Qual Life Res (2015) 24:2985–2992
DOI 10.1007/s11136-015-1025-4

Dynamic weight-bearing assessment of pain in knee osteoarthritis: a reliability and agreement study

Louise Klokker[1] · Robin Christensen[1,2] · Richard Osborne[3] · Elisabeth Ginnerup[1] · Eva E. Waehrens[1,4] · Henning Bliddal[1] · Marius Henriksen[1]

Accepted: 22 May 2015 / Published online: 6 June 2015

Abstract
Purpose To evaluate the reliability, agreement and smallest detectable change in a measurement instrument for pain and function in knee osteoarthritis; the Dynamic weight-bearing Assessment of Pain (DAP).
Methods The sample size was set to 20 persons, recruited from the outpatient osteoarthritis clinic at Frederiksberg Hospital, Copenhagen. Two physiotherapists tested all participants during two visits; at the first visit, one single DAP (including four scores) was conducted by rater one; at the second visit, DAP was conducted by both raters one and two in randomized order with concealed allocation. The time interval was approximately 1.5 h. Measurement error was estimated by standard error of measurement (SEM). The intra- and inter-rater reliability was estimated by Intra-class Correlation Coefficients for agreement based on a two-way ANOVA with random effects (single measures ICC 2.1). Smallest detectable change (SDC) and limits of agreement were calculated.
Results The pain score showed excellent reliability in terms of ICC (intra-rater 0.93, CI 0.83–0.97, inter-rater 0.91, CI 0.78–0.96), low SEM (intra-rater 0.70, inter-rater 0.86, on a scale from 0 to 10), and acceptable SDC for intra-rater test (1.95). The three knee bend scores all had ICC above 0.50, showing fair-to-good reliability. None of the knee bend scores showed acceptable SEM and SDC.
Conclusions The reproducibility of the DAP pain score meets the demands for use in clinical practice and research. The total knee bend could be useful for motivational purpose in clinical use. Testing of other psychometric properties of the DAP is pending.

Keywords Measurement · Knee osteoarthritis · Physical function · Pain

✉ Louise Klokker
Louise.Klokker.Madsen@regionh.dk

Robin Christensen
robin.christensen@regionh.dk

Richard Osborne
richard.osborne@deakin.edu.au

Elisabeth Ginnerup
elisabeth.marie.ginnerup-nielsen@regionh.dk

Eva E. Waehrens
eva.elisabet.waehrens@regionh.dk

Henning Bliddal
henning.bliddal@regionh.dk

Marius Henriksen
marius.henriksen@regionh.dk

[1] The Parker Institute, Bispebjerg & Frederiksberg University Hospitals, Copenhagen, Denmark

[2] Faculty of Health Sciences, Institute of Clinical Research, University of Southern Denmark, Odense, Denmark

[3] Faculty of Health, Population Health Strategic Research Centre, School of Health and Social Development, Deakin University, Burwood, Australia

[4] The Research Initiative for Activity studies and Occupational Therapy, Institute of Public Health, University of Southern Denmark, Odense, Denmark

Introduction

The impact of knee osteoarthritis (OA) on the individual is usually estimated by evaluating pain, physical function and the patient's global assessment of well-being [1, 2]. For these purposes, patient-reported outcome measures (PROMs) are commonly used, but also performance measures (PMs) could be considered to quantify physical function [3]. It is well established that PROMs and PMs do not capture the same aspects of physical function in musculoskeletal conditions including knee OA [4–9]; it is suspected that PROMs generally measure an overall comprehensive experience [10, 11], whereas PMs may target a more specific construct linked to impairments in body functions [11]. One study ($n = 115$) found that the sensitivity to change over a period of 2 years was better for PMs than for PROMs in a population of patients with hip and/or knee OA [12]. The same conclusion was reached in a population of patients undergoing hip or knee replacement ($n = 73$), leading the authors to suggest that PMs should be core outcome measures in knee OA [13].

Hence, both PROMs and PMs should be used, as they contribute to a comprehensive understanding of a patient's situation [14–16]. Pain triggered by activity is a characteristic feature of knee OA [17–20]. This leads to the anticipation that PMs may contribute with further valuable information if a pain score is integrated with a PM and therefore measure a specific construct of pain during an activity. In fact, it has been suggested that pain measures in knee OA should always include either performing pain-provoking activities or asking about pain during these activities [21]. Even though several PMs exist, which presumably provoke pain, there are no validated PMs with associated pain assessment for knee OA. A solution to this could be to extend an existing PM to include a pain score. However, we believe that it is possible to exceed the feasibility of existing PMs and therefore increase the incentive to use outcome measures in clinical practice.

Based on input from patients and health professionals [22], we have developed a Dynamic weight-bearing Assessment of Pain (DAP) for knee OA. The instrument combines a PM (weight-bearing knee bends) with a PROM (self-reported pain intensity). The pain intensity is measured on a 0–10 Numeric Rating Scale (NRS) as preferred by patient groups [23] and recommended for measuring pain intensity in clinical trials [19]. The psychometric properties of the DAP are yet to be established. As a first step, the objective of this study was to estimate the reliability, agreement and smallest detectable change (SDC) in the DAP to establish thresholds for detection of change between tests.

Methods

Participants

This study was nested within an assessor- and participant-blinded randomized controlled trial comparing corticosteroid injection with placebo prior to 12 weeks of supervised exercise three times weekly in people with knee OA (EudraCT: 2012-002607-18). Inclusion criteria for the trial were as follows: age above 40 years, radiologically verified diagnosed knee OA, 'pain while walking on a flat surface' of at least 4 on a 0–10 NRS, and a body mass index >20 and <35 kg/m^2. Exclusion criteria were use of intra-articular corticosteroids in the knee or participation in physiotherapeutic exercise for knee OA within the last 3 months, or severe diseases. As part of the larger trial, the participants filled in the 'Knee injury and Osteoarthritis Outcome Score' (KOOS) [24].

Data for the current reliability and agreement study were collected at the follow-up assessments in the hosting trial done 3 months after termination of study interventions. All participants in this study gave informed consent before enrolling in the hosting trial and received a copy of the consent. All participants were asked about adverse events during a rheumatologist consultancy within 2 weeks after the tests. This report follows the recommended reporting guideline [25] suggested for reliability and agreement studies (GRRAS statement) [26]. The statistical analyses follow the COSMIN standards [27, 28].

Test description

The DAP is a simple performance test with an integrated pain score, designed to provide useful information on the interaction between pain and function for monitoring treatment progress and evaluating treatment effects. The DAP is intended for use both in research and in clinical practice, primarily physiotherapy related. The patient is asked to perform as many standing knee bends as possible within 30 s. For each bend, the knees should reach approximately 90 degrees of flexion (visually inspected by the observer) and full extension (to the extent possible for the individual patient). Limited range of motion does not preclude a test and does not result in missing data. This is supervised by the rater, who decides whether the test performance is approved according to the purpose, e.g., clinical monitoring of treatment progress or scientific purposes. There are three scores in the test: (1) number of pain-free knee bends; (2) number of painful knee bends; and (3) pain during knee bends on a 0–10 NRS. Scores from (1) and (2) are added to give the total number of knee bends. The pain score is obtained immediately after the knee bend tests with

the question: 'How much pain did you feel during the knee bends, on a scale from 0 to 10, where 0 is no pain and 10 is the worst pain you can imagine?' In case the pain varies during the test, the highest pain intensity is recorded. The DAP takes about 1 min to perform including instructions and does not require any equipment besides a stopwatch/watch. The numbers of knee bends are direct measures of the patient's ability to repeat a weight-bearing activity within a short timeframe; the pain intensity scores are measures of pain during a specific weight-bearing physical activity. The purpose is to reflect the limitations of daily activities due to knee OA that involves weight-bearing knee bending (e.g., getting up from/down in a chair, gardening, cleaning).

Study design

The study design is shown in Fig. 1. Two physiotherapists tested the DAP on all participants. Rater A (LK) is the test developer and experienced in using the DAP. Rater B (EG) had no experience with the DAP, but had an introduction and one rehearsal session with a knee OA patient not included in the study. Each participant had two visits separated by minimum 2 days, and maximum 1 week. At the first visit, one single DAP was conducted by rater A. At the second visit, DAP was conducted by both raters A and B in a randomized order separated by approximately 1.5 h. Thus, each participant was tested twice by rater A and once by rater B.

Statistical analysis

The statistical analyses were performed using SAS statistical software (version 9.3) and SPSS (IBM SPSS Statistics 19). Reliability was estimated by Intra-class Correlation Coefficients (ICC) for agreement based on a two-way ANOVA with random effects (single measures ICC 2.1) [28]. ICC were calculated for both intra- and inter-rater data. We had decided a priori to interpret the ICC value using the criteria for clinical acceptability suggested by Fleiss where ICC < 0.4 represent poor, $0.4 <$ ICC < 0.75 represent fair to good and ICC > 0.75 represent excellent agreement [29]. However, as the DAP is also intended for use in clinical practice, the quality criterion for this purpose was conservatively set to an ICC of at least 0.90 [30, 31], with a lower 95 % confidence limit of at least 0.75.

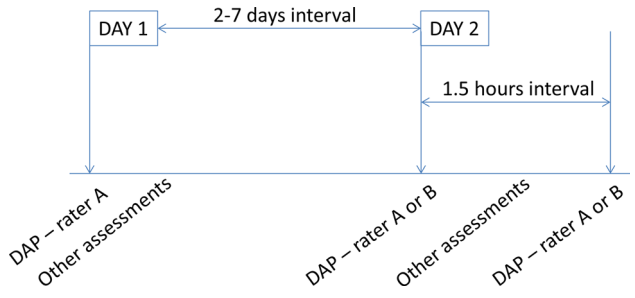

Fig. 1 Study design

We calculated the measurement error 'standard error of measurement' (SEM) that represents the standard deviation of repeated measures in one patient. SEM was calculated as the square root of the residual mean square value obtained from the two-way analysis of variance, which is used to calculate the ICC [28]. SEM was calculated for both within (intra-rater) and between raters (inter-rater). Subsequently, the smallest detectable change (SDC) was calculated (representing the minimal change that must appear to ensure that the observed change is beyond measurement error); SDC is calculated as $1.96 \times \sqrt{2} \times$ SEM for both intra- and inter-rater data [28]. Based on the SDC, the limits of agreement (LoA) were calculated ($\bar{d} \pm$ SDC) and presented in Bland and Altman plots. Acceptable SDC was set to a maximum of two points or a reduction of 30.0 % in the 11-point NRS for pain, as this is regarded to represent the minimal clinically important difference [32]. For the knee bending scores, the a priori maximum SDC was set to 2.6, based on the minimal clinically important difference in the 30-s sit-to-stand test, tested in a hip OA population [33].

Sample size considerations

The power calculation was based on estimates of the 95 % confidence interval of the ICC. Assuming that the reliability of the DAP corresponded to an ICC of 0.80, including 13 participants would result in a lower 95 % confidence limit of 0.60 [28]. Based on this analysis, the number of participants was conservatively set to 20. We recruited 20 participants among the last 20 participants enrolled in the hosting trial.

Results

A total of 20 hosting trial participants who met the eligibility criteria were invited. All accepted to participate, and all completed the study. Their characteristics are presented in Table 1. Summary statistics from tests and retests are provided in Table 2. Table 3 presents the results for ICC, SEM, SDC and LoA.

Of the 4 scores, the pain intensity score showed the best properties in terms of low SEM (0.70 for the intra-rater tests and 0.86 for the inter-rater tests on a scale from 0 to 10), acceptable SDC for the intra-rater tests (1.95) and excellent ICC (0.93, CI 0.83 to 0.97 for the intra-rater tests and 0.91, CI 0.78 to 0.96 for the inter-rater tests). SDC for the inter-rater test did not reach the a priori acceptable level

(2.39). The three knee bend scores all had ICC above 0.50, showing fair-to-good agreement. However, only for the inter-rater tests did the lower confidence limit not fall below 0.40. The SEM for knee bends varied from 2.95 to 6.85 for the intra-rater tests and 2.56 to 5.95 for the inter-rater tests, in both cases with the lowest SEM for the total knee bends scores. None of the knee bend scores fell below the a priori defined maximum SDC of 2.6. The Bland and Altman plots in Fig. 2 illustrate the differences between observers plotted against the mean value of both observers for (A) pain intensity, (B) total bends, (C) pain-free bends and (D) painful bends.

Within 2 weeks after the tests (at the rheumatologist consultancy), one participant complained about pain in the days after performing the DAP. The excess pain had disappeared at the time of the consultancy and was not considered related to the test. Otherwise, no adverse events were noted.

Discussion

This study supports the integration of a pain score with a performance measure in order to capture another perspective of pain: the interaction with function. In this population of people with symptomatic knee OA, the DAP pain score shows excellent ICC, comparable with patient-reported outcome measures [34] and other performance-based outcome measures such as walking , stair, and chair stand tests [35]. Furthermore, the DAP has the advantage of being very short and not requiring any equipment besides a (stop) watch, whereas other performance measures typically require walking lanes, stairs, or chairs of standard dimensions. This, together with the low SEM and, for the intra-rater test, acceptable SDC, supports the applicability of the DAP in research and clinical practice, with the pain score as primary indicator.

The excellent ICC (0.91 and 0.93) of the pain score suggest that measuring pain during a pain-provoking activity yields reliable results. The demands to reliability and measurement error for instruments applied on the individual level in clinical practice are higher than on group level, as there is often only one score (no averaging) [36]. Thus, the low measurement error of the DAP pain score makes the DAP useful on individual levels in clinical practice.

The knee bend scores did not show adequate reliability and agreement in this population; hence, the knee bend scores may be omitted leaving only the pain score in the test, making this even simpler. However, the number of knee bends may have a motivational effect because of the more detailed information on treatment progress provided. This remains to be evaluated.

Limitations

This population had relatively mild symptoms with a mean of 70.5 on the KOOS pain subscale, and 78.5 on the KOOS function subscale (0–100 scales; higher is better). However, a mean of 55.3 on the KOOS quality-of-life subscale (range 0–87.5) indicates that the participants were indeed affected by their knee OA. Three patients had a DAP pain score of 0 at the first visit (NRS = 0), and six patients had a DAP pain score of 0 at the second visit (regardless of the rater). This calls for attention to the risk of floor or ceiling effects of the DAP. However, as there is no reason to

Table 1 Participants' characteristics

Participants ($n = 20$)	Mean	SD	Median	Min	Max	N	%
Female (n)	na	na	na	na	na	14	70
Caucasian European (n)	na	na	na	na	na	20	100
Age	64	6.6	64.3	46	76	na	na
Weight	85.5	14.3	83.0	57.0	117.0	na	na
BMI	30.2	3.4	31.2	22.8	35.0	na	na
Pain, current (paindetect)	2.5	1.9	2	1	7	na	na
Pain, average last 4 weeks (paindetect)	3.65	2.4	3.5	1	9	na	na
Pain, worst last 4 weeks (paindetect)	2.75	1.9	2	1	8	na	na
Kellgren/Lawrence score (0–4)	2.8	0.7	3.0	2	4	na	na
KOOS function subscale (0–100)	78.5	19.7	80.9	42.6	100	na	na
KOOS pain subscale (0–100)	70.5	23.2	76.4	22.2	100	na	na
KOOS quality-of-life subscale (0–100)	55.3	20.2	56.2	0	87.5	na	na
KOOS sport and recreation (0–100)	49.3	31.0	42.5	0	100	na	na
KOOS symptoms (0–100)	70.7	20.1	73.2	32.1	100	na	na

na not applicable

Table 2 Summary statistics for intra- and inter-rater test, retest and difference

	Test					Retest					Test–retest difference			
	Mean	SD	Median	Min	Max	Mean	SD	Median	Min	Max	Mean	SD	95 %CI	
													Low	High
Intra-rater														
Pain intensity, 0–10	3.1	2.6	2.0	0.0	8.0	2.7	3.0	1.0	0.0	8.0	−0.4	2.8	−2.2	1.4
Total bends, *n*	19.3	4.4	18.5	10.0	29.0	21.4	5.3	21.5	13.0	34.0	2.2	4.9	−1.0	5.3
Pain-free bends, *n*	5.5	8.8	0.0	0.0	29.0	6.2	9.2	0.0	0.0	23.0	0.7	9.0	−5.1	6.5
Painful bends, *n*	13.8	7.7	16.5	0.0	25.0	15.3	11.5	16.5	0.0	34.0	1.5	9.8	−4.8	7.7
Inter-rater														
Pain intensity, 0–10	2.7	3.0	1.0	0.0	8.0	2.4	2.6	1.0	0.0	8.0	−0.3	2.8	−2.1	1.5
Total bends, *n*	21.4	5.3	21.5	13.0	34.0	21.1	5.4	21.0	11.0	34.0	−0.3	5.4	−3.7	3.1
Pain-free bends, *n*	6.2	9.2	0.0	0.0	23.0	7.5	10.2	0.0	0.0	29.0	1.3	9.7	−4.9	7.5
Painful bends, *n*	15.3	11.5	16.5	0.0	34.0	13.7	11.0	15.0	0.0	34.0	−1.6	11.2	−8.8	5.6

Table 3 Intra-class Correlation Coefficients (ICC) with 95 % confidence interval (CI), standard error of measurement (SEM), smallest detectable change (SDC) and limits of agreement (LoA)

	ICC(2.1.A)	95 % CI		SEM	SDC	LoA	
		Lower	Upper			Lower	Upper
Intra-rater							
Pain intensity, 0–10	0.93	0.83	0.97	0.70	1.95	−2.4	1.6
Total bends, *n*	0.59	0.21	0.82	2.95	8.18	−6.0	10.3
Pain-free bends, *n*	0.63	0.26	0.83	5.57	15.43	−14.7	16.1
Painful bends, *n*	0.52	0.11	0.78	6.85	19.00	−17.5	20.4
Inter-rater							
Pain intensity, 0–10	0.91	0.78	0.96	0.86	2.39	−2.7	2.1
Total bends, *n*	0.78	0.52	0.91	2.56	7.08	−7.4	6.8
Pain-free bends, *n*	0.77	0.50	0.90	4.70	13.03	−11.7	14.3
Painful bends, *n*	0.72	0.43	0.88	5.95	16.50	−18.1	14.9

discriminate patients reporting no pain any further, this cannot be categorized as a floor effect [28]. The same can be assumed regarding the knee bend score, as a limited knee mobility does not exclude anyone from performing the DAP. The reliability of the DAP is still unknown for populations with more severe symptoms. The small sample size, based on a priori calculations, is a possible limitation to the study. Furthermore, the lack of a stable external measure to ensure the absence of change between the two visits is a limitation to this study, as potential changes could have affected the correlation coefficients.

In general, the reliability was higher between the two raters than within the same rater, at least for the knee bends scores. This may be related to the study design, with one test by rater A on the first visit, and tests by both raters A and B on the second visit; higher mean knee bend scores and lower mean pain scores on the second visit suggest a certain learning effect. The difference could also be due to day-to-day variability. However, the SEM did not vary much between intra- and inter-rater measures. As measurement error is more a characteristic of a test in itself [27], it is expected to remain stable across populations and raters. The random sequence of the raters at the second visit may have influenced the intra-rater reliability, given that about half of the tests at the second visit were preceded by a test with the other rater. However, there was no significant difference related to the sequence of tests; mean pain score difference was 0.8 (3.0 where rater A tested first, and 2.2 where rater A tested second, $p = 0.58$); mean total knee bend score difference was 2.2 (22.4 where rater A tested first, and 20.2 where rater A testes second, $p = 0.39$).

In this study, we asked the participants to bend their knees from a standing position until reaching flexion of approximately 90°. This is a somewhat unspecific instruction and was only monitored visually by the rater; thus, certain variability is assumed. For example, the two participants who reached more than 30 in the total number

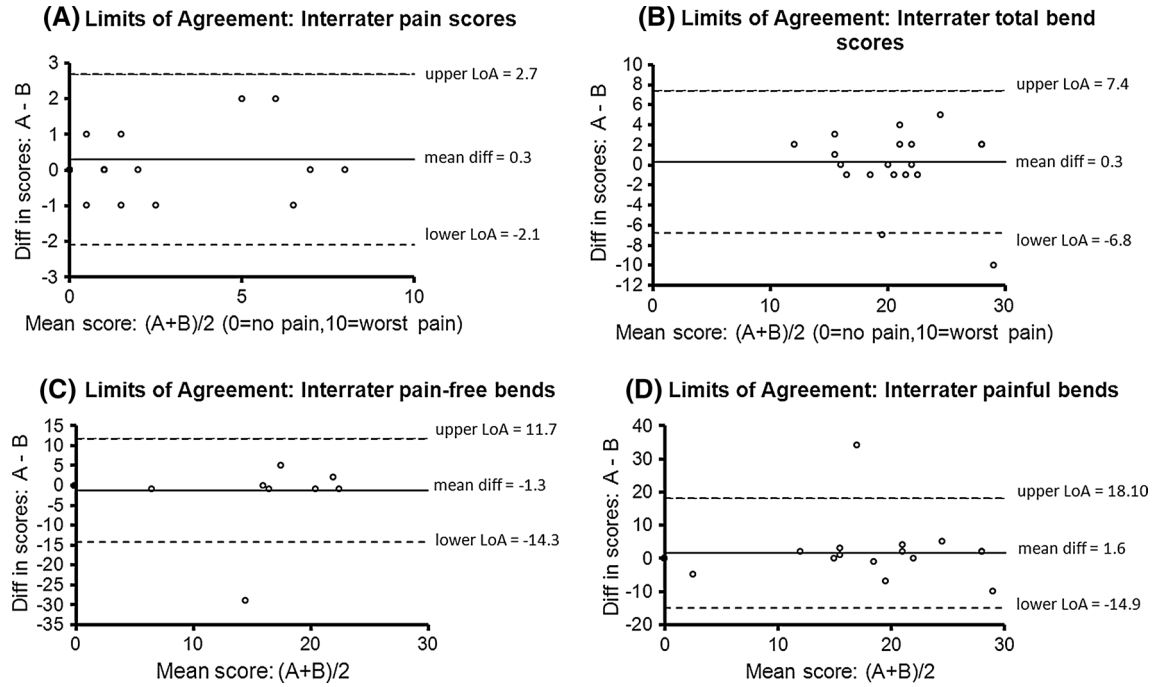

Fig. 2 Bland and Altman plot illustrating the differences between observers **a**–**b** plotted against the mean value of observers (A + B)/2 for **a** the pain intensity scores, **b** the total knee bend scores, **c** pain-free knee bends and **d** painful knee bends. *Solid line* (mean difference) is an expression of the systematic error between observers while the limits of agreement define the boundaries of random error

of knee bends on their second visit are unlikely to have reached 90°. The good results in this study despite this uncertainty yield further support to the properties of the DAP. Also, bending knees to approximately 90° is a reasonably easy task to comprehend for most patients, and we believe that a pragmatic approach to a test design facilitates feasibility and cooperation from the patients. During instructions, it was emphasized to the patients that no predefined number of squats was expected from them; the number of squats was according their personal limit of tolerance. This might result in some patients choosing to endure pain in exchange for better performance (more bends) and some choosing less pain on the expense of high performance. This is true both in everyday life and in the interaction with the healthcare system, and unpredictable pain behavior is a premise for all performance measures. The DAP is developed in an attempt to address this pain behavior; i.e., we believe that that pain score of the DAP reflect the interaction between pain and function. Hence, we do not think of this as a limitation to the DAP.

We chose to only include one rehearsal session with the non-experienced rater. We were confident with this choice because basic physiotherapy knowledge enables understanding and performing this simple test. Furthermore, we wanted to test whether this minimal instruction would be sufficient; as this seems to hold true, the feasibility of the DAP is promising in this regard. However, the low reliability of the knee bend scores suggests that more explicit instructions are warranted; this is pending. Importantly, the results of this study only apply to physiotherapist; whether the DAP can be used by other groups of health professionals remains to be examined.

All participants were asked about adverse events during a rheumatologist consultancy within 2 weeks after the tests. Only one participant complained about pain in his unaffected knee after the tests, but this was not considered related to the DAP. Thus we are confident that the DAP is safe and with no excessive risks compared to everyday activities in a population with mild knee OA.

In conclusion, the reliability, agreement and, for the intra-rater test, the smallest detectable change in the DAP pain score meet the demands for use in clinical practice and research. The total knee bend score should be kept for motivational reasons. Evaluation of other important psychometric properties of the DAP such as validity, responsiveness and feasibility is pending.

Acknowledgments We would like to thank the patients who participated in this study. The Parker Institute is grateful for the financial support received from public and private foundations, companies and private individuals over the years. This study was supported by grants from The Oak Foundation, The Danish Physiotherapy Association and The Danish Rheumatism Association. Financial support was provided from The Parker Institute and Deakin University. The sponsors and funders of the study had no role in the study design, data collection and analysis, interpretation or reporting of this work or the decision to submit the work for publication.

Conflicts of interest The Parker Institute is supported by a core grant from the Oak Foundation; The Oak Foundation is a group of philanthropic organizations that, since its establishment in 1983, has given grants to not-for-profit organizations around the world. Dr. Christensen reports: I am involved in many health-care initiatives and research that could benefit from wide uptake of this publication (including Cochrane, OMERACT, and the GRADE Working Group). The remaining authors declare that they have no conflict of interests that could influence their work and conclusions in relation to this manuscript.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

1. Bellamy, N., Kirwan, J., Boers, M., Brooks, P., Strand, V., Tugwell, P., et al. (1997). Recommendations for a core set of outcome measures for future phase III clinical trials in knee, hip, and hand osteoarthritis. Consensus development at OMERACT III. *Journal of Rheumatology, 24*(4), 799–802.
2. Dreinhofer, K., Stucki, G., Ewert, T., Huber, E., Ebenbichler, G., Gutenbrunner, C., et al. (2004). ICF core sets for osteoarthritis. *Journal of Rehabilitation Medicine, 44*(Suppl), 75–80.
3. Dobson, F., Hinman, R. S., Roos, E. M., Abbott, J. H., Stratford, P., Davis, A. M., et al. (2013). OARSI recommended performance-based tests to assess physical function in people diagnosed with hip or knee osteoarthritis. *Osteoarthritis Cartilage, 21*(8), 1042–1052.
4. Amris, K., Waehrens, E. E., Jespersen, A., Bliddal, H., & Danneskiold-Samsoe, B. (2011). Observation-based assessment of functional ability in patients with chronic widespread pain: A cross-sectional study. *Pain, 152*(11), 2470–2476.
5. Stevens-Lapsley, J. E., Schenkman, M. L., & Dayton, M. R. (2011). Comparison of self-reported knee injury and osteoarthritis outcome score to performance measures in patients after total knee arthroplasty. *PM R, 3*(6), 541–549.
6. van Dijk, G. M., Veenhof, C., Lankhorst, G. J., & Dekker, J. (2009). Limitations in activities in patients with osteoarthritis of the hip or knee: The relationship with body functions, comorbidity and cognitive functioning. *Disability and Rehabilitation, 31*(20), 1685–1691.
7. Waehrens, E., Bliddal, H., Danneskiold-Samsoe, B., Lund, H., & Fisher, A. (2012). Differences between questionnaire- and interview-based measures of activities of daily living (ADL) ability and their association with observed ADL ability in women with rheumatoid arthritis, knee osteoarthritis, and fibromyalgia. *Scandinavian Journal of Rheumatology, 41*, 95–102.
8. Wright, A. A., Hegedus, E. J., Baxter, G. D., & Abbott, J. H. (2011). Measurement of function in hip osteoarthritis: Developing a standardized approach for physical performance measures. *Physiotherapy Theory and Practice, 27*(4), 253–262.
9. van den Akker-Scheek, I., Zijlstra, W., Groothoff, J. W., Bulstra, S. K., & Stevens, M. (2008). Physical functioning before and after total hip arthroplasty: Perception and performance. *Physical Therapy, 88*(6), 712–719.
10. Moseley, G. L., & Flor, H. (2012). Targeting cortical representations in the treatment of chronic pain: A review. *Neurorehabilitation and Neural Repair, 26*, 646–652.
11. Bean, J. F., Olveczky, D. D., Kiely, D. K., LaRose, S. I., & Jette, A. M. (2011). Performance-based versus patient-reported physical function: What are the underlying predictors? *Physical Therapy, 91*(12), 1804–1811.
12. Botha-Scheepers, S., Watt, I., Rosendaal, F. R., Breedveld, F. C., Hellio le Graverand, M. P., & Kloppenburg, M. (2008). Changes in outcome measures for impairment, activity limitation, and participation restriction over two years in osteoarthritis of the lower extremities. *Arthritis Care & Research, 59*(12), 1750–1755.
13. Stratford, P. W., Kennedy, D. M., & Riddle, D. L. (2009). New study design evaluated the validity of measures to assess change after hip or knee arthroplasty. *Journal of Clinical Epidemiology, 62*(3), 347–352.
14. Bennell, K., Dobson, F., & Hinman, R. (2011). Measures of physical performance assessments: Self-Paced Walk Test (SPWT), Stair Climb Test (SCT), Six-Minute Walk Test (6MWT), Chair Stand Test (CST), Timed Up & Go (TUG), Sock Test, Lift and Carry Test (LCT), and Car Task. *Arthritis Care Research (Hoboken), 63*(Suppl 11), S350–S370.
15. Kennedy, D., Stratford, P. W., Pagura, S. M., Walsh, M., & Woodhouse, L. J. (2002). Comparison of gender and group differences in self-report and physical performance measures in total hip and knee arthroplasty candidates. *Journal of Arthroplasty, 17*(1), 70–77.
16. Zeni, J, Jr, Abujaber, S., Pozzi, F., & Raisis, L. (2014). Relationship between strength, pain, and different measures of functional ability in patients with end-stage hip osteoarthritis. *Arthritis Care & Research (Hoboken), 66*(10), 1506–1512.
17. Damsgard, E., Thrane, G., Anke, A., Fors, T., & Roe, C. (2010). Activity-related pain in patients with chronic musculoskeletal disorders. *Disability and Rehabilitation, 32*(17), 1428–1437.
18. Harding, G., Parsons, S., Rahman, A., & Underwood, M. (2005). "It struck me that they didn't understand pain": The specialist pain clinic experience of patients with chronic musculoskeletal pain. *Arthritis and Rheumatism, 53*(5), 691–696.
19. Dworkin, R. H., Turk, D. C., Farrar, J. T., Haythornthwaite, J. A., Jensen, M. P., Katz, N. P., et al. (2005). Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. *Pain, 113*(1–2), 9–19.
20. Hawker, G. A., Stewart, L., French, M. R., Cibere, J., Jordan, J. M., March, L., et al. (2008). Understanding the pain experience in hip and knee osteoarthritis–an OARSI/OMERACT initiative. *Osteoarthritis Cartilage, 16*(4), 415–422.
21. Tsai, P. F., & Tak, S. (2003). Disease-specific pain measures for osteoarthritis of the knee or hip. *Geriatric Nursing, 24*(2), 106–109.
22. Klokker, L., Osborne, R. H., Waehrens, E. E., Norgaard, O., Bandak, E., Bliddal, H., et al. (2014). A conceptual model for an activity-based pain measure to monitor and evaluate the effects of knee osteoarthritis treatment. *Osteoarthritis and Cartilage, 22*(Suppl.), 180–181.
23. Herr, K. A., Spratt, K., Mobily, P. R., & Richardson, G. (2004). Pain intensity assessment in older adults: Use of experimental pain to compare psychometric properties and usability of selected pain scales with younger adults. *Clinical Journal of Pain, 20*(4), 207–219.
24. Roos, E. M., Roos, H. P., Lohmander, L. S., Ekdahl, C., & Beynnon, B. D. (1998). Knee injury and osteoarthritis outcome score (KOOS)—Development of a self-administered outcome measure. *Journal of Orthopaedic and Sports Physical Therapy, 28*(2), 88–96.
25. Christensen, R., Bliddal, H., & Henriksen, M. (2013). Enhancing the reporting and transparency of rheumatology research: A guide to reporting guidelines. *Arthritis Research Therapy, 15*(1), 109.
26. Kottner, J., Audige, L., Brorson, S., Donner, A., Gajewski, B. J., Hrobjartsson, A., et al. (2011). Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *Journal of Clinical Epidemiology, 64*(1), 96–106.

27. de Vet, H. C., Terwee, C. B., Knol, D. L., & Bouter, L. M. (2006). When to use agreement versus reliability measures. *Journal of Clinical Epidemiology, 59*(10), 1033–1039.
28. de Vet, H. C., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011). *Measurement in medicine a practical guide*. Cambridge: Cambridge University Press.
29. Fleiss, J. L. (1986). *The design and analysis of clinical experiments*. New York: Wiley.
30. Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
31. Streiner, D. L., & Norman, G. R. (2003). *Health measurement scales: A practical guide to their development and use* (3rd ed.). Oxford: Oxford University Press.
32. Farrar, J. T., Young, J. P, Jr, LaMoreaux, L., Werth, J. L., & Poole, R. M. (2001). Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. *Pain, 94*(2), 149–158.
33. Wright, A. A., Cook, C. E., Baxter, G. D., Dockerty, J. D., & Abbott, J. H. (2011). A comparison of 3 methodological approaches to defining major clinically important improvement of 4 performance measures in patients with hip osteoarthritis. *Journal of Orthopaedic and Sports Physical Therapy, 41*(5), 319–327.
34. Howe, T. E., Dawson, L. J., Syme, G., Duncan, L., & Reid, J. (2011). Evaluation of outcome measures for use in clinical practice for adults with musculoskeletal conditions of the knee: A systematic review. *Manual Therapy, 17*, 100–118.
35. Dobson, F., Hinman, R. S., Hall, M., Terwee, C. B., Roos, E. M., & Bennell, K. L. (2012). Measurement properties of performance-based measures to assess physical function in hip and knee osteoarthritis: A systematic review. *Osteoarthritis Cartilage, 20*(12), 1548–1562.
36. Dekker, J., Dallmeijer, A. J., & Lankhorst, G. J. (2005). Clinimetrics in rehabilitation medicine: Current issues in developing and applying measurement instruments 1. *Journal of Rehabilitation Medicine, 37*(4), 193–201.