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 and Rana S. Hinman
PHYS THER. Published online February 20, 2014

The online version of this article, along with updated information and services, can be found online at: http://ptjournal.apta.org/content/early/2014/02/20/ptj.20130266

Collections
This article, along with others on similar topics, appears in the following collection(s):
- Balance
- Injuries and Conditions: Hip
- Osteoarthritis
- Osteoarthritis
- Tests and Measurements

E-mail alerts
Sign up here to receive free e-mail alerts

Online First articles are published online before they appear in a regular issue of Physical Therapy (PTJ). PTJ publishes 2 types of Online First articles:

Author manuscripts: PDF versions of manuscripts that have been peer-reviewed and accepted for publication but have not yet been copyedited or typeset. This allows PTJ readers almost immediate access to accepted papers.

Page proofs: edited and typeset versions of articles that incorporate any author corrections and replace the original author manuscript.
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, 2 Simei St 3, 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.


© 2014 American Physical Therapy Association

Published Ahead of Print: XXX
Accepted: February 14, 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.** To estimate reliability and measurement error of four common clinical standing balance tests in people with hip OA.

**Design.** Prospective study with repeated measures between two independent raters within one session and within one rater over a one-week interval.

**Methods.** Thirty people with hip OA were evaluated. Reliability was estimated for the four-square step test, step test, functional reach and test timed single leg stance using intra-class 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 were sufficiently reliable between raters (ICC 0.85-0.94, lower 1-sided 95%CI: 0.71-0.89), whereas the step test (standing on study limb) and timed single leg stance (standing on non-study limb) were sufficiently reliable within a rater over a week interval (ICC 0.91, lower 1-sided 95%CI: 0.80-0.83). The step test (study limb) and timed single leg stance (non-study limb) achieved optimal levels of reliability (ICC >0.90, lower 1-sided 95%CI >0.70), with acceptable measurement error (<10%) for clinical outcome measures. The functional reach tests were not sufficiently reliable. A ceiling effect was detected for timed single leg stance.

**Limitations.** Reliability was only assessed between two raters during a single session, and within one rater over a one-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 a reduction in quality of life.\textsuperscript{1,2} In the USA, it has been estimated that nearly 27 million adults aged 25 and over have symptoms and clinical findings of OA.\textsuperscript{3} The hip is one of the most common joints affected by OA. Epidemiological studies show that hip OA affects 7-25\% of the population aged over 55, and this prevalence is expected to increase gradually as the whole population ages.\textsuperscript{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 centre of mass remains within the base of support when balance is challenged.\textsuperscript{5,6} A variety of symptoms and physical impairments associated with hip OA can impact on the balance system, including joint pain, muscle weakness, joint stiffness and sensory dysfunction.\textsuperscript{7-9} Not surprisingly, impaired standing balance has been reported in people with hip OA compared to age-matched healthy participants\textsuperscript{10-13} and is frequently observed by clinicians managing people with hip OA. Importantly, impaired balance is recognised as a risk factor for falls in the older population\textsuperscript{14,15}, and falls are frequently reported in people with hip OA\textsuperscript{16}, with the majority of falls occurring during ambulation and stair ascent/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\textsuperscript{11,17,18}, however such equipment is expensive and impractical for
regular use in most clinical and indeed, many research, settings. For many clinicians and researchers, simple clinical tests are the most practical methods for measuring standing balance in people with hip OA.\textsuperscript{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.\textsuperscript{21} However, to date, there is insufficient evidence regarding the clinimetric properties for clinical standing balance tests in people with hip OA.\textsuperscript{22} Our recent systematic review, which synthesised evidence on clinimetric properties of observer-rated impairment tests (including balance tests) in people with hip and groin problems\textsuperscript{22} failed to identify a single study investigating the reliability (or indeed, 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.\textsuperscript{20, 23, 24}

The aim of this study was to estimate the reliability of four common clinical balance tests in people with hip OA, including the four-square step test, step test, functional reach test and the timed single leg stance. A secondary aim was to estimate the amount of measurement error associated with each test.

\textbf{Method}

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

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

**Procedure**

Participants were tested on two separate occasions (approximately one week apart). At the first test session (Session 1), participants performed the balance tests with two independent raters (Rater A and Rater B) to examine between-rater reliability. The testing order of both the raters and the balance tests was randomised using a computerized random number generator. Participants were given five minutes rest between each rater’s independent assessments. At the second test session (Session 2), participants repeated the balance tests again with the more experienced Rater A, (who was blinded to the results from Session 1) in order to examine within-rater reliability. A one-week test interval was used to provide sufficient time to limit recall of test
scores, but short enough to limit potential real change in clinical status. At Session 2, participants completed a self-reported global rating of change. This was used as a reference standard for stability and determined if 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, whilst the least painful (for bilateral disease) or non-painful hip was defined as the non-study limb. A Visual Analogue Scale (VAS) was used to assess the average level of study hip pain over the past week. Participants were asked to mark an “X” on a 100 millimetre line, anchored with “no pain” on the left and “worst pain possible” on the right. The distance from the left anchor to the X-mark was then measured in millimetres and as such, higher VAS scores indicated more severe pain.\(^{26}\) The VAS has demonstrated reliability in people with OA.\(^{27}\)

The Hip dysfunctions 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 five subscales: pain (10 items), other symptoms (5 items), function in daily living (17 items), function in sport 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 people with hip OA.\(^{30}\)

A global change scale (GCS) was used to assess self-reported change in the study hip pain and physical function across the two testing sessions. The GCS was measured on a five-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. Several studies have previously used these scales to determine changes in participants’ condition, where “minimal or slight changes” were defined as non-meaningful change. The GCS has been shown to be highly reliable in people with musculoskeletal dysfunction.

Assessment of balance

Participants were tested barefooted on the following four clinical balance tests:

Four-square step test: Four walking sticks were placed on the floor at right angles with handles outwards to form four squares. Participants started in square 1, facing square 2 and remained facing this direction for the duration of the test. Participants then stepped both feet as quickly as possible forwards into square 2, then sideways to the right into square 3, then backwards 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 two test trials. According to original published instructions for the test, the faster of the two trials was recorded to the nearest 10th of a second.

Step test: A 15 centimetre (cm) height step was used with a 5 cm width cardboard template positioned on the floor along the edge of the step to provide a standardised 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 (e.g. 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 three to four
practice steps were performed, immediately followed by one 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.

**Functional reach test:** Two types of functional reach tests were performed.

(i) *Forward reach.* Participants started in a normal relaxed stance with their dominant arm facing side-on, but not touching, a wall. A levelled 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 (i.e. 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 three 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 three test trials. According to original published instructions for the test, the mean difference between the starting and finishing points across the three trials was recorded to the nearest millimetre as the test score.

(ii) *Lateral reach test.* Participants started in a normal relaxed stance with their back facing, but not touching, a wall. A levelled measuring tape was then mounted on the wall at
the acromion height. Participants abducted one arm to 90 degrees (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 three seconds without losing their balance or taking a step or leaning on the wall. Knee flexion and trunk flexion/rotation were not permitted. Participants were instructed not to bend at the knees or at the trunk. If this 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 three test trials on each side. The mean differences between the starting and finishing points across the trials for each side was recorded to the nearest millimetre as the test score. A reach in the direction of the study hip was defined as the ipsilateral reach, whilst a reach away from the study hip was defined as the contralateral reach.

Timed single leg stance\textsuperscript{40}. Participants started with their hands on their hips, and stood on one leg for as long as possible up to a maximum of 30 seconds. The non-stance hip remained in a neutral position with the knee flexed so as the foot was positioned behind and was not permitted to touch the stance leg. Participants were encouraged to look at a non-moving target one to three meters ahead. The test ceased if participants moved their hands off their hips, touched the non-stance foot down or touched the non-stance leg onto the stance leg. A demonstration was provided followed immediately by two test trials on each leg (based on original published instructions).
The longest time, up to a maximum of 30 seconds, of the two trials on each leg was recorded to
the nearest 10\textsuperscript{th} of a second as the test score for each leg.

\textit{Data Analysis}

Data analyses were performed using the IBM SPSS 21 statistical package for Windows (USA).
Data were checked for normality and for systematic differences between test sessions.
Descriptive analyses were conducted across raters and sessions, including means, standard
deviations (SD) and the range of scores. Percentages of maximal scores (ceiling effects) were
also calculated for the timed single leg stance test since the score for this test is capped at 30
seconds.

Within- and between-rater reliability were each calculated using intra-class correlation
coefficients (ICC 2,1) with 95\% confidence intervals (CI) for a two-way random effects model
and absolute agreement. Interpretation of ICC values was based on published
recommendations\textsuperscript{21}, where values more than 0.75 indicate sufficient reliability and values more
than 0.90 indicate optimal reliability\textsuperscript{21,41}. Furthermore, confidence intervals were inspected to
ensure that lower one sided 95\% CI met a recommended \textit{minimum acceptable level}, which was
set at 0.70.\textsuperscript{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 ANOVA. The \textit{MDC}_{90} was calculated as \textit{SEM} \times 1.65 (z score of 90 \% interval) \times
√2. For both the SEM and MDC_{90}, 95% CIs were calculated according to recommended methods.\textsuperscript{44}

As the units of measurements for the four balance tests varied, SEM and MDC_{90} were also expressed as SEM percentage (SEM\%) and MDC percentage (MDC\%) in order to assist interpretation of the results. These values were defined as the SEM and MDC divided by the mean of all testing scores on the two test sessions and were calculated as SEM\% = (SEM/mean) x 100 and MDC\% = (MDC_{90}/mean) x 100\textsuperscript{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 0.90 and a minimal acceptable lower 1-sided 95% confidence limit of 0.70, at the 95% confidence level and power of 80\%.\textsuperscript{47} In this study, 30 participants were recruited in order to allow for any potential dropouts and/or the exclusion of data from participants who reported a meaningful change in their condition across sessions.

**Results**

Thirty people (mean age 63.3 years, SD 5.71 years, range 50-75 years, 18 females (60\%)) with hip OA participated. Descriptive characteristics of the participants are summarised in Table 1. In this cohort of participants, there were more women than men, and most of participants were overweight (BMI > 25 kg/m\textsuperscript{2}). One third reported bilateral symptoms. Most had not sustained a
fall in the past twelve 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 two 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 seven days for most participants (25/27), and was six days and eight days for the remaining two. 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 tests within Rater A over the one-week interval, and for the forward functional reach test between raters A and B within the single session (p < 0.05).

Between-rater reliability on two 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 were sufficiently reliable between raters (ICC 0.85-0.94, lower 1-sided 95%CI: 0.71-0.89). Further inspection of the point estimates and confidence limits demonstrated that the step test (study limb) and the timed single leg stance tests also met the optimal level of reliability (ICC >0.90, lower CI >0.70).

Within-rater reliability of repeated measures over a one-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 (non-study limb) were sufficiently reliable within one rater over a one-week interval and also met the optimal levels of reliability (ICC 0.91, lower 1-sided 95%CI 0.80-0.83).

Ceiling/Floor effects

Inspection of minimum/maximum scores (Tables 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 single leg stance test with maximal holding of 30 seconds at each test occasion.

Measurement error

The SEMs, SEM%, MDC$_{90}$ and MDC% between raters at Session 1, and within one rater over a one-week interval, are provided in Tables 2 and 3 respectively. The SEM of the tests between raters varied between 7.4 - 16.1% of the test score, whereas it varied between 9.0 - 21.2% of the test score when repeatedly measured by one rater over a one-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 four 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 were sufficiently reliable between raters within a session, whereas the step test (study limb) and timed single leg stance (non-study
limb) were sufficiently reliable within one rater over a one-week interval. The step test (study limb) and timed single leg stance (non-study 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 functional reach tests 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 that this is the first study known to the authors to estimate reliability of balance tests in hip OA, it is difficult to discuss our findings in relation to previous research. However, parallels can be drawn to a related study where our findings are generally in agreement with those from a study evaluating reliability of balance measurements in patients with hip fracture. In this previous study, Sherrington et al.\(^4^8\) reported good test-retest reliability for the step test with similar levels of reliability (ICC 0.85-0.92) and lower confidence intervals (0.71-0.83) to those found in the current study. In contrast, our findings were quite different to those from an earlier study evaluating inter-rater reliability of a battery of tests, including the timed single leg stance, in patients following surgically fixed hip fractures.\(^4^9\) In this previous study, the single leg stance test was one of the least reliable tests and reliability estimates were much lower (kappa 0.14-0.63) than those reported in the current study.
To our knowledge, no reliability estimates for the functional reach test or the four-square step test in a comparable group have been conducted.

Measurement errors associated with the four balance tests were also estimated in the current study, which have not previously been reported. This 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% than the functional reach test and timed single leg stance test. This means that, compared to the functional reach tests and timed single leg stance test, smaller amounts of change are required on the step test and four-square step test to be confident 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 three steps on the step test (standing on the affected side), two seconds on the four-square step test, 9.9 cm for the functional forward reach, 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 has a number of strengths. These include 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 are some limitations to the current study. Given a participant’s global rating of change and balance performance may not be independent and thus the potential for correlated error, it is possible our estimates of reliability are inflated somewhat. Results may have been different if participants with a change in their clinical condition were included in the analyses. As both our between- 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 tests were found over the one-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 two raters were used for evaluating between-rater reliability, which may limit generalizability of our findings to a wider pool of raters with different abilities and clinical backgrounds. We did, however, choose raters from different professional backgrounds (Rater A was a clinical physiotherapist and Rater B was a researcher with a human movement science background), with different levels of experience in assessing older patients with pathology, which helps to increase the generalizability of our findings. Additionally, only one rater was used for evaluating within-rater reliability. Although this rater was a physiotherapist, 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 those patients with end-stage disease awaiting arthroplasty.
Future research is required to provide comprehensive data about the clinimetric properties for clinical balance tests in people with hip osteoarthritis. In particular, evaluations of the validity and responsiveness of these tests are required. Information about the minimum 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) had unilateral hip OA, a post-hoc sub-analysis 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 required to examine reliability of balance tests within this subgroup.

In conclusion, this study provides estimates of reliability and measurement error of four 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 demonstrated optimal levels of reliability for clinical measurement tests. When measurement error and ceiling effects are also considered, our data suggest the step test (standing on the affected side) is the most useful clinical measures of
standing balance for people with hip OA. Further research is needed to determine the responsiveness, and in particular, the minimum 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.

References


Table 1: Participant characteristics (n=30), presented as mean (standard deviation) unless otherwise indicated.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>63.3  (5.7)</td>
</tr>
<tr>
<td>Female gender (n/%)</td>
<td>18 (60%)</td>
</tr>
<tr>
<td>BMI (kg/m²)</td>
<td>26.8 (3.9)</td>
</tr>
<tr>
<td>Duration of symptoms (years)</td>
<td>5.9 (8.1)</td>
</tr>
<tr>
<td>Right-sided study limb = (n/%)</td>
<td>17 (56.7%)</td>
</tr>
<tr>
<td>Right leg dominant (n/%)</td>
<td>26 (86.7%)</td>
</tr>
<tr>
<td>Bilateral symptoms (n/%)</td>
<td>10 (33.3%)</td>
</tr>
<tr>
<td>History of falls (n/%)</td>
<td>6 (20.0%)</td>
</tr>
<tr>
<td>Frequency of falls (n)*</td>
<td>3.0 (2.1)</td>
</tr>
<tr>
<td>Test-retest interval (days)</td>
<td>7.0 (0.3)</td>
</tr>
<tr>
<td>Hip pain (VAS) (millimetres)</td>
<td>40.9 (18.7)</td>
</tr>
<tr>
<td>HOOS: Pain</td>
<td>63.2 (13.1)</td>
</tr>
<tr>
<td>Other symptoms</td>
<td>65.5 (12.0)</td>
</tr>
<tr>
<td>Activities of daily living</td>
<td>67.0 (13.4)</td>
</tr>
<tr>
<td>Sport</td>
<td>53.8 (17.9)</td>
</tr>
<tr>
<td>Quality of life</td>
<td>50.6 (14.7)</td>
</tr>
</tbody>
</table>

SD = standard deviation, BMI = body mass index, VAS = visual analogue scale, HOOS = hip dysfunction and osteoarthritis outcome score. § self-reported leg used to kick a ball. *number of falls sustained in the last twelve months for participants with history of falls.
Table 2. Between-rater reliability: balance test scores, ICC, SEM and MDC\(_{90}\) across the two raters at Session 1 (n=30)

<table>
<thead>
<tr>
<th>Test</th>
<th>Rater A</th>
<th>Rater B</th>
<th>ICC (95%CI)</th>
<th>Lower 1-sided CI</th>
<th>SEM (95%CI)</th>
<th>SEM %</th>
<th>MDC(_{90}) (95%CI)</th>
<th>MDC %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Four-square step test (seconds)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mean (SD)</td>
<td>Mean (SD)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Min/Max</td>
<td>Min/Max</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8.97 (2.32)</td>
<td>8.56 (2.01)</td>
<td>0.86 (0.72-0.93)</td>
<td>0.75</td>
<td>0.77 s (0.65-1.04)</td>
<td>8.8</td>
<td>1.80 s (1.43-2.42)</td>
<td>20.5</td>
</tr>
<tr>
<td>Step test (number)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standing on non-study limb</td>
<td>13.40 (4.02)</td>
<td>12.83 (3.79)</td>
<td>0.85 (0.71-0.93)</td>
<td>0.74</td>
<td>1.48 steps (1.18-1.99)</td>
<td>11.3</td>
<td>4 steps (2.75-4.65)</td>
<td>26.4</td>
</tr>
<tr>
<td>Standing on study limb</td>
<td>14.63 (4.63)</td>
<td>14.13 (4.33)</td>
<td>0.94 (0.88-0.97)</td>
<td>0.89</td>
<td>1.06 steps (0.85-1.43)</td>
<td>7.4</td>
<td>3 steps (1.97-3.33)</td>
<td>17.2</td>
</tr>
<tr>
<td>Functional reach (cm)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forward reach</td>
<td>28.74 (6.84)</td>
<td>25.39 (6.96)</td>
<td>0.68 (0.29-0.85)</td>
<td>0.36</td>
<td>3.43 cm (2.73-4.61)</td>
<td>12.7</td>
<td>8.0 cm (6.37-10.76)</td>
<td>29.6</td>
</tr>
<tr>
<td>Ipsilateral reach</td>
<td>16.57 (3.20)</td>
<td>15.47 (3.81)</td>
<td>0.62 (0.34-0.80)</td>
<td>0.38</td>
<td>2.12 cm (1.69-2.85)</td>
<td>13.2</td>
<td>4.9 cm (3.94-6.65)</td>
<td>30.9</td>
</tr>
<tr>
<td>Contralateral reach</td>
<td>16.09 (4.49)</td>
<td>16.24 (4.85)</td>
<td>0.74 (0.53-0.90)</td>
<td>0.57</td>
<td>2.36 cm (1.88-3.17)</td>
<td>14.6</td>
<td>5.5 cm (4.39-7.40)</td>
<td>34.1</td>
</tr>
<tr>
<td>Timed single leg stance (seconds)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standing on non-study limb</td>
<td>22.65 (9.68)</td>
<td>21.20 (10.84)</td>
<td>0.90 (0.80-0.95)</td>
<td>0.82</td>
<td>3.12 s (2.48-4.19)</td>
<td>14.4</td>
<td>7.27 s (5.79-9.78)</td>
<td>33.6</td>
</tr>
<tr>
<td>Standing on study limb</td>
<td>21.26</td>
<td>3.38/30</td>
<td>21.65</td>
<td>2.28/30</td>
<td>0.89</td>
<td>0.80</td>
<td>3.46 s</td>
<td>16.1</td>
</tr>
<tr>
<td>-----------------------</td>
<td>-------</td>
<td>---------</td>
<td>-------</td>
<td>---------</td>
<td>------</td>
<td>------</td>
<td>--------</td>
<td>------</td>
</tr>
<tr>
<td>(10.30)</td>
<td>(50.0%)*</td>
<td>(10.15)</td>
<td>(46.7%)*</td>
<td>(0.78-0.95)</td>
<td>(2.76-4.66)</td>
<td>(6.44-10.87)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Percentage of participants scored the maximal possible score of thirty seconds in these tests. SD = standard deviation, ICC = intra-class correlation coefficient, SEM = standard error of measurement, MDC90 = minimal detectable change based on the 90% confidence interval, SEM% = standard error of measurement percentage, MDC% = minimal detectable change percentage, cm = centimetres.
Table 3. Within-rater reliability: balance test scores, ICC, SEM and MDC<sub>90</sub> across the two test sessions (n=27).

<table>
<thead>
<tr>
<th>Test</th>
<th>Session 1</th>
<th>Session 2</th>
<th>ICC (95%CI)</th>
<th>Lower 1-sided CI</th>
<th>SEM (95%CI)</th>
<th>SEM %</th>
<th>MDC&lt;sub&gt;90&lt;/sub&gt; (95%CI)</th>
<th>MDC %</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Four-square step test (seconds)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>9.07 (2.35)</td>
<td>8.31 (2.45)</td>
<td>0.83 (0.57-0.93)</td>
<td>0.62</td>
<td>0.86 s (0.68-1.17)</td>
<td>9.9</td>
<td>2.00 s (1.58-2.72)</td>
<td>23.0</td>
</tr>
<tr>
<td>Min/Max</td>
<td>4.97/15.6</td>
<td>5.12/17.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Step test (number)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standing on non-study limb</td>
<td>13.5 (4.07)</td>
<td>15.18 (4.40)</td>
<td>0.81 (0.42-0.93)</td>
<td>0.51</td>
<td>1.54 steps (1.22-2.10)</td>
<td>10.7</td>
<td>4 steps (2.84-4.89)</td>
<td>25.1</td>
</tr>
<tr>
<td>Standing on study limb</td>
<td>14.71 (4.74)</td>
<td>15.68 (4.74)</td>
<td>0.91 (0.77-0.96)</td>
<td>0.80</td>
<td>1.37 steps (1.08-1.86)</td>
<td>9.0</td>
<td>3 steps (2.52-4.34)</td>
<td>21.0</td>
</tr>
<tr>
<td><strong>Functional reach (cm)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forward reach</td>
<td>28.67 (7.05)</td>
<td>28.02 (7.88)</td>
<td>0.68 (0.42-0.84)</td>
<td>0.47</td>
<td>4.26 cm (3.36-5.79)</td>
<td>15.0</td>
<td>9.9 cm (7.85-13.52)</td>
<td>35.0</td>
</tr>
<tr>
<td>(cm)</td>
<td>15.67/45.67</td>
<td>13/43</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ipsilateral reach</td>
<td>16.58 (3.31)</td>
<td>16.86 (3.77)</td>
<td>0.64 (0.35-0.82)</td>
<td>0.41</td>
<td>2.15 cm (1.70-2.93)</td>
<td>12.9</td>
<td>5.0 cm (3.97-6.83)</td>
<td>30.0</td>
</tr>
<tr>
<td>Standing on non-study limb</td>
<td>16.14 (4.63)</td>
<td>16.85 (3.94)</td>
<td>0.73 (0.50-0.86)</td>
<td>0.54</td>
<td>2.23 cm (1.76-3.04)</td>
<td>13.5</td>
<td>5.2 cm (4.11-7.08)</td>
<td>31.5</td>
</tr>
<tr>
<td>Contralateral reach</td>
<td>5.33/26.6</td>
<td>8/26</td>
<td>0.54</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Timed single leg stance (seconds)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standing on non-study limb</td>
<td>22.12 (9.82)</td>
<td>21.48 (10.31)</td>
<td>0.91 (0.81-0.96)</td>
<td>0.83</td>
<td>3.08 s (2.43-5.67)</td>
<td>14.7</td>
<td>7.2 s (5.67-9.77)</td>
<td>34.2</td>
</tr>
<tr>
<td>(48.1%)*</td>
<td>(44.4%)*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standing on study limb</td>
<td>20.63 (10.39)</td>
<td>3.38/30 (44.4%)*</td>
<td>21.31 (10.91)</td>
<td>3/30 (44.4%)*</td>
<td>0.82 (0.64-0.91)</td>
<td>0.68 (3.65-6.29)</td>
<td>4.62 s</td>
<td>21.2</td>
</tr>
</tbody>
</table>

*Percentage of participants scored the maximal possible score of thirty seconds in these tests. SD= standard deviation, ICC= intra-class correlation coefficient, SEM= standard error of measurement, MDC= minimal detectable change based on the 90% confidence interval, SEM% = standard error of measurement percentage, MDC% = minimal detectable change percentage, cm = centimetres.
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 and Rana S. Hinman

PHYS THER. Published online February 20, 2014