Clinical Rheumatology (2021) 40:1547–1558
https://doi.org/10.1007/s10067-020-05403-x

ORIGINAL ARTICLE

Psychometric properties of the 12-item Knee injury and Osteoarthritis Outcome Score (KOOS-12) Spanish version for people with knee osteoarthritis

Gabriel Horta-Baas[1,2] · Rodrigo Vargas-Mena[3] · Erik Alejandre[3] · Ingris Peláez-Ballestas[4] · María del Socorro Romero-Figueroa[5] · Gloria Queipo[6,7]

Received: 29 July 2020 / Revised: 4 September 2020 / Accepted: 12 September 2020 / Published online: 18 September 2020

Abstract

To evaluate the validity, reliability, and responsiveness to change of the 12-item Knee injury and Osteoarthritis Outcome Score (KOOS) Spanish version questionnaire. This study was based on a questionnaire validation design. A cross-sectional survey of 199 patients with knee osteoarthritis (KOA) and ten healthy controls was studied to evaluate the validity and reliability of KOOS-12. One hundred and sixteen patients were assessed for test-retest reliability, and 38 patients were included for a responsiveness assessment. Structural validity was assessed by the confirmatory factor analysis (CFA). Item response theory-based methods were used to determine the performance of the items. Internal consistency reliability was appropriate for all scales (Cronbach's alpha = 0.85–0.94). The intra-class correlation coefficient of KOOS-12 scales ranged from 0.60 to 0.71. The CFA and generalized partial credit model showed that KOOS-12 scales presented a good overall model fit. No differential item functioning was found. Convergent validity was demonstrated by strong correlations (Spearman's rho ≥ 0.70) with KOOS, International Knee Documentation Committee subjective knee evaluation form (IKDC), and Knee Intermittent and Constant Osteoarthritis Pain (ICOAP). Known-groups validity showed that KOOS-12 well discriminated subgroups of patients (radiographic severity and nutritional status). Standardized response means for KOOS-12 scales were ≥ 0.75. Changes in KOOS-12 scales had a moderate to

Key Points
- *KOOS-12 is a short self-reported measure that assesses patient's opinions about the difficulties they experience due to problems with their knee and also covers aspects of pain, functional limitations, and knee-related quality of life.*
- *The Spanish version of KOOS-12 questionnaire is a valid instrument for measuring the patients' opinions about their knee and associated problems, and is both reliable and responsiveness to change.*

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10067-020-05403-x) contains supplementary material, which is available to authorized users.

✉ Gabriel Horta-Baas
gabho@hotmail.com

[1] Servicio de Reumatología, Hospital General Regional número 1, Delegación Yucatán, Instituto Mexicano del Seguro Social, Calle 41 número 439 x 34, Colonia Industrial, 97150. Mérida, Yucatán, Mexico

[2] Programa de Maestría y Doctorado en Ciencias Médicas, Odontológicas y de la Salud, Doctorado en Ciencias Médicas, Universidad Nacional Autónoma de Mexico, Ciudad de México, Mexico

[3] Servicio de Traumatologia y Ortopedia, Hospital General Regional número 1, Delegación Yucatán, Instituto Mexicano del Seguro Social, Mérida, Yucatán, Mexico

[4] Servicio de Reumatología, Hospital General de México "Dr. Eduardo Liceaga", Ciudad de México, Mexico

[5] Centro de Investigación en Ciencias de la Salud, Universidad Anáhuac México, Campus Norte Huixquilucan, Ciudad de México, Mexico

[6] Servicio de Genética, Hospital General de México "Dr. Eduardo Liceaga", Ciudad de México, Mexico

[7] Facultad de Medicina, Universidad Nacional Autónoma de México, Mexico City, Mexico

strong correlation (Pearson's $r \geq 0.40$) with the changes in the KOOS, ICOAP, and IKDC scales. The KOOS-12 Spanish version is a valid, reliable, and responsiveness to change questionnaire to measure patients' opinions about their knee and associated problems in Mexican subjects with KOA.

Keywords Knee pain · Osteoarthritis · Pain measurement · Patient outcome assessment · Psychometrics · Quality of life

Introduction

Patient-reported outcome measures (PROMs) reflect the perceived impact of a specific clinical condition on individuals and are extensively used to measure health care interventions [1]. A knee-specific PROM should be brief and provide a summary measure of overall knee impact, along with pain, function, and quality of life (QoL) [2]. In order to measure physical function, self-reported measures of function and testing of the execution of a specific task associated with function (performance-based tests) could be used [3]. Additionally, performance-based measures aim at quantifying what patients can actually do; the most relevant functional domains are level walking, stair negotiation, and sit-to-stand movement [3, 4]. On the other hand, PROMs assess patients' perceptions about their abilities [3].

Knee osteoarthritis (KOA) implies an enormous burden of illness for people suffering it [5], so it is necessary to have valid instruments to measure the perception of the burden illness by patients with KOA. The Knee injury and Osteoarthritis Outcome Score (KOOS) is a knee-specific instrument, developed to assess patients' opinion about their knee and associated problems [6], and is one of the most widely used PROMs to evaluate patients with KOA. However, completing the 42-item questionnaire presents a significant burden for patients, and it is often regarded as being time-consuming for routine clinical use [7].

KOOS-12 is a shortened version of KOOS and was developed using item response theory (IRT) methods, as well as patients' opinions, clinicians, and researchers on its content, clinical importance, and translatability [2]. Evaluation of the psychometric properties of KOOS-12 demonstrates that it is a valid and reliable instrument to be used in patients with KOA who had a total knee replacement (TKR) [7, 8]. Psychometric analyses of the Spanish version of the KOOS-12 questionnaire are required to assess whether the scale measures the patients' opinion about their knee and associated problems as intended in Spanish-speaking populations and populations other than KOA patients with TKR. This study aimed at assessing the reliability, construct validity, and responsiveness to change of the Spanish version of the KOOS-12 questionnaire in patients with KOA.

Methods

This study was based on a validation design. Consecutive outpatients in the orthopedic and rheumatology clinics of two secondary care public hospitals were invited to participate. Admitted subjects were those diagnosed with primary KOA as established by the American College of Rheumatology [9] with Kellgren–Lawrence (K-L) grade one to four [10]. In bilateral knee involvement, the degree of the worst knee was recorded as the K-L grade. To investigate known-groups validity, individuals without KOA (healthy people) older than 18 years were invited to participate. A prior total knee replacement surgery, joint surgery six months before, another rheumatic disease (e.g., rheumatoid arthritis, psoriatic arthritis, and fibromyalgia), diabetic neuropathy, any known malignancy or major organ failure, neurological diseases, and unwillingness to complete the questionnaire were reasons for exclusion.

The sample size calculation was based on recommendations by experts in this field. For the factor analysis, the sample size should be at least seven times the number of items (i.e., 28 patients per scale) with a minimum of 100 [11]. Concerning sample size for the Rasch modeling analysis, recent guidelines indicate that a sample of ≥ 200 patients allows robust estimates of the model parameters [11]. Some authors consider a sample size of 200 participants as the minimal sample size for estimating stable GPCM parameters [12–14]. Also, a minimum sample of 200 participants was required.

Patients who agreed to participate were asked to complete a questionnaire set (paper and pencil format). Following completion of the questionnaires, the functional capacity of each patient was measured by two performance-based tests to assess physical function. The study was in accordance with the ethical standards of the Declaration of Helsinki and was approved by the Ethics Committee of the Instituto Mexicano del Seguro Social (approval date 2019-01-02; approval number R-201-3201-085), and patients signed an informed consent to participate.

Measures

KOOS It comprises 42 items with five scales: (1) pain, frequency, and severity during functional activities; (2) symptoms; (3) function in daily living (ADL), difficulty experienced during everyday activities; (4) sport and recreational activities,

difficulty experienced with sport and recreational activities; and (5) knee-related QoL. Patients respond to each item based on their knee condition over the previous week on a five-point rating scale. Such subscales are scored separately, a total score is not recommended. Scores are transformed to a 0–100 scale; higher scores represent better outcomes [15].

KOOS-12 (electronic supplementary material) It contains three domain-specific scales that measure pain, function, and knee-specific QoL. At least half of the items in the scale must be answered to calculate a scale score, and a person-specific estimate is imputed for any missing item data. Scores are then transformed to a score from 0 to 100: 0 is the worst and 100 is the best possible score. The KOOS-12 summary knee impact score was calculated as the average of the three scales scores. A summary score is not calculated if any of the three scale scores are missing [2].

Knee intermittent and constant osteoarthritis pain It comprises 11 items. Five items evaluate constant pain, and six items consider intermittent pain. Patients respond to each of them based on their knee condition over the previous week on a five-point rating scale. Total scores are created by adding up item scores and normalizing from 0 (no pain) to 100 (extreme pain); higher scores represent worse outcomes [16].

International knee documentation committee subjective knee evaluation form It comprises 18 items, in the domains of symptoms, functioning during activity of daily living and sports, current function of the knee, and participation in work and sports. The total score is calculated as (sum of items)/(maximum possible score) × 100. Possible score goes from 0 to 100, where 100 means no limitation with daily or sporting activities and the absence of symptoms; lower scores represent worse outcomes [4].

World Health Organization Disability Assessment Schedule 2.0 It measures people's activity limitations and participation restrict ability per the constructs included in the International Classification of Functioning, Disability, and Health(ICF). The World Health Organization Disability Assessment Schedule 2.0 (WHODAS 2.0) cognition, mobility, and ADL are self-administered questionnaires based on six, five, and eight items, respectively. Subscale scores are created by summing item scores and normalizing from 0 (no disability) to 100 (total disability pain) [17].

Performance-based measures

The 30-s Chair Stand Test (30-s CST) and Timed Up and Go test (TUG) were applied using a folding chair without arms; with a seat height of 43 cm, 30-s CST is a performance-based measure that evaluates the activity "sit-to-stand movement." The test is executed by scoring the maximum amount of complete chair stand movements during 30 s [18]. Time (seconds) is taken to rise from a chair, walk 3 m, turn, walk back to the chair, and then sit down wearing regular footwear, using a walking aid if required [18].

Radiographic severity of knee osteoarthritis

Bilateral weight-bearing anteroposterior and lateral semi-flexed radiographs were recorded for both knees in each subject. They were radiologically graded according to the Kellgren–Lawrence classification [10]. The radiographs were evaluated blindly by an experienced rheumatologist (GHB).

Reliability assessment

The internal consistency of the scale was evaluated using the Cronbach's alpha coefficient, McDonald's omega coefficient, and the IRT reliability coefficient. Interpretation of the IRT reliability coefficient is similar to Cronbach's alpha; a value between 0.80 and 0.95 was considered good internal consistency [19, 20]. For the test-retest reliability evaluation, all patients were invited to a second evaluation. The clinical assessment was repeated by the same physician 14 days after the first assessment at the same study site. The test-retest reliability was calculated by using an intra-class correlation coefficient (ICC, two-way mixed-effect ANOVA model with interaction for the absolute agreement between single scores). An ICC > 0.70 was considered adequate [20]. Measurement error was obtained by evaluating the standard error of measurement (SEM), smallest detectable change (SDC), and Bland-Altman limits of agreement [20].

Validity assessment

Construct validity was evaluated by structural validity, relationships to scores of other instruments, and differences between relevant groups [21]. Structural validity was assessed by confirmatory factor analysis (CFA). The diagonal weighted least squares estimation (DWLS) method with polychoric correlations was used to estimate the factorial model parameters [22]. Factorial loadings > 0.70 are desirable [21]. The goodness of fit of the model was analyzed with the chi-square test, whose p value ≥ 0.05 indicates that the proposed model fits the data. Other indicators of good fit are the root mean square error of approximation (RMSEA) < 0.06, the standardized root mean square residual (SRMR) ≤ 0.08, the comparative fit index (CFI) ≥ 0.95, and the Tucker-Lewis index (TLI) ≥ 0.95 [22, 23]. An index RMSEA < 0.08 demonstrates an adequate fit, while an RMSEA > 0.1 is considered a poor fit [24].

Construct validity was considered adequate if expected correlations were found with existing measures assessing similar (convergent validity) and different (divergent validity)

constructs [4]. To establish convergent validity, some hypotheses were tested:

1. KOOS-12 pain should present a positive and strong correlation ($r > 0.7$) with KOOS pain, ADL, and QoL [8], and negative and moderate correlation ($r > -0.6$) with ICOAP [25–27].
2. KOOS-12 function should present a positive and strong correlation ($r > 0.7$) with KOOS ADL [8], and moderate ($0.4 < r < 0.7$) with KOOS sport, TUG, and 30-s CST.
3. KOOS-12 summary should present a strong correlation ($r > 0.7$) with KOOS pain, ADL, and QoL, and moderate to strong ($r > 0.4$) with TUG and 30-s CST.
4. IKDC score would show a positive and moderate correlation with the KOOS-12 pain and QoL, and strong correlation with KOOS-12 function and summary.
5. Correlation between KOOS-12 function and WHODAS 2.0 mobility and activities of daily living subscales should be moderate to strong ($0.4 < r < 0.8$).

To establish the divergent validity, the following hypotheses were tested:

1. KOOS-12 pain scale should present a positive and moderate correlation ($0.4 < r < 0.69$) with KOOS symptoms and sport [8].
2. KOOS-12 QoL should show a moderate correlation with KOOS pain, symptoms, and sport.
3. Correlation between KOOS-12 function scale and WHODAS 2.0 cognitive disability scale should be low ($r < 0.4$).

Known-groups validity was measured by testing a priori hypotheses about subgroups expected significant differences in mean KOOS-12 scores. Hypotheses were formulated as follows: (1) KOOS-12 scales and summary score in healthy people would be significantly higher than score in patients with KOA; (2) KOOS-12 subscales and summary score would be significantly higher in patients with KOA grade 1 compared with patients with grades 3 and 4; and finally, (3) the KOOS-12 summary score would be significantly lower in obese patients.

Item response theory analysis The G2-LD index and Q3 index were used to evaluate the local independence assumption of items [28, 29]. The partial credit model (PCM) and generalized partial credit model (GPCM) were used to obtain item and person parameters using the marginal maximum likelihood estimation with expectation-maximization (MML-EM) algorithm. Two models fitted the data, and their overall fits were compared using the likelihood ratio test (LRT). The LRT assesses whether the model with unrestricted values for the discrimination parameter is necessary to improve the model's fit. In the PCM (an extension of the Rasch model), all item response functions have the same discrimination parameter (α). In GPCM, discrimination parameter was allowed to vary across items. Difficulty parameters (β-parameters) were interpreted as standard deviations showing the range of latent trait covered by the item. The higher the β-parameters, the higher the trait level a respondent needs to endorse that response option [30]. The discrimination parameter measures the strength of the relationship between the item and the latent trait being measured [30]. Item fit was assessed using $S\text{-}X^2$, and misfit was indicated by significant results with a Benjamini and Hochberg adjusted overall alpha level of 0.05 [31]. The overall model fit was analyzed using limited information statistics (M2), along with the associated RMSEA and SRMR index [32]. A p value > 0.05 of the M2, RMSEA < 0.05, and SRMR < 0.027 demonstrates an excellent fit, while an RMSEA <0.089 and SRMR <0.05 are considered an adequate fit [32]. Differential item functioning (DIF) was investigated for age (< 64 years, 64–71 years, and > 71 years), sex (male vs. female), and education level (< 10 years vs. ≥ 10 years). DIF was declared present if significant differences in model fit between non-DIF and DIF models were observed [15]. Person fit was examined by using the standardized statistic *Zh* (Drasgow, Levine, and Williams) [33]. Person-fit statistics compare a person's observed and expected item scores across test items. Patients with *Zh*-values above or below 2 reflect participants with "atypical" or "inconsistent" response patterns [33]. Large negative *Zh*-values indicate non-fitting response patterns given the model and the trait value.

Floor and ceiling effects

Floor or ceiling effects were considered present if more than 15% of respondents achieved the lowest or highest possible score [20].

Responsiveness assessment

From 199 patients at the beginning of the study, 38 received intra-articular treatment (Hylan-GF 20, collagen-PVP, or glucocorticoids) and were included in the responsiveness assessment, which was performed 2 months after the first dose of the treatment. Three methods were used to evaluate responsiveness: the standardized response mean (SRM), effect sizes, and hypothesis testing. For the interpretation of the SRM, the following cutoff points were established: 0.20, 0.50, and 0.80 to indicate a low, moderate, and high sensitivity to change, respectively [34]. Hypothesis testing assessed whether the changes in pain intensity measured by KOOS-12 subscales and summary were correlated ($r \geq 0.70$) with changes measured by KOOS subscales, IKDC, and ICOAP.

Statistical analysis

Results are presented as *n* (%) for categorical variables and as mean and standard deviation (mean ± SD) or median (interquartile range) for continuous variables, as appropriate. To evaluate the differences among two groups, the *t* test was used, and a size effect estimation was reported with Cohen's *d*, considering 0.2, 0.5, and 0.8 as threshold values to estimate low, medium, and large size effects, respectively [35]. One-way ANOVA with multiple comparisons was conducted using the Bonferroni test to discern differences between groups [36]. Size effect estimation was reported with eta squared, considering 0.01, 0.06, and 0.14 as threshold values to estimate low, medium, and large size effects, respectively [35]. Strength and direction of the relationship between two variables were evaluated using Pearson's correlation coefficient (*r*) if both variables are measured on an interval scale and normally distributed, otherwise using Spearman's correlation coefficient (rho). Statistical analysis was performed using the R statistical program (2020, R Core Team, Vienna, Austria). CFA was approached with the lavaan package, the parameters of the GPCM were determined with the mirt package, and DIF analyses using proportional odds cumulative logistic models were run in the lordif package.

Results

A total of 199 patients with KOA and ten healthy people participated in this study. One hundred and sixteen patients were re-evaluated for reliability testing. The mean age of the participants in validation sample ($n = 209$) was 63 years (minimum 34, maximum 90 years), and 78.95% ($n = 165$) were women. The median scores on KOOS-12 pain, function, quality of life, and summary were 43.75, 37.5, 31.25, and 37.5, respectively. Rates of missing data were low (< 2%). The characteristics of all included patients at baseline, test-retest sample, and responsiveness sample are presented in Table 1.

Reliability

KOOS-12 showed appropriate internal consistency. Cronbach's alpha was 0.87, 0.91, 0.85, and 0.94 for KOOS-12 pain, function, QoL, and summary, respectively. The omega coefficient was 0.87, 0.90, and 0.86 for KOOS-12 pain, function, and QoL, respectively. The ICC of KOOS-12 was 0.63, 0.60, 0.71, and 0.71 for the pain, function, QoL, and summary, respectively. Standard error of measurement values ranged between 9.38 and 13.19. The smallest detectable change ranged from 28.32 to 36.56 (Table 2).

Structural validity

Three separate analyses were carried out on the pain, the function, and the QoL scale (Table 3). All items load strongly (factorial loadings > 0.7) onto their respective factors. CFA revealed that one-factor for KOOS-12 function and QoL models showed a good model fit. The one-factor KOOS-12 pain model had an adequate model fit.

Item response theory analyses

IRT analyses were conducted for KOOS-12 pain, function, and QoL scale separately. For all scales, the PCM (Rasch-based) was tested against the GPCM, and this one fit better than the PCM (Supplementary Table S1). The item parameter estimates from KOOS-12 subscales of the GPCM calibrations are reported in Table 4. All the items included in three scales presented a good fit at the item level. There was no item local dependence. No DIF was found between sex, age, or education level. The function and QoL scales had an overall good model fit, and the order of categories' thresholds for all items was good. Person-fit statistics detected < 4% persons with "atypical" response patterns. Within the IRT framework, all scales yielded appropriate reliability (Table 4).

KOOS-12 pain scale Item 1 "Frequency knee pain" had the lowest discrimination parameter, indicating these items did not discriminate as well between respondents as other items. The four items of the pain scale covered a wide range of difficulties ranging from − 1.61 to 1.69. Item 4 "Pain sitting or lying" was the item with the most considerable difficulty on the pain scale, that is, high levels of pain severity are required for the patient to have a higher probability of selecting the last category of the response options "extreme." In contrast, the item with the least difficulty was the item 3 "Pain up/downstairs," that is, very low levels of knee pain severity are required for the patient to have a higher probability of selecting the response category "None." Conversely, the category 4 "extreme" is the high probability for the patients with higher knee severity, but probability decreases as knee pain severity does.

Item 2 "Pain walking on flat" in KOOS-12 pain scale provided more information than the other items. The overall fit of the model for the pain subscale was acceptable (M2 = 0.01, RMSEA = 0.12, and SRMR < 0.05; Table 3). The reliability coefficient for KOOS-12 pain was 0.90, yielding appropriate reliability.

The correlation of the scale scores was calculated using the summated and transformed scoring method (0 to 100) and using the IRT-based scoring (Theta or latent trait) were very strong. The Pearson correlation coefficient between the scores obtained by these two

Table 1 Clinical and demographic characteristics of participants

Characteristics	Validity sample ($n = 199$ patients and 10 healthy subjects)	Test-retest sample ($n = 116$ patients)	Responsiveness sample ($n = 38$ patients)
Females, *n* (%)	165 (78.95)	100 (86.21)	32 (84.21)
Comorbidity, *n* (%)			
Obesity	108 (51.67)	70 (60.34)	19 (50)
Hypertension	86 (41.15)	51 (43.97)	18 (47.37)
Diabetes mellitus	46 (22.01)	25 (21.55)	7 (18.42)
Dyslipidemia	26 (12.44)	15 (12.93)	6 (15.79)
Osteoporosis	11 (5.29)	9 (7.75)	1 (2.63)
Hypothyroidism	6 (2.87)	5 (4.31)	1 (2.63)
Years of education	9 (6)	9 (6)	9 (9)
BMI	30.47 (7.23)	32.00 (6.40)	29.95 (8.69)
Age	63.36 ± 11.83	62.26 ± 10.20	64.86 ± 10.21
ICOAP	52.27 (38.63)	57.13 ± 20.03	66.14 ± 21.75
KOOS-12			
Pain	47.28 ± 24.47	43.75 ± 17.74	36.67 ± 23.97
Function	37.5 (31.25)	36.11 ± 20.77	33.38 ± 25.88
Quality of life	31.25 (37.50)	28.25 ± 17.76	22.47 ± 18.25
Summary	37.5 (29.16)	35.24 ± 16.33	30.84 ± 20.32
KOOS			
Pain	49.68 ± 24.87	44.10 ± 18.55	39.54 ± 24.33
Symptoms	50 (39.28)	45.72 ± 19.48	42.19 ± 22.94
Function, daily living	47.05 (35.29)	43.51 ± 20.66	39.31 ± 25.59
Function, sports and recreational activities	20 (27.73)	20 (30)	5 (25)
Missing data			
Pain scale			
"Frequency knee pain"	0	1 (0.86%)	0
"Pain walking on flat"	0	0	0
"Pain up/down stairs"	1 (0.48%)	1 (0.86%)	1 (2.63%)
"Pain sitting or lying"	0	0	0
Function scale			
"Rising from sitting"	1 (0.48%)	0	0
"Standing"	2 (0.96%)	1 (0.86%)	0
"Getting in/out of car"	2 (0.96%)	1 (0.86%)	0
"Twisting/pivoting"	4 (1.91%)	4 (3.45%)	1 (2.63%)
QoL scale			
"Aware of knee problem"	0	0	0
"Modified lifestyle due to knee"	4 (1.91%)	1 (0.86%)	0
"Lack of confidence in knee"	1 (0.48%)	0	0
"Overall difficulty with knee"	1 (0.48%)	1 (0.86%)	0

Mean ± standard deviation. Median (interquartile range). BMI: Body Mass Index; NRS: Numeric Rating Scale; ICOAP: Intermittent and Constant Osteoarthritis Pain scale; KOOS, Knee injury and Osteoarthritis Outcome Score

methods was −0.983 (95%CI 0.987 to −0.978), −0.982 (95%CI −0.986 to −0.976), and −0.974 (95%CI −0.980 to −0.966) for pain, function, and QoL, respectively (Supplementary Figure). Therefore, it was considered to present the results on a 0 to 100 scale.

Convergent validity, divergent validity, and validity of known groups

Seventy-nine percent of the hypotheses raised for the evaluation of the convergent validity were verified. Similarly, 83%

Table 2 Test-retest reliability and responsiveness to change of the KOOS-12 and KOOS questionnaires

	Intra-class correlation coefficient[1]	Test-retest ($n = 116$)					Responsiveness to change ($n = 38$)		
		Standard error of measurement	Smallest detectable change for individual subject	Smallest detectable change for group	Average difference (standard deviation)	95% limits of agreement (Bland-Altman)	Average difference (standard deviation)	Standardized response mean	Effect size
KOOS-12									
KOOS-12 pain	0.63 (0.51 to 0.73)	10.96	30.38	2.82	− 0.28 (15.50)	− 30.67; 30.09	20.77 (25.57)	0.81	0.86
KOOS-12 function	0.60 (0.47 to 0.71)	13.19	36.56	3.39	1.45 (18.65)	− 38.21; 35.25	10.29 (22.49)	0.85	0.74
KOOS-12 quality of life	0.71 (0.60 to 0.78)	10.22	28.32	2.63	0.66 (14.45)	− 27.66; 28.99	12.88 (17.15)	0.75	0.72
KOOS-12 summary	0.71 (0.61 to 0.79)	9.38	26.00	2.41	0.61 (13.26)	− 25.39; 26.61	17.65 (18.77)	0.94	0.87
KOOS									
KOOS pain	0.64 (0.53 to 0.74)	11.34	31.44	2.91	− 0.44 (16.04)	− 31.88; 30.99	22.24 (26.23)	0.84	0.80
KOOS ADL	0.63 (0.50 to 0.72)	12.78	35.42	3.28	0.43 (18.07)	− 34.99; 35.86	18.97 (22.41)	0.84	0.74
KOOS sport/rec	0.64 (0.52 to 0.74)	11.98	33.22	3.09	0.19 (16.94)	− 33.02; 33.41	15.76 (24.82)	0.63	0.64

[1] Two-way mixed-effect ANOVA model with interaction for the absolute agreement

of the hypothesis in the assessment of the divergent validity were confirmed. KOOS-12 pain scale showed a very strong correlation (rho ≥ 0.79) with KOOS pain and ADL scales and ICOAP scale. KOOS-12 scale function showed very strong correlations (rho ≥ 0.80) with KOOS pain, ADL and sport scale, IKDC, and ICOAP scale. KOOS-12 QoL was strongly correlated with KOOS ADL and sports scales, and IKDC scale. KOOS-12 summary scale presented very strong correlations with KOOS pain, ADL and QoL scales, IKDC, and ICOAP scale. Relationship between KOOS-12 scale and subscales was moderately correlated with TUG and the 30-s CST (0.46 < rho < 0.55; Table 5).

Digital radiographs were available for 172 patients to assess known-groups validity. As hypothesized, patients with major K-L grading reported more pain severity. Post hoc analysis demonstrated a significant decrease in KOOS-12 scales

Table 3 Results from classical item analysis and unidimensionality analysis of the KOOS-12 questionnaire. Factor loadings (standard error) and goodness of fit indices from confirmatory factor analysis

Items	KOOS-12		
	Pain scale	Function scale	Quality of life scale
Item 1	0.71 (0.00)	0.90 (0.00)	0.73 (0.00)
Item 2	0.94 (0.07)	0.90 (0.02)	0.71 (0.05)
Item 3	0.88 (0.06)	0.89 (0.02)	0.86 (0.06)
Item 4	0.77 (0.06)	0.80 (0.03)	0.92 (0.06)
Average variance extracted	0.70	0.77	0.66
Model Fit			
Chi-squared	0.12	0.60	0.78
Root mean square error of approximation, RMSEA (90% confidence interval)	0.07 (0.00–0.-17)	0.001 (0.001–0.1-1)	0.001 (0.001–0.0-8)
Comparative fit index, CFI	0.99	1.00	1.00
Tucker-Lewis index, TLI	0.99	1.00	1.02
Standardized root mean square, SRMR	0.02	0.01	0.01

and summary score among subjects with KOA vs. healthy controls ($p < 0.01$), grade 4 vs. grade 1 ($p < 0.01$), grade 4 vs. grade 2 ($p < 0.01$), and grade 4 vs. grade 3 ($p < 0.05$). Otherwise, no significant differences were found. KOOS-12 pain, function, QoL, and the summary score were significantly higher in non-obese than in obese patients. Validity results of known groups were very similar using the summed scores or the Theta levels in the comparison of the groups (Table 6).

Floor and ceiling effects

None of the KOOS-12 scales presented a floor or ceiling effect. In patients with KOA, 2.01% ($n = 4$) had the highest score (best outcome) on KOOS-12 scales and summary. Similarly, 2.01% ($n = 4$) presented the lowest score (worst result) on KOOS-12 scales and in KOOS-12 summary.

Responsiveness

Responsiveness assessment indicated that KOOS and KOOS-12 were sensitive to change. There were significant improvements in KOOS-12 scores eight weeks after intra-articular treatment. KOOS-12 summary had the highest effect size of all scales (Table 2). SRMs for KOOS-12 scores ranged from 0.75 to 0.94 (Table 2). SMR for KOOS-12 summary score was higher than the three KOOS-12 scales and the three KOOS scales evaluated. KOOS-12 summary score had strong ($r \geq 0.76$) and significant correlation with the KOOS pain and

Table 4 Estimated slope, location, and threshold parameters, model fit, and reliability coefficient for the KOOS-12

Item	Slope or item discrimination (α)	Item difficulty parameters (thresholds)[1]				Overall location or difficulty[2]	Diff	Item fit	Overall model fit	Person fit		IRT reliability coefficient
		$\beta 1$ 2 vs. 1	$\beta 2$ 3 vs. 2	$\beta 3$ 4 vs. 3	$\beta 4$ 5 vs 4			p value S-$X2$[3]		Misfitting responses[4]	Overfitting responses[5]	
Pain												
1. "Frequency knee pain"	1.23	−1.37	−1.12	−1.31	0.86	−0.73	No	0.26	M2 = 0.01; RMSEA = 0.12 (0.04–0.21); SRMR = 0.03; TLI = 0.96; CFI = 0.98	1.9%	0%	0.90
2. "Pain walking on flat"	5.39	−0.92	−0.11	0.72	1.47	0.29	No	0.68				
3. "Pain up/down stairs"	2.71	−1.61	−0.89	−0.03	0.79	−0.43	No	0.49				
4. "Pain sitting or lying"	1.55	−0.69	0.01	1.10	1.69	0.53	No	0.16				
Function												
1. "Rising from sitting"	3.20	−1.40	−0.61	0.11	1.22	−0.16	No	0.55	M2 = 0.30; RMSEA = 0.03 (< 0.01–0.14); SRMR = 0.01; TLI = 0.99; CFI 0.99	3.9%	0%	0.90
2. "Standing"	3.34	−1.25	−0.57	0.27	1.05	−0.12	No	0.57				
3. "Getting in/out of car"	2.99	−1.17	−0.60	0.20	1.22	−0.08	No	0.57				
4. "Twisting/pivoting"	1.44	−1.39	−0.60	−0.96	−0.04	−0.75	No	0.55				
Quality of life												
1. "Aware of knee problem"	1.29	−1.12	−1.70	−1.35	0.20	−0.99	No	0.12	M2 = 0.56; RMSEA = < 0.01 (< 0.01–0.11); SRMR = 0.01; TLI = 1.00; CFI 1.00	1.9%	0%	0.90
2. "Modified lifestyle due to knee"	0.85	−0.57	−1.47	1.38	−0.68	−0.33	No	0.12				
3. "Lack of confidence in knee"	2.04	−1.68	−0.94	0.17	0.05	−0.59	No	0.12				
4. "Overall difficulty with knee"	3.52	−1.55	−0.87	−0.10	0.59	−0.48	No	0.45				

[1] Each item had four thresholds ranging from $\beta 1$ to $\beta 4$, where $\beta 1$ was the first threshold (e.g., between "Mild" and "None") and $\beta 4$ was the fourth threshold (e.g., between the responses "Extreme" and "Severe")

[2] The higher locations estimates indicate more difficult items

[3] Benjamini-Hochberg adjusted

[4] Misfitting responses = Zh score lower than −2

[5] Overfitting response patterns = Zh score higher than 2

DIF, differential item functioning; RMSEA, root mean square error of approximation; SRMR, standardized root mean square; TLI, Tucker-Lewis index; CFI, comparative fit index

Table 5 Construct validity of KOOS-12 scales and summary. Spearman's rho correlation coefficients (95% confidence interval) among KOOS-12, disease-specific measures, and performance-based measures

	Knee injury and Osteoarthritis Outcome Score, 12-item scale (KOOS-12)			
	Pain	Function	Quality of life (QoL)	Summary
KOOS				
Pain	0.94 (0.93 to 0.96)	0.83 (0.78 to 0.86)	0.66 (0.58 to 0.73)	0.89 (0.86 to 0.91)
Symptoms	0.73 (0.66 to 0.79)	0.78 (0.72 to 0.83)	0.66 (0.58 to 0.73)	0.79 (0.73 to 0.83)
Activities of daily living (ADL)	0.85 (0.81 to 0.89)	0.94 (0.92 to 0.95)	0.73 (0.66 to 0.78)	0.93 (0.91 to 0.94)
Sport and recreational activities	0.61 (0.52 to 0.69)	0.80 (0.75 to 0.84)	0.70 (0.63 to 0.76)	0.77 (0.71 to 0.82)
QoL	0.63 (0.55 to 0.71)	0.73 (0.66 to 0.78)	1.00	0.87 (0.83 to 0.90)
IKDC	0.72 (0.65 to 0.78)	0.78 (0.72 to 0.83)	0.71 (0.63 to 0.77)	0.82 (0.77 to 0.86)
ICOAP	− 0.79 (− 0.83 to − 0.73)	− 0.73 (− 0.79 to − 0.66)	− 0.67 (− 0.73 to − 0.58)	− 0.80 (− 0.84 to − 0.74)
WHODAS 2.0				
Cognition	− 0.20 (− 0.30 to − 0.05)	− 0.32 (− 0.45 to − 0.19)	− 0.22 (− 0.35 to − 0.08)	− 0.28 (− 0.40 to – 0.14)
Mobility	− 0.60 (− 0.68 to − 0.50)	− 0.69 (− 0.76 to − 0.60)	− 0.62 (− 0.71 to − 0.53)	− 0.70 (− 0.77 to − 0.62)
ADL	− 0.41 (− 0.52 to − 0.28)	− 0.49 (− 0.59 to − 0.37)	− 0.47 (− 0.58 to − 0.35)	− 0.50 (− 0.60 to − 0.39)
Performance-based measures				
Timed Up and Go	− 0.46 (− 0.58 to − 0.33)	− 0.55 (− 0.65 to − 0.42)	− 0.46 (− 0.58 to − 0.32)	− 0.54 (− 0.65 to − 0.42)
30-s chair test	0.50 (0.38 to 0.61)	0.51 (0.38 to 0.61)	0.49 (0.36 to 0.59)	0.54 (0.43 to 0.64)

KOOS, Knee injury and Osteoarthritis Outcome Score; IKDC, International Knee Documentation Committee (IKDC) subjective knee evaluation form; ICOAP, Knee Intermittent and Constant Osteoarthritis Pain; WHODAS 2.0, World Health Organization Disability Assessment Schedule 2.0

ADL scales, ICOAP score, and IKDC score, and had no significant correlation with the Timed Up and Go test and 30-s CST (Supplementary Table S2).

Discussion

KOOS-12 is a short self-reported measure that assesses patient's opinions about the difficulties they experience due to problems with their knee and also covers aspects of pain, functional limitations, and knee-related QoL [8]. Therefore, there are currently three versions of KOOS in Spanish (adapted to Spain, Peru, and United States Spanish). Instead of doing another translation, this study evaluated the psychometric properties of KOOS-12 using the Spanish version for Peru. Some patients needed help clarifying some response options, so minor modifications were made to improve the understanding of the response options. Some patients had difficulty understanding the difference between "daily" to "always." Therefore, the clarification *Una vez al día* was added to the "daily" option. Similarly, to the response options "severe" and "very severe," a clarification was added *Severo/Fuerte* and *Muy severo/Extremo*.

The Spanish version of the KOOS-12 questionnaire shows appropriate internal consistency reliability for evaluating patients' knee problems [9]. This confirms previous findings, Cronbach's alpha was 0.75–0.82, 0.78–0.82, 0.80–0.84, and 0.90–0.93 for the KOOS-12 pain, function, QoL, and summary, respectively [2]. The test-retest of the KOOS-12 pain and function scales was moderate (ICC < 0.7). No previous study has reported the test-retest reliability of KOOS-12.

From KOOS-12 scales, the pain scale was the only one that did not present a good overall fit. Overall model misfit may be related to the reversed category boundaries, so the response option "monthly" will never be the most probable response for patients at any point on the trait scale. Low frequencies could explain reversed thresholds in the response options of item 1. The frequency of the second response option "monthly" was 7.18% and 11.96% for the third response option "weekly," which were much less than 46.89% of the fourth response option "daily." The low frequency of these categories may be due to the lower number of patients with early knee osteoarthritis. Therefore, evaluation of the pain scale could be appropriate with a higher number of patients with KOA grade 1 (Kellgren–Lawrence).

Correlation of KOOS-12 pain and KOOS pain was strong, indicating that the variance in the KOOS pain scale was enough as captured by the four items of the KOOS-12 pain scale. These results agree with the values of correlation coefficient reported in two previous studies ($r = 0.89$–0.93) [2, 7]. Results in the present trial have shown a very strong correlation ($r = 0.94$) between KOOS-12 function and KOOS ADL; this is consistent with previously reported data ($r = 0.81$–0.90) [2, 7]. We found a higher correlation between KOOS-12 function and KOOS sports and recreational activities ($r = 0.80$) compared with previous studies ($r = 0.61$–0.71) [2, 7]. The

Table 6 ANOVA results for estimated KOOS-12 scales and summary by the Kellgren–Lawrence classification and nutritional status using summated and transformed scores, and item response theory (IRT) scores

Kellgren–Lawrence classification	Grade 0 ($n = 10$)	Grade 1 ($n = 21$)	Grade 2 ($n = 70$)	Grade 3 ($n = 47$)	Grade 4 ($n = 24$)	p value	Eta squared
KOOS-12 (score 0–100)							
Pain	97.5 ± 4.37	51.48 ± 22.17	45.29 ± 19.97	41.35 ± 19.31	28.82 ± 17.53	< 0.001	0.37
Function	96.62 ± 6.62	47.91 ± 28.25	42.05 ± 22.06	34.70 ± 22.26	21.09 ± 16.77	< 0.001	0.34
Quality of life	88.12 ± 12.99	44.94 ± 24.97	32.67 ± 22.32	26.68 ± 15.86	10.41 ± 11.60	< 0.001	0.42
Summary	93.75 ± 6.80	48.11 ± 22.24	40 ± 19.56	34.24 ± 16.31	19.44 ± 12.67	< 0.001	0.44
KOOS-12 (Theta or latent trait level)							
Pain	− 1.90 ± 0.25	− 0.16 ± 0.84	0.05 ± 0.78	0.22 ± 0.75	0.79 ± 0.74	< 0.001	0.36
Function	− 1.84 ± 0.34	− 0.20 ± 1.04	− 0.02 ± 0.71	0.24 ± 0.80	0.71 ± 0.76	< 0.001	0.32
Quality of life	− 1.83 ± 0.43	− 0.35 ± 0.80	0.04 ± 0.80	0.22 ± 0.65	0.88 ± 0.52	< 0.001	0.40
Nutritional status		Non-obese		Obese		p value	Cohen's d
KOOS-12 (score 0–100)							
Pain		53.09 ± 25.28		41.33 ± 21.58		< 0.001	0.50
Function		49.81 ± 26.39		34.37 ± 25.21		< 0.001	0.59
Quality of life		40.79 ± 26.94		28.41 ± 23.73		< 0.001	0.49
Summary		47.90 ± 24.04		34.70 ± 21.50		< 0.001	0.57
KOOS-12 (Theta or latent trait level)							
Pain		− 0.22 ± 0.95		0.23 ± 0.85		< 0.001	0.51
Function		− 0.27 ± 0.88		0.26 ± 0.92		< 0.001	0.56
Quality of life		− 0.24 ± 0.91		0.22 ± 0.87		< 0.001	0.52

Mean ± standard deviation

discriminant validity of the KOOS-12 was demonstrated by the low correlation between the KOOS-12 scales and the WHODAS 2.0 cognitive and ADL scales. Previously, the low correlation of KOOS-12 with the mental health scale of the SF-36® instrument was demonstrated [2]. KOOS-12 summary shows evidence of construct validity. In this study, KOOS-12 summary is highly correlated with the KOOS, IKDC, and ICOAP. Similarly, Eckhard et al. [7] has reported a KOOS-12 summary correlation with KOOS, WOMAC®, and Oxford-12 [7].

Dobson et al. [37] have reported that sit-to-stand tests with the best measurement evidence included the TUG test and the 30-s CST for KOA. In this study, the KOOS-12 scales showed a moderate correlation with the TUG test and 30-s CST. Our results are consistent with previously reported; the TUG test is negatively and moderately correlated with all the KOOS scales ($r = -0.66$ to -0.45) and with the Lequesne index [38, 39], and presents weak correlations with the WOMAC® ($r < 0.3$) [39, 40]. To the best of our knowledge, no previous study has correlated KOOS or KOOS-12 with the 30-s CST. However, the available information suggests a low-to-moderate relationship between self-reported and performance-based measures. Besides, performance-based measures and self-reported PROMs assess different patient characteristics.

The responsiveness of KOOS-12 demonstrated moderate to large effects eight weeks after intra-articular therapy. Furthermore, KOOS-12 scales and summary scale performed as well as KOOS's pain, symptoms, and activities of daily living, and IKDC in terms of responsiveness. To the best of our knowledge, this is the first study examining the responsiveness of KOOS-12 in patients with KOA treated with intra-articular therapy. In patients with total TKR, the SRMs for KOOS-12 ranged from 1.62 to 2.12, and the quality of life scale was the most sensitive to change [2]. In contrast, in patients with intra-articular treatment, the pain and function scales were the most sensitive to change. The present study shows that KOOS-12 pain and summary were most sensitive to change than KOOS scales. In line with these results, KOOS-12 summary score reported high effect sizes, and standardized response means post-TKR [7, 8].

KOOS-12 is an easily accessible instrument for clinicians, it is freely available, easy to understand and score, besides, and the number of missing values is low. In clinical settings, KOOS-12 is a brief, comprehensive knee-specific PROM with good psychometric properties and provides an overall knee impact score, along with domain-specific measures.

Our study has some limitations that must be acknowledged. First, participants were recruited through secondary care clinics, suggesting that our sample may not be representative of KOA population. Second, the sample size was not equal in

OA severity; the number of patients was lower in patients with mild and severe KOA. Third, even though the GPCM well fitted the data, the total sample size could be considered "inadequate" for accurate parameter estimates [11]. However, evidence from recent simulation studies suggests that a sample size of 200 participants is enough to achieve a robust CFA solution. The Monte Carlo data simulation techniques showed that adequate sample size for a one-factor CFA with four items and factorial loadings of 0.65 (as each scale of the KOOS-12) could be as low as 90 patients [41]. Concerning robust weighted least squares (WLS) estimation, the relative bias in the estimated standard errors of factor loadings depended on sample size and factorial loading magnitude, with a sample size of 200 participants and five-categorical data, and the relative bias for a four-indicator model was < 5% with loadings of 0.70 [42]. Although our study's sample size is considered a "very good" sample size for estimating the parameters with the Rasch model [11], our results showed that KOOS-12 scales present a poor model fit to the Rasch-based partial credit model. Also, the same discrimination parameter of the items could not be assumed. The sample size of 209 patients might also be a limitation, as GPCM analysis generally requires ≥ 500 patients due to the number of parameter estimations needed [11]. These issues will be important to address in future research.

In conclusion, this study demonstrates that the Spanish version of KOOS-12 questionnaire is a valid, reliable, and sensitive to change instrument for measuring the patients' opinion about their knee and associated problems in Mexican subjects with knee O.A.

Author contributions Conceptualization and design: Gabriel Horta-Baas; Rodrigo Vargas-Mena; Erik Alejandre; Ingris Peláez-Ballestas; María del Socorro Romero-Figueroa; and Gloria Queipo.

Provision of study materials or patients: Gabriel Horta-Baas; Rodrigo Vargas-Mena; and Erik Alejandre.

Analysis: Gabriel Horta-Baas; Ingris Peláez-Ballestas; María del Socorro Romero-Figueroa; and Gloria Queipo.

Data interpretation: Gabriel Horta-Baas; Rodrigo Vargas-Mena; Erik Alejandre; Ingris Peláez-Ballestas; María del Socorro Romero-Figueroa; and Gloria Queipo.

Manuscript writing: Gabriel Horta-Baas.

Manuscript review final approval: Gabriel Horta-Baas; Rodrigo Vargas-Mena; Erik Alejandre; Ingris Peláez-Ballestas; María del Socorro Romero-Figueroa; and Gloria Queipo.

Data availability The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Compliance with ethical standards

Disclosures The authors declare that there is no conflict of interest regarding the publication of this paper.

Ethics approval and consent to participate The research ethics committee of the Instituto Mexicano del Seguro Social approved this study. Informed consent was obtained from all individual participants included in the study.

References

1. Goncalves RS, Cabri J, Pinheiro JP, Ferreira PL, Gil J (2010) Reliability, validity and responsiveness of the Portuguese version of the Knee injury and Osteoarthritis Outcome Score–physical function short-form (KOOS-PS). Osteoarthr Cartil 18(3):372–376
2. Gandek B, Roos EM, Franklin PD, Ware JE Jr (2019) 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). Osteoarthr Cartil 27(5):746–753
3. Tolk JJ, Janssen RPA, Prinsen CAC, Latijnhouwers D, van der Steen MC, Bierma-Zeinstra SMA et al (2019) The OARSI core set of performance-based measures for knee osteoarthritis is reliable but not valid and responsive. Knee Surg Sports Traumatol Arthrosc 27(9):2898–2909
4. Collins NJ, Misra D, Felson DT, Crossley KM, Roos EM (2011) Measures of knee function: International Knee Documentation Committee (IKDC) Subjective Knee Evaluation Form, Knee injury and Osteoarthritis Outcome Score (KOOS), Knee injury and Osteoarthritis Outcome Score Physical Function Short Form (KOOS-PS), Knee Outcome Survey Activities of Daily Living Scale (KOS-ADL), Lysholm Knee Scoring Scale, Oxford Knee Score (OKS), Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC), Activity Rating Scale (ARS), and Tegner Activity Score (TAS). Arthritis Care Res (Hoboken) 63(Suppl 11):S208–S228
5. Martin-Fernandez J, Garcia-Maroto R, Sanchez-Jimenez FJ, Bau-Gonzalez A, Valencia-Garcia H, Gutierrez-Teira B et al (2017) Validation of the Spanish version of the Oxford knee score and assessment of its utility to characterize quality of life of patients suffering from knee osteoarthritis: a multicentric study. Health Qual Life Outcomes 15(1):186
6. Roos EM, Toksvig-Larsen S (2003) Knee injury and Osteoarthritis Outcome Score (KOOS) - validation and comparison to the WOMAC in total knee replacement. Health Qual Life Outcomes 1:17
7. Eckhard L, Munir S, Wood D, Talbot S, Brighton R, Walter B, Baré J (2020) The KOOS-12 shortform shows no ceiling effect, good responsiveness and construct validity compared to standard outcome measures after total knee arthroplasty. Knee Surg Sports Traumatol Arthrosc
8. Gandek B, Roos EM, Franklin PD, Ware JE Jr (2019) A 12-item short form of the Knee injury and Osteoarthritis Outcome Score (KOOS-12): tests of reliability, validity and responsiveness. Osteoarthr Cartil 27(5):762–770
9. Altman R, Asch E, Bloch D, Bole G, Borenstein D, Brandt K, Christy W, Cooke TD, Greenwald R, Hochberg M, Howell D, Kaplan D, Koopman W, Longley S, Mankin H, McShane DJ, Medsger T, Meenan R, Mikkelsen W, Moskowitz R, Murphy W, Rothschild B, Segal M, Sokoloff L, Wolfe F (1986) Development of criteria for the classification and reporting of osteoarthritis. Classification of osteoarthritis of the knee. Diagnostic and Therapeutic Criteria Committee of the American Rheumatism Association. Arthritis Rheum 29(8):1039–1049
10. Kellgren JH, Lawrence JS (1957) Radiological assessment of osteo-arthrosis. Ann Rheum Dis 16(4):494–502
11. Mokkink LB, Prinsen CAC, Patrick DL, Alonso J, Bouter LM, de Vet HC, et al. COSMIN study design checklist for Patient.reported outcome measurement instruments 2019 [Available from: https://www.cosmin.nl/wp-content/uploads/COSMIN-study-designing-checklist_final.pdf
12. Reeve BR, Fayers P (2005) Applying item response theory modelling for evaluating questionnaire itemscale properties. In: Fayers P, Hays H (eds) Assessing quality of life in clinical trials: methods and practice, 2nd edn. Oxford University Press, New York, pp 55–73

13. Nguyen TH, Han HR, Kim MT, Chan KS (2014) An introduction to item response theory for patient-reported outcome measurement. Patient. 7(1):23–35
14. Edelen MO, Reeve BB (2007) Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. Qual Life Res 16(Suppl 1):5–18
15. Vaquero J, Longo UG, Forriol F, Martinelli N, Vethencourt R, Denaro V (2014) Reliability, validity and responsiveness of the Spanish version of the Knee Injury and Osteoarthritis Outcome Score (KOOS) in patients with chondral lesion of the knee. Knee Surg Sports Traumatol Arthrosc 22(1):104–108
16. Bond M, Davis A, Lohmander S, Hawker G (2012) Responsiveness of the OARSI-OMERACT osteoarthritis pain and function measures. Osteoarthr Cartil 20(6):541–547
17. Vazquez-Barquero JL, Vazquez Bourgon E, Herrera Castanedo S, Saiz J, Uriarte M, Morales F et al (2000) Version en lengua espanola de un nuevo cuestionario de evaluacion de discapacidades de la OMS (WHO-DAS-II): fase inicial de desarrollo y estudio piloto. Grupo Cantabria en Discapacidades. Actas Esp Psiquiatr 28(2):77–87
18. Dobson F, Hinman RS, Roos EM, Abbott JH, Stratford P, Davis AM, Buchbinder R, Snyder-Mackler L, Henrotin Y, Thumboo J, Hansen P, Bennell KL (2013) OARSI recommended performance-based tests to assess physical function in people diagnosed with hip or knee osteoarthritis. Osteoarthr Cartil 21(8):1042–1052
19. Dunn TJ, Baguley T, Brunsden V (2014) From alpha to omega: a practical solution to the pervasive problem of internal consistency estimation. Br J Psychol 105(3):399–412
20. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J et al (2007) Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 60(1):34–42
21. Nolte S, Coon C, Hudgens S, Verdam MGE (2019) Psychometric evaluation of the PROMIS(R) Depression Item Bank: an illustration of classical test theory methods. J Patient Rep Outcomes. 3(1):46
22. Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, Thissen D, Revicki DA, Weiss DJ, Hambleton RK, Liu H, Gershon R, Reise SP, Lai JS, Cella D (2007) Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Med Care 45(5 Suppl 1):S22–S31
23. Hu L, Bentler PM (1999) Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equ Model Multidiscip J 6(1):1–55
24. Browne MW, Cudeck R (1993) Alternative ways of assessing model fit. In: Bollen KA, J.S L, editors. Testing structural equation models. Newbury Park, CA: Sage. p. 136–62
25. Alageel M, Al Turki A, Alhandi A, Alohali R, Alsalem R, Aleissa S (2020) Cross-cultural adaptation and validation of the Arabic version of the Intermittent and Constant Osteoarthritis Pain Questionnaire. Sports Med Int Open 4(1):E8–E12
26. Panah SH, Baharlouie H, Rezaeian ZS, Hawker G (2016) Cross-cultural adaptation and validation of the Persian version of the Intermittent and Constant Osteoarthritis Pain Measure for the knee. Iran J Nurs Midwifery Res 21(4):417–423
27. Baidya O, Prabhakar R, Wadhwa M, Baidya P (2018) Cross-cultural translation and validation of the Hindi version of the intermittent and constant osteoarthritis pain scale in knee osteoarthritis patients. Ortho & Rheum Open Access 10:1–5
28. Chen W-H, Thissen D (1997) Local dependence indexes for item pairs using item response theory. J Educ Behav Stat 22(3):265–289
29. Reise SP, Rodriguez A (2016) Item response theory and the measurement of psychiatric constructs: some empirical and conceptual issues and challenges. Psychol Med 46(10):2025–2039
30. Stover AM, McLeod LD, Langer MM, Chen WH, Reeve BB (2019) State of the psychometric methods: patient-reported outcome measure development and refinement using item response theory. J Patient Rep Outcomes 3(1):50
31. Depaoli S, Tiemensma J, Felt JM (2018) Assessment of health surveys: fitting a multidimensional graded response model. Psychol Health Med 23(sup1):13–31
32. Paek I, Cole K (2020) Using R for item response theory model applications. Routledge, New York, USA
33. Felt JM, Castaneda R, Tiemensma J, Depaoli S (2017) Using person fit statistics to detect outliers in survey research. Front Psychol 8:863
34. Husted JA, Cook RJ, Farewell VT, Gladman DD (2000) Methods for assessing responsiveness: a critical review and recommendations. J Clin Epidemiol 53(5):459–468
35. Fritz CO, Morris PE, Richler JJ (2012) Effect size estimates: current use, calculations, and interpretation. J Exp Psychol Gen 141(1):2–18
36. Dinno A (2015) Nonparametric pairwise multiple comparisons in independent groups using Dunn's test. Stata J 15(1):292–300
37. Dobson F, Hinman RS, Hall M, Terwee CB, Roos EM, Bennell KL (2012) Measurement properties of performance-based measures to assess physical function in hip and knee osteoarthritis: a systematic review. Osteoarthr Cartil 20(12):1548–1562
38. Marconcin P, Espanha M, Yázigi F, Teles J (2015) Predictor of timed "Up-and-Go" test in elderly with knee osteoarthritis. In: Cabri J, Pezarat-Correia P, editors. Proceedings of the 3rd International Congress on Sport Sciences Research and Technology Support. 1. Lisbon, Portugal: Science and Technology Publications, Lda
39. de Rezende MU, de Farias FES, da Silva CAC, Cernigoy CHA, de Camargo OP (2016) Objective functional results in patients with knee osteoarthritis submitted to a 2-day educational programme: a prospective randomised clinical trial. BMJ Open Sport Exerc Med 2(1):e000200
40. Gandhi R, Tsvetkov D, Davey JR, Syed KA, Mahomed NN (2009) Relationship between self-reported and performance-based tests in a hip and knee joint replacement population. Clin Rheumatol 28(3):253–257
41. Wolf EJ, Harrington KM, Clark SL, Miller MW (2013) Sample size requirements for structural equation models: an evaluation of power, bias, and solution propriety. Educ Psychol Meas 76(6):913–934
42. Moshagen M, Musch J (2014) Sample size requirements of the robust weighted least squares estimator. Methodology. 10(2):60–70