Published in final edited form as:
Arch Phys Med Rehabil. 2010 June ; 91(6): 932–938. doi:10.1016/j.apmr.2010.02.003.

Inter-Rater Reliability and Validity of the Stair Ascend/Descend Test in Individuals With Total Knee Arthroplasty

Gustavo J. Almeida, PT, MS, **Carolyn A. Schroeder, BS**, **Alexandra B. Gil, PT, MS**, **G. Kelley Fitzgerald, PT, PhD**, and **Sara R. Piva, PT, PhD**
From the Department of Physical Therapy, University of Pittsburgh, Pittsburgh, PA (Almeida, Schroeder, Gil, Fitzgerald, Piva)

Abstract

Objective—1) To determine the inter-rater reliability and measurement error of a 11-step stair ascend/descend test (STTotal-11) and stair up (ascend) test (STUp-11); 2) to seek evidence for the STTotal-11 and STUp-11 as valid measures of physical function by determining if they relate to measures of physical function and do not relate to measures not of physical function; and 3) to explore if the STTotal-11 and STUp-11 scores relate to lower extremity muscle weakness and knee range of motion (ROM) in individuals with total knee arthroplasty (TKA).

Design—Cross-sectional study.

Setting—Academic center.

Participants—Subjects (N=43, 30 women; mean age, 68±8years) with unilateral TKA.

Interventions—Not applicable.

Main Outcome Measures—STTotal-11 and STUp-11 were performed twice and scores were compared to scores on 4 lower extremity performance-based tasks, 2 patient-reported questionnaires of physical function, 3 psychological factors, knee ROM, and strength of quadriceps, hip extensors and abductors.

Results—Intraclass correlation coefficient was 0.94 for both the STTotal-11 and STUp-11, standard error of measurements were 1.14sec and .82sec, and Minimum Detectable Change associated with 90%CI were 2.6 sec and 1.9 sec, respectively. Correlations between stair tests and performance based measures and knee and hip muscle strength ranged from r=.40 to .78. STTotal-11 and STUp-11 had a small correlation with one of the patient-reported measures of physical function. Stair tests were not associated with psychological factors and knee extension ROM, and were associated with knee flexion ROM.

Conclusions—STTotal-11 and STUp-11 have good inter-rater reliability and MDCs adequate for clinical use. The pattern of associations supports the validity of the stair tests in TKA.

Reprint requests to Gustavo J. Almeida, PT, MS, 3600 Forbes Ave and Atwood St, Ste 6035 Forbes Tower, Pittsburgh, PA 15260, gja4@pitt.edu.

Suppliers

Keywords

Arthroplasty, replacement, knee; Muscular weakness; Rehabilitation; Task performance and analysis

Individuals with TKA experience persistent difficulty managing steps 1 year after surgery.[1] Testing the ability to manage steps has been commonly used in clinical and research settings because it is an inexpensive and simple way to measure functional progress after TKA.[1–4] The frequent use of this test supports studies seeking evidence to validate stair tests as a measure of physical performance in subjects with TKA.

Validity is the degree to which evidence and theory support the interpretation of test scores for a proposed use.5 For a test to be valid it needs to have both, good technical quality, and adequate evidence to support the meaning of its score. The technical quality relates to satisfactory reliability and thresholds for interpreting changes in test score over time.6 Results of studies that investigated the reliability of stair tests are not applicable to patients with TKA because they have not included this population. For example, Jette et al reported moderate reliability (ICC=.54) for the time to ascend/descend 2 stairs in frail older adults.7 Rejeski et al reported good reliability (ICC=.93) of a 5-step ascend/descend test in subjects with knee OA.8 Kennedy et al reported good reliability (ICC=.90) of a 9-step ascend/descend test in subjects with hip and knee OA.2 Reliability should be specifically determined in patients with TKA because sample characteristics influence variability of scores and may affect the reliability. Furthermore, the application of the results of these studies is limited because they have used diverse testing protocols (i.e., number of steps varied from 2 to 9, test performed at either usual subject's speed or as quick as possible), and have not determined if the stair tests can be consistently measured by more than one rater (inter-rater reliability). Determining inter-rater reliability is relevant because after TKA patients tend to participate in both inpatient and outpatient rehabilitation, during which they are assessed by different clinicians throughout care. Lastly, the error associated with stair test scores has not been determined in patients with TKA. Determination of measurement error is helpful to establish thresholds for interpreting changes in stair test performance.

Regarding evidence to support the meaning of a test score, studies have not evaluated how well stair tests measure the construct physical function in TKA. For the construct validity to be supported, the test scores should relate to other measures in predicted ways. Specifically, the stair test scores should have a moderate correlation with other measures of lower extremity physical function (converge) and correlate poorly with measures not of lower extremity physical function (discriminate). Studies have also not explored the associations between stair tests and other physical impairments, including the strength of the lower extremity muscles and knee ROM. This information may help clinicians choose the impairments needed to be treated to improve subjects' ability to manage stairs.

The purposes of this study were: 1) to determine the inter-rater reliability and measurement error of an 11-step stair test total (ascend/descend) time (STTotal-11) and stair test up (ascend) time (STUp-11); 2) to seek evidence for the STTotal-11 and STUp-11 as valid measures of physical function by determining if they relate to measures of physical function and do not relate to psychological factors (construct validity); and 3) to explore if the STTotal-11 and STUp-11 scores relate to lower extremity muscle weakness and knee ROM in individuals with TKA.

METHODS

Subjects participating in a randomized controlled trial of the effectiveness of a balance training program for individuals with TKA were asked to participate in this study. The study took place from January/07 to May/08 in the Department of Physical Therapy, University of Pittsburgh. Inclusion criteria were unilateral TKA in the past 2–6 months and minimum age of 50 years. Individuals were excluded if they reported ≥2 falls within the past year, were unable to ambulate a distance of 31m without an assistive device, had acute illness or cardiovascular disease, severe visual impairment, lower extremity amputation, or a progressive neurological disorder. From 250 individuals informed about the study, 76 demonstrated interest to participate and were screened. From these, 16 declined participation due to schedule conflicts and 17 were not eligible, leaving 43 eligible subjects. All 43 subjects were included in the analysis of aims 2 and 3. From these, 22 participated in the reliability portion of the study (aim 1). Reasons for not including all subjects in the reliability were unavailability of 2 raters during testing day, or subject unwillingness to participate. Subjects signed a consent form approved by the University of Pittsburgh Institutional Review Board.

Procedures

This was a cross-sectional study. All measures were administered by a physical therapist and were performed in the same testing session. During the testing session, the subject first completed self-reported questionnaires of physical function and psychological factors, followed by performance-based measures of physical function, and then measures of knee ROM and muscle strength. With the exception of the stair tests described below, all the other measures are depicted in Table 1.

The 11-Stair Tests were performed after the other measures of performance-based physical function. The tests were administered on a set of 11 steps (110cm width, 30cm depth, 17cm height) with handrails on both sides, and a platform at the top and one at the bottom (110cm by 140cm). We used a regular stairwell with 11 steps because it represents the type and size of stairs that individuals likely need to manage during daily activities. The test began with the subject behind a marked line 27cm away from the first step, and one hand (self-selected side) on the handrail to increase safety. On rater's "go", the subject climbed the stairs, turned around on the top platform, and descended using the same handrail (e.g. right-hand up, left-hand down). They were instructed to complete the task as quickly as possible. There were 2 raters to concurrently record the STTotal-11 and the STUp-11. The rater who recorded the STTotal-11 stayed at the bottom platform, and started the stopwatch at his/her command "go", and stopped when both subject's feet returned to the floor at the bottom platform. The rater who recorded the STUp-11stayed at the top platform, and used another stopwatch to record the time from the other rater's command "go", to the time the subject reached both feet on the top platform.

For the reliability component of the study, subjects completed the stair tests twice in the same testing session to avoid changes in subjects' condition. Subjects were given at least a 3-minute rest between each test. There were 3 raters involved during this procedure and they formed 3 pairs of raters (rater1 and 2, rater1 and 3, rater2 and 3). Therefore, each subject was tested by 2 physical therapists (one pair of raters) trained in the stair test protocol. Each rater was blinded to the results of the other by not sharing the value on the stopwatch. From the first to the second stair test, raters rotated the order in which they tested the subjects to avoid recording the same component of the stair test (e.g., the rater who performed the STTotal-11 during trial 1 could not perform the STTotal-11 during trial 2). Rotation of pair of raters was based on rater's availability.

Data Analysis

To test for differences between subjects participating (N=22) and not participating (N=21) in the reliability component, we performed independent sample T-test for continuous, and Chi-Square, for ordinal and nominal variables. To determine inter-rater reliability, comparisons were made among all 3 pairs of raters and between each pair of raters (rater1 vs rater2, rater1 vs rater3, and rater2 vs rater3). We calculated intraclass correlation ($ICC_{2,1}$) coefficients and the 95% CI for all scores using SPSS.[19a] Formula $ICC_{2,1}$ was used to reflect that the measures were performed by 2 raters and a single measurement was taken by each one. To account for potential systematic bias between raters we used absolute agreement definition for ICCs,[20] and also constructed Bland and Altman plots[b]. Measurement error was calculated by the SEM. SEM was based on the reliability coefficient (r) and variance (SD) for the measures of 2 raters using the equation: $SEM=SD\sqrt{1-r}$. The MDC was calculated as the amount of change needed to be certain, within a defined level of statistical confidence, that change is beyond measurement error.[21] MDC was calculated as: z-score$_{(\text{level of confidence})}x\sqrt{2}$x SEM.[22] We calculated MDCs using the standard normal scores of 1.96 (associated with 95%CI-MDC_{95}) and 1.65 (associated with 90%CI-MDC_{90}). For aims 2 and 3, Pearson product moment or Spearman correlation were calculated according to data distribution.

RESULTS

Sample characteristics are reported in Table 2. The distribution of the continuous variables did not depart from normality. Subjects who participated in the reliability portion only differed in the LEFS, having 7 points higher LEFS scores (representing better function) than the ones who did not participate. Results of the ICCs and 95%CI of the STTotal-11 and STUp-11 are shown in Table 3. The ICC values for the inter-rater reliability of all raters and for the pairs of raters during stair tests were above 0.89, representing good reliability.[20] The Bland and Altman plots are shown in Figure 1. For the STTotal-11, the plot indicates systematic bias. The line of equality (zero) in the plot was not contained within the 95%CI around the mean difference between raters, indicating that the score of the second rater was 1.2sec lower than the scores of the first rater. For the STUp-11, the plot shows no systematic bias and the mean difference between raters was −0.16sec. The SD of the STTotal-11 scores among all raters was 4.5sec, resulting in a SEM of 1.1sec. The MDC_{95} was 3.2sec and MDC_{90} was 2.6sec. For the STUp-11, the SD was 1.6sec, and the SEM was 0.8sec. The MDC_{95} was 2.3sec and MDC_{90} was 1.9sec.

Table 4 shows the associations between the stair tests and measures of function, psychological factors, and physical impairments. Both stair tests demonstrated moderate to large correlations with performance-based tasks, indicating that individuals faster in the stair tests were also faster during the other performance-based tasks. The correlations with the LEFS were small and indicated that subjects reporting better physical function managed stairs faster. Although the Cohen's classic interpretation of a small correlation ranges from .11 to .30,[23] the strength of the correlations represents a continuum. Thus, we interpreted correlations in the low range of the classification of small as no association. Accordingly, the stair tests were considered not associated with the WOMAC-PF and the psychological factors.

The stair tests correlated negatively to a moderate degree with muscle strength, meaning that weaker individuals were slower during both stair tests. Knee flexion ROM had a small and moderate association with STUp-11 and STTotal-11 respectively, indicating that individuals with limited knee flexion took longer to ascend/descend the stairs. Knee extension ROM was not associated with stair tests. For measures of muscle strength and ROM, the associations

[a]SPSS Inc, 233 S Wacker Dr, 11th Fl, Chicago, IL 60606.
[b]MedCalc Software, v.10.4.5; Broekstraat 52, 9030 Mariakerke, Belgium

with the surgical and non-surgical limb were very similar; therefore, we only report results for the surgical limb.

DISCUSSION

To the best of our knowledge, this is the first study determining the inter-rater reliability of stair tests in patients with TKA. The finding of good inter-rater reliability is relevant because throughout the rehabilitation after TKA, patients may be treated and tested by more than one clinician. In addition, as inter-rater reliability is more difficult to achieve than intra-rater reliability, we believe that the reliability would have been comparable or better if the same rater performed the measure twice (intra-rater).

Although the reliability values of both tests were good, the STTotal-11 demonstrated systematic bias. The systematic bias during the STTotal-11 may have happened because rotation of raters was based on availability, rather than being counterbalanced or randomized. The order that the raters performed the 1st and 2nd repetition of the STTotal-11, while resulted in an even distribution of pairs (Table 3), resulted in an uneven distribution or raters. For example, during the first repetition of the STTotal-11, rater1 performed 15/22 tests (68%). Therefore, a limitation of our study is that we cannot reconcile if order effect, rater effect, or both, are responsible for the bias and wide CIs in the STTotal-11. We believe the systematic bias is mainly due to order effect because the bias was lower for the pairs of raters that had a balanced distribution in the order of testing (Table 3). Furthermore, for the STUp-11, in which the order of raters was well balanced, systematic bias was not observed.

In this study we provided the MDC values for the stair tests, which provide a threshold for interpreting the scores in the stair tests over time. These values are important when investigating the effect of interventions on functional performance in TKA. For example, when the score in the STTotal-11 changes more than 3.2sec (MDC_{95}) one can be 95% confident that true change has occurred. Although a MDC_{95} increases the precision of score estimation, in clinical practice, one may prefer a less stringent threshold such as the MDC_{90}, and be 90% confident that true change has occurred when STTotal-11 score changes more than 2.6sec. We found only one study that reported the MDC for a stair test. Kennedy and colleagues reported a MDC_{90} of 5.5sec for a 9-stairs ascend/descend test in subjects with end-stage hip and knee OA.[2] While MDC values can be affected by either variability (SD) or reliability, as the reliability in Kennedy and our study are similar (ICCs=.90 and .94, respectively), we believe the differences in the MDCs are likely explained by the larger variability in Kennedy study (SD=7.4sec) compared to ours (SD=4.5sec).The larger variability in Kennedy study may be explained by a combination of using 9 rather than 11-steps, and the possible heterogeneity of test results when subjects with end-stage hip and knee OA are combined.

Results of this study provide evidence to support the construct validity of the STTotal-11 and STUp-11 because their scores correlated to a moderate degree with all performance-based measures of physical function (converged), correlated to a small degree with LEFS, and did not correlate with psychological factors (discriminated). The only findings that did not support our hypotheses were the no associations between the stair tests and the WOMAC-PF. While small correlations between stair tests and self-reported function were expected, the associations between the stair tests and the WOMAC-PF were smaller than expected. We originally hypothesized small associations between stair tests and self-reported measures because: 1) stair tests represent a narrow range of physical function compared to the range of physical function represented by the WOMAC-PF or LEFS; 2) it is known that subjects with TKA tend to self-report their outcome as good even when they experience difficulty performing daily tasks.[24, 25] We believe the difference in the associations may be because LEFS scores are influenced to a lesser degree by pain than the WOMAC-PF.[26] Moreover, the TKA literature has shown

discrepancies in the pattern of associations between LEFS, WOMAC-PF, and performance-based measures of function. In a sample of patients pre- and 8 weeks post-TKA, the WOMAC-PF scores improved compared to their preoperative values while performance scores got worse. 25 Stratford and Kennedy reported that within 16 days post-TKA, while the WOMAC-PF scores did not change, the LEFS scores changed moderately (worsened function), and performance-based scores markedly worsened.[26]

The moderate to large correlation between the stair tests and other performance-based measures suggest that the stair tests may capture different aspects (indicators) of the construct physical function than the other performance-based tasks (larger correlations would likely indicate that they capture the same indicators). For example, whereas all tests measure movement control and speed to some extent, the straight walking tasks seem to capture basic agility of locomotion, the figure-of-8 walking may primarily capture skilled movement (curved paths), chair stands may capture power of quadriceps muscles and balance; while the stair tests seem to capture power of several proximal muscles and knee mobility. Larger studies in TKA should explore the indicators captured by each test and use factor analysis methods to determine what performance tests should be included in a battery of tests. A study in subjects with knee OA that validated a battery of performance-based tests, reported that a 5 or 9-step ascend/descend test, along with a 6-min walk and a lifting and carrying task, were the tests that provided sufficient profile for the performance capabilities of those subjects.[8]

The associations between the stair tests and quadriceps strength and knee flexion ROM are not new findings. Mizner and colleagues[4] found moderate association between stair test and quadriceps strength (r= −.53), and small association (r= −.30) between stair test and knee flexion ROM 2 months post-TKA[4]. To the best of our knowledge, the moderate associations between the strength of hip muscles and stair test performance are new findings. These associations may suggest that targeting the weakness of these muscles during rehabilitation could be beneficial to improve performance during ascending and descending steps.

When deciding what stair test to use, the strong associations between the scores in the STTotal-11 and the STUp-11 (r=.84), combined with the comparable pattern of associations between each stair test and the other concurrent measures, suggest that the two tests provide similar information. Since the STUp-11 had narrower confidence intervals, showed no systematic bias, and requires shorter performance time, we believe the STUp-11 is a reasonable choice to be included in a battery of performance-based tests for subjects with TKA.

Study Limitations

Our study has limitations. First, the cross-sectional design precludes ascertainment of any causal relationships in the associations. Second, as the LEFS scores were higher in the group who participated in the reliability portion, it is important to consider that the reliability was determined in a subgroup that was functioning at a higher level. Conversely, the very small differences in the WOMAC-PF scores between the ones who did and did not participate in the reliability portion argue against the clinical relevance of the difference in the LEFS. Third, because all subjects performed the stair tests at least once, with no complaints and no need to use more than 1 handrail, we believe that the stair tests investigated in this study are suitable for subjects post-TKA. Lastly, because we excluded subjects who could not walk independently, one has to consider that these tests may not be feasible to very disabled subjects. For more disabled subjects, the ability to perform the task, rather than the speed of performance, may be a more relevant outcome.

CONCLUSIONS

STTotal-11 and STUp-11 have good inter-rater reliability and MDCs adequate for clinical use. Narrower confidence intervals around reliability estimates support the use of the STUp-11. The pattern of associations supported the construct validity of both tests. Inclusion of stair tests into a more comprehensive battery of performance-based measures of lower extremity function in subjects with TKA should be considered.

List of Abbreviations

CI	confidence interval
ICC	intraclass correlation
LEFS	Lower Extremity Function Scale
MDC	minimum detectable change
MDC90	Minimum Detectable Change associated with 90% CI
MDC95	Minimum Detectable Change associated with 95% CI
OA	osteoarthritis
ROM	range of motion
SEM	standard error of the measurement
STTotal-11	11-step stair ascend/descend test
STUp-11	stair up (ascend) test
TKA	total knee arthroplasty
WOMAC-PF	Western Ontario and McMaster Universities Osteoarthritis Index-Physical Function

Acknowledgments

Supported by the Central Research Development Fund, the University of Pi Medical Center Health System Competitive Medical Research Fund, the Pepper Center Scholars Pilot Program (grant no. P30- AG024827), and from the National Center for Research Resources (grant no. UL1RR024153), a component of the National Institutes of Health (NIH), and NIH Roadmap for Medical Research.

References

1. Walsh M, Woodhouse LJ, Thomas SG, Finch E. Physical impairments and functional limitations: A comparison of individuals 1 year after total knee arthroplasty with control subjects. Phys Ther 1998;78 (3):248–258. [PubMed: 9520970]
2. Kennedy DM, Stratford PW, Wessel J, Gollish JD, Penney D. Assessing stability and change of four performance measures: a longitudinal study evaluating outcome following total hip and knee arthroplasty. BMC Musculoskeletal Disorders 2005;6:3. [PubMed: 15679884]
3. Lamb SE, Frost H. Recovery of mobility after knee arthroplasty - Expected rates and influencing factors. J Arthroplasty 2003;18(5):575–582. [PubMed: 12934208]
4. Mizner RL, Petterson SC, Snyder-Mackler L. Quadriceps strength and the time course of functional recovery after total knee arthroplasty. J Orthop Sports Phys Ther 2005;35(7):424–436. [PubMed: 16108583]
5. Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for educational and psychological testing. Washington (DC): American Educational Research Association; 2002.

6. Piva SR, Fitzgerald GK, Irrgang JJ, Bouzubar F, Starz TW. Get up and go test in patients with knee osteoarthritis. Arch Phys Med Rehabil 2004;85(2):284–289. [PubMed: 14966715]
7. Jette AM, Jette DU, Ng J, Plotkin DJ, Bach MA. Are performance-based measures sufficiently reliable for use in multicenter trials? J Gerontol A Biol Sci Med Sci 1999;54(1):M3–M6. [PubMed: 10026655]
8. Rejeski WJ, Ettinger WH, Schumaker S, James P, Burns R, Elam JT. Assessing Performance-Related Disability in Patients with Knee Osteoarthritis. Osteoarthritis Cartilage 1995;3(3):157–167. [PubMed: 8581745]
9. Van Swearingen, JM.; Brach, JS.; Hess, RJ.; Berlin, J.; Studenski, SA. Clinical Correlates of Motor Control in Walking: The Figure-of-8 Walk [Abstract]; Gerontological Society of America Annual Meeting; 2006.
10. Hardy SE, Perera S, Roumani YF, Chandler JM, Studenski SA. Improvement in usual gait speed predicts better survival in older adults. J Am Geriatrics Soc 2007;55(11):1727–1734.
11. Bellamy N, Buchanan WW, Goldsmith CH. Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. J Rheumatol 1988;15:1833–1840. [PubMed: 3068365]
12. Binkley JM, Stratford PW, Lott SA, Riddle DL. The lower extremity functional scale (LEFS): Scale development, measurement properties, and clinical application. Phys Ther 1999;79(4):371–383. [PubMed: 10201543]
13. Andresen EM, Malmgren JA, Carter WB, Patrick DL. Screening for Depression in Well Older Adults – Evaluation of A Short-Form of the CES-D. Am J Prev Med 1994;10(2):77–84. [PubMed: 8037935]
14. Spielberger, CD.; Gorsuch, RL.; Lushene, R. Manual for the State-Trait Anxiety Inventory STAI (Form Y). Palo Alto, CA: Consulting Psych Press; 1983.
15. Wang CY, Olson SL, Protas EJ. Test-Retest Strength Reliability: Hand-Held Dynamometry in Community-Dwelling Elderly Fallers. Arch Phys Med Rehabil 2002 June;Vol 83
16. Frost KL, Bertocci GE, Wassinger CA, Munin MC, Burdett RG, Fitzgerald SG. Isometric performance after THA. JRRD 2006;Vol. 43 Number 4.
17. Pua Y, Wrigley TW, Cowan SM, Bennell KL. Intrarater Test-Retest Reliability of Hip Range of Motion and Hip Muscle Strength Measurements in Persons with Hip Osteoarthritis. Arch Phys Med Rehabil 2008 June;Vol 89
18. Cibere J, Bellamy N, Thorne A. Reliability of the knee examination in osteoarthritis: effects of standardization. Arthritis Rheum 2004;50:458–468. [PubMed: 14872488]
19. Shrout PE, Fleiss JL. Intraclass Associations: Uses in Assessing Rater Reliability. Psychol Bull 1979;86(2):420–428. [PubMed: 18839484]
20. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychological Methods 1996;1(1):30–46.
21. Nunnally, JC.; Bernstein, IH. Psychometric theory. New York: McGraw-Hill; 1994.
22. Portney, LG.; Watkins, MP. Foundations of clinical research: applications to practice. Stamford: Appleton & Lange; 1993.
23. Cohen, J.; Cohen, P.; West, SG.; Aiken, LS. Applied multiple regression/correlation analysis for the behavioral sciences. 3rd ed.. Hillsdale, NJ: Lawrence Erlbaum Associates; 2003.
24. Woolhead GM, Donovan JL, Dieppe PA. Outcomes of total knee replacement: a qualitative study. Rheumatology 2005;44(8):1032–1037. [PubMed: 15870149]
25. Parent E, Moffet H. Comparative responsiveness of locomotor tests and questionnaires used to follow early recovery after total knee arthroplasty. Arch Phys Med Rehabil 2002;83(1):70–80. [PubMed: 11782835]
26. Stratford PW, Kennedy DM. Performance measures were necessary to obtain a complete picture of osteoarthritic patients. J Clin Epidemiol 2006;59 160e7.

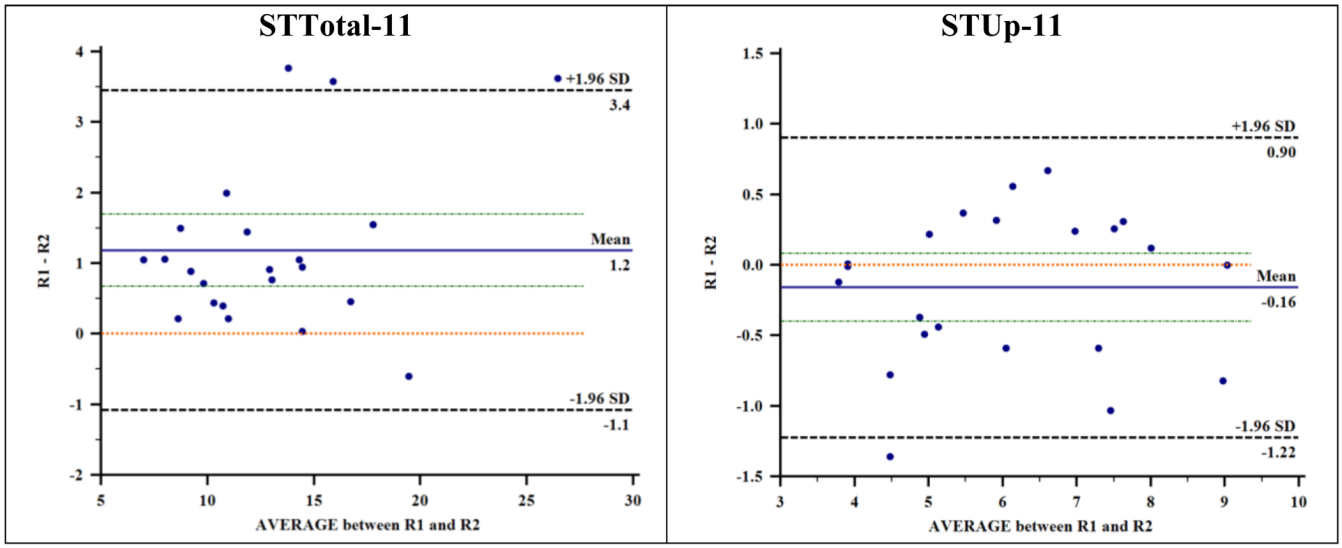

Figure 1.
Bland and Altman plots of the differences (vertical axes) versus means (horizontal axes) of rater 1 (R1) and rater 2 (R2) scores during 11-step stair ascend/descend test (STTotal-11) and stair up (ascend) test (STUp-11). NOTE. Solid lines represent the mean difference between raters. Lines with small dashes just above and below the mean difference represent the 95% CI of the mean difference. The lines over 0 (zero) are the lines of equality.

Table 1

Description of measures of performance-based and self-reported physical function, psychological factors, and physical impairments

Performance-Based Physical Function	
Figure-of-8 Walk Test	Subject stood in the middle of two markers 1.5m apart. On command "go", subject walked in one continuous figure-of-8-pattern around the markers, and ended at the starting position. Time to complete the task was recorded. Reliability of test in our department is good (ICC=.90).[9]
Timed Chair Rise	Subject was seated in a chair without armrests with arms crossed over the chest. Subject was timed while rising to a full upright position and sitting down 5 times without assistance. Reliability in older adults is good (ICC=.84 to .92).[7]
Gait Speed	Recorded time needed to pass two infra-red beams 4m apart, located in the central part of a longer path of 7m. Measured at two paces: self-selected and fast speeds. For each pace the subject was timed twice and the fastest speed was recorded (m/sec). Reliability of self-selected gait speed in old adults is good (ICC=.84).[10]
Self-Reported Physical Function	
WOMAC-PF	Western Ontario and McMaster Universities Osteoarthritis Index-Physical Function (WOMAC-PF) is a valid disease-specific measure of physical function in individuals with knee and/or hip OA.[11] It has 17 items. Subject ranks the difficulty during performance of functional activities on a 5-point scale. Scores range from 0–68. Higher scores indicate worse physical function.
LEFS	Lower Extremity Function Scale (LEFS) is a region-specific measure of lower extremity function with adequate reliability and validity.[12] The scale consists of 20 items, each scored from 0–4. Total scores range from 0 to 80. Higher scores indicate better physical function.
Psychological Factors	
Fear of Falling	Measured by the Survey of Activities and Fear of Falling in the Elderly (SAFFE). It measures 11 basic and instrumental activities of daily living, as well as more advanced mobility and social activities. Scores range from 0–3. Higher scores represent more fear of falling. SAFFE was validated in community-dwelling older adults.[9]
Depression	Measured by the Center for Epidemiological Studies Depression Scale short version including 10 questions (CES-D 10). It is a reliable and valid measure of depression symptoms.[13] Scores range from 0–30. Higher scores represent more depression.
Anxiety	Measured by the State Trait Anxiety Inventory (10-item), which is a reliable and valid measure.[14] Scores range from 4–40. Higher scores coincide with greater anxiety.
Muscle Strength	
Tested using an isokinetic dynamometer (Biodex System-3 Pro, Shirley, NY). Measures recorded in torque (Nm) and normalized to body mass (Nm/kg). Subject performed 3 warm-up trials and 5 testing trials of maximum voluntary isometric contraction. Reliability of these measures are good (ICC=.86 to .93).[15–17]	
Knee Extension	Subject was seated with the tested knee flexed at 60° and the force-sensing arm secured to the ankle. The highest torque was used in the analysis.
Hip Extension	Subject was in supine with the hip flexed at 50° and trunk flexed at 30°. The force-sensing arm was attached to the posterior thigh (proximal to the popliteal fossa). The average of 3 trials was used in the analysis.
Hip Abduction	Subject in side lying, tested hip up. The testing hip was at 0° of abduction, 5° of extension, and the knee was extended. The force-sensing arm was proximal to the lateral knee joint line. The average of 3 trials was used in the analysis.
Range of Motion	
Knee Flexion and Extension	Measured passively with a standard goniometer. Subject in supine, axis of goniometer aligned on the center of the lateral epicondyle (femur). Distal and proximal goniometer arms were aligned with the lateral malleolus and the greater trochanter respectively. For terminal knee extension, a bolster was placed under the heel to allow knee hyperextension. Technique has good reliability (ICC=.96 for flexion and .81 for extension).[18]

Table 2

Descriptive statistics of subject characteristics.

Variable	Means (SD) or Frequency (%) N = 43	Means (SD) or Frequency (%) Participated in reliability N = 22	Means (SD) or Frequency (%) Did not participate in reliability N = 21	p-value
Age	68(8)	69 (8)	67 (8)	0.54
Gender – Female n (%)	30 (70)	17 (77)	13 (62)	0.24
Height - cm	164.8 (9.9)	166.7 (8.6)	162.7 (11.1)	0.19
Weight - kg	83.6 (14)	82.2 (15.2)	85.1 (12.7)	0.51
Race – n (%)				0.19
African- American	1 (2)	1 (4)	0	
White	42 (98)	21 (96)	21 (100)	
Ethnicity - n (%)				0.35
Hispanic	1 (2)	1 (4)	0	
Non-Hispanic	42 (98)	21 (96)	21 (100)	
General Health - n (%)				0.28
Excellent	9 (21)	6 (27)	3 (14)	
Good	30 (70)	15 (68)	15 (72)	
Fair	4 (9)	1 (5)	3 (14)	
Poor or Bad	0	0	0	
Surgery on Left Side - n (%)	22 (51)	10 (46)	12 (57)	0.55
WOMAC-PF†	18.7 (8.3)	16.8 (8.1)	20.9 (8.3)	0.91
LEFS‡	49.4 (9.7)	52.7 (7.4)	45.8 (10.8)	0.02*
Stair Test - sec				
Total Time	18.1 (7.5)	17.7 (6.8)	18.6 (8.4)	0.68
Up Time	8.0 (3.4)	8.0 (2.5)	8.7 (4.5)	0.25

†Western Ontario and McMaster Universities Osteoarthritis Index-Physical Function scale

‡Lower Extremity Function Scale

*Significant at $p \leq 0.05$

Table 3

Inter-rater reliability of 11-step stair ascend/descend test (STTotal-11) and stair up (ascend) test (STUp-11) among all 3 pairs of raters and between each pair of raters.

	ICC	95% CI
Inter-rater STTotal-11. All pairs, N = 22	**0.94**	**0.55 to 0.98**
Pair 1: Rater 1 and $2_{(\text{Rater 1 was first during all 9 tests})}$. N=9	0.89	0.49 to 0.98
Pair 2: Rater 1 and $3_{(\text{Rater 1 was first during 6/7 tests})}$. N=7	0.91	0.61 to 0.98
Pair 3: Rater 2 and $3_{(\text{Rater 2 was first during 3/6 tests})}$. N=6	0.97	0.85 to 1.00
Inter-rater STUp-11. All pairs, N = 22	**0.94**	**0.87 to 0.98**
Pair 1†: Rater 1 and $2_{(\text{Rater 1 was first during 1/2 tests})}$. N=2	--	--
Pair 2†: Rater 1 and $3_{(\text{Rater 1 was first during test})}$. N=1	--	--
Pair 3: Rater 2 and $3_{(\text{Rater 2 was first during 11/19 tests})}$. N=19	0.93	0.83 to 0.97

† ICCs for pairs 1 and 2 were not calculated because measures were performed in no more than a couple of subjects.

Table 4

Descriptive statistics of measures of function, psychological factors, and physical impairments, and the correlations between these measures and STTotal-11 and STUp-11 scores.

N = 43	Mean (SD)	STTotal-11	STUp-11
Performance-Based Physical Function			
STTotal-11[†]	18.1 (7.5)	-	0.84**
STUp-11[‡]	8.0 (3.4)	0.84**	-
Figure of 8 Test Time (sec)	7.2 (1.5)	0.61**	0.54**
Timed chair rise (sec) [§]	12.6 (11.6 – 14.3)	0.59**	0.53**
Gait speed-self-selected (m/sec)	1.09 (0.17)	−0.63**	−0.58**
Gait speed-fast (m/sec)	1.53 (0.29)	−0.68**	−0.59**
Self-Report Physical Function			
WOMAC-PF [//]	18.7 (8.3)	0.12	0.07
LEFS [¶]	49.4 (9.7)	−0.22	−0.24
Psychological Factors			
Anxiety	17.2 (4.6)	0.06	−0.02
Depression	7.5 (3.3)	−0.05	0.14
Fear of falling	0.7 (0.3)	0.17	0.12
Physical Impairments			
Knee Extension (Nm/kg)	0.84 (0.34)	−0.58**	−0.66**
Hip Extension (Nm/kg)	1.49 (0.51)	−0.40*	−0.45*
Hip Abduction (Nm/kg)	0.76 (0.29)	−0.60**	−0.78**
Knee Flexion ROM (°)	123.0 (9.4)	−0.43*	−0.28
Knee Extension ROM (°)	−5.5 (3.6)	−0.01	−0.03

*Significant at $p \leq 0.01$

**Significant at $p \leq 0.001$

[†]11-step stair ascend/descend test

[‡]11-step stair up (ascend) test (STUp-11)

[§]Variable not normally distributed. Descriptive statistics represent median (Q25–Q75). Correlation coefficient represents Spearman rho

[//]Western Ontario and McMaster Universities Osteoarthritis Index-Physical Function scale

[¶]Lower Extremity Function Scale