Open Journal of
Clinical & Medical
Case Reports

Volume 8 (2022)
Issue 13

Review Article

ISSN: 2379-1039

Reliability and responsiveness of thirty-one mobility assessment tests for older adults: A systematic review

Racha Soubra*; Aly Chkeir; Jean-Luc Novella

***Corresponding Author: Racha Soubra**

Laboratoire de Modélisation et Sûreté des Systèmes (M2S), Université de Technologie de Troyes, 12 rue Marie Curie, 10004 Troyes, France.

Email: racha.soubra@utt.fr

Abstract

Nowadays, numerous measuring instruments for assessing the mobility of older adults are available in research and clinical practice. However, health care professionals and clinicians are constantly searching for a suitable and accurate measure in order to improve the thoroughness of their evaluation and prevent biased outcomes. Generally, a chosen instrument must provide accurate, valid, robust and interpretable outcomes. Consequently, the clinical feasibility as well as the psychometric properties of the measure should be taken into consideration for an informed decision. In our previous study, we provided an overview of important practicality characteristics and validity outcomes of 31 mobility assessment tests. In this study, we provide a systematic review of studies that examined the reliability and/or the responsiveness of one or more mobility tests among the 31 tests previously reviewed.

The objective of our research is to serve as a general guideline for health care professionals and scientific communities to help them select a convenient assessment tool that fulfill the purposes of their studies. From this research, we concluded that the majority of these tests have moderate to excellent reliability outcomes. Nevertheless, most of the studies show a lack of attention in evaluating the responsiveness of these tests.

Keywords

Older adults; Mobility assessment tests; Geriatric assessments; Psychometric properties; Reliability; Responsiveness.

Background and purpose

The ability to walk or transfer from one place to another is a key predictor of quality of life (QoL) among older adults [1-3]. Mobility functions such as gait, balance and transfers are initially linked to health

status. They are crucial aspects of an independent living and a successful aging. Nevertheless, about 30% of older adults encounter mobility limitations [1]. These limitations are commonly the results of cognitive conditions, osteoarthritis, muscle weakness, joint problems, pain and other natural changes [4]. In general, such mobility problems may lead to undesirable physical, cognitive and social consequences. They often cause a decline in independence, physical disability and injuries, institutionalization and an increase in hospital admissions [3,5,6]. Therefore, early interventions are significant for elderlies in order to maintain or regain the daily activity levels, achieve a healthy ageing and attain a better QoL over time [1].

A major challenge remains for clinicians and researchers to monitor the functional mobility effectively, knowing when and what type of interventions are necessary to prevent mobility loss and improving the QoL of older adults [7,8]. They mainly refer to mobility assessment tests, which play an important role in research, clinical practice and health assessment. Previously, numerous tests have been devised to evaluate gait, transfer and balance of older adults. They are often used in order to identify changes in an individual's mobility, detect early signs of decline, and assist in guiding therapeutic interventions [5,6]. However, suitable and accurate measurements are fundamental in order to ensure the thoroughness of an evaluation, correctly intervene in reducing morbidity, select appropriate plans of care, motivate elderlies and enhance communications between therapists and patients [8,9]. The selection of an unsuitable, or poor quality, outcome measurement test may introduce bias in the outcome [10]. Accordingly, it is highly significant to carefully select the correct mobility evaluation test in order to ensure the quality of results. For an appropriate selection of a mobility assessment test, it is important to understand the measurement tool in details: such as the time of administration, the required equipment; the targeted population; the assessment forms; the results interpretation, etc. In addition, the chosen instrument is supposed to provide accurate, valid, robust and interpretable outcomes [11]. Consequently, clinical feasibility, as well as the psychometric properties of the measure, must be taken into account for an informed decision. Three psychometric properties define the quality of information provided by an instrument: validity, reliability and responsiveness to change of the outcome measure [6,10-12].

In our previous study, we found thirty-one assessment tests for evaluating the mobility of older adults [13]. A general description about each test and its validity has been provided. Although the clinical feasibility and practicality information are highly significant, however, the quality of information resulting from a measurement depends partially on the test's psychometric properties. Accordingly, the purpose of this work is to investigate the reliability and responsiveness of these assessment tests. A broad review was performed in order to summarize the outcomes of all studies that estimated the two aforementioned psychometric properties.

Measurement scales

Reliability: In clinical research, the decision to use a measurement instrument depends on how accurate and meaningful the outcomes are. However, according to classical test theory, any observed score (X) is composed of two components: the true score (T) which is unknown and an error component (E) [14]. The difference between X and T values represents the measurement error or "the noise" that inhibits the

findings of the true score. As error components are unknown, reliability is evaluated to detect the degree to which a clinical test scores are free from measurement errors [15]; it estimates the amount attributed to error and the amount that represents the true value based on the statistical concept of variance. It can be expressed as the ratio between the true score (T) variance and the total score (T+E) variance. This ratio yields to what is called the reliability coefficient. This latter increases when X approaches the T, giving a maximum value of 1 when zero error is found, and decreases to 0 when X approaches the E.

Different factors and various conditions may produce systematic and random errors. Thus, reliability must be estimated to test the stability and repeatability of measurements, which can be affected by the participants, the observers, the environment, the test itself and other circumstances. In the following, estimates of reliability appear under four general approaches: Test-retest Reliability, Rater Reliability, Internal Consistency and Alternate Forms Reliability [15].

Test-Retest Reliability: refers to the stability of measurements when a test is administered two or more times on the same group of participants under the same conditions. These conditions must be as constant as possible and they include the raters, environment, equipment used, etc. Accordingly, results will show the correlation and the strength of association between the outcomes of a test performed at different points in time. Nevertheless, it is important to select a convenient time interval between repeated measurements. The time interval should be long enough to avoid fatigue or memory loss, and short enough to avoid changes in health or learning.

Rater Reliability: As many clinical measurements necessitate the intervention of a human observer/rater, it is estimated that the true source of errors originates from this intervention. Two types of rater reliability are available. First, the Inter-rater reliability, which refers to the equivalence of measurements when a test is administered once to the same group of participants but with different observers. Second, the Intra-rater reliability which refers to the stability of outcomes resulted from two or more trials of a test performed by one observer on the same group of participants. It is usually recommended to have more than two trials performed successively within a short time interval.

Several methods are commonly used to estimate Test-retest and Rater reliabilities. The list includes: Pearson's Correlation Coefficient, Spearman rho, Bland and Altman Plots and coefficient of variation. However, the Intraclass Correlation Coefficient (ICC) is the most frequently used method and considered as a preferred index since it reflects both correlation and agreement between the outputs [16].

Internal consistency: is a reliability form used to estimate the homogeneity and inter-relatedness between the items of a test. It evaluates the extent to which all items of a test are measuring the same concept. Four statistical methods are widely used to estimate internal consistency: Cronbach's Alpha Coefficient, Split Half Method adjusted with the Spearman-Brown correction, Average Inter-Item Correlation and Average Item-Total Correlation techniques.

Alternate forms reliability, also known as parallel form reliability, estimates the error developed between two or more versions of the same measurement test or instrument. It is evaluated using the

correlation coefficients and the limits of agreement measures.

In some contexts, the term reliability is replaced by repeatability and/or reproducibility depending on the degree of consistency [17]. Moreover, in 1989, Baumgarter defined reliability in two forms: relative and absolute reliability [18]. Furthermore, Rakin and Stokes declared that the ICC is unsuitable for use in isolation [19]. They showed that there is no standard acceptable level of reliability that could be provided while using this estimation technique alone. Accordingly, results must be complemented by the computation of the confidence interval construction, the standard error measurement (SEM) or the minimal detectable change (MDC).

Responsiveness

For evaluative instruments, "responsiveness" or "sensitivity to change" has been recommended as a requirement. However, there is a significant lack of clarity about the definition of this property in the literature [20]. Consequently, there is an inconsistency in the methods used to estimate responsiveness. Initially, responsiveness refers to the extent to which a measurement tool can detect a change. Hence, definitions vary according to the type of change detected, (e.g. clinically important changes, or changes due to treatment effects, etc). In a systematic review on assessing responsiveness of health-related quality of life instruments, Terwee et al [20] identified twenty-five definitions and thirty-one measures for responsiveness. As many similarities exist between the definitions, authors grouped them into 3 categories:

First group in which responsiveness is defined as the ability to detect change in general. This assembles any type of change, whether it is relevant or meaningful. This category of responsiveness is often defined to detect a statistically significant change after treatment.

Second group in which responsiveness is defined as the ability to detect a clinically important change. This requires an explicit and subjective judgment on what is to be considered important.

Third group in which responsiveness is defined as the ability to detect real changes in the concept being measured. This require, not only a judgment on what changes are important, but also a gold standard for the concept being measured.

31 different measures for evaluating responsiveness were found and summarized in Table 2 of [20]. A variety of methods exist for each of the three groups, and the same methods sometimes appear in multiple groups. The list of measures includes the effect size, the Guyatt's responsiveness statistic, the standardized response mean, the measurement of sensitivity and specificity, regression models, the correlation with overall improvement, etc.

The above-mentioned grouping is based on the kind of change that a responsive instrument must detect. However, other ways of grouping are also possible. For instance, Husted et al. discriminated between internal and external responsiveness. Internal responsiveness of a measure is defined by its ability to change over a time frame. However, external responsiveness of a measure represents the extent to which changes over a defined time frame relate to corresponding changes in a reference measure. On the other hand,

Crosby et al. discriminated between distribution-based and anchor-based methods. First, the distribution-based methods include all measures that are based on statistical significance, sample variability and measurement precision. Conversely, the anchor-based methods include both cross-sectional approaches and longitudinal approaches.

Methodology

This systematic review was completed in two consecutive phases. First, a broad research was performed to summarize all papers reporting the reliability of mobility assessment tests with their different approaches and estimation techniques. Then, second research phase was performed to detect the responsiveness of these tests. In our search, we only looked for articles published in English. However, the year of publication and the number of citations were not taken into consideration in order to gather all studies. Summary tables were constructed by a single researcher (RS) to document the attained information, supervised by (AC) and then it underwent a full review and agreement by two researchers (AC and JN).

Phase No 1: Reliability properties search

Search strategy: For each of the thirty-three mobility assessment tests, previously shown in [13], a research was performed to gather all studies reporting on at least one of the aforementioned reliability approaches. A systematic search was conducted using the databases to which we have access from the University of Technology of Troyes. The list of databases includes Science Direct, Scopus, SAGE, Springer, and Wiley. Moreover, a manual search was performed on Google Scholar in order to collect all available references. Our search was performed using the name and/or acronym of each test associated with the terms "reliability", "test-retest reliability", "intra-rater reliability", "interrater reliability", "internal consistency", "absolute reliability", "relative reliability", "repeatability" or "reproducibility". A manual database and Boolean searching were also implemented. Paper collection and data extraction were fulfilled by one author (RS) and examined by two authors (AC and JN).

Selection criteria: A study was included if the following information was provided in the full text articles: The type of studied reliability, a general description about participants (ex. number, age and health), the years of experience attained by each rater for interrater reliability, the number of trials and time intervals for intra-rater reliability, the number of repeated measurement and time intervals for test-retest reliability, if a training or trial test was performed, and the measurement techniques used for reliability estimation (ICC, Pearson's correlation…).

On the other hand, a reliability study was excluded when a mobility assessment test is performed by elderly subjects with specific disease or illness (e.g. Traumatic Brain Injury, Parkinson's Disease or Stroke), as the targeted population of our research covers healthy elderly people only. An article was also excluded when the original version of a test is translated into different languages.

Phase No 2: Responsiveness properties search

Search strategy: In this review, a second systematic review was accomplished to identify all studies

reporting the responsiveness of the thirty-one mobility assessment tests. The search strategy is similar to phase N°1. A broad search was performed using the above-mentioned databases to which we have access, and followed by a manual search on Google Scholar. Papers were screened based on their title and abstracts. The key searched terms include the name and/or acronym of each test associated with the terms "responsiveness", and "sensitivity to change".

Selection criteria: The inclusion criteria for this study is that information about responsiveness was available and revealed in the abstract of a paper. An article was considered relevant when it tackles an experiment on community-dwelling elderly people to assess the sensitivity to change of a mobility assessment test.

Results

Data collection

A total of 31 elderly mobility assessment tests had been previously found and discussed in our previous systematic review [13]. However, based on our inclusion and exclusion criteria, the reliability of 28 tests and the responsiveness of 8 tests were interpreted in this review.

No studies reporting the reliability approaches of the following 3 tests were found: Instrumented Stand and Walk (ISAW), Backward Walking and Pick Up Weight Test. Seven excluded studies were found reporting the reliability of 2 tests: BesTest and Standardized Walking Obstacle Course (SWOC). However, in these studies, the reliability approaches of both assessments were tested on elderly participants with certain diseases such as cancer, Parkinson's disease, subacute stroke and chronic hemiparesis. Additionally, three excluded articles reported the reliability of the Timed up and Go (TUG), Functional Gait Speed (FGA), Short Physical Performance Battery (SPPB) and Usual Gait Speed (UGS) tests since instructions are translated to an alternative language such as Persian.

For responsiveness outcomes, a significant number of studies reporting the responsiveness of mobility assessment tests performed on older adults with specific disease or illness (e.g. Traumatic Brain Injury, Parkinson's Disease or Stoke) were found. Furthermore, in some context, it is clear that the term "responsiveness" was replaced by "longitudinal study" or "cross-sectional study". Nevertheless, as per our inclusion and exclusion criteria, those papers were not mentioned in this review. A study was considered relevant to our review when the terms "responsiveness" or "sensitivity to change" are mentioned in the title and/or abstract and when it tackles an experimental procedure on healthy community-dwelling elderly people. From our findings, only eight papers reported the responsiveness of 8 mobility tests when applied on healthy and community-dwelling elderly people. However, to date, there have been no studies performed on control older adults to estimate the responsiveness of the twenty-three remaining mobility tests.

The Flow diagram, shown in Figure 1, documents our complete literature search.

Both psychometric criteria of the mobility evaluation tests are discussed below in details

and obviously summarized in Tables 1 and 2 of Appendix I. Furthermore, it is worth noting that some experimental procedures were settled in order to explore the reliability of several mobility assessment tests at the same time in the community dwelling elderly. The outcomes of those studies were summarized in Table 3 of Appendix I.

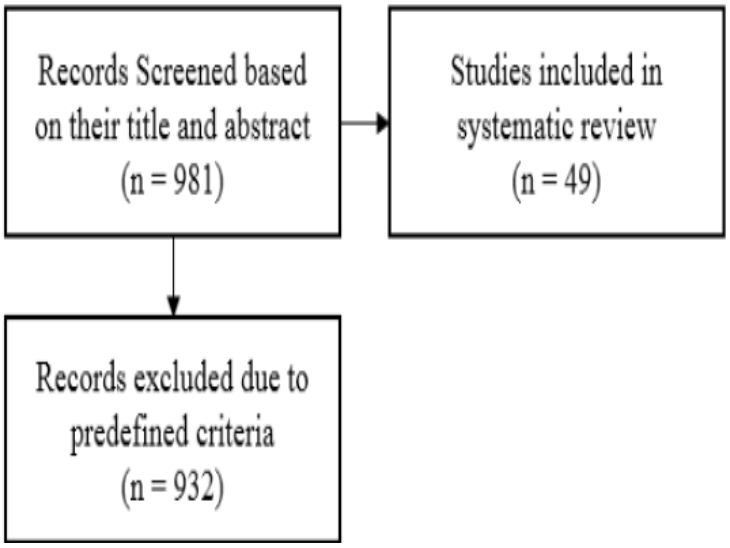

Figure 1: Flow Diagram of our Review.

Appendix I

Table 1: Summary table reporting the reliability of the mobility assessment tests.

Test	Author (Year)	Population (Mean)	Type of Reliability/ outcome measure	Measurement Description	Results
TUG	Bohannon et al (2005)	20 community dwelling elders (75 years)	Long-term test-retest	3 tests: . T1 (at instant t1) . T2 (at t1 + 6 months) . T3 (at t1 + 12 months)	Good test re-test reliability: ICCs of 0.83 between T1/T2, 0.82 betweenT2/ T3 and 0.74 between T1/T3
	McGrath et al (2011)	33 aged people	Intra-session Reliabilityof 44 parameters	6 TUG tests with 1-min rest-time	. 25 parameters shown an excellent intra-session reliability (ICC>0.75) . Turn time parameters have poor reliability
			Test-retest Reliability of44 parameters	2 tests: T1 and T2 (after 4weeks)	Turn time parameters have poor reliability
	Pernille et al (2006)	18 elderly subjects with mobility impairments	Intra-rater Reliability	2 tests performed within 1 hour	Good Reliability: ICC(2.1)=0.91-0.97 (total)and 0.7-0.97 (items)
			Inter-rater Reliability	Under the supervision of 3 raters	Good Reliability: ICC(2.1)=0.9 (total) and 0.63-0.92 (task) except for turn and sit down (ICC=0.37)
			Internal Consistency	Not Found	Cronbach's alpha of 0.74
SPPB	Ostir et al (2002)	Group of moderately to severely disabled older women	Short-term Test-retest	1 test/week (20 weeks)	Excellent reliability ICC from 0.88 to 0.92 (between weeks 5/6, 12/13 and 19/20)
			Long-term Test-retest	Test after 36 months	Slow decline of ICC = 0.77 (0.72-0.79)
8UG	Rikli & Jones (1990)	42 women and 34 men,over 60 years old	Test-retest Reliability	2 tests within 1 week	Very high reliability with ICC (CI%) = 0.90 (0.83 -0.95) across women group 0.98 (0.96 – 0.99) across men group 0.95 (0.92 – 0.97) across all participants

UGS (4- and 10-m UGS)	Denise et al (2013)	43 healthy adults (84.3 years)	Test-retest Reliability	3 consecutive walking trials with rest break given as needed	Excellent test-retest reliability with an ICC (3,1) values of 0.96-0.98, a SEM results smaller than 0.004-0.008 m/s and MDC values between 0.01 and 0.02 m/s
PPT	Reuben et al (1990) (7- and 9-items PPTs)		Inter-rater Reliability	Not Found	High reliability ICC=0.93 (PPT-7items) and ICC=0.99 (PPT-9items)
			Internal Consistency	Not Found	High reliability Cronbach's alpha of 0.79 and 0.87 for PPT 7- and 9-items respectively
	King MB et al (2000) (8-items PPT)	18 to 22 individuals with mobility impairment	Test-retest Reliability	3 tests with a period of 1-2 weeks between tests	ICC=0.88
			Inter-rater Reliability	4 testers	ICC=0.96
			Internal Consistency	Not Found	Cronbach's alpha=0.785
5TSTS	Goldberg et al (2012)	29 females (73.6 years)	Test Retest Reliability	2 trials	Excellent relative (ICC (2,1)=0.95) and absolute reliability (SEM = 0.9 sec)
	Wallmann et al (2013)	92 elderly subjects (65 years)	Inter-rater Reliability	3 clinicians with similar clinical experience evaluatingvideotapes	Excellent reliability with ICC (2,1)=1
	Matthew et al (2016)	35 volunteers (30 to 75 years)	Test-retest Reliability	3 trials	Good test-retest reliability with ICC (3,2)=0.96-0.98
L-test	Death et al (2005)	27 subjects with unilateralamputations	Intra-rater Reliability	3 trials repeated after 2 weeks	ICC(2,1)=0.97
			Inter-rater Reliability	Two raters	ICC(2,2)=0.96
	Nguyen et al (2007)	50 older adults (84 years)	Intra-rater Reliability	2 tests	2-way ANOVA ICC (95% CI) = 1 (0.99-1)
			Inter-rater Reliability	Not Found	2-way ANOVA ICC (95% CI) = 0.97 (0.95-0.98)
DEMMI	De Morton et al (2010)	Older acute medical patients	Intra-rater Reliability	Not Found	Reliable test with a maximum change score of 9 points
			Inter-rater Reliability	Test developer with another experienced physiotherapist	
F8W	Jarnlo et al (2009) (mF8W)	30 community dwelling women (76.5 years)	Test-retest reliability	Test repeated after 1 week	ICC of 0.93 for speed evaluation, and lower ICC of 0.73 for oversteps score (amplitude)
			Inter-rater Reliability	Not Found	Very high inter-rater reliability
	Hess et al (2010)	18 older adults with mobility disability (83.9 years)	Test-retest reliability	2 trials	ICC of 0.84, 0.82 and 0.61 for speed, amplitude and accuracy outcomes respectively
			Inter-rater Reliability	2 assessors	ICC of 0.90, 0.92 and 0.85 for speed, amplitude and accuracy outcomes respectively
HABAM	Chris Mac et al (1995)	15 hospitalized elderly subjects	Inter-rater Reliability	2 physicians	High reliable ICC(2,1) of 0.94
	Rockwood et al (2008)	167 frail older adults	Inter-rater Reliability	1 Geriatrician and 1 observer	High reliable ICC of 0.92
		63 inpatient older adults	Test-retest Reliability	2 tests in 2 consecutive days	Total score and mobility phase: ICC=0.91Balance phase: ICC=0.85 Transfers phase: ICC=0.82
TWT	Yamada et al (2010)	171 elderly participants (80.5 years)	Test-retest Reliability	2 tests with an interval of 2weeks	High reliability with ICC(1,1)=0.945
PWT	Lark et al (2011)	16 elderly fallers	Test-retest Reliability	2 tests; test repeated after 1 week	ICC range 0.63 to 0.90
		36 elderly fallers (81.3 years)	Inter- rater Agreement	1st and 2nd authors	High degree of reliability - ICC range of 0.93 to 0.99
CHARMI	Liebl et al (2016)	30 patients	Inter-rater Reliability	1 physician and 1 physiotherapist	Excellent reliability (Cohen's kappa=0.88) 11 items showed an exact agreement and 3 items showed a minor variation

Tinetti - POMA	Faber et al (2006)	30 elderly participants(84.9 years)	Intra-rater Reliability	2 tests within two consecutivedays	Rater 1: Spearman R= 0.86 (POMA-T), 0.78 (POMA-B) & 0.72 (POMA-G) Rater 2: Spearman R= 0.82 (POMA-T), 0.74 (POMA-B) & 0.77 (POMA-G)
			Inter-rater Reliability	2 graduate students with 8-hourtraining in scoring	Day 1: Spearman R= 0.93 (POMA-T), 0.90 (POMA-B) & 0.80 (POMA-G) Day 2: Spearman R= 0.91 (POMA-T), 0.88 (POMA-B) & 0.89 (POMA-G)
DGI	Shumway-Cook et al (1997)	5 community dwelling elderly people	Inter-rater Reliability	5 therapists	ICC=0.96
		2 community dwelling elderly people	Test-retest Reliability	2 tests, second test repeated after 1 week	Excellent reliability with ICC=0.98
	Boulgarides et al (2003)	3 subjects	Agreement between observers	Testers evaluating videotapes	80% or better agreement
FGA	Wrisley et al (2004)	6 patients with vestibulardisorder	Intra-rater Reliability	2 tests with 1-hour rest time	ICC(2,1)=0.74
			Inter-rater Reliability	10 raters	ICC(2,1)=0.86
			Internal Consistency	Not Found	Cronbach's alpha=0.79
AST	Hill et al (1996)	14 healthy elderly subjects	Retest Reliability	Not Found	High reliability with ICC>0.90
	Tiedmann et al (2008) Butler & Anne studies	30 elderly participants (80.1 years)	Test Retest Reliability	Test repeated after 2 weeks	Excellent reliability with ICC(3,1) of 0.78 with 95% CI of 0.59-0.89
EMS	Smith et al (1994)	15 Hospitalized elderly people	Inter-rater Reliability	2 therapists	No significant difference between scores
	Prosser et al (1997)	19 Hospitalized elderly people	Inter-rater Reliability	2 physiotherapists	Significant correlation between scores (Spearman R = 0.88)
PPME	Winograd et al (1994)		Inter-rater Reliability	Not Found	High reliability
			Intra-rater Reliability	Not Found	High reliability
FOC	Means et al (1996)	Elderly people*	Test-retest Reliability	Test repeated after 2 weeks	
			Inter-rater Reliability	3 independent raters scored 10videotapes	Bivariate correlations between rater pairings for the time and quality scores exceeded 0.98
	Rubenstein et al (1997)	58 community dwelling older men (75 years)	Inter-rater Reliability	Physical therapist & physician scoring videotapes	Kappa score of 0.96
TURN180	Thigpen et al (2000)	2 groups of elderly whohave and have no difficulties in turning	Intra-rater Reliability	2 tests performed within 2 weeks	Good to excellent reliability using a stopwatch ICC of 0.99 (type of turn), 0.90 (number of steps), 0.96 (time taken by processor), 0.67(time taken by stopwatch), 1 (straggling during the turn)
	Fitzpatrick et al (2005)	66 elderly people (82.45 years)	Repeatability	3 trials with 1-3 minutes rest time	Good repeatability for the number of steps (ICC of 0.828)
			Observer-agreement	1 physiotherapist and 1 therapist	Good agreement between observers with maximum of 1-step difference
DUKE	Cited by Duncan et al (1992)		Inter-rater Reliability	Not Found	High reliable ICC of 0.97
			Test-retest Reliability	Not Found	High reliable ICC of 0.97
LSMA	Baker et al (2003)	306 community dwellingelderly (75 years)	Short-term Test-retest	2 tests within 2 weeks	High degree of stability (ICC = 0.860.96)
			Long-term Test-retest	A follow up by telephone interview after 6 months	Increases and decreases in outputs (ICC between 0.49 & 0.81)
mGES	Newell et al (2012)	26 community dwelling elderly	Test-retest Reliability	2 tests within 1 month	High ICC(2,1)=0.93 and SEM=5.23
			Internal Consistency	Not Found	Cronbach's alpha=0.94

Table 2: Summary table reporting the responsiveness of the mobility assessment tests.

Test	Study	Participant	Results
SPPB	Ostir et al [25]	102 moderately to severely disabled women aged 65 and older	Highly responsive to change test
6MWT	King et al [30]	26 volunteers with early mobility impairment	Responsiveness index of 0.6
PPT	King et al [30]	26 volunteers with early mobility impairment	– Responsiveness index of 0.8 – PPT-8 is a sensitive to change test
DEMMI	de Morton et al [71]	Older acute medical population	Responsive DEMMI instrument
HABAM	Macknight et al [40]	Hospitalized patients	– relative efficiency of 3.13 – effect size of 0.59 – Responsive test
CHARMI	Liebl et al [44]	Participants from the acute care rehabilitation	a large responsiveness to change outcomes
Tinetti – POMA	Faber et al [48]	30 participants	any change in score that exceed 5 points for individual level and a mean group score greater than 0.8 for will be considered asreliable change
EMS	Spilg et al [72]	Mixed populations of inpatients and outpatients	a measurement tool that can detectimprovement in mobility

Table 3: Summary table of studies reporting the reliability of several mobility assessment tests.

Study	Methodology	Results
Tiedemann et al. [1] and Butleret al. [2]	30 elderly people undertook 6 mobility tests 2 times (after 2 weeks) in order todetermine the test-retest reliability - ICC(3.1)	Excellent test-retest reliability for the 5TSTS (ICC (3,1) = 0.89), the AST (ICC(3,1) = 0.78) and the half-turn test (ICC (3,1) = 0.75). Fair to good test-retest reliability for UGS – 6 meters (ICC (3,1) = 0.74) and 1TSTS (ICC (3,1) = 0.54). All subjects completed Pick up weight test.
Wang et al. [3]	77 community dwelling elderly performed 5 mobility tests in 2 sessions (1 weekapart) (TUG, 6MWT, UGS, FGS and 5TSTS)	Excellent test-retest reliability for all measurements ICC (2.1) = 0.80-0.95. SEMwithin 10% and smallest real difference within 26%. FGS showed the highest reliability.
Tiedemann et al. [4]	30 elderly people underwent for 2 physical assessment tests on 2 occasions 2weeks apart - ICC (2.1)	Excellent test-retest reliability for STS (ICC (2,1) = 0.89) and AST. ICC (2,1) =0.78).
Jette et al. [5]	4 performance based measurements were performed two times, afterapproximately 2 weeks	The test-retest reliability outcomes for 8UG (ICC = 0.79), the old version of TUG(ICC = 0.74), the 1TSTS (ICC = 0.25) and 5TSTS (ICC = 0.67).
Lin et al. [6]	15 elderly subjects participated in test-retest study of TUG, Tinetti POMA, and(tests were performed two times within 2 weeks)	Excellent inter and intrarater reliability for all tests. ICC ranged between 0.93 and0.99.

Reliability

Timed up and go (TUG)

Seventeen studies investigating the test-retest reliability of TUG test were previously found by Bohannon and Schaubert [21]. They comprehensively showed an acceptable short-term reliability for both older adults and for patients with specific pathologies with a correlation coefficient varying between 0.73 and 0.99. However, in order to estimate the TUG reliability over longer periods, authors compared the outcomes of 3 TUG tests obtained 6 and 12 months after the first examination among 20 community dwelling elders (mean age of 75 years). Results revealed a good test-retest reliability with ICCs of 0.83 between test 1 and 2, 0.82 between tests 2 and 3, and 0.74 between test 1 and 3. The decrease of ICC confirms the necessity of a practice trial as declared earlier by Moris et al [22]. After that, McGrath et al [23] investigated the intra-session and test-retest reliability for 44 parameters extracted while 33 elderly people performed 6 TUG trials with one-minute rest-time (intra-rater) and repeated the test after 4 weeks (test-retest). The

majority of parameters showed an excellent intra-rater and test-retest reliability (ICC (2,1) > 0.75), and the minority revealed fair to good outcomes (ICCs between 0.4 and 0.75). Besides, Pernille et al [24] aimed to study the intra-rater, interrater and internal consistency of the expanded version of TUG test. In their study, 18 elderly subjects with mobility impairment performed the test two times within one hour under the supervision of three raters. The test appeared to have a good reliability for experienced raters and acceptable internal consistency with Cronbach's alpha of 0.74.

Short Physical performance battery (SPPB)

The reliability of this measurement was firstly established by Ostir et al in 2002 [25]. ICC was reported for short-term and long-term test-retest reliability among a group of moderately to severely disabled older women. Results revealed an excellent test-retest reliability over three arbitrary chosen pairs of weeks (weeks 5 and 6, weeks 12 and 13, and weeks 19 and 20), and a slow decline in ICC outcomes appears for interviews made 36 months apart.

Six-minute walk test (6MWT)

The reliability of the 6MWT has been reviewed by Sadaria and Bohannon in 2001 [26]. As indicated, numerous studies have validated good to excellent reliability outcomes. Although each study has its own test formats and methodology, the reliability coefficients varied from 0.73 to 0.99. Additionally, it should be pointed out that these results refer to 6MWTs completed by a variety of elderly participants such as subjects with heart failure, chronic pulmonary or renal failure, lung diseases, and healthy older adults.

8-Foot up-and-go (8UG)

The test-retest reliability was reported to be very high for the 8UG test [27]. In this study, 42 women and 34 men (over the age of 60) performed the test twice within a period of one week. ICCs (CI%) were found to be 0.90 (0.83 -0.95) across the women's group, 0.98 (0.96 – 0.99) across the group of men and 0.95 (0.92 – 0.97) across all participants.

Usual or habitual gait speed (UGS/HGS)

A reliability study of 3 consecutive walking trials completed by 43 healthy older adults showed an excellent test-retest reliability for both 4- and 10-meters UGS [28]. Outcomes for both tests revealed an ICC (3,1) values of 0.96-0.98, a SEM results smaller than 0.004-0.008 m/s and MDC values between 0.01 and 0.02 m/s.

Physical performance test (PPT)

Both PPTs versions (7-item and 9-item PPT tests) showed a high internal consistency and interrater reliability. Reuben et al [29] found a Cronbach's alpha of 0.79 and 0.87, and an ICC of 0.93 and 0.99 for the 7-item and 9-item PPTs respectively. To evaluate the test-retest reliability for the PPT, King MB et al [30] suggested the use of an 8-item test and re-examining its internal consistency. In this version, the 9th item

(number of flights climbed) was dropped in order to prevent fatigue among participants. The test was performed 3 times, within a period of 1 to 2 weeks between performance, by 18 to 22 individuals with early mobility impairment under the supervision of four testers. Outcomes revealed a high ICC of 0.88 and 0.96 for test-retest and interrater reliability respectively, and a Cronbach's alpha of 0.785 for internal consistency.

5-Time sit-to-stand (TSTS) Test

In 2011, Richard W. Bohannon [31] conducted a systematic review to summarize the ICC outcomes estimated to describe the test-retest reliability for the 5TSTS test. He found 10 studies tested on older and community dwelling participants with an interval range of 2 days to 10 weeks between tests. Outcomes suggested good to high test-retest reliability with an ICC ranging between 0.64 and 0.96. Afterward, Goldberg et al [32] performed a 2 trials test on 29 females (mean age 73.6 years). They declared that the high ICC (ICC (2,1) =0.95) and low SEM (0.9 sec) outputs reveal an excellent relative and absolute reliability respectively for this test. Moreover, the good test-retest reliability has been confirmed by Matthew et al [33] with an ICC (3,2) =0.96-0.98. On the other hand, Wallmann et al [34] proved an excellent interrater reliability for this test (ICC (2,1) = 1). They conducted a study in which the videotapes of 92 elderly subjects (mean age of 65 years) performing the test were evaluated by three clinicians with similar clinical experience.

L-Test of functional mobility (L-Test)

Two studies have reported the reliability of this test. In 2005, developers conducted the reliability study across 27 people with unilateral amputations who performed 3 trials of L-test in a first instance and repeated the evaluation after 2 weeks under the supervision of two raters [35]. Results revealed an ICC (2,2) of 0.96 for interrater and ICC (2,1) of 0.97 for intra-rater reliability. Afterwards, the L-test was also found to be a reliable measure with an interrater and intra-rater 2-way ANOVA ICC of 1 (CI 0.99-1.00) and 0.97 (CI 0.95-0.98) respectively [36].

Backward walking (BW)

There have been no studies performed to estimate the reliability of this test.

De Morton mobility index (DEMMI)

In 2010, De Morton et al have reported the MDC with 90% confidence in order to estimate the interrater and intra-rater reliability during development and validation of DEMMI [37]. Additionally, Kappa statistics and absolute percentage agreement were calculated to assess item reliability. First, the test developer and another experienced physiotherapist examined older acute medical patients while performing the test. Then, using a five-point Global Rating Change (GRC), patients and therapist independently completed the rating of change between mobility at admission assessment test and at hospital discharge test. Finally, it was concluded that DEMMI is sufficiently reliable test having a maximum change score of 9 points. Nevertheless, this study was conducted for older acute medical patients, thus more examination is needed for other clinical populations.

Figure of 8 walk test (F8W)

Two studies demonstrated a high interrater and test-retest reliability of F8W and mF8W tests. First, Jarnlo et al [38] found an ICC of 0.93 for speed evaluation when 30 community dwelling women (mean age of 76.5 years) repeated the test one week after their first assessment. However, in this study, the test-retest reliability was lower for oversteps score (amplitude) showing an ICC equal to 0.73. Then, Hess et al [39] found an ICC of 0.84, 0.82 and 0.61 for test-retest reliability and an ICC of 0.90, 0.92 and 0.85 for interrater reliability of speed, amplitude and accuracy outcomes respectively among 2 trials performed by 18 older adults with mobility disability (mean age ± standard deviation of 83.9 ± 4.1).

Instrumented stand and walk (ISAW) Test

To date, the reliability of ISAW test has not been examined.

Hierarchical assessment of balance and mobility (HABAM)

Developers of HABAM found a high, reliable ICC (2,1) of 0.94 for inter-rater reliability that was evaluated by two physicians across 15 hospitalized elderly subjects [40]. Subsequently, Rockwood et al [41] affirmed the previous results with an ICC of 0.92 assessed by a geriatrician and an observer across 167 frail older adults. They also evaluated the test-retest reliability for the total HABAM and its three sub-components. The test was performed twice on 2 consecutive days by 63 inpatients. ICCs were 0.91 for total score and mobility subcomponent, 0.85 for balance subcomponent, and 0.82 for transfers subcomponent.

Trail walking test (TWT)

TWT was found to have a high test-retest reliability with an ICC (1,1) equal to 0.945 [42]. The study was completed by 171 elderly participants (mean age ± standard deviation of 80.5 ± 5.6 year) who performed the TWT two times with an interval of 2 weeks between assessments.

Parallel walk test (PWT)

Lark et al [43] conducted a study to determine the interrater agreement and test-retest reliability of the PWT. In order to evaluate the interrater agreement, 36 elderly fallers (mean ± standard deviation age of 81.3 ±5.4 years) were rated by the 1st and 2nd authors of the study. They found a high degree of reliability with an ICC range of 0.93 to 0.99 for widths of 20, 30.5 and 38 cm. On the other hand, to estimate the test-retest reliability, 15 participants repeated the test after a week. ICC outcomes were lower with a range of 0.63 to 0.90 for the 3 widths.

Charité mobility index (CHARMI)

Cohen's Kappa statistics of k=0.88 was reported for the interrater reliability of CHARMI test in a study where 30 patients were rated by a physician and a physiotherapist [44]. 11 items showed an exact agreement between the two raters, and the 3 remaining items showed a minor variation. Nevertheless, test-retest and internal consistency have not been described.

Standardized walking obstacle course (SWOC)

No studies have been found investigating the reliability of SWOC test across healthy elderly people. [45] and [46] reported the test-retest and interrater reliability of this test, however participants were elderly people with strokes and/or arthritis.

Pick-up weight test

As stated by Tiedemann et al [47], the reliability of this test was not calculable since all participants were able to reach down and pick up the object on two occasions.

Tinetti – performance oriented mobility assessment (POMA)

Intra-rater reliability, interrater reliability and responsiveness of Tinetti – POMA have been assessed by Faber et al [48]. In their study, the Tinetti test was performed twice within two consecutive days by 30 elderly participants and scored independently by 2 graduate students who received an 8 hour-training in scoring. Spearman correlations outputs and Bland-Altman plots revealed a good relative reliability for POMA-T and its subscales (POMA-Balance and POMA-Gait), though POMA-Gait showed less performance results.

Berg balance scale (BBS)

A systematic review on relative and absolute reliability of the BBS has been executed by Downs et al in 2013 [49]. It was concluded that BBS has a high relative reliability with an estimated ICC of 0.97 and 0.98. Similarly, BBS demonstrated a high absolute reliability when participants' scores exceed the 20 points out of 56: MDC varied between 2.8 and 6.6 points. However, as declared by Downs et al [49], no studies were identified studying the absolute reliability of this test among participants with a mean score below 20. and chronic hemiparesis [52-56].

Functional gait assessment (FGA)

To study the test reliability, the FGA was repeated two times with a one-hour rest-time between trials by 6 patients having vestibular disorders [57]. 10 raters examined the performance and attained an ICC (2,1) of 0.74 for intra-rater and 0.86 for interrater reliability, with a Cronbach's alpha of 0.79 for the internal consistency. Additionally, the percentage agreement and Kappa values for each item and the total FGA score were also provided in this study.

Alternate step test (AST)

The test-retest reliability of AST was firstly found to be high with an ICC greater than 0.90 across a subgroup of 14 healthy elderly subjects [58]. Later on, [59] and [60] estimated an ICC (3,1) of 0.78 (CI 0.59-0.89) when a group of 30 elderly participants (mean age ± standard deviation of 80.1 ± 4) completed the test 2 weeks after their first performance.

Elderly mobility scale (EMS)

There have been no studies of test-retest reliability performed for the original EMS. Nolan et al [61] were the first to conduct such study, however all participants were recruited from medical, acute rehabilitation and orthopedic wards. For interrater reliability, Smith et al [62] and Prosser et al [63] showed a significant correlation between the outcomes of two therapists evaluating the EMS of hospitalized elderly people.

Physical performance and mobility examination (PPME)

As investigated by Winograd et al [64], both Pass-Fail and 3-level scoring techniques showed a high interrater and intra-rater reliability across elderly people with impaired mobility. For Pass-Fail scoring system, individual tasks revealed a mean percent agreement ranged from 96 to 100 % and Kappa ranged from 0.8 to 1 for interrater reliability. For summary scales, outputs revealed a Pearson product moment coefficient and a mean ICC of 0.99 for test-retest and interrater reliability respectively. For the 3-level scoring system, individual tasks revealed a mean percent agreement ranged from 90 to 100 % and Kappa ranged from 0.75 to 1 for interrater reliability. For summary scales, outputs revealed a Pearson product moment coefficient of 0.98 for test-retest and a mean ICC of 0.99 for interrater reliability.

Functional obstacle course (FOC)

Developers of the FOC have found reliable qualitative and quantitative outcomes for this test [65]. In their study, 3 independent raters scored 10 videotaped obstacle courses performed by elderly people and they reviewed them after 2 weeks for intra-rater reliability. Similarly, Rubenstein et al demonstrated a Kappa score of 0.96 while a physical therapist and a physician scored the videotapes of 58 community dwelling older men (mean age of 75 years).

TURN180

The time taken to accomplish TURN180, the number of steps completed and the staggering during the turn demonstrated good to excellent intra-rater reliability among two groups of elderly people with and without difficulties in turning while performing the TUG test [66]. In addition, Fitzpatrick et al [67] revealed a good repeatability of the number of steps in TURN180 test (ICC = 0.828) estimated across 66 elderly people (mean age 82.45 years) who performed three trials with 1 to 3 minutes of rest time between each.

Duke progressive mobility skills test

Duncan et al [68] cited a high reliable ICC of 0.97 for both interrater and test-retest reliability of Duke Mobility Skills test.

Life space mobility assessment (LSMA)

Baker et al [69] conducted a study to evaluate the test-retest reliability and stability of LSMA over a

short and long period of follow-up. Scores from 306 community dwelling elderly (mean ± standard deviation of 75 ± 6.8 years) were reported within 2 weeks and after 6-month follow up by telephone interview. It showed a high degree of stability over the short period (ICC The responsiveness of this test was assessed by Ostir et al. [25] on 102 moderately to severely disabled women aged 65 and older. The outcomes approved that SPPB test is highly responsive to change.

Six-minute walk test (6MWT)

A brief review of literature was provided by Sadaria and Bohannon [26] overviewing the responsiveness of the 6MWT. For all studies, the test was performed by elderly patients. However, only one study reported the responsiveness of this test when it was carried out by older adults with early mobility impairment. 26 volunteers were assigned to an exercise program and then 19 subjects were assigned to a group control [30]. The responsiveness index was 0.6 showing that there is no change in 6MWT distance in the intervention group when compared with the control group.

Physical performance test (PPT)

The responsiveness of this test was assessed by King et al. [30]. 26 volunteers were assigned to an exercise program and then 19 subjects were assigned to a group control. An improvement of 2.4 and 0.7 points were detected for the intervention and control groups respectively. The responsiveness index was 0.8 showing that PPT-8 is a sensitive to change test.

De Morton mobility index (DEMMI)

In [71], a distribution based index (Effect Size Index – ESI), and a criterion based index (Guyatt's Responsiveness Index – GRI) were assessed to evaluate the responsiveness of this test for older acute medical population. Outcomes revealed a responsive DEMMI instrument.

Hierarchical assessment of balance and mobility (HABAM)

HABAM measurement was proved to be a responsive test when applied to hospitalized patients [40]. This was demonstrated by a relative efficiency of 3.13 and an effect size of 0.59.

Charité mobility index (CHARMI)

Liebl et al. [44] initiated a study in which the CHARMI test was assessed during admission and at discharge. The outcomes demonstrated a large responsiveness to change between both phases ($|d| = 1.7$, $p < 0.001$).

Tinetti – performance oriented mobility assessment (POMA)

Faber et al. [48] interpreted the responsiveness of the POMA-T at both individual and group levels of 30 participants. MDC values were 4 to 4.2 and 0.7 to 0.8 for individual and group assessments respectively. These values indicate that any change in score that exceed 5 points for individual level and a mean group

score greater than 0.8 for will be considered as reliable change.

Elderly mobility scale (EMS)

Spilg et al. [72] demonstrated that EMS is a measurement tool that can detect improvement in mobility for mixed populations of inpatient and outpatients.

Discussion

In this review, a literature study was performed searching for studies reporting the reliability and responsiveness of elderly mobility assessment tests. Our main goal remains to support health care professionals, clinicians and researchers with a valuable reference guideline in order to wisely select an appropriate assessment test for elderly mobility evaluation.

In the domain of gerontology, human gait analysis has become an important area of research due its numerous interests and valuable prognosis. The evaluation of the way or manner an older person walks helps in identifying mobility impairments, detecting early signs of mobility decline and selecting convenient therapeutic interventions [5]. Nowadays, several mobility assessment tests have been devised. However, these evaluation tests differ from each other in various characteristics such as the performance steps, the format of assessment (i.e. performance based, judgment based or self-report), the methods of scoring and results interpretation, etc. It is known that, the choice of which test to use is based on the user's objectives as well as the properties of the test itself. As declared by [73], there is a lack of consensus on which assessment test to use. Nevertheless, it is important to select an accurate and appropriate one in order to improve the thoroughness of evaluations, determine precise plans of care, monitor progress better, motivate elderly people and enhance the communication between the geriatrician and the patient [8,9,74].

The findings of our previous systematic review revealed the existence of thirty-one assessment tests that have been developed to evaluate the mobility of healthy older adults [13]. We aimed to provide a general information set about each measurement test, its important practical characteristics and validity outcomes, if available. Accordingly, health care professionals, clinicians and researchers can refer to our reference guide to find more easily the information necessary to select a form of assessment based on their needs and the purpose of their study. We believe that a suitable test selection could be achieved by knowing the measurement tool in details. It is worth noting the clear description of each test, such as the administration time, the equipment required, the targeted population, the assessment format, the results interpretation, etc. In addition, the chosen instruments is intended to provide accurate, valid, robust and interpretable outcomes [11]. Consequently, both clinical feasibility and psychometric properties of the measurement must be taken into account for an informed decision. Three psychometric properties define the quality of information provided by an instrument: validity, reliability and responsiveness or sensitivity to change of the outcome measure [6,10-12]. First, evaluation of validity is used to determine the extent to which a test is measuring what it is supposed to measure. This concept is known to be one of the most important criteria for the quality of a test as it evaluates its accuracy and trustworthiness [75]. Second, evaluation of reliability is used to detect the degree to which a clinical test scores are free of measurement errors [15]. It

estimates the amount attributed to error and the amount that represents the true value. Finally, evaluation of responsiveness refers to the test's ability to detect a change over time. This is a very important characteristic of a measure since it detects improvements or worsening in mobility and determines the therapeutic effectiveness of a rehabilitation treatment [6].

As explained previously, the validity of the thirty-one mobility assessment tests has already been reported. Nevertheless, we recognize the importance of reviewing the two psychometric properties: reliability and responsiveness. Accordingly, our new contribution aims to provide a systematic review for health care professionals and the scientific community of these two aforementioned psychometric properties of all the available tests that are used to evaluate the mobility of older adults.

Information about the software information, its 'Model', 'Type' and 'Definition', with their 95% confidence intervals. Furthermore, it is worth mentioning that several adjectives have been devised to describe ranges of reliability values and to judge between adequate and inadequate reliability. For instance, the Landis and Koch [80] adjectives are proposed as follows: (0.0,0.10) virtually none; (0.11,0.40) slight; (0.41,0.60) fair; (0.61,0.80) moderate; and (0.81,1.0) substantial. These labels have been utilized in several studies, however they have been criticized by others [81,82]. Therefore, the ICC is unsuitable for use in isolation and should be completed by confidence intervals.

On the other hand, a significant number of studies were found reporting the responsiveness of mobility assessment tests performed on older adults with a specific disease or illness (e.g. Traumatic Brain Injury, Parkinson's Disease or Stoke). However, few studies were applied to healthy and community-dwelling elderly people. In this review, we aimed to gather studies with elderly control subjects only in order to provide a general guide for a suitable test selection. Our review demonstrated an important gap in examining the responsiveness of most of the mobility assessment tests. Our findings revealed that eight papers reported the responsiveness of 8 mobility tests. However, to date, no studies have been performed on older adult control groups adults to estimate the responsiveness of the twenty-three remaining mobility tests, even though, this psychometric property acts as a significant criterion for a suitable test selection.

The results obtained revealed that mobility assessment tests are responsive. Nevertheless, as declared by Terwee et al. [20], since several definitions and measures of responsiveness exists, different conclusions may be found when applying a responsiveness concept depending on the index selected. Hence, several papers ensure the use of multiple indices in order to specify a responsive instrument. Many authors also apply both anchor-based and distribution-based methods as both have To conclude, studying the three psychometric properties (i.e. validity, reliability and responsiveness) is very crucial for an appropriate test selection.

Conclusion

This research is a complementary study to our systematic review published in BioMed Research International Journal. After collecting all available mobility assessment tests that are used to evaluate gait, transfer and balance of healthy older adults, providing a clear description set for every tool and reporting

their validity, we aimed to summarize all papers studying the reliability and responsiveness of the tests. Our objective is to provide an overview of what is available, valid, reliable and sensitive to changes, which could serve as a general guide for health care professionals and scientific communities, enabling them easily to select a convenient assessment tool for their purposes and study.

Conflicts of interest: The authors declare that they have no conflicts of interest.

Acknowledgments: This work was supported by Troyes Champagne Métropole - (TCM) France, Fonds Européen de développement régional (FEDER), et Région Grand EST.

References

1. Musich S, et al. The impact of mobility limitations on health outcomes among older adults. Geriatric nursing. 2018; 39: 162-169.

2. Sakari-Rantala R, Heikkinen E, Ruoppila I. Difficulties in mobility among elderly people and their association with socioeconomic factors, dwelling environment and use of services. Aging Clinical and Experimental Research. 1995; 7: 433-440.

3. Webber SC, Porter MM, Menec VH. Mobility in older adults: A comprehensive framework. The Gerontologist. 2010; 50: 443-450.

4. Sakari R. Mobility and its decline in old age: determinants and associated factors. Studies in sport, physical education and health. 2013.

5. Macri E, et al. The de morton mobility index: normative data for a clinically useful mobility instrument. Journal of aging research. 2012. 2012.

6. Menezes KVRS, et al. Instruments to evaluate mobility capacity of older adults during hospitalization: A systematic review. Archives of gerontology and geriatrics. 2017; 72: 67-79.

7. Zijlstra W, Aminian K. Mobility assessment in older people: New possibilities and challenges. European Journal of Ageing. 2007; 4: 3-12.

8. Van Swearingen JM, Brach JS. Making geriatric assessment work: Selecting useful measures. Physical Therapy. 2001; 81: 1233-1252.

9. Middleton A, Fritz SL. Assessment of gait, balance, and mobility in older adults: considerations for clinicians. Current Translational Geriatrics and Experimental Gerontology Reports. 2013; 2: 205-214.

10. Mokkink, LB, et al. The COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) and how to select an outcome measurement instrument. Brazilian journal of physical therapy. 2016; 20: 105-113.

11. Souza ACd, Alexandre NMC, Guirardello EDB. Psychometric properties in instruments evaluation of reliability and validity. Epidemiologia e Serviços de Saúde. 2017; 26: 649-659.

12. Freiberger E, et al. Performance-based physical function in older community-dwelling persons: a systematic review of instruments. Age and ageing. 2012; 41: 712-721.

13. Soubra R, Chkeir A, Novella JL. A Systematic Review of Thirty-One Assessment Tests to Evaluate Mobility in Older Adults. BioMed Research International. 2019. 2019.

14. Crocker L, Algina J. Introduction to classical and modern test theory. 1986: ERIC.

15. Portney LG. Foundations of Clinical Research: Applications to Practice 3th (third) Edition. 2009.

16. Bruton A, Conway JH, Holgate ST. Reliability: what is it, and how is it measured? Physiotherapy. 2000; 86: 94-99.

17. Bartlett J, Frost C. Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2008; 31: 466-475.

18. Baumgartner TA. Norm-referenced measurement: reliability. Measurement concepts in physical education and exercise science. 1989; 20: 45-7.

19. Rankin G, Stokes M. Reliability of assessment tools in rehabilitation: an illustration of appropriate statistical analyses. Clinical rehabilitation. 1998; 12: 187-199.

20. Terwee C, et al. On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. Quality of life research. 2003; 12: 349-362.

21. Bohannon RW, Schaubert K. Long-term reliability of the timed up-and-go test among community-dwelling elders. Journal of Physical Therapy Science. 2005; 17: 93-96.

22. Morris S, Morris ME, Iansek R. Reliability of measurements obtained with the Timed "Up & Go" test in people with Parkinson disease. Physical therapy. 2001; 81: 810-818.

23. McGrath D. et al. Reliability of quantitative TUG measures of mobility for use in falls risk assessment. in 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2011. IEEE.

24. Botolfsen P, et al. Reliability and concurrent validity of the Expanded Timed Up-and-Go test in older people with impaired mobility. Physiotherapy Research International. 2008; 13: 94-106.

25. Ostir GV, et al. Reliability and sensitivity to change assessed for a summary measure of lower body function: results from the Women's Health and Aging Study. Journal of clinical epidemiology. 2002; 55: 916-921.

26. Sadaria K. Bohannon R. The 6-minute walk test: A brief review of literature. Clinical exercise physiology. 2001; 3: 127-132.

27. Rikli RE, Jones CJ. Development and validation of a functional fitness test for community-residing older adults. Journal of aging and physical activity. 1999; 7: 129-161.

28. Peters DM, Fritz SL, Krotish DE. Assessing the reliability and validity of a shorter walk test compared with the 10-Meter Walk Test for measurements of gait speed in healthy, older adults. Journal of geriatric physical therapy. 2013; 36: 24-30.

29. Reuben DB, Siu AL. An objective measure of physical function of elderly outpatients: the Physical Performance Test. Journal of the American Geriatrics Society. 1990; 38: 1105-1112.

30. King MB, et al. Reliability and responsiveness of two physical performance measures examined in the context of a functional training intervention. Physical therapy. 2000; 80: 8-16.

31. Bohannon RW. Test-retest reliability of the five-repetition sit-to-stand test: a systematic review of the literature involving adults. The Journal of Strength & Conditioning Research. 2011; 25: 3205-3207.

32. Goldberg A. et al. The five-times-sit-to-stand test: validity, reliability and detectable change in older females. Aging clinical and experimental research. 2012; 24: 339-344.

33. Northgraves MJ. et al. The test-retest reliability of four functional mobility tests in apparently healthy adults. Isokinetics and Exercise Science. 2016; 24: 171-179.

34. Wallmann HW. et al. Interrater reliability of the five-times-sit-to-stand test. Home Health Care Management & Practice. 2013; 25: 13-17.

35. Deathe AB, Miller WC. The L test of functional mobility: measurement properties of a modified version of the timed "up & go" test designed for people with lower-limb amputations. Physical therapy. 2005; 85: 626-635.

36. Nguyen VC. et al. Measurement properties of the L test for gait in hospitalized elderly. American journal of physical medicine & rehabilitation. 2007; 86: 463-468.

37. de Morton NA, Davidson M, Keating JL. Reliability of the de Morton mobility index (DEMMI) in an older acute medical population. Physiotherapy Research International. 2011; 16: 159-169.

38. Jarnlo GB, Nordell E. Reliability of the modified figure of eight--a balance performance test for elderly women. Physiotherapy Theory and Practice. 2003; 19: 35-43.

39. Hess RJ. et al. Walking skill can be assessed in older adults: validity of the Figure-of-8 Walk Test. Physical therapy. 2010; 90: 89-99.

40. Macknight C, Rockwood K. A hierarchical assessment of balance and mobility. Age and Ageing. 1995; 24: 126-130.

41. Rockwood K, et al. Reliability of the hierarchical assessment of balance and mobility in frail older adults. Journal of the American Geriatrics Society. 2008; 56: 1213-1217.

42. Yamada M, Ichihashi N. Predicting the probability of falls in community-dwelling elderly individuals using the trail-walking test. Environmental health and preventive medicine. 2010; 15: 386.

43. Lark SD, PW. McCarthy, Rowe DA. Reliability of the parallel walk test for the elderly. Archives of physical medicine and rehabilitation. 2011; 92: 812-817.

44. Liebl ME. et al. Introduction of the Charité Mobility Index (CHARMI)–A novel clinical mobility assessment for acute care rehabilitation. PloS one. 2016; 11: e0169010.

45. Ng SS, et al. Reliability and concurrent validity of standardized walking obstacle course test in people with stroke. Journal of rehabilitation medicine. 2017; 49: 705-714.

46. Taylor M. Standardized Walking Obstacle Course: Reliability and Validity of a Functional Measurement Tool. Journal of Neurologic Physical Therapy. 1997; 21: 167.

47. Tiedemann A, et al. The comparative ability of eight functional mobility tests for predicting falls in community-dwelling older people. Age and ageing. 2008; 37: 430-435.

48. Faber MJ, Bosscher RJ, van Wieringen PC. Clinimetric properties of the performance-oriented mobility assessment. Physical therapy. 2006; 86: 944-954.

49. Downs, S., J. Marquez, and P. Chiarelli, The Berg Balance Scale has high intra-and inter-rater reliability but absolute reliability varies across the scale: A systematic review. Journal of physiotherapy. 2013; 59: 93-99.

50. Shumway-Cook A, et al. Predicting the probability for falls in community-dwelling older adults. Physical therapy. 1997; 77: 812-819.

51. Boulgarides LK, et al. Use of clinical and impairment-based tests to predict falls by community-dwelling older adults. Physical therapy. 2003; 83: 328-339.

52. Huang MH, et al. Reliability, validity, and minimal detectable change of Balance Evaluation Systems Test and its short versions in older cancer survivors: a pilot study. Journal of Geriatric Physical Therapy. 2016; 39: 58-63.

53. Löfgren N, et al. The Mini-BESTest-a clinically reproducible tool for balance evaluations in mild to moderate Parkinson's disease? BMC neurology. 2014; 14: 235.

54. Chinsongkram B, et al. Reliability and validity of the Balance Evaluation Systems Test (BESTest) in people with subacute stroke. Physical therapy. 2014; 94: 1632-1643.

55. Leddy AL, Crowner BE, GM Earhart. Utility of the Mini-BESTest, BESTest, and BESTest sections for balance assessments in individuals with Parkinson disease. Journal of neurologic physical therapy: JNPT. 2011; 35: 90.

56. Rodrigues LC, et al. Reliability of the Balance Evaluation Systems Test (BESTest) and BESTest sections for adults with hemiparesis. Brazilian journal of physical therapy. 2014; 18: 276-281.

57. Wrisley DM, et al. Reliability, internal consistency, and validity of data obtained with the functional gait assessment. Physical

therapy. 2004; 84: 906-918.

58. Hill KD, et al. A new test of dynamic standing balance for stroke patients: reliability, validity and comparison with healthy elderly. Physiotherapy Canada. 1996; 48: 257-262.

59. Tiedemann A, Lord SR, Sherrington C. The development and validation of a brief performance-based fall risk assessment tool for use in primary care. Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences. 2010; 65: 896-903.

60. Butler AA, et al. Age and gender differences in seven tests of functional mobility. Journal of neuroengineering and rehabilitation. 2009; 6: 31.

61. Nolan JS, Remilton LE, Green MM. The reliability and validity of the Elderly Mobility Scale in the acute hospital setting. Internet Journal of Allied Health Sciences and Practice. 2008; 6: 5.

62. Smith R, Validation and reliability of the Elderly Mobility Scale. Physiotherapy. 1994; 80: 744-747.

63. Prosser L, Canby A. Further validation of the Elderly Mobility Scale for measurement of mobility of hospitalized elderly people. Clinical Rehabilitation. 1997; 11: 338-343.

64. Winograd CH, et al. Development of a physical performance and mobility examination. Journal of the American Geriatrics Society. 1994; 42: 743-749.

65. Means KM. The obstacle course: A tool for the assessment of functional balance and mobility in the elderly. Journal of Rehabilitation Research and Development. 1996; 33: 413-428.

66. Thigpen MT, et al. Turning difficulty characteristics of adults aged 65 years or older. Physical Therapy, 2000; 80: 1174-1187.

67. Fitzpatrick C, et al. The measurement properties and performance characteristics among older people of TURN180, a test of dynamic postural stability. Clinical rehabilitation. 2005; 19: 412-418.

68. Teresi JA, Annual Review of Gerontology and Geriatrics, Volume 14, 1994: Focus on Assessment Techniques. 1994: Springer Publishing Company.

69. Baker PS, Bodner EV, Allman RM. Measuring life-space mobility in community-dwelling older adults. Journal of the American Geriatrics Society. 2003; 51: 1610-1614.

70. Newell AM, et al. The modified gait efficacy scale: establishing the psychometric properties in older adults. Physical therapy. 2012; 92: 318-328.

71. de Morton NA, Davidson M, Keating JL. Validity, responsiveness and the minimal clinically important difference for the de Morton Mobility Index (DEMMI) in an older acute medical population. BMC geriatrics. 2010; 10: 72.

72. Spilg EG, et al. A comparison of mobility assessments in a geriatric day hospital. Clinical Rehabilitation. 2001; 15: 296-300.

73. Rubio FC, et al. Mobility assessment in elderly people. Description of measuring instruments for mobility: A review. Revista espanola de salud publica. 2015; 89: 545-561.

74. Metz DH, Mobility of older people and their quality of life. Transport policy. 2000; 7: 149-152.

75. Kimberlin CL, Winterstein AG. Validity and reliability of measurement instruments used in research. American Journal of Health-System Pharmacy. 2008; 65: 2276-2284.

76. Kottner, J, et al. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. International journal of nursing studies. 2011; 48: 661-671.

77. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychological bulletin. 1979; 86: 420.

78. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychological methods. 1996; 1: 30.

79. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of

chiropractic medicine. 2016; 15: 155-163.

80. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977: 159-174.

81. Dunn G, Design and analysis of reliability studies: The statistical evaluation of measurement errors. 1989: Edward Arnold Publishers.

25

82. Shrout PE. Measurement reliability and agreement in psychiatry. Statistical methods in medical research. 1998; 7: 301-317.

Manuscript Information: Received: June 22, 2022; Accepted: August 29, 2022; Published: August 31, 2022

Authors Information: Racha Soubra*; Aly Chkeir; Jean-Luc Novella
Laboratoire de Modélisation et Sûreté des Systèmes (M2S), Université de Technologie de Troyes, 12 rue Marie Curie, 10004 Troyes France.

Citation: Soubra R, Chkeir A, Novella JL. Reliability and responsiveness of thirty-one mobility assessment tests for older adults: A systematic review. Open J Clin Med Case Rep. 2022; 1899.