Research Report

Interrater and Intrarater Reliability of Common Clinical Standing Balance Tests for People With Hip Osteoarthritis

Yik Ming Choi, Fiona Dobson, Joel Martin, Kim L. Bennell, Rana S. Hinman

Y.M. Choi, DClinPhysio, Centre for Health, Exercise and Sports Medicine, Department of Physiotherapy, School of Health Sciences, The University of Melbourne, Carlton, Victoria, Australia, and Department of Rehabilitative Services, Changi General Hospital, Singapore.

F. Dobson, PhD, Centre for Health, Exercise and Sports Medicine, Department of Physiotherapy, School of Health Sciences, The University of Melbourne.

J. Martin, BAppSc, Centre for Health, Exercise and Sports Medicine, Department of Physiotherapy, School of Health Sciences, The University of Melbourne.

K.L. Bennell, PhD, Centre for Health, Exercise and Sports Medicine, Department of Physiotherapy, School of Health Sciences, The University of Melbourne.

R.S. Hinman, PhD, Centre for Health, Exercise and Sports Medicine, Department of Physiotherapy, Melbourne School of Health Sciences, The University of Melbourne, Alan Gilbert Building, 161 Barry St, Carlton, Victoria, 3053, Australia. Address all correspondence to Dr Hinman at: ranash@unimelb.edu.au.

[Choi YM, Dobson F, Martin J, et al. Interrater and intrarater reliability of common clinical standing balance tests for people with hip osteoarthritis. *Phys Ther.* 2014;94:696–704.]

Published Ahead of Print: February 20, 2014
Accepted: February 13, 2014
Submitted: June 25, 2013

Background. Hip osteoarthritis (OA) is a common musculoskeletal condition affecting older individuals. Clinical balance tests are frequently used to assess standing balance in these people. There is insufficient information regarding the reliability of these tests.

Objective. The aim of this study was to estimate reliability and measurement error of 4 common clinical standing balance tests in people with hip OA.

Design. A prospective study was conducted with repeated measures between 2 independent raters within 1 session and within 1 rater over a 1-week interval.

Methods. Thirty people with hip OA were evaluated. Reliability was estimated for the Four-Square Step Test, Step Test, Functional Reach Test, and Timed Single-Leg Stance Test using intraclass correlation coefficients (ICC [2,1]). Measurement error was expressed as standard error of measurement and minimal detectable change.

Results. The Four-Square Step Test, Step Test, and Timed Single-Leg Stance Test were sufficiently reliable between raters (ICC=.85–.94, lower 1-sided 95% confidence interval [95% CI]=.71–.89), whereas the Step Test (standing on study limb) and Timed Single-Leg Stance Test (standing on nonstudy limb) were sufficiently reliable within a rater over a 1-week interval (ICC=.91, lower 1-sided 95% CI=.80–.83). The Step Test (standing on study limb) and Timed Single-Leg Stance Test (standing on nonstudy limb) achieved optimal levels of reliability (ICC >.90, lower 1-sided 95% CI >.70), with acceptable measurement error (<10%) for clinical outcome measures. The Functional Reach Test was not sufficiently reliable. A ceiling effect was detected for the Timed Single-Leg Stance Test.

Limitations. Reliability was assessed only between 2 raters during a single session and within 1 rater over a 1-week interval, which limits generalizability.

Conclusions. The Step Test (standing on study limb) is recommended as a highly reliable test with acceptable measurement error for assessing standing balance in people with hip OA.

Osteoarthritis (OA) is a common musculoskeletal condition affecting many individuals, especially older people. It typically causes joint pain and a decrease in physical function, thus limiting individual participation in society and leading to a reduction in quality of life.[1,2] In the United States, it has been estimated that nearly 27 million adults aged 25 years and older have symptoms and clinical findings of OA.[3] The hip is one of the most common joints affected by OA. Epidemiological studies show that hip OA affects 7% to 25% of the population aged over 55 years, and this prevalence is expected to increase gradually as the whole population ages.[1,4]

Standing balance is essential for many daily activities such as lower body dressing, ambulating, and stair climbing. Control of balance depends upon sensory input, central processing of afferent input, and coordinated neuromuscular responses to ensure the center of mass remains within the base of support when balance is challenged.[5,6] A variety of symptoms and physical impairments associated with hip OA, including joint pain, muscle weakness, joint stiffness, and sensory dysfunction, can affect balance.[7–9] Not surprisingly, impaired standing balance has been reported in people with hip OA compared with age-matched participants who were healthy[10–13] and is frequently observed by clinicians treating people with hip OA. Importantly, impaired balance is recognized as a risk factor for falls in the older population,[14,15] and falls are frequently reported in people with hip OA,[16] with the majority of falls occurring during ambulation and stair ascent and descent. Thus, assessment of standing balance is an integral component of hip OA management.

Balance may be measured using complex and sophisticated equipment, such as force platforms or posturography systems[11,17,18]; however, such equipment is expensive and impractical for regular use in most clinical settings and in many research settings. For many clinicians and researchers, simple clinical tests are the most practical methods of measuring standing balance in people with hip OA.[19,20] To ensure judicious use of clinical standing balance tests, it is essential to confirm that these tests are reliable, as well as understand the measurement error associated with their use, in the population of interest.[21] However, to date, there is insufficient evidence regarding the clinimetric properties of clinical standing balance tests in people with hip OA.[22] Our recent systematic review, which synthesized evidence on clinimetric properties of observer-rated impairment tests (including balance tests) in people with hip and groin problems,[22] failed to identify a single study investigating the reliability (or any clinimetric property) of balance tests for hip OA. This remarkable dearth of literature evaluating measurement properties of balance tests in people with hip OA is concerning, given that such tests are frequently used in the clinical setting and to assess treatment outcomes in clinical trials.[20,23,24]

The primary aim of this study was to estimate the reliability of 4 common clinical balance tests in people with hip OA: Four-Square Step Test, Step Test, Functional Reach Test (FRT), and Timed Single-Leg Stance Test. A secondary aim was to estimate the amount of measurement error associated with each test.

Method

In this study, *between-rater reliability* refers to repeated measures between 2 independent raters within a session, and *within-rater reliability* refers to repeated measures by a rater over a 1-week interval. As such, both designs also include an element of test-retest reliability.

Participants

Volunteers were sourced from a database of research volunteers from the community maintained by the Centre for Health, Exercise and Sports Medicine, Department of Physiotherapy, The University of Melbourne. To be eligible, participants were required to fulfill the following inclusion criteria based on clinical diagnostic criteria for hip OA established by the American College of Rheumatology[25]: (1) age >50 years; (2) hip pain on most days of the previous month; and (3) at least one of the following radiological or clinical presentations: presence of joint space narrowing and osteophytes on hip radiographs taken in the previous year, hip internal rotation of <15 degrees and hip flexion of ≤115 degrees, and hip internal rotation of ≥15 degrees in the presence of pain and morning stiffness of the hip for ≤60 minutes. Participants also were required to be able to ambulate independently in the community and read and follow instructions in English. Participants were not eligible if they: (1) had previous hip or knee joint replacement; (2) had any hip surgery in the previous 6 months; (3) had other muscular, joint, or neurological conditions causing pain and dysfunction of lower limbs; or (4) used any form of walking aid. All participants provided written informed consent.

Procedure

Participants were tested on 2 occasions (approximately 1 week apart). At the first test session, participants performed the balance tests with 2 independent raters (rater A and rater B) to examine between-rater reliability. The testing order of both the raters and the balance tests was ran-

domized using a computerized random number generator. Participants were given 5 minutes' rest between each rater's independent assessments. At the second test session, participants repeated the balance tests with the more experienced rater A (who was blinded to the results from session 1) to examine within-rater reliability. A 1-week test interval was used to provide sufficient time to limit recall of test scores, but it was short enough to limit potential real change in clinical status. At session 2, participants completed a self-report global rating of change. This measure was used as a reference standard for stability and determined whether any substantial change in the participant's hip condition had occurred between test sessions.

Assessment of hip OA symptoms. As both lower limbs were assessed during the balance testing, the most painful hip was defined as the *study limb*, and the least painful (for bilateral disease) or nonpainful hip was defined as the *nonstudy limb*. A visual analog scale (VAS) was used to assess the average level of hip pain over the previous week. Participants were asked to mark an "X" on a 100-mm line, anchored with "no pain" on the left and "worst pain possible" on the right. The distance (in millimeters) from the left anchor to the X mark was then measured, with higher VAS scores indicating more severe pain.[26] The VAS has demonstrated reliability in people with OA.[27]

The Hip Dysfunction and Osteoarthritis Outcome Score (HOOS) was used to assess patient-reported symptoms and disability related to hip OA.[28] It consists of 40 items over 5 subscales: pain (10 items), other symptoms (5 items), function in daily living (17 items), function in sports and recreation (4 items), and hip-related quality of life (4 items).[29,30] All items are answered on a 5-point Likert scale, and a total score is calculated, ranging from 0 ("no disability") to 100 ("extreme disability").[29,30] The HOOS has demonstrated reliability in people with hip OA.[30]

A global change scale (GCS) was used to assess self-reported change in hip pain and physical function across the 2 testing sessions. The GCS was measured on a 5-point adjectival scale ("much worse," "slightly worse," "no change," "slightly better," and "much better"). Participants who recorded "much better" or "much worse" were excluded from the within-rater analyses. Some studies have previously used these scales to determine changes in participants' conditions, where "minimal or slight changes" were defined as nonmeaningful change.[31–33] The GCS has been shown to be highly reliable in people with musculoskeletal dysfunction.[34,35]

Assessment of balance. Participants were tested barefooted on each of the 4 clinical balance tests.

In the Four-Square Step Test,[36] 4 walking sticks were placed on the floor at right angles with handles outward to form 4 squares. Participants started in square 1, facing square 2, and remained facing this direction for the duration of the test. Participants then stepped forward with both feet as quickly as possible into square 2, then sideways to the right into square 3, then backward into square 4, and finally sideways to the left back into square 1. They then reversed the sequence back to the starting position. A demonstration was provided, and an initial practice was performed, immediately followed by 2 test trials. According to original published instructions for the test, the faster of the 2 trials was recorded to the nearest 10th of a second.

For the Step Test,[37] a 15-cm height step was used with a 5-cm-wide cardboard template positioned on the floor along the edge of the step to provide a standardized starting position. The test was performed standing on the study leg the entire time, while the other leg was moved back and forth from the step to the floor (eg, the stepping foot was placed flat up onto the step, then back down flat onto the ground) as many times as possible in 15 seconds without overbalancing (moving the stance leg from the start position). A demonstration was provided, and 3 or 4 practice steps were performed, immediately followed by 1 test trial standing on each leg. The number of whole steps (up and back down to a flat position on the floor) performed in 15 seconds was recorded for each standing leg. If participants overbalanced, the test was concluded, and the number of completed steps and the time taken were recorded.

The FRT consisted of 2 types of tests: (1) forward reach and (2) lateral reach. In the forward reach test,[38] participants started in a normal relaxed stance with their dominant arm facing side-on, but not touching, a wall. A leveled measuring tape was then mounted on the wall at the acromion height. Participants made a fist with the dominant hand and elevated the arm to 90 degrees (ie, shoulder level). The position of the third knuckle (metacarpophalangeal joint) along the tape was recorded as the starting point. Keeping the contralateral arm by the side and both heels on the floor, participants reached as far forward as possible to maintain a maximal reach position for 3 seconds without losing balance (such as taking a step, leaning on the wall, or needing to be assisted by the rater). The final reach position of the third knuckle along the tape was

recorded as the finishing point. A demonstration was provided, immediately followed by 3 test trials. According to original published instructions for the test, the mean difference between the starting and finishing points across the 3 trials was recorded to the nearest millimeter as the test score.

In the lateral reach test,[39] participants started in a normal relaxed stance with their back facing, but not touching, a wall. A leveled measuring tape was then mounted on the wall at the acromion height. Participants abducted 1 arm to 90 degrees (ie, shoulder level) with all fingers extended. The position of the tip of the third finger along the tape was recorded as the starting point. Keeping the contralateral arm by the side and both heels on the floor, participants reached as far sideways as possible to maintain a maximal reach position for 3 seconds without losing their balance, taking a step, or leaning on the wall. Knee flexion and trunk flexion and rotation were not permitted. Participants were instructed not to bend at the knees or at the trunk. If bending at the knees or trunk occurred during testing, the test was stopped immediately and corrected. A re-trial was then conducted. The final position of the tips of the third fingers along the tape was recorded as the finishing point. A demonstration was provided, immediately followed by 3 test trials on each side. The mean difference between the starting and finishing points across the trials for each side was recorded to the nearest millimeter as the test score. A reach in the direction of the study hip was defined as the *ipsilateral reach*, and a reach away from the study hip was defined as the *contralateral reach*.

Participants started the Timed Single-Leg Stance Test[40] with their hands on their hips and stood on 1 leg for as long as possible up to a maximum of 30 seconds. The nonstance hip remained in a neutral position with the knee flexed so that the foot was positioned behind and was not permitted to touch the stance leg. Participants were encouraged to look at a nonmoving target 1 to 3 m ahead. The test was stopped if participants moved their hands off their hips, touched the nonstance foot down on the floor, or touched the stance leg with the nonstance leg. A demonstration was provided, followed immediately by 2 test trials on each leg (based on original published instructions). The longest time, up to a maximum of 30 seconds, of the 2 trials on each leg was recorded to the nearest 10th of a second as the test score for each leg.

Data Analysis

Data analyses were performed using the IBM SPSS 21 statistical package for Windows (IBM Corp, Armonk, New York). Data were checked for normality and for systematic differences between test sessions. Descriptive analyses were conducted across raters and sessions, including means, standard deviations, and ranges of scores. Percentages of maximal scores (ceiling effects) also were calculated for the Timed Single-Leg Stance Test because the score for this test is capped at 30 seconds.

Within-rater and between-rater reliability were each calculated using intraclass correlation coefficients (ICC [2,1]) with 95% confidence intervals (95% CIs) for a 2-way random effects model and absolute agreement. Interpretation of ICC values was based on published recommendations,[21] where values higher than .75 indicate sufficient reliability and values higher than .90 indicate optimal reliability.[21,41] Furthermore, 95% CI values were inspected to ensure that lower 1-sided 95% CI values met a recommended minimum acceptable level, which was set at .70.[41–43]

Measurement error was expressed as the standard error of measurement (SEM) and minimal detectable change (MDC). The SEM was calculated as the square root of the mean square error term from the analysis of variance. The MDC at the 90% confidence level (MDC_{90}) was calculated as SEM $\times$ 1.65 (z score of 90% interval) $\times \sqrt{2}$. For both the SEM and MDC_{90}, 95% CIs were calculated according to recommended methods.[44]

As the units of measurement for the 4 balance tests varied, SEM and MDC_{90} also were expressed as SEM percentage (SEM%) and MDC percentage (MDC%) to assist with interpretation of the results. These values were defined as the SEM and MDC divided by the mean of all testing scores on the 2 test sessions and were calculated as SEM% = (SEM/mean) $\times$ 100 and MDC% = (MDC_{90}/mean) $\times$ 100.[42,45,46]

Sample Size

Sample size calculations were based on a priori set levels of optimal and minimal acceptable limits of reliability for clinical measurement. As such, a minimum of 19 participants were required to achieve an optimal ICC of .90 and a minimal acceptable lower 1-sided 95% CI of .70 at a power of 80%.[47] In this study, 30 participants were recruited to allow for any potential dropouts and the exclusion of data from participants who reported a meaningful change in their condition across sessions.

Results

Thirty people with hip OA (18 female [60%], 12 male [40%]; mean age=63.3 years, SD=5.71, range=50–75) participated. Descriptive characteristics of the participants are summarized in Table 1. In this cohort of participants, there were

Table 1.
Participant Characteristics (N=30)[a]

Characteristic	Data
Age (y)	63.3 (5.71)
Sex, n (%)	
Female	18 (60)
Male	12 (40)
BMI (kg/m^2)	26.8 (3.9)
Duration of symptoms (y)	5.9 (8.1)
Right-sided study limb, n (%)	17 (56.7)
Right leg dominant, n (%)[b]	26 (86.7)
Bilateral symptoms, n (%)	10 (33.3)
History of falls, n (%)	6 (20.0)
Frequency of falls (n)[c]	3.0 (2.1)
Test-retest interval (d)	7.0 (0.3)
Hip pain (VAS) (mm)	40.9 (18.7)
HOOS	
Pain	63.2 (13.1)
Other symptoms	65.5 (12.0)
Activities of daily living	67.0 (13.4)
Sports	53.8 (17.9)
Quality of life	50.6 (14.7)

[a] Data are presented as mean (SD), unless otherwise indicated. BMI=body mass index, VAS=visual analog scale, HOOS=Hip Dysfunction and Osteoarthritis Outcome Score.
[b] Self-reported leg used to kick a ball.
[c] Number of falls sustained in the previous 12 months for participants with history of falls.

more women than men, and most of the participants were overweight (body mass index >25 kg/m^2). One-third reported bilateral symptoms. Most had not sustained a fall in the previous 12 months. In addition, most participants reported a moderate level of hip pain and disability according to VAS and HOOS scores.

Within-rater reliability was based on data from 27 participants, as 2 participants were unable to return for session 2 and a further participant reported substantial change in hip pain ("much worse") at session 2 and was excluded from further analysis. The within-rater reliability test interval was 7 days for most participants (25/27) and was 6 days and 8 days for the remaining 2 participants. There was no missing data, and no adverse events occurred at any testing occasion. The majority of data were normally distributed. There were systematic differences for the Four-Square Step Test and Step Test within rater A over the 1-week interval and for the forward reach part of the FRT between raters A and B within the single session ($P<.05$).

Between-Rater Reliability on 2 Test Occasions Within a Single Session

Balance test scores between raters for all 30 participants at session 1, along with the percentages of maximal scores for the Timed Single-Leg Stance Test and ICCs, are presented in Table 2. The Four-Square Step Test, Step Test, and Timed Single-Leg Stance Test were sufficiently reliable between raters (ICC=.85–.94, lower 1-sided 95% CI=.71–.89). Further inspection of the point estimates and confidence limits demonstrated that the Step Test (study limb) and the Timed Single-Leg Stance Test also met the optimal level of reliability (ICC >.90, lower 1-sided 95% CI >.70).

Within-Rater Reliability of Repeated Measures Over a 1-Week Interval

Balance test scores for 27 participants assessed by rater A during session 1 and session 2, along with the percentages of maximal scores for the Timed Single-Leg Stance Test and ICCs, are presented in Table 3. The Step Test (study limb) and Timed Single-Leg Stance Test (nonstudy limb) were sufficiently reliable within 1 rater over a 1-week interval and met the optimal levels of reliability (ICC=.91, lower 1-sided 95% CI=.80–.83).

Ceiling and Floor Effects

Inspection of minimum and maximum scores (Tabs. 2 and 3) showed a consistent ceiling effect for the Timed Single-Leg Stance Test. Approximately half of the participants (44%–57%) were able to perform the Timed Single-Leg Stance Test with maximal holding of 30 seconds at each test occasion.

Measurement Error

The SEM, SEM%, MDC_{90}, and MDC% between raters at session 1 and within 1 rater over a 1-week interval are provided in Tables 2 and 3, respectively. The SEM of the tests between raters varied between 7.4% and 16.1% of the test score, whereas it varied between 9.0% and 21.2% of the test score when repeatedly measured by 1 rater over a 1-week interval. The Step Test (study limb) and Four-Square Step Test had sufficiently low measurement error (<10% of the test score) for both situations, whereas the Timed Single-Leg Stance Test showed the largest measurement error (>14%) in both situations.

Discussion

In this study, we aimed to estimate the reliability and measurement error associated with 4 clinical standing balance tests in a cohort of people with symptomatic hip OA. We found that the Four-Square Step Test, Step Test, and Timed Single-Leg Stance Test were sufficiently reliable between raters within a session, whereas the Step Test (study limb) and Timed Single-Leg Stance Test (nonstudy limb) were sufficiently reliable within 1 rater over a 1-week interval. The Step Test (study limb) and Timed Single-Leg Stance Test (nonstudy limb) achieved optimal levels of reliability in both situations, but only the Step Test (study limb) also had sufficiently low measurement error to be confident of a measured value in the clinical situation. In view of the larger amount of measurement error and our observed ceiling effect for the Timed Single-Leg Stance Test, this test may be a

less useful measure of standing balance for people with hip OA, despite being a reliable test. Furthermore, the FRT subtests were not sufficiently reliable either between or within raters, and the larger amount of measurement error associated with these tests limits the confidence in a measured value and the usefulness of these tests in the clinical setting. Thus, our findings suggest that the Step Test (standing on most affected limb) is the most useful clinical test of standing balance in hip OA, as it is highly reliable with sufficiently low measurement error.

Due to the paucity of earlier research in this area, and because this is the first study, to our knowledge, to estimate reliability of balance tests in hip OA, it is difficult to discuss our findings in relation to previous research. However, our findings are generally in agreement with those of a study that evaluated the reliability of balance measurements in patients with hip fracture.[48] In that study, Sherrington and Lord[48] found good test-retest reliability for the Step Test, with similar levels of reliability (ICC=.85–.92) and lower 95% CI values (.71–.83) compared with those found in the current study. In contrast, our findings are quite different from those of an earlier study that evaluated interrater reliability of a battery of tests, including the Timed Single-Leg Stance Test, in patients following surgically fixed hip fractures.[49] In that study, the Timed Single-Leg Stance Test was one of the least reliable tests, and reliability estimates were much lower (kappa=.14–.63) than those found in the current study. To our knowledge, no reliability estimates for the FRT or the Four-Square Step Test in a comparable group have been conducted.

Measurement errors associated with the 4 balance tests, which have not previously been reported, also were estimated in the current study. This

Table 2.
Between-Rater Reliability: Balance Test Scores, Intraclass Correlation Coefficients (ICCs), Standard Errors of Measurement (SEMs), and Minimal Detectable Change at the 90% Level of Confidence (MDC_{90}) Across the 2 Raters at Session 1 (N=30)[a]

Test	Rater A		Rater B		ICC (95% CI)	Lower 1-Sided 95% CI	SEM (95% CI)	SEM %	MDC_{90} (95% CI)	MDC %
	$\bar{X}$ (SD)	Range	$\bar{X}$ (SD)	Range						
Four-Square Step Test (s)	8.97 (2.32)	4.97–15.69	8.56 (2.01)	5.84–15.69	.86 (.72–.93)	.75	0.77 (0.65–1.04)	8.8	1.80 (1.43–2.42)	20.5
Step Test (no. of steps)										
Standing on nonstudy limb	13.40 (4.02)	5–21	12.83 (3.79)	4–20	.85 (.71–.93)	.74	1.48 (1.18–1.99)	11.3	4 (2.75–4.65)	26.4
Standing on study limb	14.63 (4.63)	5–27	14.13 (4.33)	5–24	.94 (.88–.97)	.89	1.06 (0.85–1.43)	7.4	3 (1.97–3.33)	17.2
Functional Reach Test (cm)										
Forward reach	28.74 (6.84)	15.67–45.67	25.39 (6.96)	10.67–38.83	.68 (.29–.85)	.36	3.43 (2.73–4.61)	12.7	8.0 (6.37–10.76)	29.6
Ipsilateral reach	16.57 (3.20)	8.67–21.5	15.47 (3.81)	8.67–22.17	.62 (.34–.80)	.38	2.12 (1.69–2.85)	13.2	4.9 (3.94–6.65)	30.9
Contralateral reach	16.09 (4.49)	5.33–26.67	16.24 (4.85)	4.83–26.67	.74 (.53–.90)	.57	2.36 (1.88–3.17)	14.6	5.5 (4.39–7.40)	34.1
Timed Single-Leg Stance Test (s)										
Standing on nonstudy limb	22.65 (9.68)	3.40–30 (53.3%)[b]	21.20 (10.84)	3.09–30 (56.7%)[b]	.90 (.80–.95)	.82	3.12 (2.48–4.19)	14.4	7.27 (5.79–9.78)	33.6
Standing on study limb	21.26 (10.30)	3.38–30 (50.0%)[b]	21.65 (10.15)	2.28–30 (46.7%)[b]	.89 (.78–.95)	.80	3.46 (2.76–4.66)	16.1	8.08 (6.44–10.87)	37.7

[a] ICC=intraclass correlation coefficient, SEM=standard error of measurement, MDC_{90}=minimal detectable change at the 90% level of confidence, SEM%=standard error of measurement percentage, MDC%=minimal detectable change percentage.
[b] Percentage of participants who scored the maximum possible score of 30 seconds in these tests.

Table 3.
Within-Rater Reliability: Balance Test Scores, Intraclass Correlation Coefficients (ICCs), Standard Errors of Measurement (SEMs), and Minimal Detectable Change at the 90% Level of Confidence (MDC_{90}) Across the 2 Test Sessions (n=27)[a]

Test	Session 1		Session 2		ICC (95% CI)	Lower 1-Sided 95% CI	SEM (95% CI)	SEM %	MDC_{90} (95% CI)	MDC %
	$\bar{X}$ (SD)	Range	$\bar{X}$ (SD)	Range						
Four-Square Step Test (s)	9.07 (2.35)	4.97–15.69	8.31 (2.45)	5.12–17.19	.83 (.57–.93)	.62	0.86 (0.68–1.17)	9.9	2.00 (1.58–2.72)	23.0
Step Test (no. of steps)										
Standing on nonstudy limb	13.5 (4.07)	5–21	15.18 (4.40)	5–24	.81 (.42–.93)	.51	1.54 (1.22–2.10)	10.7	4 (2.84–4.89)	25.1
Standing on study limb	14.71 (4.74)	5–27	15.68 (4.74)	5–25	.91 (.77–.96)	.80	1.37 (1.08–1.86)	9.0	3 (2.52–4.34)	21.0
Functional Reach Test (cm)										
Forward reach	28.67 (7.05)	15.67–45.67	28.02 (7.88)	13–43	.68 (.42–.84)	.47	4.26 (3.36–5.79)	15.0	9.9 (7.85–13.52)	35.0
Ipsilateral reach	16.58 (3.31)	8.67–21.5	16.86 (3.77)	11–23	.64 (.35–.82)	.41	2.15 (1.70–2.93)	12.9	5.0 (3.97–6.83)	30.0
Contralateral reach	16.14 (4.63)	5.33–26.67	16.85 (3.94)	8–26	.73 (.50–.86)	.54	2.23 (1.76–3.04)	13.5	5.2 (4.11–7.08)	31.5
Timed Single-Leg Stance Test (s)										
Standing on nonstudy limb	22.12 (9.82)	3.40–30 (48.1%)[b]	21.48 (10.31)	3–30 (44.4%)[b]	.91 (.81–.96)	.83	3.08 (2.43–4.19)	14.7	7.2 (5.67–9.77)	34.2
Standing on study limb	20.63 (10.39)	3.38–30 (44.4%)[b]	21.31 (10.91)	3–30 (44.4%)[b]	.82 (.64–.91)	.68	4.62 (3.65–6.29)	21.2	10.78 (8.52–14.67)	51.4

[a] ICC=intraclass correlation coefficient, SEM=standard error of measurement, MDC_{90}=minimal detectable change at the 90% level of confidence, SEM%=standard error of measurement percentage, MDC%=minimal detectable change percentage.
[b] Percentage of participants who scored the maximum possible score of 30 seconds in these tests.

information assists with the interpretation of and confidence in an obtained measure. For a measure to be clinically useful, it must have a sufficiently high ICC and sufficiently low SEM. We also calculated the SEM% and MDC% so that tests could be compared, given that the units of measurement varied across the tests. In the current study, the Step Test and Four-Square Step Test were found to have lower SEM% and MDC% values than the FRT and Timed Single-Leg Stance Test. This finding means that, compared with the FRT and Timed Single-Leg Stance Test, smaller amounts of change are required on the Step Test and Four-Square Step Test to be confident that a real change in balance has occurred. To be confident of real change in balance when applying these tests in individuals with hip OA, clinicians and researchers should aim to see a change of 3 steps on the Step Test (standing on the affected side), 2 seconds on the Four-Square Step Test, 9.9 cm on the forward reach component of the FRT, (5.0 and 5.2 cm for ipsilateral and contralateral functional reach, respectively), and 10.8 seconds on the Timed Single-Leg Stance Test.

Our study had a number of strengths, including the robust sample size that was adequately powered to detect our a priori optimal level of reliability, inclusion of a range of commonly used clinical balance tests, and exclusion of participants with a change in clinical state from the within-rater analysis. Importantly, we also determined the measurement error associated with the balance tests, which will enable clinicians and researchers to interpret change in balance scores across time with respect to real change.

There were some limitations to the current study. Given a participant's global rating of change and balance performance may not be indepen-

dent, and thus the potential for correlated error, it is possible our estimates of reliability were inflated somewhat. Results might have been different if participants with a change in their clinical condition were included in the analyses. As both our between-rater and within-rater analyses also included a component of test-retest reliability, the additional source of error resulting from potential differences in participants' performance across the repeated measures may have increased the measurement error estimates for these clinical tests. Indeed, as systematic differences for the Four-Square Step Test and Step Test were found over the 1-week interval, it is possible these errors were not only due to rater error but also represent altered performance by the participant between sessions.

Only 2 raters were used for evaluating between-rater reliability, which may limit the generalizability of our findings to a wider pool of raters with different abilities and clinical backgrounds. However, we did choose raters from different professional backgrounds (rater A was a clinical physical therapist, and rater B was a researcher with a human movement science background) and with different levels of experience in assessing older patients with pathology, which helps to increase the generalizability of our findings. Additionally, only 1 rater was used for evaluating within-rater reliability. Although this rater was a physical therapist, and thus improves the generalizability of the findings to clinicians, inclusion of additional raters would have strengthened the study. Although our cohort of participants with hip OA were all community recruits, representing at most a moderate level of disease severity based on symptomatic data, it is not clear whether the present findings apply to participants who are not community-dwelling or to patients with end-stage disease awaiting arthroplasty.

Future research is needed to provide comprehensive data about the clinimetric properties for clinical balance tests in people with hip OA. In particular, evaluations of the validity and responsiveness of these tests are needed. Information about the minimal clinically important difference is needed so that researchers and clinicians can determine what amount of change in the balance tests is required with interventions in order to achieve meaningful clinical improvements in health status for the patient. Although we have determined the MDC, which tells clinicians and researchers the amount of change needed to be sure of a *real* change beyond that associated with measurement error, it is not necessarily the same as the minimal clinically important difference. Although a third of our participants in this study had bilateral hip OA, a subgroup analysis of these participants was not performed because the study was not powered sufficiently for such an analysis. However, as two-thirds (n=20) of the participants had unilateral hip OA, a post hoc subanalysis with sufficient power revealed that reliability estimates for unilateral hip OA were approximately the same as those for the entire sample. Furthermore, interpretation of these values based on a priori criteria was no different from the interpretation of the values of the group as a whole. As estimates may differ for those with bilateral disease, we recommend that future research is needed to examine the reliability of balance tests within this subgroup.

In conclusion, this study provides estimates of reliability and measurement error of 4 clinical standing balance tests in a cohort of 30 participants with hip OA. Only the Step Test (standing on the affected side) and the Timed Single-Leg Stance Test demonstrated optimal levels of reliability for clinical measurement tests. When measurement error and ceiling effects also are considered, our data suggest the Step Test (standing on the affected side) is the most useful clinical measure of standing balance for people with hip OA. Further research is needed to determine the responsiveness and, in particular, the minimal clinically important difference, for these tests.

Dr Choi, Dr Dobson, Dr Bennell, and Dr Hinman provided concept/idea/research design and writing. Dr Choi and Mr Martin provided data collection. Dr Choi, Dr Dobson, and Dr Hinman provided data analysis and project management. Dr Dobson, Mr Martin, and Dr Bennell provided consultation (including review of manuscript before submission).

This prospective reliability study received ethics approval from The University of Melbourne Ethics Committee.

This research was funded by National Health and Medical Research Council Program Grant 631717. Dr Bennell was partly funded by an Australian Research Council Future Fellowship. Dr Choi was funded by the Singapore Ministry of Health Reinvestment Fund.

DOI: 10.2522/ptj.20130266

References

1 Dagenais S, Garbedian S, Wai EK. Systematic review of the prevalence of radiographic primary hip osteoarthritis. *Clin Orthop Relat Res.* 2009;467:623–637.

2 Salaffi F, Carotti M, Stancati A, Grassi W. Health-related quality of life in older adults with symptomatic hip and knee osteoarthritis: a comparison with matched healthy controls. *Aging Clin Exp Res.* 2005;17:255–263.

3 Lawrence RC, Felson DT, Helmick CG, et al. Estimates of the prevalence of arthritis and other rheumatic conditions in the United States, part II. *Arthritis Rheum.* 2008;58:26–35.

4 Zhang Y, Jordan JM. Epidemiology of osteoarthritis. *Rheum Dis Clin North Am.* 2008;34:515–529.

5 Horak FB, Shupert CL, Mirka A. Components of postural dyscontrol in the elderly: a review. *Neurobiol Aging.* 1989;10:727–738.

6 Massion J. Postural control system. *Curr Opin Neurobiol.* 1994;4:877–887.

7 Kosek E, Ordeberg G. Abnormalities of somatosensory perception in patients with painful osteoarthritis normalize following successful treatment. *Eur J Pain.* 2000;4:229–238.

8 Loureiro A, Mills PM, Barrett RS. Muscle weakness in hip osteoarthritis: a systematic review. *Arthritis Care Res (Hoboken).* 2013;65:340–352.

9 Bijlsma JW, Berenbaum F, Lafeber FP. Osteoarthritis: an update with relevance for clinical practice. *Lancet.* 2011;377: 2115–2126.

10 Kiss R. Effect of the degree of hip osteoarthritis on equilibrium ability after sudden changes in direction. *J Electromyogr Kinesiol.* 2010;20:1052–1057.

11 Giemza C, Ostrowska B, Matczak-Giemza M. The effect of physiotherapy training programme on postural stability in men with hip osteoarthritis. *Aging Male.* 2007; 10:67–70.

12 Nantel J, Termoz N, Centomo H, et al. Postural balance during quiet standing in patients with total hip arthroplasty and surface replacement arthroplasty. *Clin Biomech (Bristol, Avon).* 2008;23:402–407.

13 Tateuchi H, Ichihashi N, Shinya M, Oda S. Anticipatory postural adjustments during lateral step motion in patients with hip osteoarthritis. *J Appl Biomech.* 2011;27: 32–39.

14 Robbins AS, Rubenstein LZ, Josephson KR, et al. Predictors of falls among elderly people: results of two population-based studies. *Arch Intern Med.* 1989;149:1628–1633.

15 Stalenhoef PA, Diederiks JP, Knottnerus JA, et al. A risk model for the prediction of recurrent falls in community-dwelling elderly: a prospective cohort study. *J Clin Epidemiol.* 2002;55:1088–1094.

16 Arnold CM, Faulkner RA. The history of falls and the association of the timed up and go test to falls and near-falls in older adults with hip osteoarthritis. *BMC Geriatr.* 2007;7:17.

17 Rasch A, Dalen N, Berg HE. Muscle strength, gait, and balance in 20 patients with hip osteoarthritis followed for 2 years after THA. *Acta Orthop.* 2010;81:183–188.

18 Arokoski JP, Leinonen V, Arokoski MH, et al. Postural control in male patients with hip osteoarthritis. *Gait Posture.* 2006;23:45–50.

19 Arnold CM, Faulkner RA. The effect of aquatic exercise and education on lowering fall risk in older adults with hip osteoarthritis. *J Aging Phys Act.* 2010;18:245–260.

20 Hinman RS, Heywood SE, Day AR. Aquatic physical therapy for hip and knee osteoarthritis: results of a single-blind randomized controlled trial. *Phys Ther.* 2007;87:32–43.

21 Portney LG, Watkins MP. *Foundations of Clinical Research: Applications to Practice*. 3rd ed. Upper Saddle River, NJ: Pearson/Prentice Hall; 2009.

22 Dobson F, Choi YM, Hall M, Hinman RS. Clinimetric properties of observer-assessed impairment tests used to evaluate hip and groin impairments: a systematic review. *Arthritis Care Res (Hoboken).* 2012;64:1565–1575.

23 Hale LA, Waters D, Herbison P. A randomized controlled trial to investigate the effects of water-based exercise to improve falls risk and physical function in older adults with lower-extremity osteoarthritis. *Arch Phys Med Rehabil.* 2012;93:27–34.

24 Bennell KL, Egerton T, Pua YH, et al. Efficacy of a multimodal physiotherapy treatment program for hip osteoarthritis: a randomised placebo-controlled trial protocol. *BMC Musculoskelet Disord.* 2010;11:238.

25 Altman RD, Alarcon G, Appelrouth D, et al. The American College of Rheumatology criteria for the classification and reporting of osteoarthritis of the hip. *Arthritis Rheum.* 1991;34:505–514.

26 Kahl C, Cleland JA. Visual analogue scale, numeric pain rating scale and the McGill Pain Questionnaire: an overview of psychometric properties. *Phys Ther Rev.* 2005;10:123–128.

27 Bellamy N. Osteoarthritis clinical trials: candidate variables and clinimetric properties. *J Rheumatol.* 1997;24:768–778.

28 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; 44:1186–1196.

29 Nilsdotter AK, Lohmander LS, Klassbo M, Roos EM. Hip Disability and Osteoarthritis Outcome Score (HOOS): validity and responsiveness in total hip replacement. *BMC Musculoskelet Disord.* 2003;4:10.

30 Klassbo M, Larsson E, Mannevik E. Hip disability and osteoarthritis outcome score: an extension of the Western Ontario and McMaster Universities Osteoarthritis Index. *Scand J Rheumatol.* 2003; 32:46–51.

31 Cleland JA, Childs JD, Whitman JM. Psychometric properties of the Neck Disability Index and Numeric Pain Rating Scale in patients with mechanical neck pain. *Arch Phys Med Rehabil.* 2008;89:69–74.

32 Perera S, Mody SH, Woodman RC, Studenski SA. Meaningful change and responsiveness in common physical performance measures in older adults. *J Am Geriatr Soc.* 2006;54:743–749.

33 Wright AA, Cook CE, Baxter GD, et al. A comparison of 3 methodological approaches to defining major clinically important improvement of 4 performance measures in patients with hip osteoarthritis. *J Orthop Sports Phys Ther.* 2011;41: 319–327.

34 Costa LO, Maher CG, Latimer J, et al. Clinimetric testing of three self-report outcome measures for low back pain patients in Brazil: which one is the best? *Spine.* 2008; 33:2459–2463.

35 Kamper SJ, Ostelo RW, Knol DL, et al. Global perceived effect scales provided reliable assessments of health transition in people with musculoskeletal disorders, but ratings are strongly influenced by current status. *J Clin Epidemiol.* 2010;63: 760–766, e761.

36 Dite W, Temple VA. A clinical test of stepping and change of direction to identify multiple falling older adults. *Arch Phys Med Rehabil.* 2002;83:1566–1571.

37 Hill KD, Bernhardt J, McGann AM, et al. A new test of dynamic standing balance for stroke patients: reliability, validity and comparison with healthy elderly. *Physiother Can.* 1996;48:257–262.

38 Duncan PW, Weiner DK, Chandler J, Studenski SA. Functional reach: a new clinical measure of balance. *J Gerontol.* 1990; 45:M192–M197.

39 Brauer S, Burns Y, Galley P. Lateral reach: a clinical measure of medio-lateral postural stability. *Physiother Res Int.* 1999;4:81–88.

40 Bohannon RW, Larkin PA, Cook AC, et al. Decrease in timed balance test scores with aging. *Phys Ther.* 1984;64:1067–1070.

41 de Vet HC, Terwee CB, Mokkink LB, Knol DL. *Measurement in Medicine: A Practical Guide*. New York, NY: Cambridge University Press; 2011.

42 Goldberg A, Casby A, Wasielewski M. Minimum detectable change for single-leg-stance-time in older adults. *Gait Posture.* A2011;33:737–739.

43 Scholtes VA, Terwee CB, Poolman RW. What makes a measurement instrument valid and reliable? *Injury.* 2011;42:236–240.

44 Stratford PW, Goldsmith CH. Use of the standard error as a reliability index of interest: an applied example using elbow flexor strength data. *Phys Ther.* 1997;77: 745–750.

45 Flansbjer UB, Holmback AM, Downham D, et al. Reliability of gait performance tests in men and women with hemiparesis after stroke. *J Rehabil Med.* 2005;37:75–82.

46 Huang SL, Hsieh CL, Wu RM, et al. Minimal detectable change of the Timed "Up & Go" Test and the Dynamic Gait Index in people with Parkinson disease. *Phys Ther.* 2011;91:114–121.

47 Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. *Stat Med.* 1998;17:101–110.

48 Sherrington C, Lord SR. Reliability of simple portable tests of physical performance in older people after hip fracture. *Clin Rehabil.* 2005;19:496–504.

49 Fox KM, Felsenthal G, Hebel JR, et al. A portable neuromuscular function assessment for studying recovery from hip fracture. *Arch Phys Med Rehabil.* 1996;77: 171–176.