EDITOR'S CHOICE

The Copenhagen Hip and Groin Outcome Score (HAGOS): development and validation according to the COSMIN checklist

K Thorborg,[1] P Hölmich,[1] R Christensen,[2,3] J Petersen,[1] E M Roos[2]

► Additional appendices are published online. To view these files please visit the journal online (http://bjsm.bmj.com)

[1]Arthroscopic Centre Amager, Amager Hospital, University of Copenhagen, Copenhagen, Denmark
[2]Research Unit for Musculoskeletal Function and Physiotherapy, Institute of Sports Science and Clinical Biomechanics, University of Southern Denmark, Odense, Denmark
[3]The Parker Institute: Musculoskeletal Statistics Unit, Copenhagen University Hospital, Frederiksberg, Copenhagen, Denmark

Correspondence to
Kristian Thorborg, Faculty of Health Sciences, Department of Orthopaedic Surgery, University of Copenhagen, DK-2300 Copenhagen S, Denmark;
kristianthorborg@hotmail.com

Accepted 2 March 2011

ABSTRACT

Background Valid, reliable and responsive Patient-Reported Outcome (PRO) questionnaires for young to middle-aged, physically active individuals with hip and groin pain are lacking.

Objective To develop and validate a new PRO in accordance with the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) recommendations for use in young to middle-aged, physically active patients with long-standing hip and/or groin pain.

Methods Preliminary patient interviews (content validity) included 25 patients. Validity, reliability and responsiveness were evaluated in a clinical study including 101 physically active patients (50 women); mean age 36 years, range 18–63 years.

Results The Copenhagen Hip and Groin Outcome Score (HAGOS) consists of six separate subscales assessing Pain, Symptoms, Physical function in daily living, Physical function in Sport and Recreation, Participation in Physical Activities and hip and/or groin-related Quality of Life (QOL). Test–retest reliability was substantial, with intraclass correlation coefficients ranging from 0.82 to 0.91 for the six subscales. The smallest detectable change ranged from 17.7 to 33.8 points at the individual level and from 2.7 to 5.2 points at the group level for the different subscales. Construct validity and responsiveness were confirmed with statistically significant correlation coefficients (0.37–0.73, $p < 0.01$) for convergent construct validity and for responsiveness from 0.56 to 0.69, $p < 0.01$.

Conclusion HAGOS has adequate measurement qualities for the assessment of symptoms, activity limitations, participation restrictions and QOL in physically active, young to middle-aged patients with long-standing hip and/or groin pain and is recommended for use in interventions where the patient's perspective and health-related QOL are of primary interest.

Trial registration ClinicalTrials.gov NCT00716729

INTRODUCTION

Pain in the hip and groin region is a common musculoskeletal complaint in the young to middle-aged population[1] affecting physical function and health-related quality of life (QOL).[2] Furthermore, hip and groin pain can be a long-standing condition, being difficult to fully recover from.[3 4] Musculoskeletal disorders such as long-standing hip and groin complaints, therefore, have a large impact on healthcare expenditure, sick leave and work disability,[5] resulting in substantial social and economic costs.[6]

Novel treatment methods, such as hip arthroscopy, incipient groin hernia repair, ultrasound-guided corticosteroid injections and specific exercise regimens, are advancing rapidly in the management of young and middle-aged physically active patients with hip and groin pain.[7–15] There is a general consensus that Patient-Reported Outcomes (PROs) should serve as the gold standard in the assessment of musculoskeletal conditions, where the patient's perspective and health-related QOL are of primary interest.[16–19] However, valid, reliable and responsive PRO questionnaires for physically active patients with long-standing hip and/or groin pain are lacking.[20] The need for reliable and valid instruments is emphasised in a study by Marshall *et al*,[21] who demonstrated that clinical trials using unpublished measurement instruments were more likely to report positive effects of treatment than clinical trials using published instruments. Therefore, in order to properly evaluate the large spectrum of treatment strategies and regimens for young to middle-aged physically active patients with hip and groin pain, a valid, reliable and responsive PRO questionnaire is needed.[20]

In a recent international consensus process, including leading experts in the fields of psychology, epidemiology, statistics and clinical medicine from all over the world, a consensus on the taxonomy, terminology and definitions of measurement properties for health-related PROs was reached[22] and formulated in a COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) checklist.[23]

The objective of this study was to develop and validate a new PRO questionnaire aimed at young to middle-aged physically active people with long-standing hip and/or groin pain by following the COSMIN recommendations on terminology and research agenda.[22 23]

METHODS

Development of the questionnaire

The methodological framework for developing and evaluating a PRO questionnaire included the following steps: (1) identification of a specific patient population, (2) item generation, (3) item reduction and (4) determination of the validity, reliability and responsiveness. Steps 1 and 2 involved developing a preliminary version of the questionnaire, which is described in the Methods section. Step 3 involved testing the individual items and subscales of the preliminary version by analysing

patient responses. Based upon these analyses, a final version of the questionnaire was decided upon. Step 4 involved testing the final version of the questionnaire for validity, reliability and responsiveness. Steps 3 and 4 are described in the Results section. A flowchart of the complete study process is shown in figure 1.

Population identification

The goal of this instrument is to evaluate hip and/or groin disability related to impairment (body structure and function), activity (activity limitations) and participation (participation restrictions) according to the International Classification of Functioning, disability and health (ICF),[24] in young to middle-aged physically active patients with hip and/or groin pain. Disability in this study encompasses the health dimensions within the methodological framework of ICF as categorised in one of three levels: impairment (body structure and function), activity limitations (activities) and participation restrictions (participation).[24] The objective would be to achieve a quantitative measure of the patient's hip and groin disability according to the different levels of the ICF. The measure should reflect the patient's perception of his/her disability as well as his/her actual disability. Physically active patients refer to any patient who is physically active at least 2.5 h a week.[25]

The groin is anatomically located in the anterior-medial part of the hip region, and the hip and groin region share vascular and neural supply.[26] The pathologies of the hip joint and the groin often present simultaneously and the symptoms can be overlapping.[27–30] This makes the hip and groin a complex anatomical region where validated diagnostic tools for differentiation of musculoskeletal diagnoses are lacking.[31–34] We, therefore, chose not to restrict our measurement instrument to be evaluated in a patient group with a specific diagnosis, but

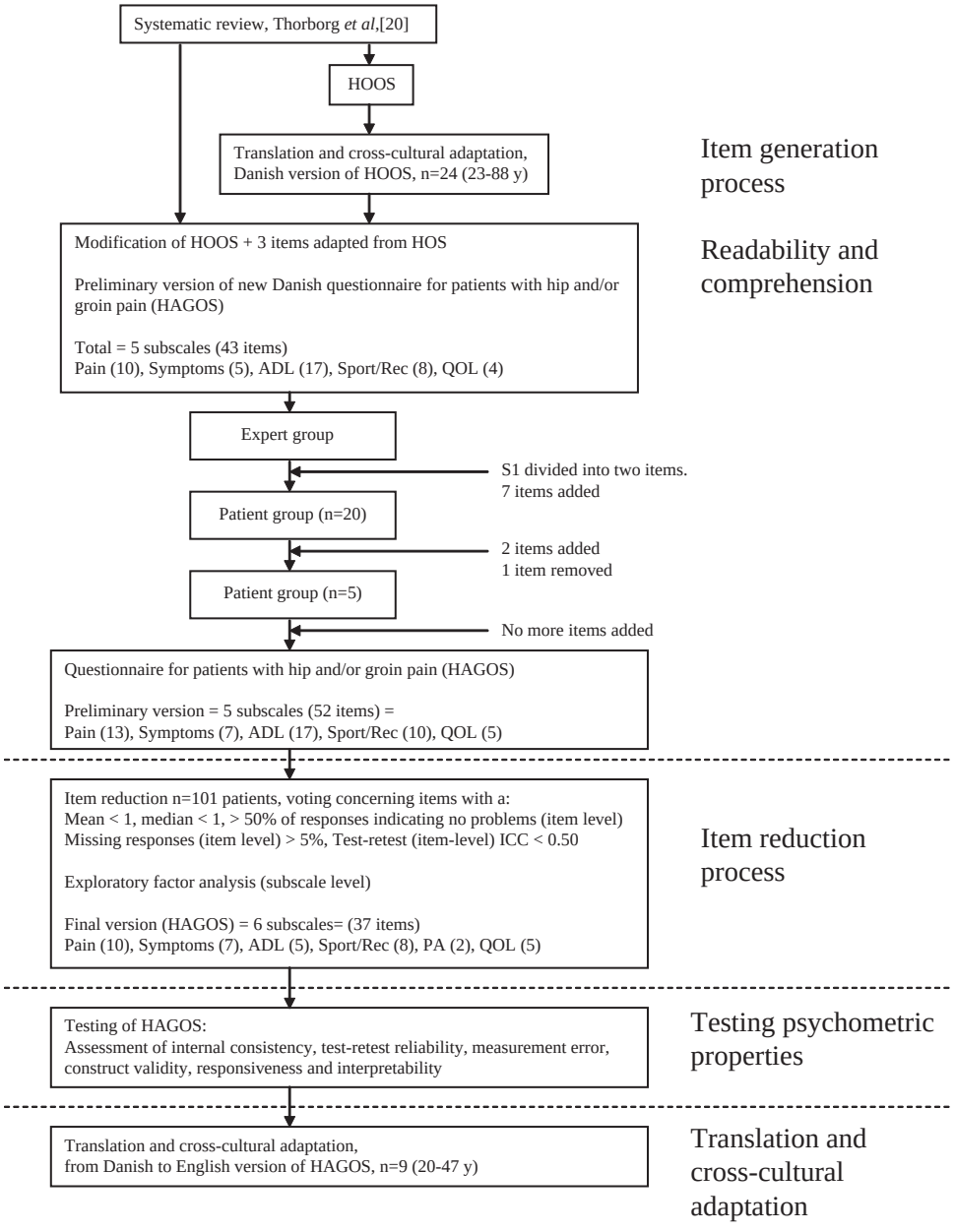


Figure 1 Flowchart of the study process.

instead we wanted to focus on the commonalities of hip and/groin pain in physically active patients.

The patient flow is presented in figure 2. Patients with hip and/or groin pain, from primary and secondary care, who were at least 18 years of age, were recruited from January 2009 to February 2010. Patients were screened by a specialist (orthopaedic surgeon or sports physiotherapist) within the area of musculoskeletal examination of hip and/or groin pain in younger physically active patients. If the specialist suspected that hip and/groin pain was not of musculoskeletal origin, the patient was referred for further investigation and was not invited to participate in the study. All other patients presenting with hip and/or groin pain were considered eligible for the study and were invited to participate. These patients were informed about the purpose of the research by the people responsible for the study, and written consent was obtained from those who agreed to participate. A self-reported questionnaire was used to screen for inclusion and exclusion of the patients who agreed to participate in the study. Patients seeking medical care presenting with hip and/or groin pain were included if they fulfilled all the following criteria: (1) had received treatment for their hip and/or groin pain, (2) were restricted in their activities due to hip and/or groin pain, (3) had hip and/or groin pain in the previous 14 days, (4) had hip and/or groin pain of more than 6 weeks' duration, (5) had hip and/or groin pain located in one of five predefined regions in a pain drawing (region 3, 6, 7, 8 or 9, figure 3) and (6) were physically active for at least 2.5 h per week. Patients with self-reported limiting comorbidities[35] were excluded from the study. The pain drawing (figure 3) was adapted from methods for determining location of pain used in previous studies,[36 37] and pain of more than 6 weeks' duration has previously been defined as long-standing in nature concerning the population under study.[9]

Item generation

The item generation phase included the following steps: a systematic review of the literature,[20] a focus group involving experts and individual patient interviews. The systematic review identified existing PROs that showed adequate measurement qualities or promise concerning validity, reliability and responsiveness when assessing patients with hip and/

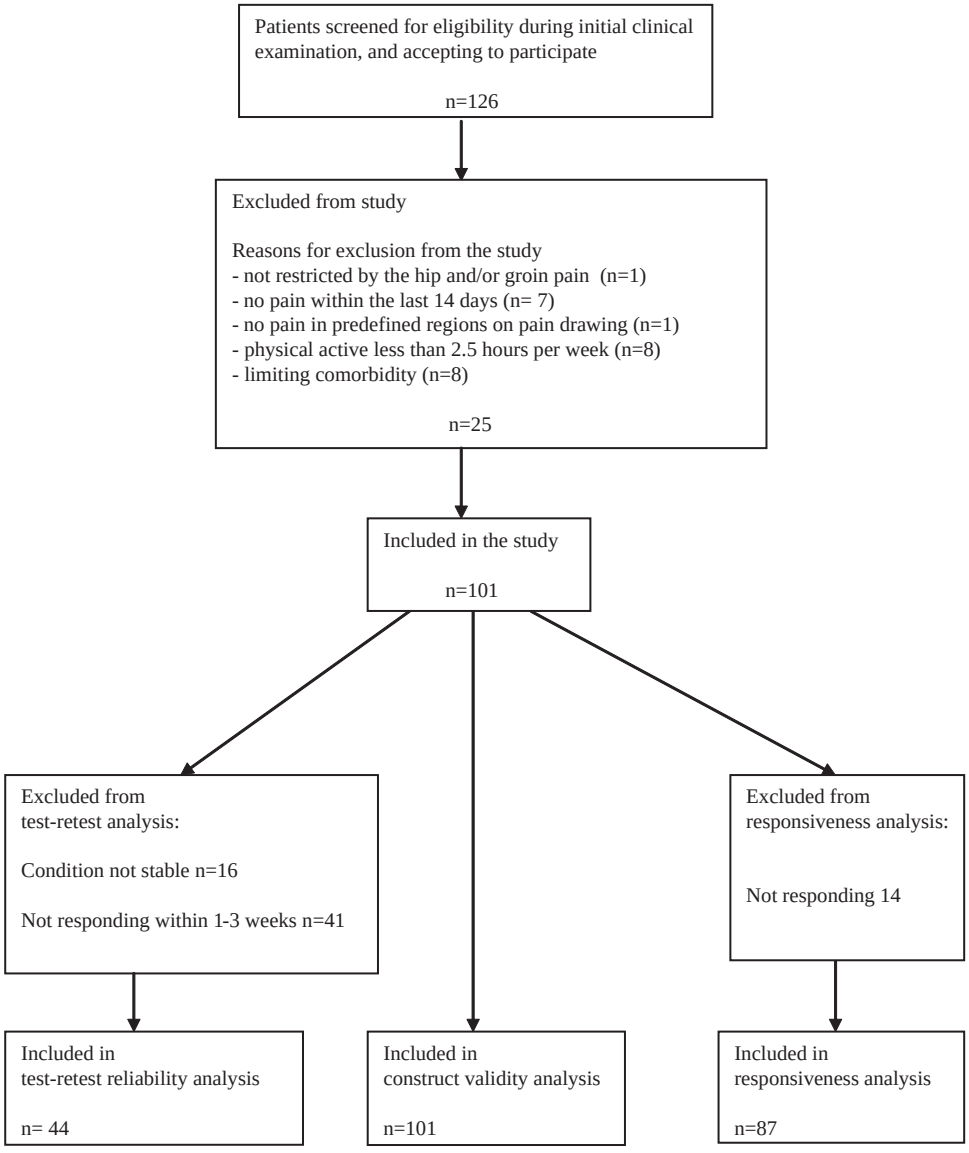


Figure 2 Clinical study profile.

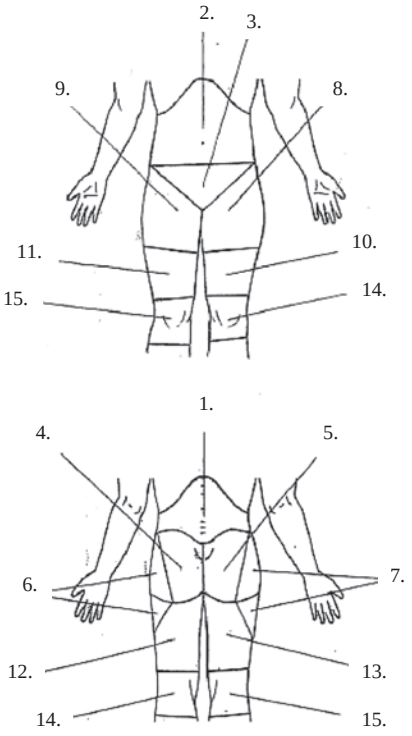
or groin disability.[20] The Hip disability and Osteoarthritis Outcome Score (HOOS) and the Hip Outcome Score (HOS) were found to be promising tools for patients with hip and/or groin disability; however, the HOOS questionnaire had only been validated in patients with hip osteoarthritis or following total hip replacement, and the HOS in patients following hip arthroscopy. Therefore, the items were not necessarily addressing our target group of young to middle-aged physically active patients with hip and/or groin pain.[20]

The HOOS was chosen as a template for the development of a new PRO questionnaire because HOOS consists of items and subscales related to body structure and function, activity and participation according to the ICF classification. It shows excellent measurement qualities in patients with hip disability for all dimensions. HOOS consists of five subscales: Pain, Symptoms, Function in daily living (ADL), Sport and Recreation function (Sport/Rec) and hip-related QOL.[38] Furthermore, HOOS includes a format that is user friendly, self-explanatory and is already adopted in hip rehabilitation research worldwide.[20] We, therefore, decided to translate and cross-culturally adapt the HOOS from the original Swedish version to a Danish version according to existing guidelines[39 40] in a process that included 24 patients with hip disability.[41] We then incorporated and adapted three items that seemed relevant from the HOS – Sports subscale that were not present in HOOS.[42–44] The items from the HOS were named SP7, SP9 and SP10 (table 2).

Groin problems are common in physically active people and HOOS and HOS address dimensions, such as sport, that are relevant to young to middle-aged physically active people.[20] However, HOOS and HOS do not include groin-related questions, only questions related to the hip. This is problematic because young to middle-aged physically active patients often report groin symptoms[27 28 30] and often do not describe their symptoms as being located in the hip.[20] All questions in the new outcome questionnaire were therefore rephrased so that they referred to the term 'hip and/or groin', instead of the term 'hip' alone, to improve the face validity of the questionnaire. We found this appropriate based on the existing data that have shown that patients with hip and groin pathology often report symptoms that do not seem to be restricted to one of these anatomical regions,[27 28 30] recognising that these regions have never been precisely defined anatomically, and therefore merely reflect individual and cultural beliefs.[37] By using the term 'hip and/or groin', we believe that the questionnaire covers a body region that also refers to the frontal and medial part of the hip region (the groin) that patients often refer to as a separate region.[37] The new questionnaire was therefore named the Copenhagen Hip and Groin Outcome Score, abbreviated to HAGOS (appendices 1 and 2).

Expert focus group

The second step involved interviewing experts in the field. Three doctors (two orthopaedic surgeons and one physician) and four physiotherapists (four sports physiotherapists, one also being a musculoskeletal physiotherapist) with extensive experience and special expertise in treating physically active patients with hip and/or groin pain were interviewed. The experts underwent a semi-structured interview in which they were asked to fill out the preliminary version while commenting on issues related to questions they felt were missing, the questionnaire's readability and its ease of comprehension. The purpose of the interview was to identify relevant items that were missing and to improve the readability and comprehension of the questionnaire.



1.	Back region	(16%)
2.	Upper abdominal region	(0%)
3.	Lower abdominal region	(11%)
4.	Left buttock	(9%)
5.	Right buttock	(11%)
6.	Left hip/buttock	(27%)
7.	Right hip/buttock	(18%)
8.	Left hip/groin/thigh	(52%)
9.	Right hip/groin/thigh	(57%)
10.	Left anterior thigh	(5%)
11.	Right anterior thigh	(7%)
12.	Left posterior thigh	(4%)
13.	Right posterior thigh	(4%)
14.	Left knee	(11%)
15.	Right knee	(7%)

Figure 3 Pain drawing showing percentages of included patients (n = 101) indicating pain in 15 predefined regions at baseline.

The experts commented that the introductory information on the questionnaire, where patients were asked to report disability related to the previous week, was problematic. The experts stated that many patients with hip and groin disability have had the problem for a long time and due to their disability, may not have performed these activities at all during the previous week, and therefore would not be able to answer this question in a valid way. It was therefore decided to add the following introductory information: *If an item does not pertain to you or you have not experienced it in the past week please make your 'best guess' as to which response would be the most accurate*. This solution has previously been used in the format of The Western Ontario Rotator Cuff Index and the Western Ontario Instability Score.[45 46] Because the current outcome questionnaire is not only a measure of actual disability but also perceived disability, we found this solution appropriate. Based upon the focus group involving the experts, item S1 from the original HOOS[38] was divided into S1 and S2 as discomfort and clicking were considered to be different symptomatic aspects. Furthermore, six items, named P12, P13, SP5, SP6, Q4 and Q5, were added after suggestions by the experts (table 2).

Patient interviews

The final step in the item generation process was to interview patients with hip and/or groin disability individually. Individual patients were specifically chosen for an interview so that there would be representation of sex, age, type of injury, time from initial injury and severity of symptoms. The preliminary questionnaire was piloted on patients until data saturation was achieved. The patients underwent a semi-structured interview in which they were asked to fill out the preliminary version while commenting on issues related to questions they felt were missing, the questionnaire readability and its ease of comprehension. This process included 25 patients, 12 men and 13 women (34 ± 11 years) recruited from the Artroscopic Centre Amager, Amager Hospital. Twenty patients were interviewed individually before data saturation was achieved and two items were added, P2 and SP8 (table 2). Furthermore, several patients mentioned that they did not understand the meaning of Q3 from the original HOOS: *How much are you troubled with lack of confidence in your hip?*[38] Even though the main purpose of this process was not to omit items, we decided that the item had to be removed because too many patients did not understand the meaning of the question. This new preliminary version was piloted on five patients and did not require further modification. The preliminary questionnaire consisted, after item generation, of 52 items in five subscales (Symptoms (7), Pain (13), ADL (17), Sport/Rec (10) and QOL (5)).

Methodological testing and evaluation of measurement qualities of the new patient-reported questionnaire using the COSMIN checklist

Internal consistency

Internal consistency is the degree of interrelatedness among the items.[47] A principal component factor analysis was performed on the individual subscales to assess their structural validity. Failure to load on a single major factor suggests that the items do not all measure the same construct. Cronbach's α was calculated per subscale and a score above 0.70 was taken as an indication of sufficient homogeneity of the items in the subscale.[48 49]

Test–retest reliability

Test–retest reliability is the extent to which scores for the same patients are unchanged for repeated measurements over time.[47] Intraclass correlation coefficients (ICCs) were reported and test–retest ICC should be ≥0.70 for all subscales.[48 49] Test–retest reliability was evaluated after 1–3 weeks in 44

Table 1 Baseline characteristics

	Total (n = 101)	Men (n = 51)	Women (n = 50)
Age, years (mean (SD), range)	36 (11), 18–63	33 (8), 18–53	39 (12), 18–63
Weight, kg (mean (SD), range)	74 (13), 32–104	81 (10), 62–104	67 (12), 32–96
Height, cm (mean (SD), range)	176 (9), 159–198	182 (7), 166–198	169 (5), 159–180
Pain duration			
>6 Weeks	1 (1%)	1 (2%)	0 (0%)
>12 Weeks	11 (11%)	9 (18%)	2 (4%)
>6 Months	14 (14%)	8 (16%)	6 (12%)
>12 Months	75 (74%)	33 (65%)	42 (84%)
Pain medication use			
None	80 (80%)	47 (92%)	33 (66%)
Paracetamol/NSAIDs	18 (18%)	4 (2%)	14 (28%)
Opioids	3 (3%)	0 (0%)	3 (6%)
Physical activity			
≥2.5 h/Week	27 (27%)	11 (22%),	16 (32%)
≥5 h/Week	40 (40%)	22 (43%)	18 (36%)
≥10 h/Week	34 (34%)	18 (35%)	16 (32%)
BMI, kg/m^2 (mean (SD), range)	23.78 (2.97), 17–31.05	24.51 (2.13), 20–31.05	23.4 (3.49), 17.7–31
Primary physical activity form			
Cycling	26 (26%)	8 (16%)	18 (36%)
Soccer	18 (18%)	18 (35%)	0 (0%)
Running	15 (15%)	10 (20%)	5 (10%)
Strength training/fitness	13 (13%)	8 (16%)	5 (10%)
Other(s)	29 (29%)	8 (16%)	22 (44%)

BMI, body mass index; m, mean; n, number of patients; NSAIDs, non-steroidal anti-inflammatory drugs; %, percentage of patients.

Table 2 Preliminary items and subscales in HAGOS

HAGOS (preliminary version)	Missing responses	Frequecies (%) answer option: 0, 1, 2, 3 and 4. 0 indicates no problem and 4 extreme problem	Median	Mean	Intraclass correlation coefficient (ICC)
Symptoms					
S1. Do you feel discomfort in your hip and/or groin?	0	2, 2, 29.7, 36.6, 29.7	3	2.90	0.71
S2. Do you hear clicking or any other type of noise from your hip and/or groin?	0	29.7, 18.8, 23.8, 24.8, 3	2	1.52	0.80
S3. Do you have difficulties stretching your legs far out to the side?	1	19, 18, 29, 27, 7	2	1.85	0.69
S4. Do you have difficulties taking full strides when you walk?	0	40.6, 29.7, 16.8, 7.9, 5	1	1.07	0.76
S5. Do you experience sudden twinging/stabbing sensations in your hip and/or groin?	0	18.8, 13.9, 32.7, 30.7, 4	2	1.87	0.74
S6. How severe is your hip and/or groin stiffness after first awakening in the morning?	0	21.8, 30.7, 32.7, 12.9, 2	1	1.43	0.78
S7. How severe is your hip and/or groin stiffness after sitting, lying or resting later in the day?	0	23.8, 26.7, 34.7, 13.9, 1	1	1.42	0.72
Pain					
P1. How often is your hip and/or groin painful?	1	1, 2, 27, 57, 13	3	2.79	0.56
P2. How often do you have pain in areas other than your hip and/or groin that you think may be related to your hip and/or groin problem?	0	31.7, 10.9, 25.7, 26.7, 5	2	1.62	0.73
P3. Straightening your hip fully	2	34.3, 31.3, 26.3, 7.1, 1	1	1.09	0.66
P4. Bending your hip fully	0	20.8, 29.7, 20.8, 23.8, 5	1	1.62	0.68
P5. Walking on a flat surface	0	42.6, 31.7, 15.8, 8.9, 1	1	0.94	0.73
P6. Walking up or down stairs	0	20.8, 34.7, 28.7, 11.9, 4	1	1.44	0.78
P7. At night while in bed (pain that disturbs your sleep)	0	48.5, 20.8, 15.8, 1.9, 3	1	1.00	0.88
P8. Sitting or lying	0	24.8, 36.6, 26.7, 7.9, 4	1	1.30	0.81
P9. Standing upright	0	28.7, 39.6, 23.8, 6.9, 1	1	1.19	0.74
P10. Walking on a hard surface (asphalt, concrete, etc)	1	34, 33, 21, 11, 1	1	1.11	0.74
P11. Walking on an uneven surface	1	25, 40, 17, 16, 2	1	1.30	0.70
P12. Coughing and/or sneezing	0	68.3, 14.9, 8.9, 7.9, 0	1	0.56	0.58
P13. Squeezing your legs together	0	27.7, 34.7, 20.8, 12.9, 4	1	1.31	0.79
Activities of Daily Living (ADL)					
A1. Walking down stairs	0	31.7, 39.6, 14.9, 6.9, 6.9	1	1.18	0.79
A2. Walking up stairs	0	44.6, 34.7, 11.9, 5, 4	1	0.89	0.79
A3. Rising from sitting	0	31.7, 32.7, 22.8, 8.9, 4	0	0.76	0.77
A4. Standing	0	52.5, 24.8, 16.8, 5.9, 0	1	0.78	0.78
A5. Bending down, eg, to pick something up from the floor	0	33.7, 32.7, 18.8, 12.9, 2	1	1.17	0.81
A6. Walking on a flat surface	0	47.5, 36.6, 8.9, 4, 3	1	0.78	0.78
A7. Getting in/out of car	0	25.7, 34.7, 26.7, 7.9, 5	1	1.32	0.73
A8. Going shopping	1	54, 25, 12, 8, 1	0	0.77	0.78
A9. Putting on socks/stockings	0	44.6, 32.7, 10.9, 8.9, 3	1	0.93	0.66
A10. Rising from bed	0	42.6, 31.7, 20.8, 3, 2	1	0.90	0.60
A11. Taking off socks/stockings	0	53.5, 26.7, 10.9, 6.9, 2	0	0.77	0.69
A12. Lying in bed (turning over or maintaining the same hip position for a long time)	0	28.7, 28.7, 24.8, 14.9, 3	1	1.35	0.82
A13. Getting in/out of bath	0	67.3, 20.8, 8.9, 1, 2	0	0.50	0.70
A14. Sitting	0	41.6, 32.7, 18.8, 5, 2	1	0.93	0.69
A15. Getting on/off toilet	0	55.4, 26.7, 9.9, 7.9, 0	0	0.70	0.72
A16. Heavy domestic duties (scrubbing floors, vacuuming, moving heavy boxes, etc)	0	30.7, 23.8, 31.7, 5, 8.9	1	1.38	0.80
A17. Light domestic duties (cooking, dusting, etc)	0	58.4, 21.8, 15.8, 2, 2	0	0.67	0.66
Sport/Recreation					
SP1. Squatting	0	22.8, 27.7, 23.8, 19.8, 5.9	1	1.58	0.79

Continued

Table 2 Continued

HAGOS (preliminary version)	Missing responses	Frequecies (%) answer option: 0, 1, 2, 3 and 4. 0 indicates no problem and 4 extreme problem	Median	Mean	Intraclass correlation coefficient (ICC)
SP2. Running	0	8.9, 12.9, 26.7, 24.8, 26.7	3	2.47	0.86
SP3. Twisting/pivoting on a weight bearing leg	0	21.8, 17.8, 28.7, 17.8, 13.9	2	1.84	0.77
SP4. Walking on an uneven surface	0	34.7, 30.7, 21.8, 8.9, 4	1	1.17	0.83
SP5. Running as fast as you can	1	9, 9, 18, 18, 46	3	2.83	0.88
SP6. Bringing the leg forcefully forward and/or out to the side, such as in kicking, skating, etc	0	10.9, 16.8, 18.8, 24.8, 28.7	3	2.44	0.79
SP7. Sudden explosive movements that involve quick footwork, such as accelerations, decelerations, change of directions, etc	0	7.9, 12.9, 18.8, 29.7, 30.7	3	2.62	0.83
SP8. Situations where the leg is stretched into an outer position (such as when the leg is placed as far away from the body as possible)	0	9.9, 14.9, 20.8, 24.8, 29.7	3	2.50	0.78
SP9. Are you able to participate in your preferred physical activities for as long as you would like?	1	5, 14, 17, 18, 46	3	2.86	0.81
SP10. Are you able to participate in your preferred physical activities at your normal performance level?	1	4, 8, 17, 18, 53	4	3.08	0.74
Hip and/groin-related Quality of Life (QOL)					
Q1. How often are you aware of your hip and/or groin problem?	0	0, 0, 11.9, 67.3, 20.8	3	3.09	0.61
Q2. Have you modified your life style to avoid activities potentially damaging to your hip and/or groin?	0	5.9, 19.8, 22.8, 40.6, 10.9	3	2.31	0.67
Q3. In general, how much difficulty do you have with your hip and/or groin?	0	1, 5.9, 36.6, 40.6, 15.8	3	2.64	0.73
Q4. Does your hip and/or groin problem affect your mood in a negative way?	0	3, 14.9, 36.6, 42.6, 3	3	2.28	0.49
Q5. Do you feel restricted due to your hip and/or groin problem?	0	0, 4, 21.8, 45.5, 28.7	3	2.99	0.87

stable patients. This time interval between test and retest was chosen because we believe it is long enough to prevent recall of previous answers, though short enough to assume that the condition in most cases will not change.[49] Patients reported at the retest whether their hip and/or groin pain was 'better', 'not changed' or 'worse' since the initial test. Patients reporting scores as 'unchanged' were considered stable and included in test–retest reliability analysis.[22 23]

Measurement error

Measurement error is the systematic and random error of a patient's score that is not attributed to true changes in the construct to be measured.[47] The smallest detectable change (SDC), which is the threshold for determining clinical changes beyond measurement error, was calculated on the basis of the SEM of the test–retest reliability.[49 50]

Construct validity

Construct validity is the degree to which the scores of a PRO instrument are consistent with a priori hypotheses, based on the assumption that the PRO instrument validly measures the construct to be measured.[47] Construct validity was studied by correlating the subscale scores of the HAGOS with the subscales of the Short Form-36 items (SF-36). SF-36 (Acute version, 1.1, Health Assessment Lab, Hillerød, Denmark, 1993) was used because it is a PRO measure that contains relevant domains for assessing physically active patients with reduced physical function and pain.[51–53] SF-36 is a generic measure of health status comprising eight subscales: Physical Functioning (PF), Role-Physical (RP), Bodily Pain (BP), General Health (GH), Vitality (VT), Social Functioning (SF), Role-Emotional (RE) and Mental Health (MH). The SF-36 is a valid and reliable instrument also when used in the Danish population.[54–56] Convergent and divergent evidence was examined by assessment of the associations between the HAGOS and SF-36 by the use of Spearman correlation. This construct validity was determined by cross-sectional comparison of the questionnaires when first administered.

A priori hypotheses were formulated.[22 23] We expected the highest correlations when comparing the scales that are supposed to measure similar constructs. Since the HAGOS is designed to measure physical health in patients with hip and/or groin pain rather than mental health, we expected to observe generally higher correlations between the HAGOS subscales and the SF-36 subscales of PF, RP and BP (convergent construct validity) than between the HAGOS subscales and the SF-36 subscales of MH, VT, RE, SF and GH (divergent construct validity).

Furthermore, we hypothesised that the correlation between the HAGOS subscales ADL and Sport/Rec and the SF-36 subscale PF was at least 0.5, and higher than for the other HAGOS subscales. The correlation between the SF-36 subscale Pain and HAGOS subscales Pain and Symptoms should be at least 0.5 and 0.4, respectively, and higher than for the other HAGOS subscales. At last, for the subscale QOL, which hypothetically relates to both physical and mental health, we expected a correlation of at least 0.4 to the SF-36 subscale MH.

Responsiveness

Responsiveness is defined as the ability of a PRO instrument to detect change over time in the construct to be measured.[47] For evaluating responsiveness, a Global Perceived Effect (GPE) score, where the patients rate their condition in one of seven categories was used. At a 4-month administration (follow-up), patients were asked to rate possible change in their condition

since the initial administration (baseline) in relation to their hip and/or groin pain. A 4-month follow-up was chosen since this was a reasonably long timeframe to expect clinical improvement to occur in patients with long-standing hip and/or groin pain,[57] though still short enough to assume that patients would be able to recall whether any changes in their condition had occurred during this period. The GPE had the following answer options: much better (3), better (2), somewhat better (1), no change (0), somewhat worse (–1), worse (–2) and much worse (–3). A priori hypotheses were formulated for responsiveness.[22 23] We hypothesised that the change in scores of the six subscales of the HAGOS between the initial administration and the 4-month administration would correlate with the GPE score, and that the correlation was at least 0.4 for all subscales. Furthermore, standardised response mean (SRM) and effect size (ES) should be higher for patients who reported their condition to be better or much better, than patients reporting no change, only somewhat better or worse on the GPE score. SRM and ES should also be lower for patients reporting worse or much worse than patients reporting no change or only somewhat better or worse on the GPE score.

Interpretability

Interpretability is the degree to which one can assign qualitative meaning to an instrument's quantitative scores or change in scores.[47] Interpretability includes the distribution of total scores and change scores in the study sample and in relevant subgroups, floor and ceiling effects, estimates of minimal important change (MIC) and/or minimal important difference (MID).[58] Floor and ceiling effects are present if the questionnaire fails to demonstrate a worse score in the patients demonstrating signs of clinical deterioration and an improved score in patients who show clinical improvement as this can be an indication that a scale is not sufficiently comprehensive. In this study, floor and ceiling effects were defined to be present if more than 15% of the patients were reporting worst (0) or best (100) possible score.[49 59]

Statistical analyses

A sample size ≥100 patients and 7 times the number of items in the scale has been recommended for factor analysis.[49] Unidimensionality of the different subscales was assessed by exploratory factor analysis using principal component analysis with varimax rotation in SPSS statistics (version 17.0).[60] Median values were imputed in situations where missing values existed. Eigenvalues and factor loading patterns were used to identify and extract factors.[61] Items with the lowest factor loading were sequentially deleted until only one eigenvalue above 1 was produced. The relative test–retest reliability has been calculated based on a linear mixed model (with participants handled as random effects). To estimate the test–retest reliability of the HAGOS subscales, ICCs (3.1, two-way mixed effects model absolute agreement) with 95% CIs were calculated.[61]

Measurement error was expressed as the SEM, which was calculated as SD × $\sqrt{1}$ – ICC, where SD is the standard deviation of all scores from the participants.[61 62] The SEM was used for calculating the SDC at the individual level, calculated as SEM × 1.96 × $\sqrt{2}$, and at the group level calculated as SEM × 1.96 × $\sqrt{2}$ / $\sqrt{n}$.[63 64] Internal consistency, or interitem correlation, was assessed by calculation of Cronbach's α of the baseline values.[61] A 95% CI for the SDC was calculated using the upper and lower confidence limits of the ICC used to derive the SEM.

Convergent and divergent validity of the HAGOS and the SF-36 were investigated by Spearman's correlation coefficient. Likewise, associations on responsiveness were then measured by correlating the GPE with the change scores of each HAGOS subscale at the 4-month assessment, using Spearman's correlation coefficients. Correlations of 0.5 are considered large, 0.3 is moderate and 0.1 is small.[65] Furthermore, to evaluate the responsiveness of the HAGOS, two distribution-based statistics were evaluated concerning different groups of GPE: (1) the SRM, calculated as the mean change in score divided by the SD of the change and (2) the ES, equal to the mean change in score divided by the SD of the baseline score.[61] Both SRM and ES are calculated at the 4-month assessment, compared with baseline.

RESULTS

Prospective clinical study

A prospective clinical study was designed to assess validity, reliability and responsiveness. The study was conducted at the Arthroscopic Centre Amager, Amager Hospital, Copenhagen. The Danish ethics committee of the capital region, and the Danish Data Protection Agency approved the study. Patients were recruited from primary and secondary care. One hundred and twenty-six patients were screened for eligibility during a clinical consultation by a specialist (an orthopaedic surgeon or a sports physiotherapist). One hundred and one patients were included in the study and they completed the HAGOS and SF-36 questionnaires at the initial consultation. Patients were sent the HAGOS after 1 week and asked to complete the questionnaire a second time and return it by mail as soon as possible. At the 4-month follow-up, the HAGOS and the GPE scores were sent by mail, and completed at home. At the 4-month follow-up, patients who did not respond within 3 weeks received one reminder via email or telephone. Eighty-seven patients (87%) responded at the 4-month follow-up (figure 2).

The clinical study included 50 women and 51 men, mean age 36 years, range 18–63 years. Patient characteristics including age, height, weight, body mass index, physical activity level, pain duration and pain medication use are shown in table 1. Localisation of pain according to body region was reported by all patients and the results are shown in figure 3.

Content validity

Item reduction

Based upon the first and second administration of the preliminary HAGOS version (table 2), item reduction was performed using the following strategy, which incorporated both quantitative and qualitative components. Individual items at the first administration (baseline) that had a median score of <1, and/or a mean score of <1, and/or where more than 50% of the respondents reported no problems, and/or more than 5% of patients had a missing response to an item, and/or a test–retest reliability (ICC 3.1, agreement) coefficient of less than 0.50 were considered possibly irrelevant for the population under study. For all 14 items identified as possibly irrelevant, four members (KT, PH, RC and EMR) of the study group voted about whether these individual items should be removed or not. Each member was told to consider the feasibility of each item based upon content, relevance, patient response and measurement qualities. Each member had one vote and items were removed if at least three of four voted for their removal. If two were for and

two were against, consensus was sought by further discussion concerning the relevance of the item. Based upon this, 13 of the 14 items deemed possibly irrelevant were removed. Items P5 and P12 were removed from the Pain subscale. From the ADL subscale, items A1, A3, A4, A6, A8, A9, A10, A11, A13, A14, A15 and A17 were removed. Q4 was also considered for removal due to an ICC below 0.5 (table 2), but it was decided to keep this item, since only one person in the study group voted for its removal. After this process, the questionnaire consisted of 38 items in five subscales (Symptoms (7), Pain (11), ADL (5) Sport/Rec (10) and QOL (5)).

Internal consistency

Factor analysis of the five individual subscales showed that the items in the Symptoms, ADL and QOL subscales loaded on one factor with eigenvalues of 3.2 (46% of the variance), 3.3 (66% of the variance) and 2.9 (58% of the variance), respectively. Factor analysis of the Pain subscale showed that two factors with an eigenvalue greater than 1 were produced. Factor analysis was repeated sequentially omitting item 13 'Do you have any pain when squeezing your legs together?' and the subscale only loaded on one factor, with an eigenvalue of 5.6 (56% of the variance), and item P13 was therefore removed from the questionnaire. Factor analysis of the Sports subscale showed that two factors with an eigenvalue greater than 1 were produced. Items 9 and 10 seemed to form a separate subscale and these were omitted from the Sports subscale and further tested as a separate subscale. Items 1–8 in the Sports scale loaded on a single factor, with an eigenvalue of 5.3 (66% of the variance) and items 9 and 10 loaded on a single factor, with an eigenvalue of 1.8 (89% of the variance) and this new subscale was named Participation in Physical Activity (PA). The final version of the HAGOS then held 37 items in six separate subscales: Pain (10 items), Symptoms (7 items), ADL (5 items), Sport/Rec (8 items), PA (2 items) and QOL (5 items) (appendix 1). For each of the six HAGOS subscales, Cronbach's α were above 0.78, indicating a sufficient homogeneity of all items in the subscales (table 3).

Testing the final version of HAGOS

Missing data

HAGOS: Few individual items were missing. At baseline, 9 items from a total of 101 patients × 37 items = 0.2% were missing. A total score could be calculated for all subjects for all subscales except for PA, where a total score could be calculated for all but one subject. At retest, 1 item of 44 patients × 37 items = 0.1% was missing. Test–retest analyses could be performed for 44 subjects for all subscales except for PA, where test–retest analysis could be calculated for 43 subjects. At the 4-month follow-up, 21 items of 87 patients × 37 items = 0.7% were missing.

SF-36: Few individual items were missing. At the baseline measurement, 7 items of 101 patients × 36 items = 0.2% were missing. A total score could be calculated for all subjects for all subscales.

Test–retest reliability and measurement error

Table 3 shows ICCs, SEM and SDC of all subscales of the HAGOS. Retest was completed within a mean of 11 days, and a range of 7–21 days. For all subscales of the HAGOS, the ICCs were between 0.82 and 0.92 indicating good test–retest reliability. The SDC at the individual level ranged from 17.7 to 33.8 points and at the group level from 2.7 to 5.2 points for the different subscales.

Construct validity

Generally higher correlations were found between the HAGOS subscales and the SF-36 subscales of PF, RP and BP (convergent construct validity) than between the HAGOS and the SF-36 subscales of MH, VT, RE, SF and GH (divergent construct validity) (table 4). As hypothesised, the correlations between the HAGOS subscales ADL and Sport/Rec and the SF-36 subscale PF were at least 0.5, and higher than for the other HAGOS subscales (Pain, Symptoms, PA and QOL). The correlations between the HAGOS subscales Pain and Symptoms and the SF-36 subscale BP were at least 0.5 and 0.4, respectively, and as hypothesised, higher than for the HAGOS subscales PA and QOL, but not higher than for the HAGOS subscales ADL and Sport/Rec. The subscale QOL was moderately correlated to the SF-36 subscale MH, at 0.38 but did not reach the hypothesised threshold of being at least 0.4.

Responsiveness

As hypothesised, change in the six subscales of the HAGOS correlated with the GPE score, and the correlation was at least 0.4 for all subscales. As hypothesised, ES and SRM were lower for patients reporting worse or much worse than patients reporting somewhat worse, no change or somewhat better on the GPE score, for all subscales. Furthermore, ES and SRM for all subscales were higher for patients who reported their condition to be better or much better than patients reporting no change or only somewhat better or worse on the GPE score (table 6).

Interpretability

Floor and ceiling effects, predefined as present if more than 15% of the patients were reporting worst (0) or best (100) possible score, were found for the HAGOS subscales PA and ADL at some time points. Much larger floor and ceiling effects (40–80%) were seen for some of the SF-36 subscales. The distributions of total scores and change scores in the study sample and in relevant subgroups are presented in tables 5 and 6, and floor and ceiling effects of the HAGOS and SF-36 are presented in table 5.

DISCUSSION

The HAGOS is, to our knowledge, the first patient-reported questionnaire developed for young to middle-aged physically active patients with long-standing hip and groin pain, using a prospective research design. Furthermore, this is one of the first studies following the full COSMIN checklist in the development and testing of a PRO instrument – a checklist based on the recent international consensus process involving leading experts in the development and testing of PRO questionnaires.[22 23] The current study therefore stringently follows the mandatory steps concerning reliability, validity and responsiveness.[22 23]

We found the checklist easy to use and helpful when designing the current study. The purpose of the COSMIN checklist is to evaluate the methodological quality of studies concerning measurement properties of PRO instruments. However, it is important to be aware that the COSMIN checklist is not yet aimed for a specific evaluation of the quality of the PRO instruments themselves.[22 23] In the current study, we therefore had to rely on criteria for what constitutes adequate measurement qualities previously proposed by different authors.[48 49] In order to assess the quality of PRO instruments, we agree with the COSMIN panel that future consensus regarding criteria for what constitutes adequate measurement qualities

should be included in the COSMIN recommendations[58] to ensure methodological standardisation of this part of the process as well.

Content validity

In contrast to the development of many previous PROs concerning hip disability,[20] the HAGOS meets the standards for the development of a PRO instrument by including patients in the development process.[49 61] A study by Martin *et al*,[66] involving patients comparable with the patients in the current study, showed that large discrepancies exist between clinicians and patients when they are asked to rate the importance of different questions related to hip problems.[66] This study by Martin *et al*[66] indicates that these patients perceive questions related to sports and recreation and social-emotional aspects to be of most importance. This seems to be in accordance with the results of the current study, where the lowest baseline scores existed in the subscales Sport/Rec, PA and hip and/or groin-related QOL.

Internal consistency

Unidimensionality of a (sub)scale indicates that all the items measure the same aspect.[61] The factor structures of the preliminary HAGOS subscales Pain and Sport/Rec were not unidimensional. Therefore, remodelling the factor structure of these subscales and creating a new subscale (PA) seemed warranted. In the process of remodelling the factor structure, we removed one item in the Pain subscale, since this item did not conceptually fit under any of the other factors. This item asks about pain when 'squeezing your legs together' and may be difficult for patients to comprehend, since this is not a frequent activity or movement that all patients perform. This item was included by the expert panel and may represent a more clinical way of thinking, since the adductor squeeze is an important clinical test performed in this population.[27 28 67 68] The factor analysis revealed that two items formed a separate subscale concerning the ability to participate in physical activity (PA). The PA subscale seems highly relevant for the population that it is intended for because the inability to fully participate in sports and other physical activities often is one of the most frustrating aspects for these individuals.

Test–retest reliability and measurement error

The ICC values were adequate for all subscales indicating adequate test–retest reliability at the group level.[48 49] The SDC for the subscales ranged from 15 to 18 points for the subscales Pain, Symptoms, ADL, Sport/Rec and QOL. For the PA subscale, the SDC was 34 points. Changes above SDC values can be considered real changes at the individual level. Large SDC values at the individual level ($SDC_{individual}$) in the current study are common findings concerning patient-reported questionnaires,[69 70] indicating that patient-reported questionnaires can be problematic for use at the individual level, due to their incapacity to detect minimal but still clinically important changes.[50] At the group level, the SDC (SDC_{group}) ranged from 2.7 to 5.2 for the different subscales, which means that changes above 5 points in group mean scores can be detected with 95% confidence. The fact that the SDC_{group} is much smaller than the corresponding $SDC_{individual}$ implies that the HAGOS is much better at detecting changes at a group level.

Construct validity

Validation of instruments assessing PROs is a challenge since no gold standard is available for comparisons.[58] Instead, construct validity has been assessed by correlating the new measure with already existing well-validated measures for similar constructs (convergent construct validity) and dissimilar constructs

Table 3 Descriptive statistics and test–retest reliability of HAGOS

Patients with hip and/or groin pain (n = 44)	Test, mean (SD)	Retest, mean (SD)	Difference test–retest, mean (SD)	SEM (95% CI)	$SDC_{(ind)}$ (95% CI)	$SDC_{(group)}$ (95% CI)	ICC (95% CI)	Cronbach's α
Pain	62.3 (20.6)	64.8 (20.8)	2.6 (9.6)	6.8 (5.0–9.2)	18.8 (13.8–25.4)	2.8 (2.1–3.8)	0.89 (0.80–0.94)	0.91
Symptoms	56.5 (16.7)	58.6 (17.9)	2.1 (9.0)	6.4 (5.1–8.4)	17.7 (14.1–23.2)	2.7 (2.1–3.5)	0.86 (0.76–0.92)	0.79
ADL	68.6 (23.5)	68.8 (24.7)	0.1 (10.1)	7.2 (5.4–9.3)	20.0 (14.9–25.7)	3.0 (2.2–3.9)	0.91 (0.85–0.95)	0.87
Sport/Rec	45.0 (26.0)	44.9 (27.5)	−0.1 (11.6)	8.0 (6.0–10.7)	22.2 (16.6–29.6)	3.3 (2.5–4.5)	0.91 (0.84–0.95)	0.93
PA*	25.9 (30.7)	26.2 (27.7)	0.3 (17.8)	12.2 (9.2–16.2)	33.8 (25.4–44.8)	5.2 (3.9–6.8)	0.82 (0.69–0.90)	0.87
QOL	33.4 (15.8)	37.3 (15.9)	3.9 (8.4)	6.4 (4.8–9.0)	17.7 (13.3–24.9)	2.7 (2.0–3.8)	0.84 (0.68–0.91)	0.81

A normalised score (100 indicating no symptoms and 0 indicating extreme symptoms) is calculated for each subscale.
*n = 43.
ADL, Activities of Daily Living; ICC, intraclass correlation coefficient (3.1, agreement); PA, Participation in Physical Activity; QOL, Quality of Life; $SDC_{(ind)}$, smallest detectable change at the individual level; $SDC_{(group)}$, smallest detectable change at group level.

Table 4 Spearman's correlation coefficients (r) determined when comparing the six dimensions in HAGOS to the eight different subscales in SF-36, N = 101

HAGOS	SF-36 Physical Function	SF-36 Physical Role	SF-36 Bodily Pain	SF-36 General Health	SF-36 Vitality	SF-36 Social Functioning	SF-36 Emotional Role	SF-36 Mental Health
Pain	0.67*	0.32*	0.64*	0.34*	0.22*	0.25*	0.08	0.17
Symptoms	0.57*	0.22*	0.56*	0.34*	0.18	0.10	0.07	0.17
ADL	0.76*	0.42*	0.68*	0.31*	0.19	0.35*	0.18	0.23*
Sport/recreation	0.73*	0.32*	0.57*	0.29*	0.27*	0.35*	0.15	0.31*
PA	0.37*	0.34*	0.23*	0.23*	0.30*	0.15	0.05	0.31*
QOL	0.56*	0.36*	0.45*	0.32*	0.34*	0.32*	0.10	0.38*

*Significant correlation, $p < 0.01$.
ADL, Activities of Daily Living; HAGOS, Copenhagen Hip and Groin Outcome Score; PA, Participation in Physical Activity; QOL, Quality of Life; SF-36, Short Form-36 items.

Table 5 HAGOS score, baseline and 4-month assessment and SF-36 score, baseline assessment

	Mean	SD	Median	Range	Floor effects	Ceiling effects
HAGOS – baseline (n = 101)						
Pain	64.0	19.7	68	10–95	0	0
Symptoms	56.9	18.5	61	11–89	0	0
ADL	68.1	23.2	70	0–100	1 (1)	9 (8.9)
Sport/Rec	45.5	25.9	44	0–100	1 (1.0)	2 (2.0)
PA*	25.8	29.0	13	0–100	39 (39)	3 (3.0)
QOL	33.5	16.1	35	5–75	0	0
HAGOS – 4 months (n = 87)						
Pain	73.4	19.4	75	30–100	0	5
Symptoms	67.8	20.2	68	18–100	0	4
ADL	75.8	22.9	80	15–100	0	19 (18.8)
Sport/Rec†	56.9	27.2	56	3–100	0	7 (7)
PA†	36.1	34.2	25	0–100	28 (28)	7 (6.9)
QOL	45.6	23.4	45	5–95	0	0
SF-36 – baseline (n = 101)						
SF-36 PF	70.5	19.7	75	20–100	0	3 (3.0)
SF-36 RP	65.6	35.2	75	0–100	13 (12.9)	40 (39.6)
SF-36 BP	54.3	20.0	61	0–84	3 (3)	0
SF-36 GH	74.5	18.3	77	20–100	0	7 (6.9)
SF-36 VT	62.2	19.3	65	5–100	0	1 (1)
SF-36 SF	90.1	18.2	100	12.5–100	0	67 (66.3)
SF-36 RE	86.8	28.3	100	0–100	6 (5.9)	79 (78.2)
SF-36 MH	77.5	15.3	80	28–100	0	3 (3)

*n = 100; †n = 86.
ADL, Activities of Daily Living; BP, Bodily Pain; GH, General Health; HAGOS, Copenhagen Hip and Groin Outcome Score; MH, Mental Health; PA, Participation in Physical Activity; PF, Physical Functioning; QOL, Quality of Life; RE, Role-Emotional; RP, Role-Physical; SF, Social Functioning; SF-36, Short Form-36 items; Sport/Rec, Sport and Recreation function; VT, Vitality.

Table 6 Responsiveness

	GPE score, total (n = 87)	'Much worse' and 'worse', total (n = 7)	'Somewhat worse' and 'not changed' and 'somewhat better', total (n = 46)	'Much better'and 'better', total (n = 34)
HAGOS	**Spearman r**	**SRM, ES**	**SRM, ES**	**SRM, ES**
Pain	0.59*	−0.81, −0.63	0.23, 0.19	1.13, 1.12
Symptoms	0.68*	−0.77, −0.60	0.27, 0.16	1.27, 0.90
ADL	0.58*	−1.10, −0.89	0.08, 0.05	0.90, 0.77
Sport/Rec†	0.61*	−0.96, −0.95	0.16, 0.10	1.01§, 1.00§
PA†	0.56*	−0.88, −1.29	0.01‡, 0.01‡	1.08, 1.18
QOL	0.69*	−1.51, −0.84	0.21, 0.19	1.46, 1.78

*Significant Spearman (r) correlation, p < 0.01. †n = 86; ‡n = 45; §n = 33.
ADL, Activities of Daily Living; ES, effect size; GPE, global perceived effect; HAGOS, Copenhagen Hip and Groin Outcome Score; n, number of patients; PA, Participation in Physical Activity; QOL, Quality of Life; Sport/Rec, Sport and Recreation function; SRM, standardised response mean.

(divergent construct validity).[58] Being the first PRO for physically active patients with hip and/or groin pain, obviously no ideal instrument for comparison existed. We therefore chose to use the SF-36, since this is a well-validated measure,[54–56] with adequate measurement qualities, which has been used in similar populations with similar musculoskeletal complaints from other anatomical regions.[51–53]

Responsiveness

Responsiveness is a very important measurement quality in an outcome score,[48] because it is an indication of the PRO's ability to detect when patients are undergoing relevant clinical changes.[48 49] In the COSMIN process, it was recommended that appropriate measures to evaluate responsiveness are the same as those for hypotheses testing and construct validity, with the only difference being that the hypotheses should focus on the change score of the instrument.[58] The GPE score is only based on one transition question and has therefore been assumed to be less reliable than a multi-item instrument.[71] However, despite its possible lack of measurement precision, all a priori hypotheses concerning responsiveness of all the HAGOS subscales were confirmed in the current study and showed high correlations between the GPE score and the change scores of the HAGOS subscales ranging between 0.56 and 0.69. ESs for the different subscales for patients reporting to be 'better' or 'much better' ranged from 0.9 to 1.2 for Symptoms, Sport/Rec and PA, whereas it was 0.77 for ADL and 1.78 for QOL. This indicates that more patients are needed for a clinical trial when the ADL subscale is the primary outcome, and fewer patients are needed when QOL is the primary outcome, compared with when using the subscales Symptoms, Sport/Rec and PA as primary outcomes.

Interpretability

Few patients reported a floor or ceiling score for the HAGOS, indicating a possibility to measure both improvement and

deterioration over time. The exception was the subscale PA where 39 subjects reported worst possible score (floor effect) at the initial administration and 28 patients reported worst possible score at the 4-month administration. A floor effect of the PA subscale was, however, not surprising considering the response options in these items. The answer options to the questions concerning the ability to participate in physical activities ranges from 'always' to 'never'. It is not possible to participate to a degree less than 'never', and therefore the high number of patients answering 'never' to these questions does not seem problematic for the subscale because further deterioration is not possible. Instead we believe that the floor effects in this subscale emphasise the relevance of these items for the population under study. The floor effect could most likely be avoided in the future if easier items are added to the PA scale. However, items concerning PA should be patient derived (in order for it to have true content validity), and thus should be based on further patient interviews focusing on this particular issue. For the ADL subscale, a ceiling effect was present at the 4-month assessment. Again, this is hardly surprising since the items concerning function and ADL are usually not the most important for the population under study.[66] However, for patients with severe hip and groin pain assessing their limitations in daily activities may still be relevant.

Large ceiling effects were seen in the SF-36 for the subscales RP, SF and RE, indicating that these subscales may not be very relevant for the population in the current study. However, for the subscales PF and BP, which were primarily used for testing convergent validity in the current study, no floor and ceiling effects existed.

The MIC or the MID has been proposed for establishing cut-points for minimal but still patient-relevant clinical improvements. The MIC is the smallest change in score (within a patient) in the construct that can be measured that patients still perceive as important.[58] The MID is the smallest difference in the construct that can be measured (between patients) that is considered important.[58] There is an ongoing debate in the literature, about which methods should be used to determine the MIC and/or the MID of a PRO instrument.[58] Within the COSMIN Delphi process, no consensus on standards for assessing MIC or MID could be reached,[58] which is also reflected in the large variation in reporting and interpretation of these concepts in the literature.[71] However, it has been shown that under many circumstances, when patients with a chronic disease are asked to identify minimal change, the estimates fall very close to half an SD.[72] The MIC of the HAGOS subscales would fall between 10 and 15 points for the six subscales, using this approach (table 5). We recognise that future research on the interpretability of PRO instruments may provide new evidence which necessitates a different approach. Until then, we agree with Norman *et al*[72] that applying the rule of thumb that the estimates of the MIC fall very close to half an SD does not seem inappropriate in the absence of more specific information.

Methodological limitations

For practical reasons, the second and third administration of the questionnaire was done by the patients at home, and therefore performed in an environment different from the hospital setting. Since the administration of all the questionnaires used in this study is completely self-administered, we do not believe that this poses a methodological problem. However, whether this approach has any impact on the results remains uncertain.

Item response theory (IRT) is a relatively new method to evaluate questionnaires in healthcare and has some potential advantages over classical test theory.[61 73] The Rasch model, a mathematical model applied in IRT, has been used to develop and internally validate measures, and it uses a logistic function that creates an interval-scaled measure.[61 74] The sample size of the current study was too small for Rasch analysis since we needed a sample size of at least 200 patients for analysing this kind of instrument.[75] However, Rasch analysis should certainly be considered for possible improvements of the HAGOS in the future when a larger sample size can be included. Moreover, testing of reliability, validity and responsiveness of PROs should be an ongoing process and the most optimal and constructive approach concerning the HAGOS is to modify the scale if new knowledge about its psychometric properties emerges. We are, however, confident that HAGOS in its present form will improve the current evaluation of physically active patients with hip and groin pain.

Another limitation of the HAGOS is that it was only tested in Denmark. However, based upon the experiences of HOOS which was originally developed in Swedish[38] this should not be a barrier to translation into other languages. Since Danish is not a world language, we decided to translate and cross-culturally adapt the HAGOS to an English version according to existing guidelines.[39 40] This version is given in online appendices 1 and 2. HAGOS can be downloaded from http://www.koos.nu/.

CONCLUSION

The HAGOS questionnaire has adequate measurement qualities for the assessment of symptoms, activity limitations, participation restrictions and QOL in physically active young to middle-aged patients with long-standing hip and/or groin pain. The HAGOS should be implemented in the evaluation of treatment strategies and regimens for physically active patients with long-standing hip and/or groin pain in relevant situations where the patient's perspective and health-related QOL are of primary interest.

Acknowledgements The authors would like to thank all the people involved in the study: patients, doctors, nurses and physiotherapists at the Arthroscopic Centre Amager, Amager Hospital for participating or helping out during the study; the expert group who contributed to the development of the HAGOS: Physiotherapist Niels Bo Schmidt from the Sportsmedicine Clinic, Amager Hospital, Physiotherapists Pernille Mogensen and Theresa Bieler from the Department of Physiotherapy, Bispebjerg Hospital; orthopaedic surgeons Torsten Warming from the Sportsmedicine Clinic, Hamlet, Frederiksberg, Claus Ol Hansen and Otto Kraemer from the Arthroscopic Centre Amager, Amager Hospital for assisting in screening patients for the study; Professor Peter Magnusson and associate professor Nina Beyer, from the Musculoskeletal Research Unit, Department of Physiotherapy, Bispebjerg Hospital and Senior Research Fellow Anthony Schache and PhD student Joanne Kemp from the Department of Engineering, Melbourne University for assisting in the translation and cross-cultural adaptation of the HAGOS from Danish to English.

Funding This work was funded by the Arthroscopic Centre Amager, Department of Orthopaedic Surgery, Amager University Hospital, Denmark, The Association of Danish Physiotherapists, Danish Regions, The Lundbeck Foundation and the Danish Rheumatism Association. RC is funded by grants from the OAK foundation.

Competing interests None.

Contributors KT drafted the manuscript. KT, EMR and PH were responsible for the study concept and design. Acquisition of data was performed by KT and JP. KT, EMR, PH and RC were responsible for analysis and interpretation of data. KT and RC were responsible for statistical analysis. All the authors critically revised the manuscript.

Ethics approval The Danish ethics committee of the capital region approved the trial protocol (H-C-2007-0129), which was registered with the Danish Data Protection Agency (2007-41-1606).

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Picavet HS,** Schouten JS. Musculoskeletal pain in the Netherlands: prevalences, consequences and risk groups, the DMC(3)-study. *Pain* 2003;**102**:167–78.
2. **Picavet HS,** Hoeymans N. Health related quality of life in multiple musculoskeletal diseases: SF-36 and EQ-5D in the DMC3 study. *Ann Rheum Dis* 2004;**63**:723–9.
3. **Fricker PA,** Taunton JE, Ammann W. Osteitis pubis in athletes. Infection, inflammation or injury? *Sports Med* 1991;**12**:266–79.
4. **van der Waal JM,** Bot SD, Terwee CB, *et al.* The course and prognosis of hip complaints in general practice. *Ann Behav Med* 2006;**31**:297–308.
5. **Elliott AM,** Smith BH, Penny KI, *et al.* The epidemiology of chronic pain in the community. *Lancet* 1999;**354**:1248–52.
6. **Hoffman C,** Rice D, Sung HY. Persons with chronic conditions. Their prevalence and costs. *JAMA* 1996;**276**:1473–9.
7. **Caudill P,** Nyland J, Smith C, *et al.* Sports hernias: a systematic literature review. *Br J Sports Med* 2008;**42**:954–64.
8. **Choi H,** McCartney M, Best TM. Treatment of osteitis pubis and osteomyelitis of the pubic symphysis in athletes: a systematic review. *Br J Sports Med* 2011;**45**:57–64.
9. **Jansen JA,** Mens JM, Backx FJ, *et al.* Treatment of longstanding groin pain in athletes: a systematic review. *Scand J Med Sci Sports* 2008;**18**:263–74.
10. **Machotka Z,** Kumar S, Perraton LG. A systematic review of the literature on the effectiveness of exercise therapy for groin pain in athletes. *Sports Med Arthrosc Rehabil Ther Technol* 2009;**1**:5.
11. **Muschaweck U,** Berger L. Minimal Repair technique of sportsmen's groin: an innovative open-suture repair to treat chronic inguinal pain. *Hernia* 2010;**14**:27–33.
12. **Robertson WJ,** Kadrmas WR, Kelly BT. Arthroscopic management of labral tears in the hip: a systematic review of the literature. *Clin Orthop Relat Res* 2007;**455**:88–92.
13. **Schilders E,** Bismil Q, Robinson P, *et al.* Adductor-related groin pain in competitive athletes. Role of adductor enthesis, magnetic resonance imaging, and entheseal pubic cleft injections. *J Bone Joint Surg Am* 2007;**89**:2173–8.
14. **Standaert CJ,** Manner PA, Herring SA. Expert opinion and controversies in musculoskeletal and sports medicine: femoroacetabular impingement. *Arch Phys Med Rehabil* 2008;**89**:890–3.
15. **Swan KG,** Wolcott M. The athletic hernia: a systematic review. *Clin Orthop* 2007;**455**:78–87.
16. **Dawson J,** Doll H, Fitzpatrick R, *et al.* The routine use of patient reported outcome measures in healthcare settings. *BMJ* 2010;**340**:c186.
17. **Patrick DL,** Burke LB, Powers JH, *et al.* Patient-reported outcomes to support medical product labeling claims: FDA perspective. *Value Health* 2007;**10**(Suppl 2):S125–37.
18. **Speight J,** Barendse SM. FDA guidance on patient reported outcomes. *BMJ* 2010;**340**:c2921.
19. **Timmins N.** NHS goes to the PROMS. *BMJ* 2008;**336**:1464–5.
20. **Thorborg K,** Roos EM, Bartels EM, *et al.* Validity, reliability and responsiveness of patient-reported outcome questionnaires when assessing hip and groin disability: a systematic review. *Br J Sports Med* 2010;**44**:1186–96.
21. **Marshall M,** Lockwood A, Bradley C, *et al.* Unpublished rating scales: a major source of bias in randomised controlled trials of treatments for schizophrenia. *Br J Psychiatry* 2000;**176**:249–52.
22. **Mokkink LB,** Terwee CB, Patrick DL, *et al.* The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol* 2010;**63**:737–45.
23. **Mokkink LB,** Terwee CB, Knol DL, *et al.* The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. *BMC Med Res Methodol* 2010;**10**:22.
24. World Health Organization. International Classification of Functioning, Disability and Health. Geneva, 2001.
25. Global Recommendations on Physical activity for Health. http://www.who.int/dietphysicalactivity/factsheet_recommendations/en/index.html (accessed Mar 2011).
26. **Falvey EC,** Franklyn-Miller A, McCrory PR. The groin triangle: a patho-anatomical approach to the diagnosis of chronic groin pain in athletes. *Br J Sports Med* 2009;**43**:213–20.
27. **Bradshaw CJ,** Bundy M, Falvey E. The diagnosis of longstanding groin pain: a prospective clinical cohort study. *Br J Sports Med* 2008;**42**:851–4.
28. **Hölmich P.** Long-standing groin pain in sportspeople falls into three primary patterns, a "clinical entity" approach: a prospective study of 207 patients. *Br J Sports Med* 2007;**41**:247–52.
29. **Lesher JM,** Dreyfuss P, Hager N, *et al.* Hip joint pain referral patterns: a descriptive study. *Pain Med* 2008;**9**:22–5.
30. **Philippon MJ,** Maxwell RB, Johnston TL, *et al.* Clinical presentation of femoroacetabular impingement. *Knee Surg Sports Traumatol Arthrosc* 2007;**15**:1041–7.
31. **Jansen JA,** Mens JM, Backx FJ, *et al.* Diagnostics in athletes with long-standing groin pain. *Scand J Med Sci Sports* 2008;**18**:679–90.
32. **Leibold MR,** Huijbregts PA, Jensen R. Concurrent criterion-related validity of physical examination tests for hip labral lesions: a systematic review. *J Man Manip Ther* 2008;**16**:E24–41.
33. **Martin RL,** Enseki KR, Draovitch P, *et al.* Acetabular labral tears of the hip: examination and diagnostic challenges. *J Orthop Sports Phys Ther* 2006;**36**:503–15.
34. **Martin RL,** Irrgang JJ, Sekiya JK. The diagnostic accuracy of a clinical examination in determining intra-articular hip pain for potential hip arthroscopy candidates. *Arthroscopy* 2008;**24**:1013–18.
35. **Sangha O,** Stucki G, Liang MH, *et al.* The Self-Administered Comorbidity Questionnaire: a new method to assess comorbidity for clinical and health services research. *Arthritis Rheum* 2003;**49**:156–63.
36. **Benjamin S,** Morris S, McBeth J, *et al.* The association between chronic widespread pain and mental disorder: a population-based study. *Arthritis Rheum* 2000;**43**:561–7.
37. **Birrell F,** Lunt M, Macfarlane GJ, *et al.* Defining hip pain for population studies. *Ann Rheum Dis* 2005;**64**:95–8.
38. **Nilsdotter AK,** Lohmander LS, Klässbo M, *et al.* Hip disability and osteoarthritis outcome score (HOOS) – validity and responsiveness in total hip replacement. *BMC Musculoskelet Disord* 2003;**4**:10.
39. **Beaton DE,** Bombardier C, Guillemin F, *et al.* Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine* 2000;**25**:3186–91.
40. **Guillemin F,** Bombardier C, Beaton D. Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. *J Clin Epidemiol* 1993;**46**:1417–32.
41. **Beyer N,** Thorborg K, Vinther A. Translation and Cross-Cultural Adaptation of the Danish Version of the Hip Dysfunction and Osteoarthritis Outcome Score 2.0 (HOOS 2.0). http://www.koos.nu/ (accessed Mar 2011).
42. **Martin RL,** Kelly BT, Philippon MJ. Evidence of validity for the hip outcome score. *Arthroscopy* 2006;**22**:1304–11.
43. **Martin RL,** Philippon MJ. Evidence of validity for the hip outcome score in hip arthroscopy. *Arthroscopy* 2007;**23**:822–6.
44. **Martin RL,** Philippon MJ. Evidence of reliability and responsiveness for the hip outcome score. *Arthroscopy* 2008;**24**:676–82.
45. **Kirkley A,** Griffin S, McLintock H, *et al.* The development and evaluation of a disease-specific quality of life measurement tool for shoulder instability. The Western Ontario Shoulder Instability Index (WOSI). *Am J Sports Med* 1998;**26**:764–72.
46. **Kirkley A,** Alvarez C, Griffin S. The development and evaluation of a disease-specific quality-of-life questionnaire for disorders of the rotator cuff: The Western Ontario Rotator Cuff Index. *Clin J Sport Med* 2003;**13**:84–92.
47. **Mokkink LB,** Terwee CB, Patrick DL, *et al.* The COSMIN Checklist Manual, 2009.
48. **Lohr KN,** Aaronson NK, Alonso J, *et al.* Evaluating quality-of-life and health status instruments: development of scientific review criteria. *Clin Ther* 1996;**18**:979–92.
49. **Terwee CB,** Bot SD, de Boer MR, *et al.* Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol* 2007;**60**:34–42.
50. **Terwee CB,** Roorda LD, Knol DL, *et al.* Linking measurement error to minimal important change of patient-reported outcomes. *J Clin Epidemiol* 2009;**62**:1062–7.
51. **Ashby E,** Grocott MP, Haddad FS. Outcome measures for orthopaedic interventions on the hip. *J Bone Joint Surg Br* 2008;**90**:545–9.
52. **Beaton DE,** Hogg-Johnson S, Bombardier C. Evaluating changes in health status: reliability and responsiveness of five generic health status measures in workers with musculoskeletal disorders. *J Clin Epidemiol* 1997;**50**:79–93.
53. **Patel AA,** Donegan D, Albert T. The 36-item short form. *J Am Acad Orthop Surg* 2007;**15**:126–34.
54. **Bjorner JB,** Damsgaard MT, Watt T, *et al.* Tests of data quality, scaling assumptions, and reliability of the Danish SF-36. *J Clin Epidemiol* 1998;**51**:1001–11.
55. **Bjorner JB,** Thunedborg K, Kristensen TS, *et al.* The Danish SF-36 Health Survey: translation and preliminary validity studies. *J Clin Epidemiol* 1998;**51**:991–9.
56. **Bjorner JB,** Kreiner S, Ware JE, *et al.* Differential item functioning in the Danish translation of the SF-36. *J Clin Epidemiol* 1998;**51**:1189–202.
57. **Hölmich P,** Uhrskou P, Ulnits L, *et al.* Effectiveness of active physical training as treatment for long-standing adductor-related groin pain in athletes: randomised trial. *Lancet* 1999;**353**:439–43.
58. **Mokkink LB,** Terwee CB, Patrick DL, *et al.* The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Qual Life Res* 2010;**19**:539–49.
59. **McHorney CA,** Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? *Qual Life Res* 1995;**4**:293–307.
60. **de Vet HC,** Adèr HJ, Terwee CB, *et al.* Are factor analytical techniques used appropriately in the validation of health status questionnaires? A systematic review on the quality of factor analysis of the SF-36. *Qual Life Res* 2005;**14**:1203–18.

61. **Streiner DL,** Norman GR. *Health Measurement Scales: A Practical Guide to Their Development and Use*. New York, New York: Oxford University Press 2003.
62. **Weir JP.** Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J Strength Cond Res* 2005;**19**:231–40.
63. **Busija L,** Osborne RH, Nilsdotter A, *et al.* Magnitude and meaningfulness of change in SF-36 scores in four types of orthopedic surgery. *Health Qual Life Outcomes* 2008;**6**:55.
64. **de Vet HC,** Bouter LM, Bezemer PD, *et al.* Reproducibility and responsiveness of evaluative outcome measures. Theoretical considerations illustrated by an empirical example. *Int J Technol Assess Health Care* 2001;**17**:479–87.
65. **Cohen J.** *Statistical Power Analysis for the Behavioral Sciences.* Hillsdale, New Jersey: Lawrence Erlbaum 1988.
66. **Martin RL,** Mohtadi NG, Safran MR, *et al.* Differences in physician and patient ratings of items used to assess hip disorders. *Am J Sports Med* 2009;**37**:1508–12.
67. **Hölmich P,** Hölmich LR, Bjerg AM. Clinical examination of athletes with groin pain: an intraobserver and interobserver reliability study. *Br J Sports Med* 2004;**38**:446–51.
68. **Verrall GM,** Slavotinek JP, Barnes PG, *et al.* Description of pain provocation tests used for the diagnosis of sports-related chronic groin pain: relationship of tests to defined clinical (pain and tenderness) and MRI (pubic bone marrow oedema) criteria. *Scand J Med Sci Sports* 2005;**15**:36–42.
69. **de Boer MR,** de Vet HC, Terwee CB, *et al.* Changes to the subscales of two vision-related quality of life questionnaires are proposed. *J Clin Epidemiol* 2005;**58**:1260–8.
70. **Quintana JM,** Escobar A, Bilbao A, *et al.* Responsiveness and clinically important differences for the WOMAC and SF-36 after hip joint replacement. *Osteoarthr Cartil* 2005;**13**:1076–83.
71. **Terwee CB,** Roorda LD, Dekker J, *et al.* Mind the MIC: large variation among populations and methods. *J Clin Epidemiol* 2010;**63**:524–34.
72. **Norman GR,** Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Med Care* 2003;**41**:582–92.
73. **McHorney CA,** Cohen AS. Equating health status measures with item response theory: illustrations with functional status items. *Med Care* 2000;**38**:II43–59.
74. **Davis AM,** Perruccio AV, Canizares M, *et al.* The development of a short measure of physical function for hip OA HOOS-Physical Function Shortform (HOOS-PS): an OARSI/OMERACT initiative. *Osteoarthr Cartil* 2008;**16**:551–9.
75. **Comins J,** Brodersen J, Krogsgaard M, *et al.* Rasch analysis of the Knee injury and Osteoarthritis Outcome Score (KOOS): a statistical re-evaluation. *Scand J Med Sci Sports* 2008;**18**:336–45.

Corrections

Thorborg K, Hölmich P, Christensen R, *et al*. The Copenhagen Hip and Groin Outcome Score (HAGOS): development and validation according to the COSMIN checklist (*Br J Sports Med* 2011;**45**:478–491). In table 2, in rows A1 Walking down stairs and A2 Walking up stairs the data was inadvertently swapped. The journal apologises for this error.

Br J Sports Med 2011;**45**:742. doi:10.1136/bjsm.2010.080937corr1