Physiotherapy 105 (2019) 244–253

Reliability and validity of the Four Square Step Test in patients with hip osteoarthritis before and after total hip replacement

M. Batting [a,b,*], K.L. Barker [a,c]

[a] *Nuffield Orthopaedic Centre, Physiotherapy Research Unit, Headington, Oxford, UK*
[b] *University of Southampton, Health Sciences, Southampton, UK*
[c] *Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Science, University of Oxford, Oxford, UK*

Abstract

Objective To determine the validity and inter- and intra-rater reliability of the Four Square Step Test (FSST) in assessing gait performance, balance and physical function for patients with hip osteoarthritis before and after total hip replacement (THR).
Design Observational, repeated measures.
Setting A specialist orthopaedic hospital.
Participants Fifty-eight participants with moderate to severe hip osteoarthritis scheduled to receive primary hip replacement within 4 months from recruitment.
Main outcome measure Time to complete the FSST, time and steps to complete the Figure of 8 Walk Test (F8W) and Berg Balance Scale score (BBS).
Results The Bland and Altman limits of agreement for intra-rater measurements of the FSST were −3.2 s to 3.5 seconds before THR and −1.5 to 2.0 seconds after THR. Limits of agreement for two different raters were −2.2 to 3.4 seconds, all with small mean differences indicating little bias between raters or replications. Concurrent validity was assessed, and the FSST correlated highly with the F8W ($r = 0.7$, $P < 0.001$) and moderately with the BBS ($r = 0.6$, $P < 0.001$). Only one participant was rated as being at moderate risk of falls on the BBS, with the other participants scoring low; only one participant failed to complete the F8W. This is in contrast to the FSST, which 21 people failed to complete pre-operatively.
Conclusions The FSST is a valid and reliable measure of multi-directional stepping speed and balance, giving a more informative measure of gait performance than the F8W and BBS, and is feasible for use in a clinical population of patients both before and after THR.

Keywords: Outcome assessments; Gait; Balance; Osteoarthritis; Total hip replacement; Reproducibility of results

Introduction

Our living communities are complex environments that continue to challenge our balance control and walking patterns to help avoid obstacles, change directions, carry loads, and plan a path to a destination based on prior cognitive maps [1]. However, most measures of walking ability and static balance provide limited insight to everyday mobility, as they are constrained to situations of minimal environmental challenge [2]. Patients with hip osteoarthritis have been shown to have reduced toe clearance, impaired obstacle avoidance, and gait and balance disorders [3]. The same is true post operatively, where patients with a total hip replacement (THR) are reported to have a slower gait speed and shorter stride length [4] as well as reduced postural balance [5].

One measure that does challenge obstacle avoidance and change of direction is the Four Square Step Test (FSST).

* Corresponding author at: Physiotherapy Research Unit, Nuffield Orthopaedic Centre, Oxford University Hospitals NHS Foundation Trust, Windmill Road, Headington, Oxford OX3 7HE, UK.
E-mail address: Martha.moore@ouh.nhs.uk (M. Batting).

https://doi.org/10.1016/j.physio.2018.07.014
0031-9406/

The test requires a person to step forwards, backwards and sideways over obstacles in a specified sequence.

Developed by Dite and Temple [6], the FSST has been shown to provide a measure of dynamic standing balance and mobility. Both balance and mobility are the most consistently identified risk factors linked to falls [7–9]. Most falls occur during movement itself, with trips and slips accounting for a large proportion of falls [10–12], indicating that the ability to take a rapid step may help prevent some of these falls. The FSST could therefore be used to identify an increased risk of falls in those who have slower test times. Furthermore, the small testing space and low equipment requirements of the FSST make it an appealing outcome measure for the clinical setting.

The FSST has been shown to discriminate between healthy and non-ambulatory populations ($P < 0.01$), displaying high inter-rater reliability [intraclass correlation coefficient (ICC) = 0.99] and retest reliability (ICC = 0.98) [6]. It has also been shown to be a valid and reliable measure to assess balance deficits in community dwelling older adults, and patients with Parkinson's disease, Huntington's disease, multiple sclerosis, vestibular disorders, post stroke, post transtibial amputation and knee pain [13]. The reliability of the FSST has also been gauged in participants with hip osteoarthritis (OA) [14]; however, the validity for this population has yet to be determined.

It is therefore pertinent that the FSST be validated in this population so that effective rehabilitation to address any deficits can occur. To provide a measure of concurrent validity, the FSST was evaluated against the Figure of 8 Walk Test (F8W) and the Berg Balance Scale (BBS), which are both measures of dynamic balance that are reliable and used within the orthopaedic setting [15,16]. It is also essential to determine the reliability of this population before and after THR as, in clinical situations, multiple assessors may need to use the test or the same assessor will need to repeat the FSST to measure change.

The objectives of this study were to explore the validity of the FSST by comparing its agreement with the F8W and BBS in patients with hip OA before and after THR, and to establish the inter- and intra-rater reliability of the FSST by comparing the limits of agreement within and between two separate assessors.

Methods

This study was reported according to the Standards for the Reporting of Diagnostic Accuracy Studies reporting guidelines.

Participants

Eighty participants were recruited from a specialist orthopaedic hospital, with a total of 58 participants completing all three study assessments. To ensure a sufficient sample size, the recommendations from the consensus-based standards for the selection of health measurement instruments guidelines were followed [17]. Individuals were eligible if they had moderate to severe hip OA, were due to undergo a primary THR and were aged over 55 years. Individuals were not eligible if they were found to have severe cardiovascular or pulmonary disease, severe dementia or communication difficulties, rheumatoid arthritis or a neurological condition that would affect their ability to take part in the balance tests, registered with a visual impairment, or due to have further planned treatment on the hip within the next 4 weeks. A consecutive series of participants were enrolled on to the study. Ethical approval was obtained for the study from the Office for Research Ethics Committees, Northern Ireland. All participants were informed about the purpose and procedures of the study, and gave written consent. Fig. 1 shows a flow diagram of participants throughout the study.

Measures

Standardized protocols were developed according to the procedures stated in the original validation study for each measure. Background information on an individual's relevant medical history and physical characteristics (including height, weight and any previous lower limb surgery) was collected. Participants also completed the Oxford Hip Score, a 12-item questionnaire that measures patient-reported outcomes of hip function [18], and the Activities Specific Balance Confidence Scale, which is a subjective measure of confidence in performing various ambulatory activities [19].

Index test—Four Square Step Test

Four 90-cm canes were placed in a square resting flat on the floor. Each square was numbered one to four. The bottom left square was labelled Number 1 and the participant was asked to stand in this square facing forwards. The patient was required to step into each square in the following sequence: 2, 3, 4, 1, 4, 3, 2 and 1 (see Fig. 2). The patient was instructed to try to complete the sequence as quickly as possible without touching the canes. Both feet had to make contact with the floor in each square, and the participant had to face forwards during the entire sequence. One practice trial was performed, followed by two test trials; the fastest time was taken.

Reference standard—Figure of 8 Walk Test

Participants were instructed to stand mid-way between two cones (1.52 m apart) and to face towards one of the cones. An unmarked 0.6-m boundary around the testing area was noted by the assessor to check for walking accuracy (see Fig. 3). Participants were asked to walk at a comfortable, self-selected speed and direction following a figure-of-8 path around the cones, stopping when they had returned to their starting position. The timer began when the participant moved to take their first step, and the timer stopped once both feet

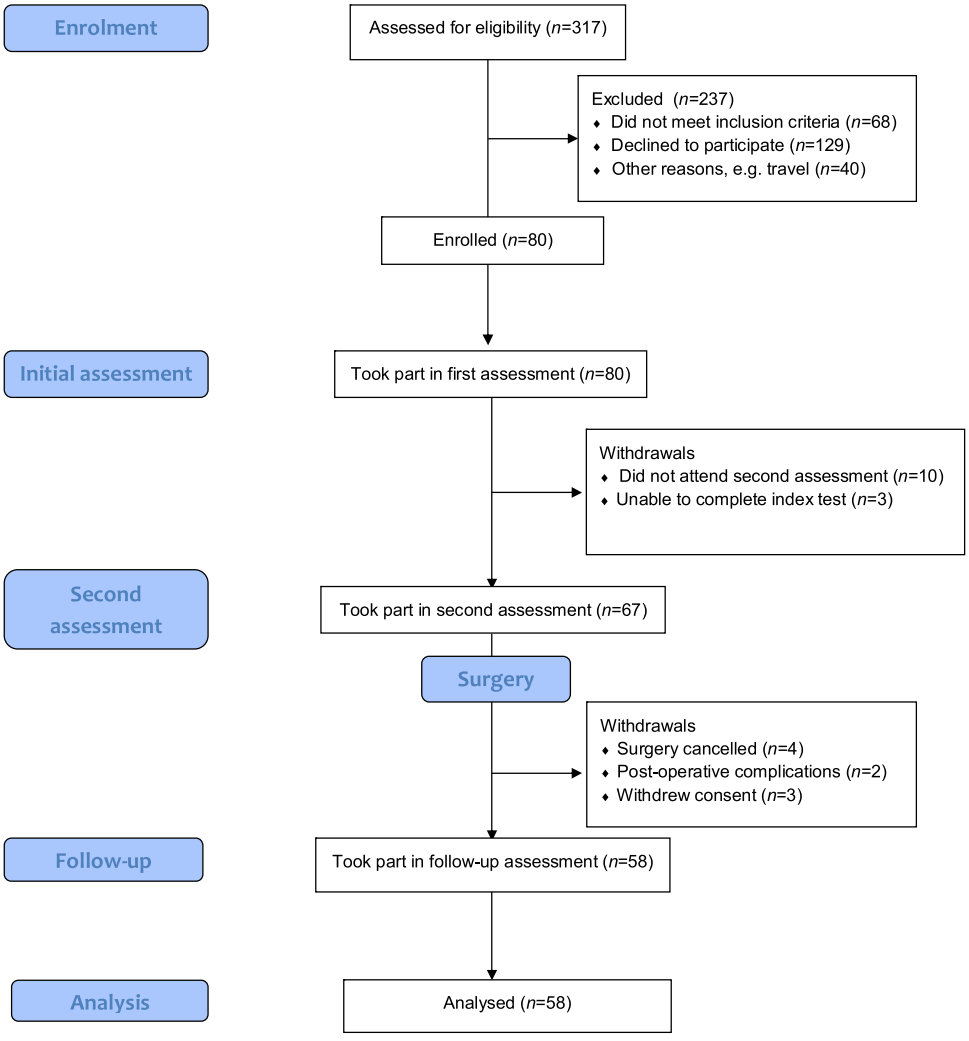

Fig. 1. Flow of participants through study.

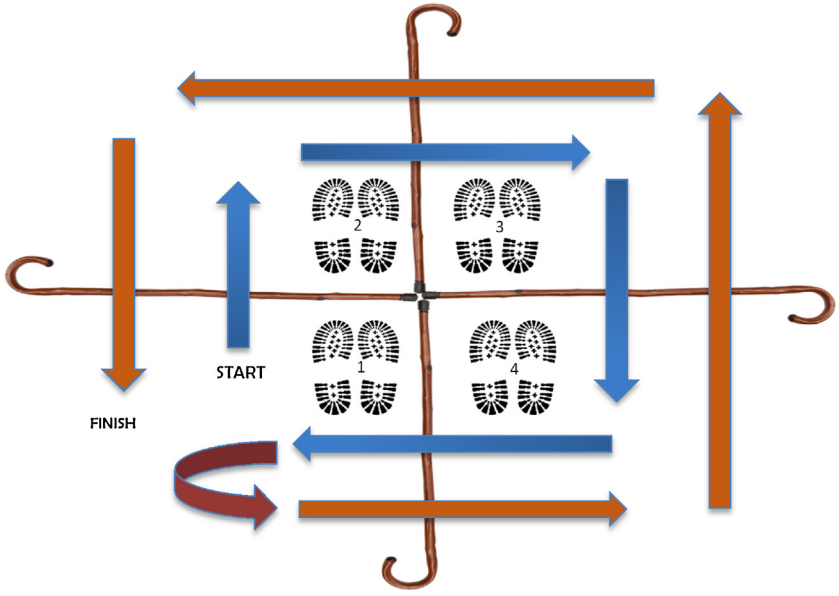

Fig. 2. Four Square Step Test set-up.

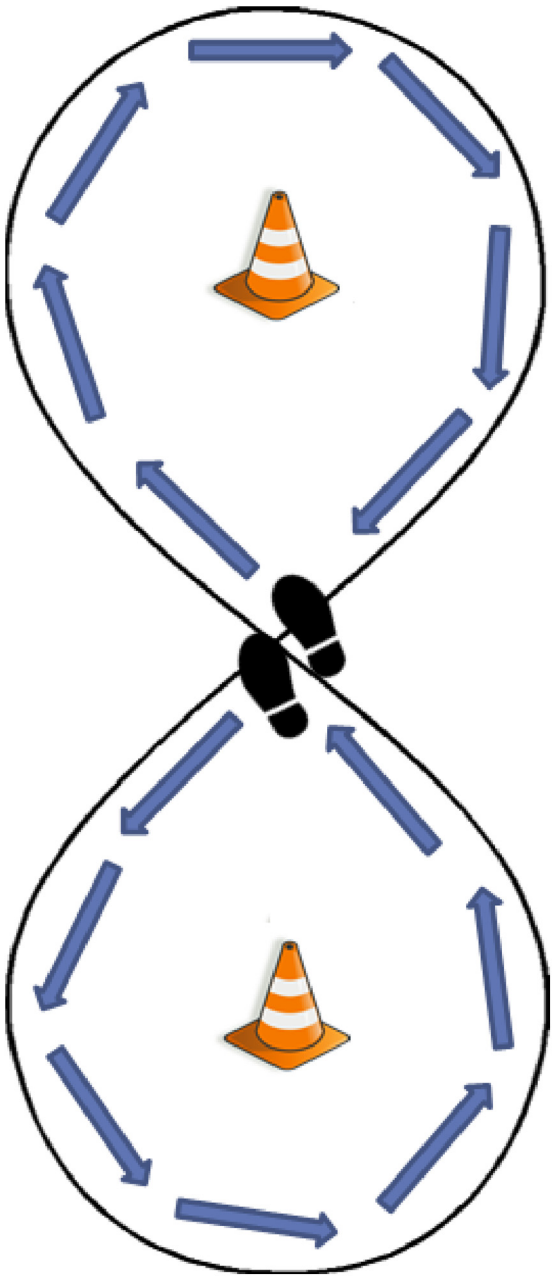

Fig. 3. Figure of 8 Walk Test diagram.

were back at the starting position. The time and number of steps taken were recorded. The assessor noted whether the participant kept within the unmarked 0.6-m boundary, and also rated the participant on a three-item walking smoothness score which evaluated stopping, change of pace and hesitation.

Reference standard—Berg Balance Scale

The participant was asked to complete 14 different balance tests including sit-to-stand, stand-to-sit, unsupported standing, unsupported sitting, transfers, standing with eyes closed, standing with feet together, reaching forwards, picking up an object from the floor, turning to look behind, 360° turn, step on to stool, tandem stance and single leg stance. Each task was marked from zero to four, with four being the best score. The test was totalled out of 56, with higher scores indicating lower risk of falls.

Procedures

The FSST was conducted a total of five-times over three separate visits for each participant before they had their surgery and then again post-operatively. At the initial assessment, all three outcome measures were completed once and carried out in a pre-set sequence picked at random. To establish inter-rater reliability, a second assessment was conducted by two separate assessors: Assessor A, who performed the initial assessment, and Assessor B. Both Assessor A and Assessor B carried out the FSST once and allowed the patient 1 minute of rest in between the tests. Both assessors were blinded to each other's result. The second assessment was scheduled 7 to 35 days after the first assessment in order to minimize learning effects. The final assessment was conducted by Assessor A 6 months after THR. The FSST was carried out twice with 1 minute of rest in between each test to establish intra-rater reliability after THR.

Statistical analysis

Descriptive statistics were used to characterize the study population. Means and standard deviations (SD) were calculated for continuous variables, and frequencies and percentages were calculated for binary variables. An analysis of intra- and inter-rater reliability for the FSST was performed for all three assessment time points using SPSS Version 24 (IBM Corp., Armonk, NY, USA) and based on the fastest FSST for each assessment. Bland and Altman plots with 95% limits of agreement were produced to provide a clinically meaningful picture of the size and range of the raters' scores [20]. In the plots, the differences between each pair of measurements are plotted against the mean of each pair of measurements. If the differences follow a standard normal distribution, then 95% of the differences will lie between 2 SD.

To investigate concurrent validity between the FSST, F8W and BBS, a Pearson's correlation coefficient was calculated. Scatter plots were also produced to show the relationship between each measure. The outcomes produced in each of the three tests are different. Consequently, when the limits of agreement between differences and averages were measured for the three tests, the raw measures of time and points were standardized by creating z-scores based on the sample's distribution. Gait/step speed and balance were the outcomes of interest, which were compared between measures.

Table 1
Characteristics of participants.

	n	Mean	SD	Range
Age (years)	58	70.6	7.1	56 to 94
BMI (kg/m^2)	58	28.3	4.9	20.2 to 40.6
ABC scale score[a]	55	69.8	20.9	17 to 99
OHS questionnaire[b]	56	21.4	8.2	5 to 40
		n	%	
BMI (kg/m^2)				
Normal weight (18 to 24.99)		12	21	
Overweight (25 to 29.99)		27	47	
Obese (≥30)		19	33	
Gender				
Female		32	55	
Male		26	45	
Previous lower limb surgery				
Yes		31	53	
No		27	47	
Other musculoskeletal conditions				
Yes		22	38	
No		36	62	
Falls history in previous year				
Yes		16	28	
No		42	72	

ABC scale, Activity-Specific Balance Confidence Scale; OHS, Oxford Hip Score; SD, standard deviation.
[a] Excludes three participants because of incomplete ABC forms (missing data).
[b] Excludes two participants because of incomplete OHS forms (missing data).

Table 2
Outcome measurements summary.

	n	Mean (SD)
First assessment (Assessor A alone)		
Four Square Step Test (seconds)	58	13.7 (5.6)
Figure of 8 Walk Test (seconds)	58	10.0 (3.4)
Figure of 8 Walk Test steps	58	15.8 (4.1)
Berg Balance Scale total score	58	52.6 (3.7)
Second assessment		
Assessor A		
Four Square Step Test (seconds)	57	14.2 (6.5)
Assessor B		
Four Square Step Test (seconds)	55	13.6 (5.6)
Follow-up assessment (Assessor A alone)		
Four Square Step Test (seconds) time 1	58	11.1 (3.2)
Four Square Step Test (seconds) time 2	58	10.8 (3.0)
	n	%
Four Square Step Test fails		
Assessment 1	7	12.1
Assessment 2	14	24.1
Assessment 3	4	6.9
Figure of 8 Walk Test boundary fails	*n*	
Assessment 1	1	1.7
Berg Balance Scale score breakdown		
Low risk of falls (41 to 56)	57	98.3
Medium risk of falls (scores 21 to 40)	1	1.7
High risk of falls (scores to 20)	0	0

Results

In total, 58 participants were included in the final review of the study from the initial 80 patients recruited (see Fig. 1). The participant characteristics are reported in Table 1.

On average, participants were assessed 18 days (SD 8.8) between the first and second assessments, and 6 months (SD 0.9) between surgery and the final assessment. A summary of the outcome measures scores is provided in Table 2. Reasons and frequency of test fails are noted in Table 3.

Intra- and inter-rater reliability

Before THR, on 95% of occasions, the second measurement made by Assessor A was within 6.1 seconds of the first measurement (above or below). When outliers were removed, the Bland and Altman limits of agreement reduced to 3.5 to −3.2, with a mean difference of 0.14.

Measurements of the FSST made by Assessor A were slightly higher than those made by Assessor B (mean difference 0.6 seconds), and the latter were within 2.8 seconds of the former on 95% of occasions. Finally, repeated measurements carried out by Assessor A at follow-up had narrower limits of agreement, where the second measurement was within 1.8 seconds of the first measurement on 95% of occasions.

Three notable outliers are presented within Fig. 4A. When these were removed, the upper limits of agreement were 3.5, lower limits of agreement were −3.2 and mean was 0.14. When the difference between measurements for intra-rater reliability before THR was recalculated with these outliers removed, a narrower agreement of 3.1 seconds on 95% of occasions was found.

Concurrent validity

Times taken to perform the FSST and F8W were strongly correlated ($r = 0.7$, $P < 0.01$). Moderate negative correlations were also found between the FSST and BBS ($r = -0.6$, $P < 0.01$). Scatter plots showed a strong positive linear relationship between the FSST and F8W (Fig. 5A). A strong negative relationship was also seen between the FSST and BBS (Fig. 5B).

Fig. 6 depicts Bland and Altman plots for the sample showing limits of agreement between the FSST, F8W and BBS *z*-scores. It confirms the correlation findings to show good agreement between the FSST, F8W and BBS with a mean difference close to zero, and almost all observations spread within narrow limits of agreement for the F8W (1.6 to −0.8 *z*-scores) and BBS (3.9 to −2 *z*-scores).

Discussion

To the authors' knowledge, this is the first study to assess the validity and reliability of the FSST before and after

Table 3
Four Square Step Test—reasons for test fails.

	First assessment	Second assessment: Assessor A	Second assessment: Assessor B	Follow-up assessment; Trial 1	Follow-up assessment; Trial 2
Failed test due to wrong sequence	4	4	1	1	1
Failed test as touched canes	3	5	3	2	1
Failed test due to loss of balance	0	1	0	0	0
Total fails	7	10	4	2	2

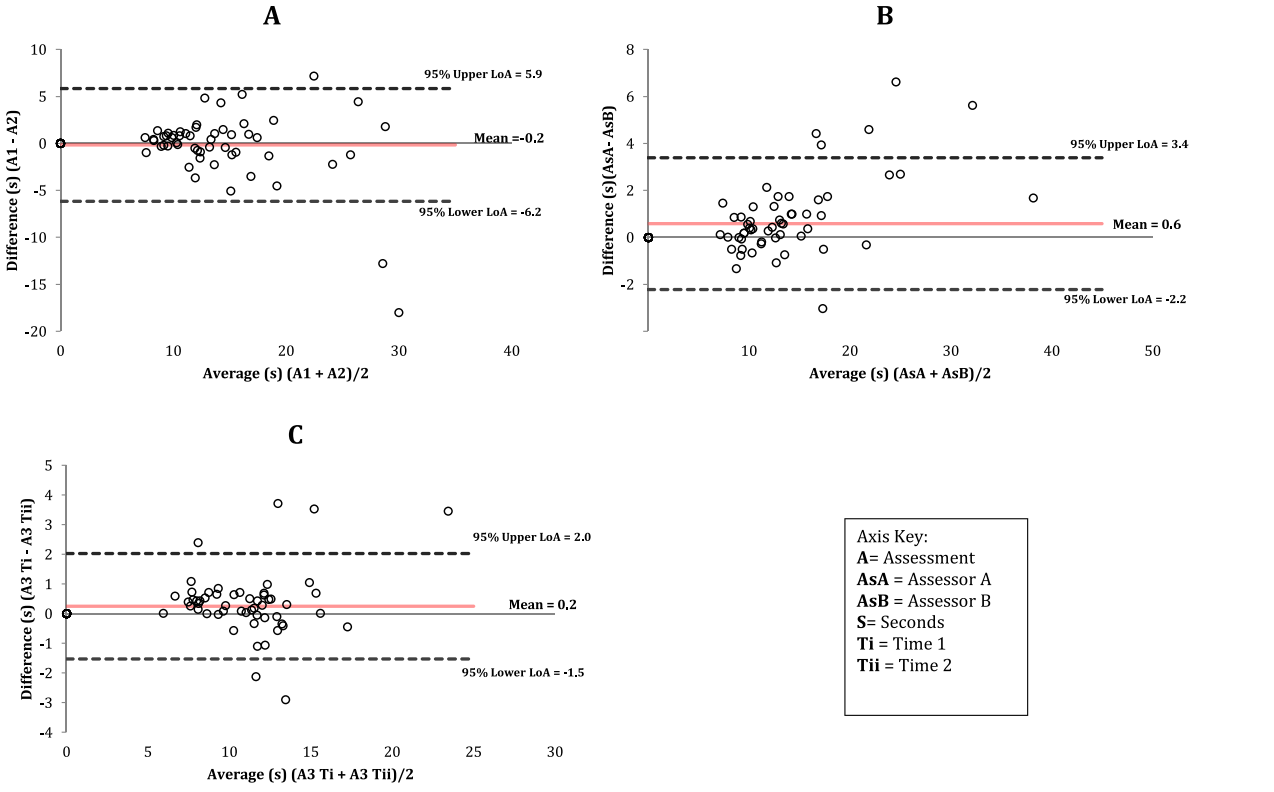

Fig. 4. Bland and Altman reliability plots. (A) Intra-rater reliability for Assessor A for first and second assessments. (B) Inter-rater reliability between Assessors A and B for second assessment. (c) Intra-rater reliability for Assessor A at the 6-month follow-up assessment.

hip replacement surgery. The study found the FSST to be a valid and reliable measure of balance for this population. Both intra- and inter-rater reliability were demonstrated with relatively small variation in scoring between and within Assessors A and B. Strong correlation was found between the FSST and F8W, and moderate negative correlation was found between the FSST and BBS.

It is important to be able to determine a patient's stepping speed and ability to help assess their risk of falls, and provide a clinically meaningful measure of balance to ensure appropriate and timely rehabilitation. This is particularly true in patients who have hip OA as they are known to have reduced stepping speed and obstacle clearance [21]. The ability to step at speed in multi-directions is required in everyday walking when responding to forward, backward and lateral perturbations (e.g. when walking in a busy street). It was noted by the assessors that, pre-operatively, participants struggled to step sideways in particular due to the known reduction in hip abductor strength [22]. Following surgery, many patients reported finding it easier to take a side step, which may, in part, explain the reduction in mean FSST step test time with a mean difference of 2.9 seconds post operatively. This correlates with the known increase in gait speed after hip replacement surgery [23].

This study reported a slower mean test time for the FSST (13.73 seconds for first assessment, 14.2 seconds for second assessment and 11.13 seconds for third assessment) compared with the results of Hess et al. [15] (8.97–8.56 seconds). These differences may, in part, be explained by the mean age of their participants, who were significantly younger. The present study also had a larger sample size with a greater age range of participants, which is more representative of the current THR population. However, there exists a dearth of literature assessing the FSST in musculoskeletal conditions, making it difficult to conclude with confidence that the present findings are reflective of the population as a whole.

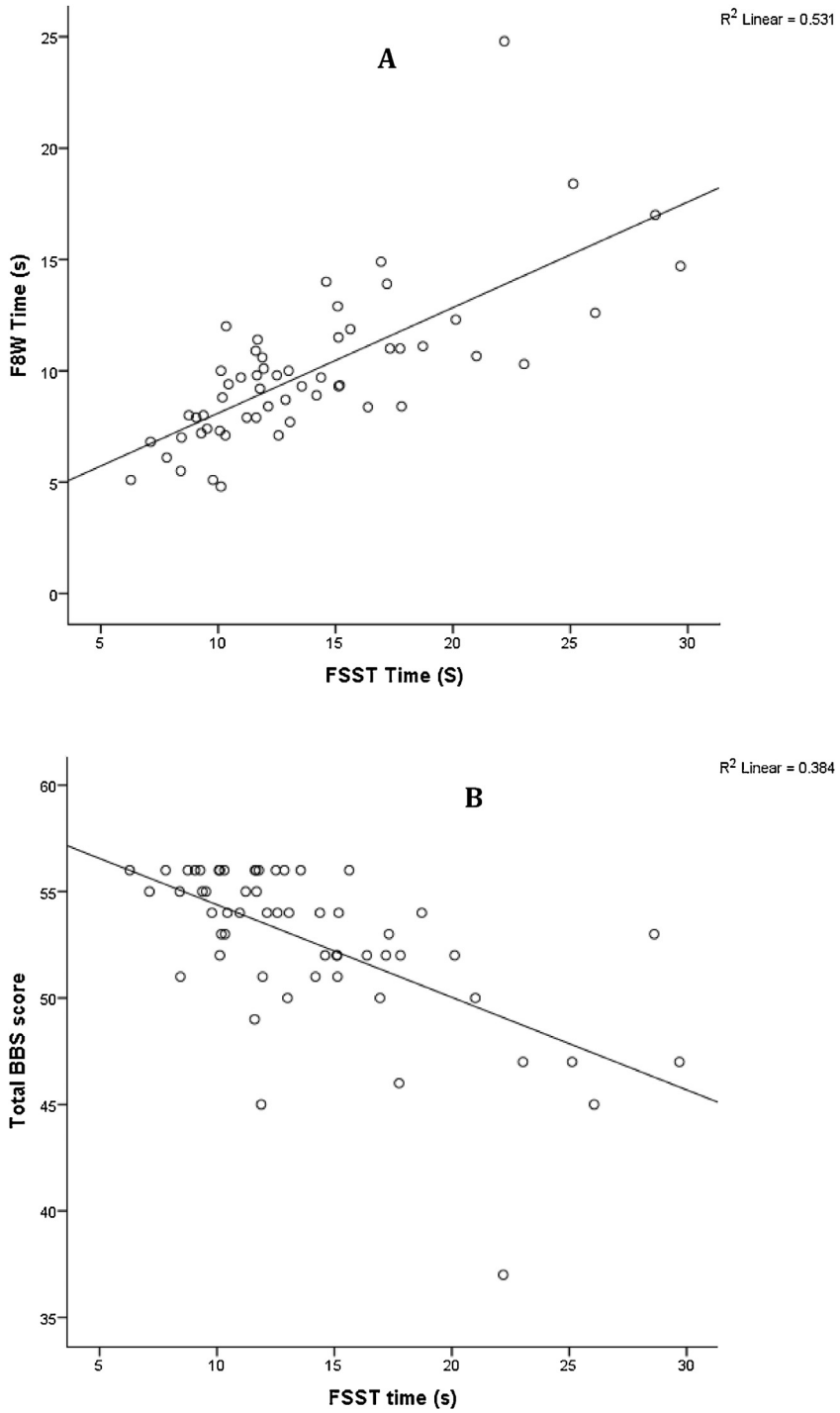

Fig. 5. Scatter plots showing correlation between the Four Square Step Test (FSST) and the Figure of 8 Walk Test (F8W) (A) and Berg Balance Scale (BBS) (B) in the first assessment.

The Bland and Altman plots reinforce the finding that the FSST is a reliable measure showing relatively narrow limits of agreement. Three significant outliers were highlighted when analysing the agreement within Assessor A, with a difference of more than 7 seconds between the two measurement times. If these outliers are not removed, a variation of up to 6 seconds is shown for 95% of occasions. Clinically, variations of up to 6 seconds would not be a reliable predictor of a patient's risk of falls, given that previous studies have found cut-off scores as low as 9.68 seconds for the FSST [24].

Additionally, when looking at two of the outliers, their second assessment time is markedly slower than the first for both Assessors A and B, which may indicate that their mobility had declined between the first and second assessments

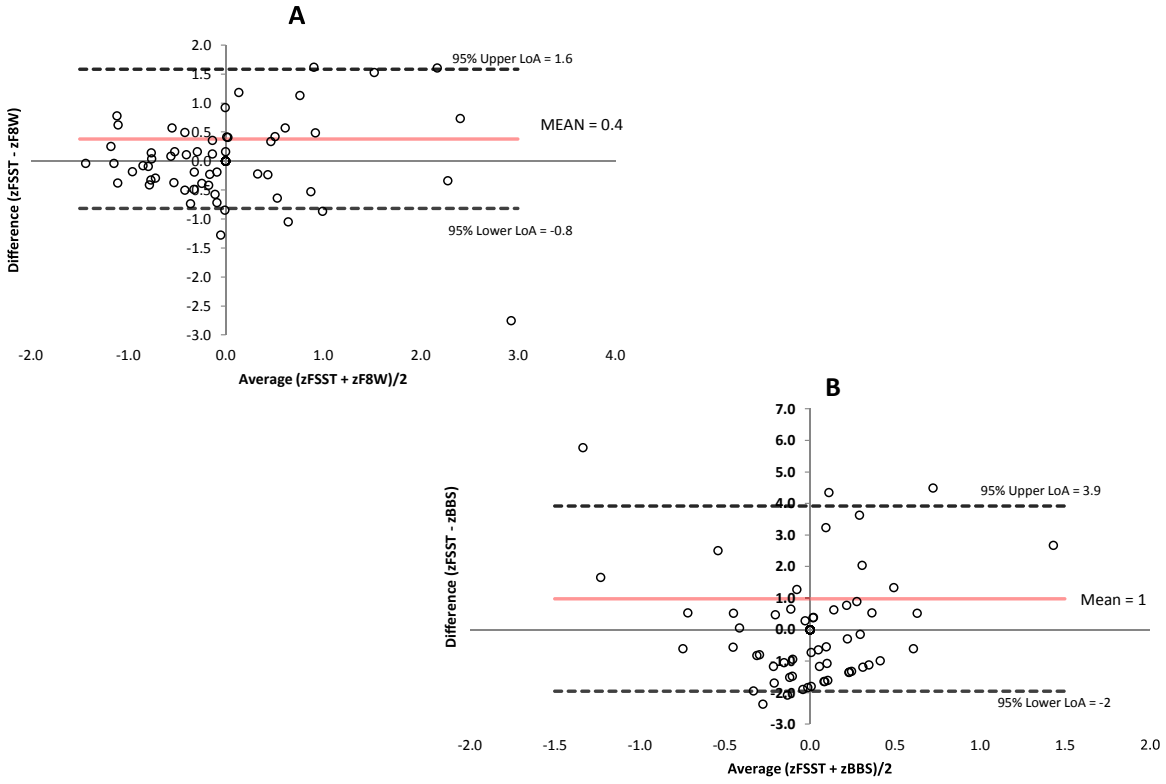

Fig. 6. Bland and Altman plots showing z-scores. (A) Difference in score between the Four Square Step Test (FSST) and Figure of 8 Walk Test (F8W). (B) Difference in score between the FSST and Berg Balance Scale (BBS).

and the result is not due to measurement error alone. One of the outliers was seen to produce faster scores at the second assessment for both assessors, and this could also indicate a potential learning effect that was not accounted for within the study.

It is also important to discuss the length of time between tests, and how this may account for variations in intra-rater reliability before and after THR. The pre-operative retest times were, on average, 18 days apart; however, at the third assessment, the two retest scores were measured just 1 minute apart. This would account for the 1.8-second difference within rater scores from before to after THR, as there are fewer confounding variables if the measures are taken within 1 minute of each other compared with multiple days apart.

When comparing the results of the FSST with the BBS, those with a higher overall score were found to be quicker at completing the FSST. For the BBS, those at low risk of falls are scored from 41 to 56, those at moderate risk of falls are scored from 21 to 40, and those at high risk of falls are scored from 0 to 20 [25]. However, only one participant in the study cohort was rated as being at moderate risk of falls, scoring 37 on the BBS, whereas 16 of the participants had reported a fall in the past year, indicating that perhaps the BBS was not sufficiently challenging and therefore sensitive to changes in this population.

Furthermore, the FSST was seen to correlate well with the F8W, with those who had faster F8W times also achieving faster FSST times. The mean F8W time of 9.99 seconds and step count of 15.8 was also similar to that shown in community dwelling older adults (10.49 seconds and 17.51 steps), with a similar mean age of participants sampled [16]. However, the F8W only showed that one person was unable to stay within the boundary of the test, which indicates a fail. This is in contrast to the FSST, for which 21 people failed pre-operatively, leading to the conclusion that the FSST is more receptive to limitations in participants' balance and gait speed. Some may argue that these findings may demonstrate that the FSST is not suitable in this population due to the failure rate in completing the test; however, the improvement of failure rate post-operatively to just four participants indicates that the test is measuring participants' balance accurately and giving a greater indication of improvement over the other measures.

In the study cohort, only five participants performed the FSST with an aid pre-operatively, reducing to just three participants post-operatively. Given the relatively small number of walking aid users, it remains difficult to draw conclusions about whether the use of a walking aid would impact on the FSST time.

The FSST is a valid and reliable measure compared with the F8W and BBS. Furthermore, the challenging nature of the FSST highlights those participants who have poor balance or gait speed, and can identify the potential causes such as poor obstacle clearance or reduced motor planning. This provides a clear advantage for the use of the FSST over the reference

standards, making it a more appealing and accurate measure for use in the hip replacement population.

Study limitations

This study had some limitations which may influence how the results are interpreted. Initially, 80 patients were recruited to the baseline cohort; there was a relatively high dropout rate of participants after the first assessment (16.25%), which, although not outside the expected range for observational studies, when combined with the further 11.25% who were unable to complete all three assessments reduced the study sample to just 58 participants. This can significantly affect the generalizability of the findings given that those participants who were unable to complete the assessments usually differ to other subjects in the sample. Furthermore, it may have been of clinical interest to retest those subjects who were unable to perform the FSST pre-operatively after their THR to see if there were any improvements in their ability to perform the test.

There was also a large variation in time between the first and second assessments, and the second and third assessments. The unpredictability of surgery dates and fitting around participant schedules meant that it was not possible to see each patient at the same 3-week time point. This resulted in some patients being seen much sooner for the retest than others, and may have affected the intra-rater reliability findings. The shorter time period for retesting between the second and third assessments also makes it difficult to conclude that there were no learning or fatigue effects from the inter-rater and postoperative intra-rater reliability. In addition, the results may not be generalizable to current clinical practice where testing intervals will typically be longer.

Finally, the purpose of this study was not to assess the predictive value of the FSST in measuring the risk of falls, nor was it to assess the responsiveness of the FSST to change. Further studies could look to provide information about the cut-off scores for risk of falls in this population, and the minimum clinically important difference so that the test can effectively be used as a measure both before and after THR.

Conclusions

The FSST is a valid and reliable measure of multidirectional stepping speed and balance giving a measure of gait performance that is feasible for use in a clinical population of patients both before and after THR. The FSST provides more informative and clinically useful data on balance and mobility compared with the F8W and BBS. The study is limited in generalizing the pre-THR inter-rater and post-THR intra-rater reliability findings due to the shorter timing interval between measurements which may not account for learning or fatigue effects, nor reflect the current practice of retesting periods. Further research needs to be conducted to assess the predictive validity of the test in this population, and its responsiveness to change if it is to provide meaningful outcomes of a participant's physical function and risk of falls.

Acknowledgements

The authors would like to acknowledge the contribution of Bronagh Walsh and Sean Ewings from the University of Southampton for their general supervision and statistical support. The authors would also like to thank Erin Hannink from the Physiotherapy Research Unit for her help in conducting the reliability assessments.

Ethical approval: Ethical approval was obtained from the Office for Research Ethics Committees Northern Ireland (Reference: 16/NI/0049).

Funding: The corresponding author is in receipt of an NIHR funded masters in Clinical Research at the University of Southampton. The work was also supported by the NIHR Biomedical Research Centre, Oxford.

Conflict of interest: None declared.

References

[1] Patla A. Mobility in complex environments: implications for clinical assessment and rehabilitation. Neurol Rep 2001;25:82–90.
[2] Frank J, Patla A. Balance and mobility challenges in older adults: implications for preserving community mobility. Am J Prev Med 2003;25:157–63.
[3] Ng C, Tan M. Osteoarthritis and falls in the older person. Age Ageing 2013;42:1–6.
[4] Sliwinski MM, Sisto SA, Batavia M, Chen B, Forrest GF. Dynamic stability during walking following unilateral total hip arthroplasty. Gait Posture 2004;19:141–7.
[5] Nantel J, Termoz N, Centomo H, Lavigne M, Vendittoli PA, Prince F. Postural balance during quiet standing in patients with total hip arthroplasty and surface replacement arthroplasty. Clin Biomech 2008;23:402–7.
[6] Dite W, Temple V. A clinical test of stepping and change of direction to identify multiple falling older adults. Arch Phys Med Rehabil 2002;83:1566–71.
[7] Studenski S, Duncan PW, Chandler J, Samsa G, Prescott B, Hogue C, *et al.* Predicting falls: the role of mobility and nonphysical factors. J Am Geriatr Soc 1994;42:297–302.
[8] Anacker SL, Di Fabio RP. Influence of sensory inputs on standing balance in community-dwelling elders with a recent history of falling. Phys Ther 1992;72:575–81.
[9] Altug F, Isik E, Cavalak U. Reliability and validity of four step square test in older adults. Turk J Geriatr 2015;18:151–5.
[10] Berg WP, Alessio HM, Mills EM, Tong C. Circumstances and consequences of falls in independent community-dwelling older adults. Age Ageing 1997;26:261–8.
[11] Blake AJ, Morgan K, Bendall MJ, Dallosso H, Ebrahim SB, Arie TH, *et al.* Falls by elderly people at home: prevalence and associated factors. Age Ageing 1988;17:365–72.
[12] Hill K, Schwarz J, Flicker L, Carroll S. Falls among healthy, community-dwelling, older women: a prospective study of frequency, circumstances, consequences and prediction accuracy. Aust NZ J Public Health 1999;23:41–8.

[13] Moore M, Barker K. The validity and reliability of the four square step test in different adult populations: a systematic review. BMC Syst Rev 2017;6:187.
[14] Choi YM, Dobson F, Martin J, Bennell KL, Hinman RS. Interrater and intrarater reliability of common clinical standing balance tests for people with hip osteoarthritis. Phys Ther 2014;94:696–704.
[15] Hess R, Brach J, Piva A, Van Swearingen J. Walking skill can be assessed in older adults: validity of the figure-of-8 walk test. J Phys Ther 2010;90:89–98.
[16] Downs S, Marquez J, Chiarelli P. The Berg balance scale has high intra and inter-rater reliability but absolute reliability varies across the scale: a systematic review. J Physiother 2013;59:93–9.
[17] Terwee CB, Mokkink LB, Knol DL, Ostelo RWJG, Bouter LM, de Vet HCW. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res 2012;21:651–7.
[18] Dawson J, Fitzpatrick R, Carr A, Murray D. Questionnaire on the perceptions of patients about total hip replacement. J Bone Joint Surg Br 1996;78:185–90.
[19] Powell LE, Myers AM. The activities-specific balance confidence (ABC) scale. J Gerontol Ser A 1995;50:28–34.
[20] Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, *et al*. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. J Clin Epidemiol 2011;64:96–106.
[21] Watelain E, Dujardin F, Babier F, Dubois D, Allard P. Pelvic and lower limb compensatory actions of subjects in an early stage of hip osteoarthritis. Arch Phys Med Rehabil 2001;82:1705–11.
[22] Zacharias A, Pizzari T, English DJ, Kapakoulakis T, Green RA. Hip abductor muscle volume in hip osteoarthritis and matched controls. Osteoarthr Cartil 2016;24:1727–35.
[23] Van den Akker-Scheek I, Stevens M, Bulstra SK, Groothoff JW, van Horn JR, Zijlstra W. Recovery of gait after short-stay total hip arthroplasty. Arch Phys Med Rehabil 2007;88:361–7.
[24] Duncan RP, Earhart GM. Four square step test performance in people with Parkinson disease. J Neurol Phys Ther 2013;37:2–8.
[25] Berg K, Wood-Dauphinee S, Williams JI, Maki B. Measuring balance in the elderly: validation of an instrument. Can J Public Health 1992;2:7–11.