ORIGINAL ARTICLE

Reliability of the Parallel Walk Test for the Elderly

Sally D. Lark, PhD, Peter W. McCarthy, PhD, David A. Rowe, PhD

ABSTRACT. Lark SD, McCarthy PW, Rowe DA. Reliability of the parallel walk test for the elderly. Arch Phys Med Rehabil 2011;92:812-7.

Objective: To determine interrater agreement and test-retest reliability of the parallel walk test (PWT), a simple method of measuring dynamic balance in the elderly during gait.

Design: Cohort study.

Setting: Outpatient clinic.

Participants: Elderly fallers (N=34; mean ± SD age, 81.3±5.4y) registered at a falls clinic participated in this study based on Mini-Mental State Examination and Barthel Index scores.

Interventions: Subjects were timed as they walked 6m between 2 parallel lines on the floor at 3 different widths (20, 30.5, 38cm) wearing their own footwear. They were scored for foot placement on (1 point) or outside the lines (2 points) by 2 separate raters. Fifteen subjects were retested 1 week later.

Main Outcome Measures: Footfall score and time to complete the PWT. Intraclass correlation coefficients (ICCs) and 95% limits of agreement were calculated for interrater and test-retest reliability.

Results: For widths of 20, 30.5, and 38cm, interrater reliability ICC range was .93 to .99 and test-retest ICC range was .63 to .90.

Conclusions: The PWT was implemented easily by 2 raters with a high degree of interrater reliability. Test-retest reliability was not as high, possibly because of the high susceptibility of variation from 1 week to the next for frail elderly subjects. The 20- and 30.5-cm widths are recommended for future use of the PWT.

Key Words: Aged; Balance; Gait; Rehabilitation; Reliability.

THE PARALLEL WALK test was validated previously as a quick and simple quantitative measure of balance during gait, which would allow direct comparisons after an intervention.[1] It is based on the premise of increased lateral movement during gait that corresponds to decreased dynamic stability.[2] The patient walks between 2 parallel lines of a designated width and is scored if he or she steps on or outside the lines. A higher score denotes a lack of stability. The PWT was reported to have been optimal in correctly classifying fallers and nonfallers at a distance of 20 to 30.5cm. Validity coefficients were .70 to .84, (.75 at 20cm with a score cutoff of 12; .70 at 30.5cm) and higher for time at .82 to .87.[1]

The PWT was developed because existing tests are temporal in assessment, such as the TUG test,[3] or are qualitative, extensive, and time consuming for a public clinic setting, such as Tinetti balance performance,[4,5] dynamic gait index,[6] or functional gait assessment.[7] The TWT commonly is used as a measure of dynamic balance during gait. However, conclusions are made about a person's balance during gait and risk for falling even when it is not attempted.[8] It has been reported previously that more than 40% would not attempt the TWT,[8,9] and similarly, all elderly fallers and 44% of nonfaller subjects would not attempt it in comparison to the PWT.[1]

In considering the PWT as a tool for assessing dynamic balance during gait, it remains to determine interrater reliability, particularly to show whether a rater who is not familiar with testing elderly balance parameters or has had limited or no formal training can achieve the same scores during the test as a more experienced trained examiner. Furthermore, test-retest reliability of the PWT needs to be determined.

Therefore, the purpose of this study was to examine the comparability of scoring between raters (interrater) and test-retest reliability of the PWT.

METHODS

Participants

Elderly fallers (N=36; mean ± SD age, 81.3±5.4y) who had had a recent fall (within the previous 6 months) and were referred to an outpatient falls clinic of a city hospital initially accepted an invitation to voluntarily participate. The study was approved by the local National Health Service Ethics Committee. All participants signed a consent form after the tests had been verbally explained and shown to them. All subjects were living independently; they were not in a nursing home, but may have been in sheltered accommodation, and they were mobile with or without the aid of a walking stick. They subsequently had recovered from any injuries (which may or may not have resulted in hospitalization) sustained by their fall before attending the falls clinic.

Recruitment criteria included a score of 23 or higher of a possible 30 points on the Folstein MMSE[10] for mental cognitive ability and higher than 10 (of 20) for the Barthel Index[11,12] for independent activities of daily living. Subjects were excluded if they had serious pathologic states that might have been exacerbated on exertion or be deemed to make the participant unsafe. These included unstable cardiovascular disease

From the Faculty of Health, Sport, and Science, Glamorgan University, Pontypridd, Wales (Lark, McCarthy); Institute of Food, Nutrition, and Human Health, Massey University, Wellington, New Zealand (Lark); and Department of Sport, Culture, and the Arts, University of Strathclyde, Jordanhill Campus, Glasgow, Scotland (Rowe).

No commercial party having a direct interest in the results of the research supporting this article has or will confer a benefit on the authors or on any organization with which the authors are associated.

Correspondence to Sally D. Lark, PhD, Institute of Food, Nutrition, and Human Health, Massey University, Private Bag 756, Wellington 6140, New Zealand, e-mail: *s.lark@massey.ac.nz.* Reprints are not available from the author.

0003-9993/11/9205-00649$36.00/0
doi:10.1016/j.apmr.2010.11.028

List of Abbreviations

ANOVA	analysis of variance
ICC	intraclass coefficient
LOA	limits of agreement
MMSE	Mini-Mental State Examination
PWT	parallel walk test
TUG	Timed Up and Go
TWT	tandem walk test

(severe hypertension, unstable angina), stroke, severe breathing problems, Parkinson's disease, peripheral neuropathy (eg, diabetic), or rheumatism/arthritis of the lower limbs that was painful on the day of examination. In the final interrater analysis, 34 patients were included, and 15 of these were retested 1 week later.

Parallel Walk Test

For the PWT, participants walked at their normal gait (step and stride length) and speed for 6m between 2 parallel lines placed at a width of either 20, 30.5, or 38cm (8, 12, and 15 inches). Each participant achieved a total footfall score (SC) based on +1 when any part of the foot was placed on the line and +2 when the footfall was outside the line or they reached for something to maintain balance (eg, wall or railing that was ~1m away and required the person to step outside the lines to reach it). Higher scores denoted worse performance and therefore more unstable gait. Time taken to complete the test also was recorded for comparison and to calculate velocity. For scores, lower score denoted better performance and therefore more stable gait.

Each subject carried out the tests in random order. Participants wore their own footwear, which generally was low-heel rubber-soled shoes, and were allowed to use their walking stick, if required, of which 13 of the 34 initial subjects and 6 of the 15 retested participants used a cane. All participants performed a short walk as a warm up (~20m) and 1 familiarization session for the 6-m length and starting instructions. They started walking on a verbal cue, and in each case, subjects were asked to look directly ahead and not at their foot placement. Raters 1 and 2 stood at opposite ends of the 6m. Both raters independently recorded the time and footfall scores for each subject during every test. This was repeated a week later for 15 subjects.

Raters

Raters were the first and second authors of this study. Rater 1 (S.D.L.) had more than 10 years experience working with elderly fallers and assessing balance and gait and had administered the PWT on more than 70 occasions. Rater 2 (P.W.M.) had not previously administered the PWT on an elderly population before this study and was given a short training session with instructions and practice on 4 patients.

Data Processing and Analysis

Footfall scores from each rater and test administration were transferred from a data collection sheet into an SPSS[a] data file. For each participant, test administration resulted in 12 scores; SC and time scores for each of the 3 conditions (8, 12, and 15 inches) from each of the 2 raters. This was repeated for the second test administration, adding a second set of 12 scores for the 15 participants for whom data were analyzed for the test-retest reliability analysis. Interrater agreement was estimated using scores from the 34 patients who participated in test administration 1. To estimate test-retest reliability for the 15 patients who participated in both test administrations, individual rater scores were averaged to minimize any test-retest variability caused by rater variability from test administration 1 to test administration 2 and focus on test-retest variability due to patient performance variability. Before the reliability analyses, descriptive statistics were calculated, including skewness and kurtosis, to evaluate the data for normality of distribution and the presence of outliers.

After data checking by using descriptive statistics and inspection of individual data points when indicated, a similar analysis plan was used to investigate interrater agreement and test-retest reliability. Disagreements between raters (or differences between test administrations) were evaluated by using methods described by Bland and Altman.[13] Systematic bias was assessed by using t tests for the significance of mean differences and calculations of Cohen's d, a standardized effect size indicating meaningfulness of any differences. Cohen suggested standards for d of .20 as small, d of 0.5 as medium, and d of 0.8 or higher as large.[14] Proportional bias was evaluated by using the correlation between differences between raters (or between test administrations) for each participant and the mean score of both raters (or both test administrations) for each participant. Additionally, the size of individual difference scores (ie, difference between raters for each participant or difference between test administrations for each participant) was evaluated by calculating 95% LOAs. Association between observations was evaluated by using ICCs from the 1-way ANOVA model, which were adjusted for a single rater or test administration by using the Spearman-Brown formula. A standard of .70 was used to indicate a minimally acceptable level of reliability.[15] All significance tests were conducted using α of .05.

RESULTS

Patient data, including Barthel Index (range, 16–20) and MMSE scores, are listed in table 1. There were twice as many men as women, and the subgroup included in the test-retest analysis (n=15) was representative of the patient group as a whole (N=36), as indicated by similar demographic data (see table 1).

Data Checking

From descriptive statistics, scores for most PWT subtests appeared to be relatively normally distributed, denoted by skewness and kurtosis values less than 2.0. The main exception was the footfall score of the 15-in width PWT (PWT15SC) subtest, which had especially high kurtosis values. Normality of distribution is an important assumption underlying the use of such parametric analyses as t tests and ICCs, although such analyses can be robust to insubstantial violations of the underlying assumptions.[16] One participant was missing data for the 8-in width for the footfall scores (PWT8SC) and recorded time at the same width (PWT8t1) from rater 2 at test administration 1 because of external factors interfering with test conditions, and 1 participant was deemed to be an outlier for PWT8t1 at test administration 1 (based on a scatterplot of data), and their data were removed from subsequent analyses. This resulted in a smaller sample size for the PWT8-in width subtest, as noted in the results tables 2-5.

Table 1: Patient Demographic Data

Sample	Men/Women	Age (y)	Mass (kg)	Barthel Index Score	MMSE Score
Interrater (n=34)	12/24	81.3±5.4	72.8±11.7	18.3±1.3	26.1±2.4
Test-retest (n=15)	5/10	80.3±5.3	73.0±12.0	18.2±1.4	26.1±2.2

NOTE. Values are mean ± 1 SD for patients included in the interrater and test-retest analyses.