowest minimum thresholds; this activity also was selected frequently in Function CAT simulations. Rising from sitting was frequently selected in Function CATs for respondents with higher levels of function and was ranked as highly important by both patients and clinicians. Getting in/out of a car was frequently selected by CAT at all function levels, covered a wide (nearly 3SD) range of measurement, and also was ranked highly by patients. Reliability, validity and responsiveness of candidate 4-item Function scales that included *kadl03*, *kadl04* and *kadl07* plus one item with a high maximum threshold (*kadl16, ksp1 or ksp4*) were similar (data not presented). Twisting/pivoting (*ksp4)* was selected for a number of reasons, including being rated as very important or important by 85% of post-knee surgery patients in a Dutch study[19] and its performance in factor analysis.

HOOS-12 Pain

Item-level missing data was low (1–2% pre-THR; 1–4% post-THR). A 1-factor CFA model was sufficiently unidimensional (factor loadings = 0.87–0.95) to support calibration of the 10 pain items using one IRT model. There was no item local dependence, and no items showed DIF. In the IRT model (Table IV), less difficult activities such as pain sitting/lying had the lowest thresholds, while more difficult activities such as climbing stairs had the highest thresholds. In CAT simulations, the stair item (*hp05*) and all walking items were selected frequently for respondents at all pain levels, while the pain frequency item (*hpn01*) was selected frequently for respondents who had lower levels of pain. In addition, pain frequency had the highest t-statistics (t = 4.68–9.19) of all pain and function items in regressions to predict the HOOS QOL scale (data not presented).

Items selected for the HOOS-12 Pain scale were pain frequency (*hpn01*) and pain severity during increasingly difficult activities (*hpn07* (sitting/lying), *hpn04* (walking on flat surface), *hpn05* (stairs)) (Table IV). Pain frequency was the best predictor of hip-specific QOL, was selected frequently by CAT for respondents with lower levels of pain, and is a different concept than pain during activities. Pain sitting/lying had the lowest IRT threshold and was used more frequently in CAT simulations than the item with the next lowest threshold (*hpn06*). Reliability, validity and responsiveness of candidate 4-item scales (*hpn01, hpn07,* plus two other items measuring walking, stairs or standing) were compared and found to be similar (data not presented). Pain on stairs was selected for the HOOS-12 because it had a high maximum threshold, was selected most frequently in Pain CAT simulations and was rated as important by patients; it also measures stair climbing with one item vs two items in the ADL scale. While all three walking items were used frequently by CAT, *hpn04* (walking on flat) was selected for HOOS-12 because it had the highest slope and best filled in remaining gaps between short form item thresholds.

HOOS-12 Function (ADL and Sport/Recreation)

Item-level missing data generally was low (1–4%), but was 7–9% for *hsp2* (running) post-THR. A 1-factor CFA model supported unidimensionality of the 21 items (factor loadings = 0.83–0.92). The correlation of ADL and Sport/Recreation factors was 0.91 in a 2-factor model, and a bifactor model also supported unidimensionality (general factor OmegaH = 0.97, ADL and Sport/Recreation group factor OmegaH = 0.02, 0.01). Thus all 21 items could be calibrated using one IRT model. There was no item local dependence, and no items showed DIF. Item slopes generally were lower for Sport/Recreation items, but these items had among the highest thresholds (Table V).

As with the KOOS, translation developers raised concerns about international differences in some HOOS items, including shopping (*hadl08*), rising from bed (*hadl10*) and getting on/off the toilet (*hadl15*),[31,32] as well as gender differences in doing light domestic activities (*hadl17*). These items were not included in the HOOS-12. Some HOOS Function items also were never selected in CAT simulations.

Table IV
HOOS pain items: Results of confirmatory factor analysis, IRT analysis and CAT simulations

Item Label	Abbreviated content	Loading	IRT Slope	IRT Thresholds		Model Prob.	% of Times Item Used in CAT for Respondents at Different Pain Levels			
				Min	Max		Most Pain	More	Less	Least
hpn01*	Frequency hip pain	0.886	2.37	−0.68	0.46	0.177	0.0	0.0	11.7	16.3
hpn02	Pain straighten hip fully	0.896	2.88	−1.57	0.34	0.229	16.4	5.3	0.1	0.0
hpn03	Pain bend hip fully	0.888	2.60	−1.34	0.57	0.236	0.1	0.0	0.5	0.8
hpn04*	Pain walking on flat	0.941	5.44	−1.63	0.52	0.702	16.5	16.5	16.5	16.6
hpn05*†	Pain up/down stairs	0.907	3.34	−1.34	0.79	0.802	17.3	28.5	33.2	33.2
hpn06	Pain at night in bed	0.872	1.78	−1.67	0.45	0.206	0.2	0.0	0.0	0.0
hpn07*	Pain sitting or lying	0.886	2.20	−1.97	0.55	0.289	4.2	0.1	0.0	0.0
hpn08	Pain standing upright	0.902	3.50	−1.62	0.45	0.246	16.3	16.5	5.0	0.1
hpn09	Pain walking on hard	0.950	4.47	−1.47	0.72	0.378	16.5	16.6	16.6	16.5
hpn10†	Pain walking uneven	0.954	4.20	−1.17	0.91	0.037	12.4	16.5	16.4	16.4

Item Selection Sample, N = 1,281.
hpn#, HOOS Pain item number; Loading, loading in one factor CFA; Model Prob., *P*-value for S-X2 fit statistic in IRT model. Item thresholds are on a scale where 0 is the mean and 1 is the SD; lower thresholds indicate activities that are easier to do and thus should be less painful. CAT results are the percent of the total number of times that the item was selected in a 5-item CAT for respondents with different levels of pain: Most pain, theta score < −1.0; More pain, theta score = −1.0 to 0; Less pain, theta score>0 to 1.0; Least pain, theta score>1.0; a lower theta score indicates that respondent has higher levels of hip pain. Each column sums to 100%. hpn01 responses: 1 = Always,2 = Daily,3 = Weekly, 4 = Monthly,5 = Never. hpn02-hpn10 responses: 1 = Extreme,2 = Severe,3 = Moderate,4 = Mild,5 = None.
*HOOS-12 item. †HOOS, JR item.
CFI = 0.986, TLI = 0.982, RMSEA = 0.187 (90% CI 0.179, 0.195) in one factor CFA.

Table V
HOOS function items: Results of confirmatory factor analysis, IRT analysis and CAT simulations

Item Label	Abbreviated content	Loading	IRT Slope	IRT Thresholds		Model Prob.	% of Times Item Used in CAT for Respondents at Different Function Levels			
				Min	Max		Lowest	Lower	Higher	Highest
hadl01‡	Descending stairs	0.894	2.78	−1.57	0.55	0.170	0.1	0.0	0.0	0.0
hadl02	Ascending stairs	0.913	3.19	−1.36	0.78	0.094	0.1	3.9	3.6	0.2
hadl03*†	Rising from sitting	0.895	3.17	−1.44	0.86	0.723	0.0	0.1	10.7	7.3
hadl04*	Standing	0.892	3.17	−1.72	0.46	0.804	9.8	0.2	0.0	0.0
hadl05†	Bending to floor	0.868	2.45	−1.22	0.97	0.305	0.0	0.0	0.0	0.2
hadl06	Walking on flat surface	0.922	4.06	−1.71	0.46	0.101	19.7	19.7	15.5	12.7
hadl07*	Getting in/out of car	0.910	3.65	−1.28	1.03	0.172	4.2	16.2	16.8	17.5
hadl08	Going shopping	0.924	3.90	−1.33	0.65	0.782	19.7	19.7	19.8	17.5
hadl09	Put on socks/stockings	0.913	1.99	−0.98	1.01	0.407	0.0	0.0	0.0	0.0
hadl10	Rising from bed	0.912	3.76	−1.55	0.65	0.440	17.7	19.6	16.9	1.0
hadl11	Take off socks/stockings	0.907	2.06	−1.17	0.90	0.192	0.0	0.0	0.0	0.0
hadl12†	Lying in bed	0.857	2.34	−1.42	0.72	0.233	0.0	0.0	0.0	0.0
hadl13‡	Get in/out of bath/shower	0.861	2.32	−1.25	0.38	0.422	0.0	0.0	0.0	0.0
hadl14†‡	Sitting	0.846	2.36	−2.13	0.32	0.028	2.3	0.0	0.0	0.0
hadl15	Getting on/off toilet	0.893	3.15	−1.65	0.51	0.211	6.8	0.7	0.0	0.0
hadl16	Heavy domestic duties	0.897	2.54	−0.82	1.16	0.526	0.0	0.0	4.8	16.9
hadl17	Light domestic duties	0.911	3.78	−1.75	0.46	0.392	19.6	19.8	9.0	0.1
hsp1	Squatting	0.838	1.42	−0.25	1.25	0.010	0.0	0.0	0.0	5.2
hsp2‡	Running	0.829	1.23	0.40	1.58	0.122	0.0	0.0	0.1	9.6
hsp3‡	Twisting/pivoting on leg	0.874	1.85	−0.53	1.00	0.024	0.0	0.0	0.0	0.3
hsp4*	Walk on uneven surface	0.897	2.62	−1.05	0.97	0.085	0.0	0.0	2.6	11.6

Item Selection Sample, N = 1,281.
hadl#, HOOS ADL item number; hsp#, HOOS Sport/Recreation item number; Loading, loading in one factor CFA; Model Prob., *P*-value for S-X2 fit statistic in IRT model. Item thresholds are on a scale where 0 is the mean and 1 is the SD; lower thresholds indicate activities that are easier to do. CAT results are the percent of the total number of times that the item was selected in a 5-item CAT for respondents with different levels of function: Lowest function, theta score < −1.0; Lower function, theta = −1.0 to 0; Higher function, theta>0 to 1.0; Highest function, theta>1.0; a lower theta score indicates lower levels of function. Each column sums to 100%. All responses: 1 = Extreme, 2 = Severe, 3 = Moderate, 4 = Mild, 5 = None.
*HOOS-12 item. †HOOS, JR item. ‡HOOS-PS item.
CFI = 0.973, TLI = 0.970, RMSEA = 0.140 (90% CI 0.137, 0.144) in one-factor CFA.
CFI = 0.988, TLI = 0.984, RMSEA = 0.102 (0.098, 0.106) in bifactor CFA.

Items selected for the HOOS-12 Function scale were *hadl03* (rising from sitting), *hadl04* (standing), *hadl07* (getting in/out of a car) and *hsp4* (walking on uneven surface) (Table V). Getting in/out of car was frequently selected in CAT simulations, covered a wide threshold range and was ranked as important during HOOS development. Walking on an uneven surface had a high maximum IRT threshold and was selected frequently in the Function CAT for respondents at the highest function level. Reliability, validity and responsiveness of candidate HOOS-12 scales that included *hadl07* and *hsp4* plus two other function items (*hadl03*, *hadl04, hadl05, hadl16*, *hsp1*, or *hsp3*) were similar (data not presented). Item *hadl04* (standing) was selected because it had one of the lowest thresholds and was frequently selected in the Function CAT at the lowest function level. Item *hadl03* (rising from sitting) best filled in remaining gaps between short form item thresholds and was ranked as important during HOOS development.

KOOS-12 and HOOS-12 Quality of Life scales

KOOS QOL item missing data was low (1–3%), and a 1-factor CFA model supported unidimensionality of the QOL scale (CFI = 0.999, TLI = 0.996, RMSEA = 0.120, factor loadings = 0.84–0.95)). Similarly, HOOS QOL item missing data was low (1–3%), and a 1-factor CFA supported unidimensionality (CFI = 0.999, TLI = 0.998, RMSEA = 0.088, factor loadings = 0.85–0.97). There was no item local dependence, and no items showed DIF.

Discussion

This paper described selection of Pain, Function and QOL items for KOOS-12 and HOOS-12. It demonstrated the benefit of examining full-length PROMs using both qualitative and modern psychometric methods, to create short forms with diverse content that spans a wide range of measurement. The observation that different items were selected most often by CAT for respondents at differing levels of pain and function underscores the advantages of adaptive item selection; i.e., the importance of matching items to where respondents score when constructing a short form. In addition, this study showed the importance of considering multiple quantitative and qualitative criteria when selecting items for a short form. While items were selected independently for KOOS-12 and HOOS-12, their overlap in item content is notable, indicating that key functional limitations are quite similar despite differences in knee and hip OA.

KOOS-12 and HOOS-12 Pain scales were constructed to include items measuring both pain at rest and pain during movement,[33] while also giving more proportional weight to the pain frequency item than the full-length KOOS and HOOS. While not a consideration in item selection, activities in the short form Pain scales include those in the OARSI-recommended minimum performance-based tests to assess physical function in people with knee or hip OA (walking, stair climbing, standing from a chair).[34] That many of the same activities were selected for performance and self-report measures, using different selection criteria, supports the content validity of the KOOS-12 and HOOS-12 Pain scales.

The most difficult Function items, such as running and jumping, were not selected for KOOS-12 and HOOS-12 because they had relatively high rates of missing data and described tasks not performed by some patients, particularly older OA patients. The extent to which function should be measured with activities of daily living or extend to sport/recreational activities that are more important to younger and more active patients continues to be a matter of debate.[35] Instead of extending the range of measurement of a Function scale by asking about difficulty in performing activities that are often not applicable, its measurement range also could be extended by asking whether respondents find it “easy” or “very easy" to do more common activities. Noteworthy measurement improvements using this approach have been demonstrated with general physical function measures.[36,37]

This study benefitted from large samples of data from patients who ranged from having end-stage OA to few functional limitations (if any) post-TJR. However, all data came from English-speaking U.S. patients. While translation difficulty was a criteria used in item selection, analysis of data from other countries, using methods similar to those used here, may reach different conclusions about item selection. In addition, this analysis used two-parameter IRT models to inform item selection. These models generally fit PRO response data better than one-parameter Rasch models, which do not allow the discrimination parameter (slope) to vary;[27] slopes did vary within each domain in this analysis. However, for interested researchers, it could be informative if Rasch models were used and led to different Pain and Function item selections. If this were done, we would recommend that relative validity testing, such as was implemented in this study,[13,14] be used to compare the resulting scales and KOOS-12/HOOS-12. Finally, patients with knee and hip disorders other than OA were not sampled. To address this limitation, administration of the Sport/Recreation scale along with KOOS-12/HOOS-12 to patients who aspire to high-level function is recommended, to supplement the one Sport/Recreation item in KOOS-12 and in HOOS-12.[13,14]

KOOS-12 and HOOS-12 items differed to some extent from items in KOOS-PS/HOOS-PS and KOOS, JR/HOOS, JR. Criteria used to select items for these forms differed, including use of two-parameter IRT models and CAT simulations for KOOS-12/HOOS-12 vs one-parameter Rasch models for the other instruments. KOOS-PS and HOOS-PS had a wider measurement range than the KOOS-12 and HOOS-12 Function scales, but many KOOS-PS and HOOS-PS items were not as informative in estimating function in TJR patients, as indicated by their low usage in CAT simulations. KOOS, JR Pain, KOOS, JR Function and HOOS, JR Pain items did not cover as wide a measurement range as corresponding KOOS-12 and HOOS-12 items. Implications of these differences for the comparative validity and responsiveness of these alternative forms will be reported in companion papers, which describe scoring algorithms for KOOS-12[13] and HOOS-12[14] and evaluate their reliability, validity and responsiveness in different samples of FORCE-TJR data, to independently evaluate KOOS-12 and HOOS-12 psychometric properties.

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.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.joca.2018.11.011.

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.

Conflict of interest

Professor Roos is developer of the KOOS and HOOS, which are 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 (UMMS). 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

KOOS-12 and HOOS-12 are available free of charge from www.koos.nu. This website also contains user's guides that provide additional information about these questionnaires. No licensing or permission to use KOOS-12, HOOS-12, or other questionnaires available from www.koos.nu is required.

References

1. Rolfson O, Eresian Chenok K, Bohm E, Lubbeke A, Denissen G, Dunn J, *et al*. Patient-reported outcome measures in arthroplasty registries. Acta Orthop 2016;87(Suppl 1):3–8.
2. Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt LW. Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. J Rheumatol 1988;15:1833–40.
3. Roos EM, Roos HP, Lohmander LS, Ekdahl C, Beynnon BD. Knee injury and osteoarthritis outcome score (KOOS)–development of a self-administered outcome measure. J Orthop Sports Phys Ther 1998;28:88–96.
4. Nilsdotter AK, Lohmander LS, Klässbo M, Roos EM. Hip disability and osteoarthritis outcome score (HOOS)–validity and responsiveness in total hip replacement. BMC Musculoskelet Disord 2003;4:10.
5. Perruccio AV, Stefan Lohmander L, Canizares M, Tennant A, Hawker GA, Conaghan PG, *et al*. The development of a short measure of physical function for knee OA KOOS-Physical Function Shortform (KOOS-PS) - an OARSI/OMERACT initiative. Osteoarthritis Cartilage 2008;16:542–50.
6. 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.
7. Lyman S, Lee YY, Franklin PD, Li W, Cross MB, Padgett DE. Validation of the KOOS, JR: a short-form knee arthroplasty outcomes survey. Clin Orthop Relat Res 2016;474:1461–71.
8. 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.
9. Bellamy N, Kirwan J, Boers M, Brooks P, Strand V, Tugwell P, *et al*. Recommendations for a core set of outcome measures for future phase III clinical trials in knee, hip, and hand osteoarthritis. Consensus development at OMERACT III. J Rheumatol 1997;24:799–802.

10. 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.
11. 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.
12. Rolfson O, Wissig S, van Maasakkers L, Stowell C, Ackerman I, Ayers D, *et al*. Defining an international standard set of outcome measures for patients with hip or knee osteoarthritis: consensus of the international consortium for health outcomes measurement hip and knee osteoarthritis working group. Arthritis Care Res (Hoboken) 2016;68:1631–9.
13. 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.
14. Gandek B, Roos EM, Franklin PD, Ware Jr JE. A 12-item short form of the Hip disability and Osteoarthritis Outcome Score (HOOS-12): tests of reliability, validity and responsiveness. Osteoarthritis Cartilage 2019;27:754–61, https://doi.org/10.1016/j.joca.2018.09.017.
15. Roos EM, Toksvig-Larsen S. Knee injury and Osteoarthritis Outcome Score (KOOS) - validation and comparison to the WOMAC in total knee replacement. Health Qual Life Outcome 2003;1:17.
16. Liebs TR, Herzberg W, Gluth J, Rüther W, Haasters J, Russlies M, *et al*. Using the patient's perspective to develop function short forms specific to total hip and knee replacement based on WOMAC function items. Bone Joint Lett J 2013;95-B:239–43.
17. Tubach F, Baron G, Falissard B, Logeart I, Dougados M, Bellamy N, *et al*. Using patients' and rheumatologists' opinions to specify a short form of the WOMAC function subscale. Ann Rheum Dis 2005;64:75–9.
18. Whitehouse SL, Lingard EA, Katz JN, Learmonth ID. Development and testing of a reduced WOMAC function scale. J.Bone Joint Surg Br. 2003;85:706–11.
19. 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.
20. Embretson SE, Reise SP. Item Response Theory for Psychologists. Mahwah,NJ: Lawrence Erlbaum Associates; 2000.
21. Wainer H, Dorans NJ, Eigor D, Flaugher R, Green BF, Mislevy RJ, *et al*. Computerized Adaptive Testing: A Primer. Hillsdale, NJ: Lawrence Erlbaum Associates; 2000.
22. 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.
23. Thissen D, Reeve BB, Bjorner JB, Chang CH. Methodological issues for building item banks and computerized adaptive scales. Qual Life Res 2007;16(Suppl 1):109–19.
24. Bjorner JB, Kosinski M, Ware Jr JE. Calibration of an item pool for assessing the burden of headaches: an application of item response theory to the Headache Impact Test (HIT). Qual Life Res 2003;12:913–33.
25. Kosinski M, Bayliss MS, Bjorner JB, Ware Jr JE, Garber WH, Batenhorst A, *et al*. A six-item short-form survey for measuring headache impact: the HIT-6. Qual Life Res 2003;12:963–74.
26. Ware Jr JE, Kosinski M, Bjorner JB, Bayliss MS, Batenhorst A, Dahlöf CG, *et al*. Applications of computerized adaptive testing (CAT) to the assessment of headache impact. Qual Life Res 2003;12:935–52.
27. 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.
28. Cook KF, Kallen MA, Amtmann D. Having a fit: impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. Qual Life Res 2009;18:447–60.
29. Nakamura N, Takeuchi R, Sawaguchi T, Ishikawa H, Saito T, Goldhahn S. Cross-cultural adaptation and validation of the Japanese knee injury and osteoarthritis outcome score (KOOS). J Orthop Sci 2011;16:516–23.
30. Seo SS, Chung KC, Kim YB. Assessment of validity, reliability and responsiveness of Korean knee injury and osteoarthritis outcome score(KOOS) for the knee injury. J Korean Orthop Assoc 2006;41:441–53.
31. Lee YK, Chung CY, Koo KH, Lee KM, Lee DJ, Lee SC, *et al*. Transcultural adaptation and testing of psychometric properties of the Korean version of the hip disability and osteoarthritis outcome score (HOOS). Osteoarthritis Cartilage 2011;19:853–7.
32. Satoh M, Masuhara K, Goldhahn S, Kawaguchi T. Cross-cultural adaptation and validation reliability, validity of the Japanese version of the Hip disability and Osteoarthritis Outcome Score (HOOS) in patients with hip osteoarthritis. Osteoarthritis Cartilage 2013;21:570–3.
33. Sayers A, Wylde V, Lenguerrand E, Beswick AD, Gooberman-Hill R, Pyke M, *et al*. Rest pain and movement-evoked pain as unique constructs in hip and knee replacements. Arthritis Care Res (Hoboken) 2016;68:237–45.
34. Dobson F, Hinman RS, Roos EM, Abbott JH, Stratford P, Davis AM, *et al*. OARSI recommended performance-based tests to assess physical function in people diagnosed with hip or knee osteoarthritis. Osteoarthritis Cartilage 2013;21:1042–52.
35. Singh JA, Dohm M, Sprowson AP, Wall PD, Richards BL, Gossec L, *et al*. Outcome domains and measures in total joint replacement clinical trials: can we harmonize them? An OMERACT Collaborative Initiative. J Rheumatol 2015;42:2496–502.
36. Fisher Jr WP, Eubanks RL, Marier RL. Equating the MOS SF36 and the LSU HSI physical functioning scales. J Outcome Meas 1997;1:329–62.
37. 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;9(1):66.