Fontys Paramedic University of Applied Sciences

Programme: Physiotherapy English Stream

Are Berg Balance Scale and 10 Meters Walking Test reliable and valid clinimetric tools to measure the ADL in the rehabilitation of elderly persons, and what is the responsiveness of these tests?

Name: Chiara Beltrami
Student number: 2167693
Phone number: +31 0647221948
E-mail: c.beltrami@student.fontys.nl

Supervisor:

Name: Annelies Simons
E-mail: a.simons-ad@fontys.nl
Phone: 0885078936

Version: 1.0

Date: 01/06/2014

Table of content

1. Abstract

Objective: The main goal of this thesis is to evaluate the clinimetric properties of Berg Balance Scale (BBS) and 10 Meters Walking Test (10MWT), when used with persons aged 60 years or older.
Method: Articles published from 1999 to 2014 were researched on PubMed, CINAHL, MEDLINE and SPORTDiscus. The studies were evaluated through the use of the COSMIN checklist to get the best evidence possible.
Results: Twenty eligible studies about reliability, validity and/or responsiveness of BBS and 10MWT have been found and compared to each other. The BBS showed good internal consistency (Cronbach alpha = 0.77) and excellent intrarater, interrater and test-retest reliability, (ICCs ranging from 0.87 to 0.998). Its MDC is 5-7 points, depending on the baseline scores. It has been validated through construct and criterion validity, which revealed that BBS predicts the risk of falls. Its responsiveness is limited, due to ceiling effects. The 10MWT showed excellent test-retest reliability (ICC = 0.95 for GCS and ICC = 0.97 for FGS). There is no clear agreement about the MDC of 10MWT. It is a valid instrument, it has high correlations with the Twenty-Meter Walk Test, the Barthel Index and the Instrumental Activity of Daily Living. The responsiveness of 10MWT is not known.
Conclusion: BBS and 10MWT are valid and reliable measurement for quality of life in elderly people, but their responsiveness is not clear.

Keywords: Berg Balance Scale, Ten Meters Walking Test, reliability, validity, responsiveness, older age.

2. Introduction

2.1. Background

It is known that the population in Europe is getting older. According to the World Health Organization (WHO), the population of the EU is estimated to reach 517 million in 2060 and nearly one third of it will be aged 65 or more.[1] In 2013, 14% of the population above 60 years was aged 80 or more. The oldest range of the aged population is showing the fastest growing trend.[2] There is not a clear definition about the age when somebody can start to be called "old", but the general agreement in Europe is that the threshold is around 60 years of age.[1]

The elderly population is heterogeneous, older people belonging to the same age range can indeed show a wide variety of health conditions. In general, in the old age, the assessment of the daily functioning plays an important role in the mainteinance of physical health.

This review is meant to focus on the measurement of some of the components that influence the activities of daily life (ADL) of individuals aged 60 years and over; the components chosen are balance and gait speed.

Balance is defined as the ability to keep the centre of gravity of the body under control in static postures and during movement. It is composed by three aspects: steadiness, symmetry and dynamic stability. The first aspect, "steadiness", is the ability to maintain a given posture for some time, with minimal or without uncontrolled movements. The second, "symmetry", deals with the distribution of load in the body, that should be equally distributed between the supporting limbs. The third, "dynamic stability" is the ability to move without uncontrolled sways or falls.[3] The aging process can lead to balance deficits and the reasons are various: diminished time spent in sport activities or general lack of movement; age-related changes in the vestibular, visual or somatosensory system; pathologies such as arthritis, neuropathies, cerebrovascular accidents.[3] Balance disorders, together with gait impairments, are the second major causes for falls in the elderly, after accidents related to the enviroment.[4] Among older adults, falls can have serious consequences, or even be fatal.[5] It is therefore important to assess the balance performance of elderly patients, in order to plan an intervention, if required.[3] The American Geriatrics Society/British Geriatrics Society clinical practice guideline for prevention of falls in older persons advises one balance/gait assessment every year for older persons who fell once in the past year and multiple fall risk assessment in those who fell more than twice.[6]

The Berg Balance Scale (BBS) is one of the most common performance tests used to assess the balance of individuals in physiotherapy, especially among elderly patients.[7]The test was originally described in 1989 by Berg et al.[8] It has been widely used to assess balance and the effect of exercise programs in various individuals, especially elderly individuals and patients with brain injuries or Parkinson's disease.[9] BBS has been frequently used as a criterion standard to assess balance and validate balance measures.[10,11]

It takes less than 15 minutes to be performed and less than 5 minutes to be scored. The equipment required are: two chairs, a stopwatch, ruler, a stool, shoes or slippers.

The items listed in the BBS are: 1.Sitting to standing, 2.Standing unsupported, 3.Sitting unsupported, 4. Standing to sitting, 5.Transfers, 6.Standing with eyes closed, 7.Standing with feet together, 8.Reaching forward with outstretched arm, 9.Retrieving object from floor, 10.Turning to look behind, 11.Turning 360 degrees, 12.Placing alternate foot on stool, 13.Standing with one foot in front, 14.Standing on one foot.

Each item can be scored from 0, if the performance was not sufficient, to 4, if the exercise did not present any problem. The minimum score is therefore 0, that means very poor balance, while the maximum score is 56, which means optimal balance for the ADL of an average elderly person. The whole description of the test can be found in Appendix 1.

Impaired balance can influence various activities, for example the ability to walk.

Walking is a crucial activity for outdoor mobility in the whole world. In Europe and USA it is the second most common mean to move from one place to another, after private cars.[12]

Gait ability involves different physiological systems in the human body the neurological system, the muscular system, the cardiovascular system, the respiratory system, the skeleton and the joints, for example. Because so many elements are required to work together for a coordinated gait, the walking speed is a good indicator of the health of those physiological systems. [13]

In general, walking speed starts decreasing around the age of 60 and it declines 1-2% per year.[14] This is due to change in coordination, strength, increased onset time for various muscle groups to start moving, joint stiffness, longer reaction time caused by cognitive impairments.[15] According to Purser et al.[16]every 0.10 m/s reduction of walking speed in frail elderly male veterans is associated with poorer health, poorer physical functioning, longer hospitalisation and higher costs, while improvement of 0.10 m/s has the opposite effects.

Physiotherapists and other clinicians use walking speed tests to assess the physical conditions of patients within a great range of pathologies, for example Parkinson's disease,[17] stroke or dementia.[18] The gait speed is also a good predictor measure for health conditions that can appear unrelated to gait, for example cognitive impairments, hospitalisation and institutionalisation.[19]

Perturbations in gait velocity are also related with fear of falling. However, it is not clear if the deceleration in walking speed is a direct consequence of the aging process, which leads to fear of falling, or if the fear of falling is the cause of such deceleration.[15] Brouwer et al.[20] reported that slow walkers are more often prone to fear of falling than fast walkers. Decreased walking velocity has indeed been associated with higher risk of falls and lower score on balance tests.[15]

For these reasons, it is important to have adequate tools to assess and record improvements of the gait velocity. Gait speed tests are safe, quick and inexpensive measurements, they can be administered by clinicians and non-clinicians and the results are easy to interpret.[13]They have shown to be highly reliable and they can be good surrogate of other more comprehensive and more time-consuming tests to assesses the quality of life.[16]

The 10 Meter Walking Test (10MWT) is a common test used by physiotherapists in Europe to assess the walking speed. It has been used by different sort of subjects, for example patients with traumatic brain injuries,[21]hip fracture,[22] multiple sclerosis,[23]spinal cord injury,[24] stroke.[25] The reason of the choice of ten meters is not completely clear, but this is probably the minimum distance that is functionally relevant in the ADL.[26] It should be performed in two phases: one at habitual walking speed and the other at maximum walking speed. Several trials are performed in both phases. [27]
There are different versions of the 10MWT. The distance of 10 m can be the total distance covered or the distance measured. In the first version, the patient walks on a straight path of 10 m, the measurement of walking speed begins after 2 m and ends 2 m before the end of the path. In the second version, the walkway is longer, in general 14 m, but the total distance may vary, and only the walking speed in the middle 10 m is measured. The speed is calculated per m/s.
Rydwik et al. reported lack of studies evaluating the reliability of walking speed in elderly people, especially regarding the maximum walking speed. This can be due to the fact that maximum walking speed is likely to diminish after the first try, especially by the weakest individuals.[12] Habitual and maximum walking speeds are reported as valid instruments.[12]

Gait speed slower than 0.17 m/s has shown to be non-functional for the ADL.[16]

As already mentioned, BBS and 10MWT are commonly used performance measurements.
In practice, it is important to use high quality measurements to evaluate the patients' performances.
The quality of an instrument is determined by its clinimetric properties, which are: reliability, validity and responsiveness.
Figure 2 shows how the clinimetric properties are located within the set of quality of a clinimetric measurement instrument.

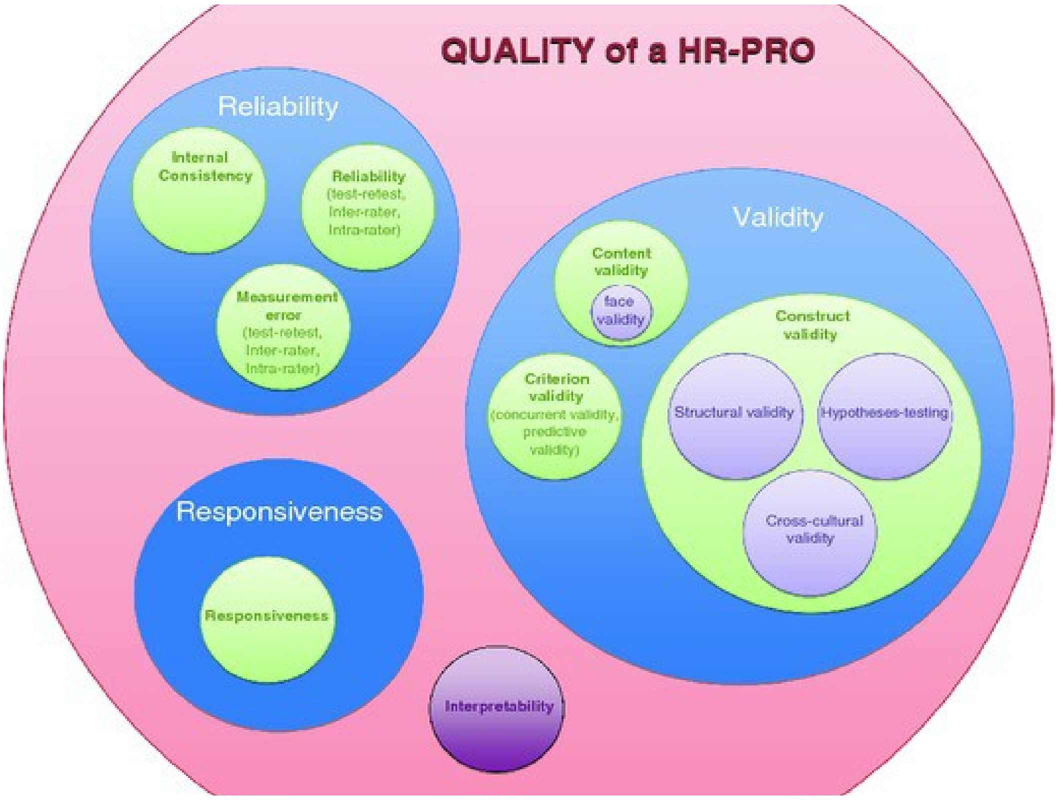

Figure 2. Clinimetric characteristics of a clinimetric measurement instrument. Source: Mokkink et al.[30] p. 542.

"Reliable" means that repeated measurements of a performance are consistent. The components of reliability are: internal consistency, interrater reliability, intrarater reliability, test-retest reliability and measurement error. Internal consistency was defined by the COSMIN panel as the degree of intercorrelations among the items of a measurement. In a unidimensional scale, it is a measure of the extent to which all the items concern the same construct.[28]Internal consistency is measured through Cronbach alpha. According to de Vet et al, the value of Cronbach Alpha is to be considered acceptable between 0.70 and 0.90. If the value is smaller, there is no correlation between the items and if the value is higher there is a redundancy.[29]

The interrater reliability is calculated when more than one person assesses the same performance, independently. The intrarater reliability is calculated when the same person performs all the assessments.[31] Test-retest reliability is evaluated by repeating the test in different occasions, with some time between one assessment and another. The time between the assessments should be long enough to avoid learning effects, but short enough to make sure that the patient is stable. It is inevitable that test-retest reliability includes intrarater error; sometimes the terms intrarater and test-retest reliability are used interchangeably.[32]

Reliability contains also the evaluation of the measurement error. This is calculated to find out which is the minimum difference in score that a patient should reach, to be sure that a true improvement was achieved, free from possible measurement errors.[28] If the studies that describe the reliability of an instrument have good methodological quality and consistent results, there is strong evidence that the instrument is reliable. The evidence is weaker if the studies concerning the test are inconsistent or have a fair methodological quality.[30]

Validity is the degree to which an instrument measures the construct(s) it purports to measure. In this case, the constructs are “balance” and “walking speed”. Content validity is the degree to which the content of an instrument is adequate to measure the construct. It includes face validity, that is the degree to which a measurement instrument looks adequate to the construct to be measured. Construct validity is the degree to which the scores of an instrument are consistent with the hypotheses. It includes structural validity, hypothesis testing and cross-cultural validity. Structural validity is the degree to which the items of a measurements are an adequate reflection to the construct to be measured. Hypothesis testing is based on the formulation and test of hypothesis; in hypothesis testing the measurement is often compared to other measurements. Cross-cultural validity is the validity of the instrument in different countries. When assessing cross-cultural validity, it is taken into account not only the process of translation of the instrument, but also if the test is appropriate to be used in a certain cultural environment. Criterion validity can be subdivided into concurrent and predictive validity. Concurrent validity is assessed through a comparison with a gold standard, while predictive validity is the ability of the instrument to predict the gold standard.[33]

Responsiveness means the ability of an instrument, in this case BBS and 10MWT, to detect change over time in the construct to be measured, in this case balance and walking speed. Reliability and responsiveness are dependent on the characteristics of the population tested.[28]

Interpretability is not a clinimetric property, because it does not refer to the quality of an instrument, but to the meaning of these results. It concerns the distribution of scores over the scale in different subgroups, the presence of floor and/or ceiling effects, the minimal important change (MIC) and minimal important difference (MID).[34]

2.2. Problem Definition

As described in the previous paragraph, it is known that BBS and 10MWT are in general valid and reliable instruments. These data are sparse and concerns different population groups. The aim of this review is to examine the clinimetric properties of BBS and 10MWT when they are used specifically in the assessment of elderly persons.

The questions to be answered are: when applied to people aged 60 years and over, is BBS a reliable tool? Is it valid? What is its responsiveness?

Is the 10MWT reliable, when performed by people aged 60 years or more? Is it valid? What is its responsiveness?

2.3. Research question

What are the clinimetric properties of BBS and 10MWT, in order to be reliable and valid tools to measure the ADL in the rehabilitation of elderly persons? What is the responsiveness of these tests?

2.4. Working Definitions

Elderly: People aged 60 or more. The patients target of this thesis are groups of healthy or non healthy persons aged at least 60 years.

Berg Balance Scale (BBS): 14-items test aimed at the assessment of balance. The whole description is to be found in appendix 1.

Ten Meters Walking Test (10MWT): Test aimed at the measurement of walking speed on a ten-metres straight path.

Clinimetric properties: Elements that contribute to the quality of the measurement instrument in terms of reliability, validity, responsiveness.

3. Method

3.1. Search Strategy

Researches in databases such as PubMed, CINAHL, MEDLINE and SPORTDiscus were performed to find the most relevant articles. The strings used to find the articles were :

- (Berg Balance Scale OR BBS) AND (reliability OR validity OR responsiveness)
- (Ten Meter Walk Test OR 10MWT) AND (reliability OR validity OR responsiveness)

The publication year and the age of the study sample were selected. If it was not possible to select the age, the thesaurus term “older people” was used. If it was not possible to use a thesaurus term, the strings were changed into:

- (((Berg Balance Scale) OR (BBS) AND (reliability OR validity OR responsiveness)) AND aged [MeSH Terms]
- ((Ten Meter Walk Test) OR 10MWT) AND (reliability OR validity OR responsiveness) AND aged [MeSH Terms]

3.2. Selection Criteria

Inclusion criteria:

- Studies about the 10 Meters Walking Test
- Studies about Berg Balance Scale
- Studies about the clinimetric properties of these tests

- Articles published in English, Italian or Dutch
- Articles published between 1999 and 2014

Exclusion criteria:

- Studies those full text is not available online
- Studies which participants are aged less than 60 years
- Studies that have other measurements as an outcome
- Studies that do not concern the clinimetric properties of Berg Balance Scale and/or 10 Meter Walk Test

3.3. Assessment of Literature

Once the articles have passed the inclusion and exclusion criteria, the studies were assessed using the COSMIN checklist.[30] This method allows the reader to estimate the quality of results in comparison with other studies.
The full COSMIN checklist consists of 12 boxes, containing 4 to 18 items per box (119 items total). Ten boxes (boxes A to J) are meant for the assessment of the methodological quality of a study. Boxes A to G contain checklists to evaluate a measurement properties, box J contains standards for the interpretability of the instrument. The IRT box is meant for studies that made use of IRT methods. In addition, the Generalisability box contain requirements for the generalisability of the results. The quality of a study is considered adequate if all items in a box are adequate. All the questions and the extraction forms are available in Appendix 2.

3.4 Method of Extraction

The COSMIN checklist was completed for every study, following a four-step procedure.
Step 1: it was determined which measurement properties were evaluated in the article and consequently, which COSMIN boxes needed to be completed.
Step 2: if Item Response Theory (IRT) methods were used in a study, these methods were assessed , according to the requirements in the IRT box.
Step 3: the boxes identified in Step 1 were completed.
Step 4: the characteristic of the study population were extracted to determine the generalizability of the study findings. This was done for each measurement property identified in Step 1.
Figure 4 shows an overview of the procedure described.

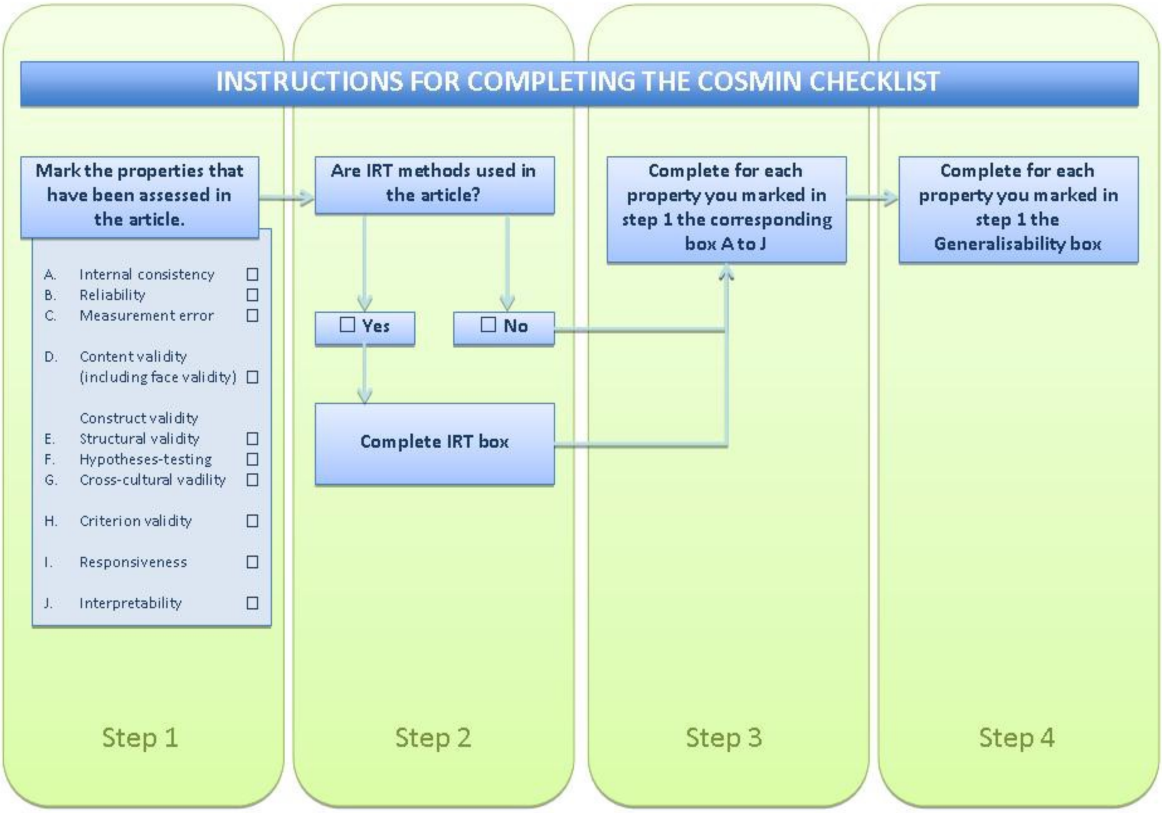
INSTRUCTIONS FOR COMPLETING THE COSMIN CHECKLIST
Mark the properties that have been assessed in the article.
A. Internal consistency □
B. Reliability □
C. Measurement error □
D. Content validity (including face validity) □
Construct validity
E. Structural validity □
F. Hypotheses-testing □
G. Cross-cultural vadility □
H. Criterion validity □
I. Responsiveness □
J. Interpretability □
Are IRT methods used in the article?
□ Yes
□ No
Complete IRT box
Complete for each property you marked in step 1 the corresponding box A to J
Complete for each property you marked in step 1 the Generalisability box
Step 1
Step 2
Step 3
Step 4

Figure 4. Four-step procedure for completing the COSMIN checklist. Source: Mokkink et al.[30]p. 544

3.5 Best evidence synthesis

It was developed by the COSMIN group a 4-point rating scale to classify each assessment as "excellent", "good", "fair" or "poor", based on the single scores in the boxes previously described.
The first step of the rating procedure was to answer every question in the box required. Then, each single item was evaluated as "excellent", "good", "fair" or "poor", according to the standards of evaluation of the COSMIN manual. An item was scored as "excellent" if there was evidence that the methodological quality of the study to which the item was referred was adequate. An item was scored as "good" when relevant information was not reported, but it could be assumed that the quality aspect was adequate. An item was scored as "fair" if it was not clear whether the methodological aspect was adequate. An item was scored as "poor" when evidence was provided that the methodological quality aspect was not adequate.
An overall score for the assessment of a given measurement property was obtained by taking the lowest score that was achieved in the box.
Each clinimetric property tested in a study was rated. Depending on the number of properties assessed in a study, some studies received one quality evaluation, while others received several. All the standards for the evaluations were not reported because of practical reasons, but the tables can be found in the COSMIN manual.[35]

After the grading procedure, the results were combined, taking into account the number of studies on a certain measurement property, their methodological quality and the consistency of their results. Table 1 presents a list of the levels of evidence possible.

Table 1. Levels of evidence for the overall quality of the measurement property. Source: van Tulder et al.[36]*p.1296.*

Level	Rating	Criteria
Strong	+++ or ---	Consistent findings in multiple studies of good methodological quality OR in one study of excellent methodological quality
Moderate	++ or --	Consistent findings in multiple studies of fair methodological quality OR in one study of good methodological quality
Limited	+ or -	One study of fair methodological quality
Conflicting	+/-	Conflicting findings
Unknown	?	Only studies of poor methodological quality

The possible overall rating for a measurement property was “positive”, “indeterminate”, or “negative”. To assess the results of the measurement properties, criteria based on Terwee et al.[37] were used, as shown in Table 2.

Table 2. Quality criteria for measurement properties..Source: Terwee et al.[37] *p.39*

Property	**Rating**	**Quality Criteria**
Reliability		
Internal consistency	+ ? - 0	Factor analysis performed on adequate sample size AND Cronbach's alpha(s) calculated per dimension AND Cronbach alpha(s) between 0.70 and 0.95; No factor analysis OR doubtful design or method; Cronbach's alpha(s) < 0.70 or > 0.95, despite adequate design and methods; No information found on internal consistency.
Intrarater, interrater, test-retest reliability	+ ? - 0	ICC/weighted Kappa ≥ 0.70 OR Pearson's *r* ≥ 0.80; Doubtful design or method (e.g., time interval not mentioned); ICC/weighted Kappa < 0.70 OR Pearson's *r* < 0.80, despite adequate design and methods; No information found on intrarater, interrater or test-retest reliability.
Measurement error	+ ? - 0	MIC>MDC OR MIC outside LoA OR convicing arguments that agreement is acceptable; Doubtful design or method OR MIC not defined AND no convicing arguments that agreement is acceptable; MIC≤MDC OR MIC equals or inside LoA, despite adequate design and methods; No informationfound on measurement error.
Validity		
Content validity	+ ? - 0	The target population considers all items to be relevant AND the measurement to be complete; No target population involvement; The target population considers items to be irrelevant OR the measurement to be incomplete; No informationfound on content validity.
Criterion validity	+ ? - 0	Convincing arguments that gold standard is "gold" AND correlation with gold standard ≥ 0.70; No convincing arguments that gold standard is "gold" OR doubtful design or method; Less than 75% of hypothesis were confirmed; No information found on criterion validity.
Construct validity	+ ? - 0	Specific hypothesis were formulated AND at least 75% of the results are in accordance with these hypotheses; Doubtful design or method (e.g., no hypothesis); Less than 75% of hypothesis were confirmed, despite adequate design and methods; No information found on construct validity.
Responsiveness		
Responsiveness	+ ? - 0	MDC or MDC<MIC or MIC outside the LoA OR AUC≥ 0.70 Doubtful design or method; MDC or MDC≥ MIC or MIC equals or inside LoA OR AUC<0.70,despite adequate design and methods; No information found on responsiveness.

MIC = minimal important change; MDC = minimum detectable change; LoA = limits of agreement; ICC = intraclass correlation.
+ = positive rating; ? = indeterminate rating; - = negative rating; 0 = no information available.
Doubtful design or method = lacking of a clear description of the design or methods of the study, sample size smaller than 50 subjects, or any methodological weakness in the design or execution of the study.

4. Results

4.1 Database research

The research was performed in four databases and resulted in 20 suitable studies.
The procedure is presented in figure 5, while the whole research strings and the list of articles used to collect references can be found in Appendix 3.

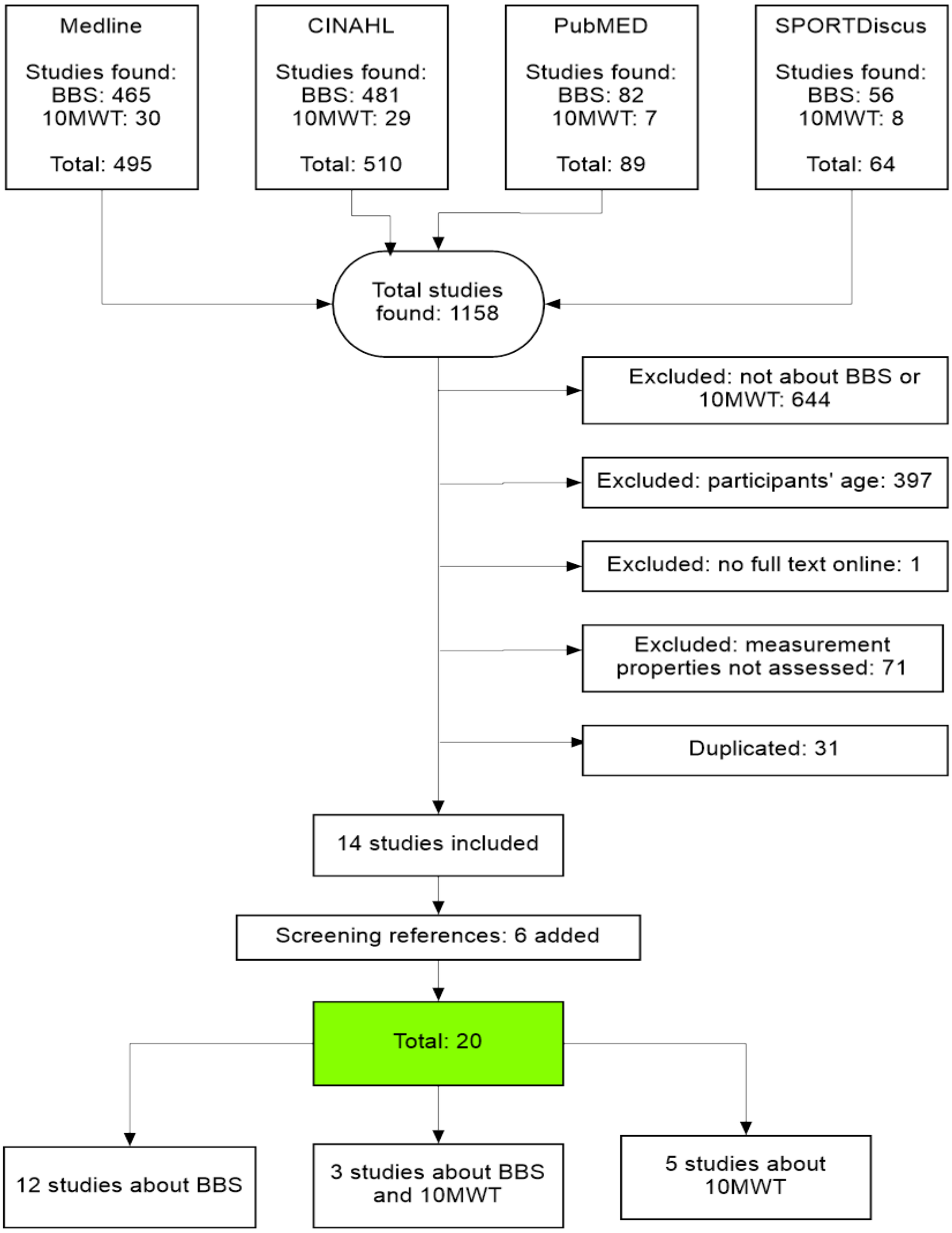

Figure 5. Flowchart: procedure used in the database research.

4.2 Description of the studies included

The research included studies performed on a wide range of participants types.
The properties of BBS and 10MWT were evaluated in heterogeneous groups of elderly people aged 60 – 90 years. The groups included independent people living in the community, as well as individuals in need for assistance. Some studies considered only subjects within a certain health status: healthy persons, stroke survivors, patients with Parkinson's disease or parkinsonism. In some studies, the participants were recruited consecutively or random in a certain community, for example a geriatric department in a hospital, and they had a mix of different health conditions.
The studies took place in USA, Norway, Canada, Brazil, Taiwan, Sweden, Ireland, Japan.
Table 3 shows the list of the articles included, with a description of the participants' number, age and characteristics, the tests measured and the clinimetric properties described.

Table 3. Studies included, description of participants' main characteristics and measures described.

Author (Year)	**Age ± SD (range)**	**Number of participants/ Characteristics**	**Measures (*)**	**Properties tested**
Boulgarides (2003) [38]	65-90	99/Community-dwelling	BBS, DGI, TUG, mCTSIB, 100%LOS	Predictive validity
Brusse et al. (2005) [39]	76 (61-86)	25/Parkinson	BBS, 10MWT, FFR, BFR,TUG, UPDRS	Interrater reliability Hypothesis testing
Chiu et al. (2003) [40]	83 ± 8 82 ± 8	29/Public falls clinic 39/Community-dwelling	BBS, TMS, TUG, EMS	Predictive validity
Conradsson et al. (2007) [31]	82.3 ± 6.6 (68 – 96)	45/Assisted in care facilities	BBS	Intrarater reliability Measurement error
Donoghue et al. (2009) [41]	65+	118/Physiotherapy rehabilitation	BBS	Measurement error
De Figueiredo et al. (2009) [42]	63-87	37/Non institutionalized	BBS (Portuguese version)	Interrater reliability
Halsaa et al. (2007) [43]	82 (69-95)	83/Geriatric rehabilitation	BBS (norvegian version)	Interrater reliability Internal consistency
Holbein-Jenny et al. (2005) [44]	74-92	26/Residents of personal care homes	BBS, MDRT, ABCS	Interrater reliability Test-retest reliability
Lajoie (2003) [45]	75.50 73.80	125/Nursing homes and senior residences	BBS, ABCS	Predictive validity
Leeraar et al. (2002) [46]	60-92	64/Healthy	10MWT, 20MWT, 3mWT, 6mWT, 12mWT	Hypothesis testing
Maeda et al. (2000) [47]	69.6 ± 8.3 72.1 ± 7.4	40/Stroke 40/Healthy	10MWT,Barthel Index, IADL, Sit and Reach	Test-retest reliability Concurrent validity
Miyamoto et al. (2004) [48]	72 (65 – 83)	36/Geriatric rehabilitation	BBS (brazilian version)	Intrarater Reliability Interrater Reliability Cross-cultural validity
Montero-Odasso (2005)[49]	78.9 ± 3	102/Community-dwelling	10MWT, TUG, POMA	Predictive validity
Pardasaney et al. (2012) [50]	75.9 ± 7.0	111/Functional limitations	BBS, POMA-T, POMA-B, DGI	Responsiveness
Perera et al. (2006) [51]	77.6 ± 7.6	100/Mobility disabilities, 100/ Subacute stroke, 492/Community-dwelling	10MWT, SPPB, 6mWT, self-reported mobility	Responsiveness Measurement error
Peters et al. (2013) [52]	84.3 ± 6.9	43/Healthy	10MWT, 4MWT	Test-retest reliability Measurement error Hypothesis testing
Steffen et al. (2002) [53]	73.0 ± 8.0 (60 – 90+)	96/Without major disabilities	BBS,10MWT, 6mWT,TUG	Test-retest reliability
Steffen et al. (2008) [54]	71 ± 12	37/Parkinsonism	BBS, Gait Speed, Fr, RT,SRT, ABCS, 6mWT, TUG, SF-36, UPDRS	Internal consistency Test-retest reliability Measurement Error
Stevenson (2001) [55]	73.5 ± 7.0	48/Stroke	BBS	Measurement Error Responsiveness
Wang et al(2006) [3]	73.8 ± 5.2	286/Community- dwelling	BBS, TUG,Usual gait speed	Internal consistency Interrater reliability Hypothesis testing

(*) Meaning of the abbreviations, in alphabetical order: 100% LOS = 100% Limits of Stability Test; 10MWT = Ten-Meter Walking Test; 12mWT = 12 minutes Walk Test; 20MWT = Twenty-Meter Walking Test; 3mWT = 3 minutes Walk Test; 4MWT = Four-Meter Walking Test; 6mWT = 6 minutes Walk Test; ABCS = Activities-specific Balance Confidence Scale; BBS = Berg Balance Scale; BFR = Backward Functional Reach test; DGI = Dynamic Gait Index; EMS = Elderly Mobility Scale; Fr = Functional reach; FFR = Forward Functional Reach test; IADL = Instrumental Activity Daily Living; mCTSIB = modified Clinical Test for Sensory Interaction on Balance; MDRT: Multi-Directional Reach Test; POMA = Performed Oriented Mobility Assessment; POMA-B = Performance-Oriented Mobility Assessment balance sub-scale; POMA-T = Performance-Oriented Mobility Assessment total scale; RT= Romberg Test; SF-36 = Medical Outcome Study Short-Form Health Survey; SPPB = Short Physical Performance Battery; SRT = Sharpened Romberg Test; TMS = Tinetti Mobility Score; TUG = Timed Up and Go Test; UPDRS = Unified Parkinson Disease Rating Scale

4.3 Assessment of the studies included

The studies included were assessed with the use of the COSMIN checklist and a grade was given for every box relevant for that particular study.

The rating was given according to the COSMIN standards. The boxes completed were:

- Box A. Internal consistency – 3 times
- Box B. Reliability – 12 times
- Box C. Measurement error – 6 times
- Box F. Hypotheses testing – 6 times
- Box G. Cross-cultural validity – 1 time
- Box H. Criterion validity – 5 times
- Box I. Responsiveness – 3 times
- Box J. Interpretability – 2 times
- Generalizability box – 20 times

Content validity (Box D) and structural validity (Box E) were not found to be relevant in any of the studies included. The whole grading procedure is described in details in Appendix 4. The interpretability and generalizability boxes have not been graded, because there are no evaluation standards for them.

The results about the BBS are going to be presented first, the results about the 10MWT will follow. The following tables show the final grades given to the studies about BBS (Table 4) and the results found (Table 5). Some properties are not listed in the tables because these were not assessed in any of the studies.

Table 4. Overview of the evidence available for the properties of BBS.

Test: BBS (*)									
Author	**Reliability**					**Validity**			**Responsiveness**
	Internal Consistency	Intrarater	Interrater	Test-retest	Measurement Error	Hypothesis testing	Cross-cultural	Predictive	Responsiveness
Boulgarides et al.[38]								Poor	
Brusse et al.[39]						Poor			
Chiu et al.[40]								Good	
Conradsson et al.[31]		Fair			Fair				
Donoghue et al.[41]					Excellent				
De Figueiredo et al.[42]			Poor						
Halsaa et al.[43]	Good		Good						
Holbein-Jenny et al.[44]			Poor	Poor		Poor			
Lajoie et al.[45]								Poor	
Miyamoto et al.[48]		Fair	Fair				Fair		
Pardasaney et al.[50]									Good
Steffen et al. (2008)[54]	Poor			Fair	Fair				
Stevenson[55]					Fair				Fair
Wang et al.[3]	Excellent		Good			Fair			

(*) When a cell is empty, it means that results about that specific property were not available in the study.

Table 4. Clinimetric properties of BBS: list of results.

Test: BBS								
Reliability					**Validity**			**Responsiveness**
Internal Consistency	Intrarater	Interrater	Test-retest	Measurement Error	Hypothesis testing	Cross-cultural	Predictive	Responsiveness
α = 0.77;[3] α = 0.86-0.87;[54] α = 0.87[43]	ICC = 0.97;[31] ICC = 0.99[48]	ICC = 0.87;[3] ICC = 0.88;[44] ICC = 0.98;[48] ICC = 0.996;[42] ICC = 0.998[43]	ICC = 0.77;[44] ICC= 0.94[54]	MDC[a] : 4-7 points;[41] MDC: 5 points; [54] MDC: 8 points; [31] MDC: 7 points [55]	Correlation with: UPDRS: r = 0.64;[39] CGS: r = 0.73[39] r= 0.46;[3] FGS: r = 0.64;[39] FFR: r = 0.50;[39] BFR: r = 0.51;[39] TUG: r = -0.78;[39] r = -0.53;[3] MDRT: r = 0.53 – 0.80;[44] ABC: r = 0.50[44]	Validated in the Brazilian-Portugues version.[48]	Predicts falls, cut-off: 46 points [45] Predicts falls, cut-off: 47 points [40] It does not predict falls [38]	Limited [50] MDC: 7 points [55]

[a] MDC = minimum detectable change; all the MDCs are reported at their 95% CI.

In the articles included, information about reliability, validity and responsiveness of the Berg Balance Scale was found. Internal consistency was evaluated in three studies of excellent, good and poor quality, respectively. Wang et al.[3]calculated the Cronbach α to be 0.77 in healthy people, and it was showed that if one of the item was deleted, the α would vary from a minimum of 0.7427 to a maximum of 0.7760. The statistics analysis performed by Steffen et al.[54] also resulted in a high internal consistency of BBS in subjects with parkinsonism, often associated with other diseases. Cronbach α was found to be 0.86 during the first measurement and 0.87 during the second, that took place 7 days later. Halsaa et al.[43] reported a Cronbach α coefficient of 0.87 in the Norwegian version of BBS, applied to geriatric patients.

Intrarater reliability was found to be excellent by two studies of fair quality, with ICC ranging from 0.97 to 0.99.[39,48]

Of the five studies evaluating interrater reliability of BBS, two of them were good quality studies. All of them demonstrated that BBS is a reliable measure, with a good to excellent interrater reliability; the ICCs found ranged from 0.87 to 0,998.[3,42,43,18]

Inter- and intrarater reliability was also found to be excellent in the Brazilian-Portuguese version of BBS,[48] but the study had fair quality due to the small number of participants (37 in total).

Test-retest reliability was assessed in three studies of poor or fair quality. The results of the study of poor quality showed that BBS has acceptable test-retest reliability with ICC = 0.77,[44] while the other two studies demonstrated high ICC values, ICC = 0.94 and ICC = 0.97.[54,31]

Four studies calculated the MDC. Of the four studies evaluated, one demonstrated excellent quality[41]and the other three studies were graded as "fair".[31,42,55] The lower grade of these studies was due to the number of participants.

1.The excellent quality study[41] showed that MDC (95% CI) for the BBS is between 4 and 7 points in elderly people who are patients in a physiotherapy practice, depending on the baseline scores of BBS. The MDCs reported were:

- 3.3 points for BBS baseline score ranging 45-65 points (35 participants).
- 4.9 points for BBS baseline score ranging 35-44 points (45 participants).
- 6.3 points for BBS baseline score ranging 25-34 points (27 participants).
- 4.6 points for BBS baseline score of 25 points or less (11 participants).

2.Fair evidence[31] demonstrated MDC of 7.7 points for elderly people dependant in activities of daily living, with a wide range of baseline scores and symptoms. Half of the patients had cognitive disabilities, which, according to the authors, may have influenced the test results.

3.A study of fair quality[42] demonstrated that patients with parkinsonism with baseline scores of 50 ± 7 points need to gain 5 points to show a genuine change.

4.Stevenson, in a study of fair quality[55] indicated a MDC of 6.9 points in stroke patients with BBS baseline mean score of 35.5 points (interquartile 25.5 – 43.0). This study three subgroups, divided in: living independently, in need for standby assistance, fully assisted. The MDCs resulting were:

- 6.3 points for people living independently (15 participants), with BBS baseline score of 47 points (interquartile 45.3-51.8).
- 6 points for people in need for standby assistance (17 participants), with BBS baseline score of 40 points (interquartile 38.0 -45.3)
- 8.1 points for people needing assistance (16 participants), with BBS baseline score of 35.5 points (interquartile 25.5 – 43.0).

Construct validity was evaluated in three studies. There is good evidence that BBS has a moderate correlation with TUG and usual gait speed, measured on a 50-foot distance.[3]
Poor evidence showed fair to good correlations with Forward and Backward Functional Reach test, good correlation with the Unified Parkinson Disease Rating Scale, the 10MWT, the Timed Up and Go test [39], the Multi-Directional Reach Test and fair to moderate correlation with the Activities-specific Balance Confidence Scale.[44]

Predictive validity was assessed in three studies. One study of good quality[40] reported that BBS is a predictor for falls, with a cut-off score of 47 points. Two studies of poor quality reported that BBS predicts falls, with a cut-off score of 46 points,[45] and that BBS does not predict falls.[38]

Stevenson,[55] after the evaluation of the MDC, calculated the responsiveness of BBS to be 7 point to show a true change over time, with 95% confidence. Pardasaney et al., in an article of good quality, demonstrated limited responsiveness in community-dwelling older people at higher functional levels,[50] but no specific change values (MDC or MIC) were given. The limited responsiveness was due to a ceiling effect: 60% of the participants started with a baseline score of 50 points or more. There is, indeed, good evidence that BBS shows a ceiling effect, especially in healthy, independent people, or groups with mild disabilities.[3,54,50] No floor effect was reported.
The most challenging items have been demonstrated to be number 13 (standing with one foot in front) and 14 (standing on one foot).[3]
The best evidence synthesis was performed from the collection of all the data, as reported in Table 5.

Table 5. Findings about the clinimetric properties of BBS: best evidence synthesis.

	Test: BBS								
	Reliability					Validity			Responsiveness
	Internal Consistency	Intrarater	Interrater	Test-retest	Measurement Error	Hypothesis testing	Cross-cultural	Predictive	Responsiveness
Level of evidence	+++	++	++	+	+/-	+	+	++	--

Grading:
+++ or --- = Strong
++ or -- = Moderate
+ or - = Limited
+/- = Conflicting
? = Unknown

Strong positive evidence was found in the area of internal consistency. Moderate positive evidence was found in the intrarater and interrater reliability. Limited positive evidence was found about test-retest reliability, construct and cross-cultural validity. Conflicting evidence were found about measurement error. Moderate negative evidence was found regarding responsiveness. These findings are going to be presented in the discussion section.

The same procedure of data collection, grading and synthesis was performed for the 10MWT.The following tables show the final grades given to the studies about 10MWT (Table 7) and the results (Table 8). Some properties are not listed in the tables because these were not assessed in any of the studies.

Table 7.Overview of the evidence available for the properties of 10MWT.

Test: 10MWT (*)							
Author	**Reliability**			**Validity**			**Responsiveness**
	Interrater	Test-retest	Measurement Error	Hypothesis testing	Concurrent	Predictive	Responsiveness
Brusse et al.[39]	Poor			Poor			
Leerar et al.[46]				Good			
Maeda et al.[47]		Fair			Good		
Montero-Odasso et al.[49]						Poor	
Perera et al.[51]			Poor				Poor
Peters et al.[52]		Fair	Fair	Fair			
Steffen et al(2002)[53]		Good					
Steffen et al. (2008)[54]		Fair	Fair				

(*) When a cell is empty, it means that results about that specific property were not available in the study.

Table 8. Clinimetric properties of the 10MWT: list of results.

Test: 10MWT						
Reliability			**Validity**			**Responsiveness**
Interrater	Test-retest	Measurement Error	Hypothesis testing	Concurrent	Predictive	Responsiveness
CGS[a]: ICC = 0.90;[39] FGS[b]: ICC = 0.94[39]	CGS[a]: ICC = 0.96-0.98;[52] ICC = 0.95;[53] ICC = 0.96;[54] FGS[b]: ICC = 0.97;[53] ICC = 0.97;[54] Spearman's r = 0.88 – 0.93[47]	CGS: MDC[c]= 0.1m/s;[52] MDC= 0.18 m/s;[54] FGS: MDC= 0.25 m/s;[54]	Correlation with BBS: CGS: r = 0.73;[39] FGS: r = 0.64;[39] UPDRS: CGS: r = -0.12;[39] FGS: r = -0.18;[39] Correlation with 20MWT: r = 0.922;[46] Low correlation with 4MWT (discrepancy of ± 0.15 to ± 0.17 m/s)[52]	Correlation with: Barthel Index: r = -0.78;[47] IADL: r = -0.76;[47] Quadriceps strength: r = -0.33 to -0.44;[47] Sit and Reach: r = - 0.06 to - 0.44[47]	Predicts adverse events, cut-off : CGS: 0.7m/s[49]	MDC= 0.06 m/s[51]

[a] CGS = comfortable gait speed; [b] FGS = fast gait speed[c] MDC = minimum detetable change (95% CI).

In the literature included, a study of poor quality found that the interrater reliability of 10MWT is excellent (CGS: ICC = 0.90; FGC: ICC = 0.94).

There is good evidence that CGS and FGS have high test-retest reliability (ICC= 0.95 – 0.97) in community-dwelling older people.[53] Fair evidence demonstrated high test-retest reliability also in healthy subjects and people with parkinsonism or stroke. In these samples, ICC was calculated to be between 0.96 and 0.98 for comfortable gait and 0.97 for fast gait.[65,69]

MDC was calculated only in articles of fair quality. MCD of comfortable walking speed was showed to be 0.1 m/s in healthy subjects[52] or 0.18 m/s in patients with parkinsonism.[54] MCD of fast walking speed had a value of 0.25 m/s in patients with parkinsonism. [54]

Construct validity was assessed through comparison of 10MWT with other tests.

There is fair evidence that 10MWT and 4MWT show a low degree of concurrent validity.[52]

There is poor evidence[39] that comfortable and fast gait speed do not correlate with UPDRS and that correlation of BBS with comfortable and fast gait speed is good.

Good evidence demonstrated very high concordance between 10MWT and 20MWT.[46]

Criterion validity of 10MWT was also evaluated. A study of good quality assessed the validity of 10MWT in 40 healthy subjects and 40 people with stroke.[47] In the article, it was demonstrated that there is a significant correlation between 10MWT and the number of centimetres reached during the Sit and Reach Test in healthy older adults, but not in individuals with stroke. It was also described a high correlation between 10MWT and Barthel Index score and between 10MWT and the Instrumental Activity of Daily Living. In both groups, it was demonstrated a moderate correlation between time needed to walk 10 meters and maximum isometric strength of quadriceps.

According to a study of poor quality,[49] 10MWT is predictor of new falls, hospitalisation, fractures and death, with a cut-off score of 0.7 m/s.

One study of good quality investigated the responsiveness of 10MWT.[51] It was calculated that MDC is 0.06 m/s for aged people with mobility disabilities. The minimum clinically important change was showed to be 0.05 m/s (small meaningful change) to 0.13 m/s (moderate meaningful change).

The protocol of the 10MWT showed some differences in the studies involved. In the articles described, gait speed was measured in four different ways and the test was called with different names. Both studies of Steffen et al.[53,54] measured comfortable and fast gait speed for 6 meters on a total distance of 10 meters; the authors called the tests “comfortable gait speed” and “fast gait speed”, respectively. Maeda et al.[47] performed the test on a distance of 11 meter, but only 10 meters were measured; the measurement was called “walking ability”. Both Brusse et al.[39] and Perera et al.[51] talked about “gait speed”. Brusse et al. used the same method as Steffen et al., while the participants of Perera's study walked for 10 meters, but the measurement procedure used was not described in details.

Only Peters et al.[52] and Leerar et al.[46] called the measurement “10 Meters Walking test”. Peters et al. used a 20 m path, with 5 m for acceleration and 5 m for deceleration, while Leerar et al. measured the speed on ten meters, without considering a space for acceleration and deceleration.
After the data collection and the grading, the best evidence synthesis was performed, based on all the findings about the clinimetric properties of the 10MWT

Table 9. Best evidence synthesis of the findings about the clinimetric properties of 10MWT.

	Test: 10MWT					
	Reliability		**Validity**			**Responsiveness**
	Test-retest	Measurement Error	Hypothesis testing	Concurrent	Predictive	Responsiveness
Level of evidence	**++**	**+/-**	**++**	**++**	**?**	**?**

Grading:
+++ or --- = Strong evidence
++ or -- = Moderate evidence
+ or - = Limited evidence
+/- = Conflicting findings
? = Unknown

It was found moderate evidence for test-retest reliability and construct validity, limited evidence for criterion validity and conflicting findings in the area or measurement error. Interrater reliability and responsiveness are unknown, because only poor evidence was found.

5.Discussion

5.1 Clinimetric properties of the Berg Balance Scale

The BBS, when used in older adults, is reliable and valid. Its reponsiveness is not known.
The results collected showed that:

- All the studies found reported a Cronbach alpha in the range suggested by the guidelines, namely 0.70 – 0.90;[29]
- All the studies reported high to excellent intrarater, interrater and test-retest reliability, with ICCs ranging from 0.77 to 0.998;
- There are conflicting findings about the MDC;
- There is moderate evidence that the BBS is valid (construct and cross-cultural validity)
- There is moderate evidence that BBS predicts falls, with a cut-off score of 47 points;
- There is moderate evidence that the responsiveness of the BBS is limited.

These results are going to be compared to other studies, in order to check if something new has been discovered in this review.

Reliability

Internal consistency:

There is strong evidence that BBS has high internal consistency. This information is supported also by literature older than 1999.[58]

Intrarater, interrater, test-retest reliability:

There is moderate evidence that intrarater reliability of BBS is excellent, with ICC ranging from 0.97 to 0.99.

There is moderate evidence that intrarater reliability of BBS is excellent, with ICC ranging from 0.87 to 0.998.

There is limited evidence that test-retest reliability is excellent, with an ICC of 0.94.

High intrarater, interrater and test-retest reliability of BBS is also confirmed by the results of the literature published before 1999. The first evaluation of the properties of BBS was performed by Berg et al. in 1989.[8] The scale showed an excellent reliability resulting in a ICC = 0.98.

In a later study by Berg et al.,[58] test-retest reliability was calculated in a sub-group of 35 acute CVA patients and it produced a ICC of 0.98. Liston et al.[59] also reported excellent test-retest reliability in 22 CVA patients, with a ICC of 0.98. Thorbahn et al. [60] found good interrater reliability from a Spearman rho of 0.88 for 17 elderly subjects. Berg et al.[58] described a smaller inter-observer agreement in the middle third of the scale.

Because of the high number of studies about reliability of BBS, and because all of them agree, regardless of the publication year and the sample considered, it can be stated that BBS has high intrarater, interrater and test-retest reliability.

Measurement error:

Measurement error of BBS is not so simple to be established. According to the literature found, minimum detectable change is somewhere between 4 and 9 points.

From all data collected it can be assumed that:

- It is not likely that MDC is at 9 points, because this value appeared only in a very small sample of people. Furthermore, the exact score calculated was 8.1, which is closer to 8 than to 9;
- It is not likely that MDC is at 4 points, because this value was demonstrated in a small sample of people. Furthermore, this sample had already high baseline score, and this means that the minimum detectable change could have been lowered because some of the participants started almost at the maximum score possible;
- MDC might be 8 points in individuals with severe cognitive disabilities, but this is demonstrated by fair evidence;
- It is likely that MDC is situated between 5 and 7 points, this values are supported by one excellent study. Another study found the MDC to be at 7 points, but this was rated as “fair”.

The reason for this grade was the number of participants, that did not reach the number of 50, but in fact the participant number was 48, which approaches very closely the number required to have good evidence. Furthermore, considering that the range found from all the studies collected was 4-9 points, 5 and 7 are in the middle range.

- Baseline scores seem influence the MDC, this can be caused by the fact that some of the items are more difficult then others to be interpreted, especially items 9 and 10. This means that, unless the performance is scored 0 (impossible to be done) or 4 (done without any problem), the score could vary between 1 and 3 points, even if the performance does not change. This might be an explanation why the MDC in the middle of range scores is higher than at the extremes.
- According to the hypothesis that MDC is situated between 5 and 7 points, if there is the need to have only one value, 7 points would be the most logical choice, because it is the highest value in the range.

According to the quality criteria presented in the introduction, the MCD should be compared with the MIC, but no evidence was found about the MIC of the BBS.
Older literature described lower MDCs (95% CI), but the samples examined were small: it was found to be 2 points for twenty-six subjects with Parkinson's Disease,[17] 3 points for twenty people with hemiparesis,[59] 4 points for five people with TBI[61] and 5 points for twenty-four elderly people with or without stroke.[58]
From the small sample size of these old studies it can be immediately understood that the newer literature has stronger weight in the estimation of the MDC.
To conclude this argumentation, strong evidence supports that MDC is situated between 5 and 7, according to the baseline scores; 7 points can be chosen as a reference MDC if one specific value is needed, but the evidence that supports this choice is lower.
Compared to the older literature, this is a new finding.

Validity

Content Validity:
In recent studies, no information was found about content validity of BBS.
The content validity was assessed by Berg in 1989,[8] that found the scale to be valid. Further validation of the scale was performed through construct and criterion validity, according to the literature found.

Construct Validity:
According to the recent literature found, there is limited evidence that BBS has a moderate correlation with TUG and usual gait speed .
In the past, a considerable amount of literature evaluated the construct validity of BBS.

Construct validity of the scale has been suppported by various studies performed on patients with stroke[62,63]

Liston et al.[59] reported a high *r* (Pearson correlation) between BBS scores and Balance Master. Berg et al.[64] calculated a correlation of *r* = 0.80 between Barthel Index and BBS in CVA patients and a moderate correlation between BBS scores and gait speed. The same authors found correlation of *r* = 0.70 between BBS and Fugl-Meyer Scale. In another study, Berg et al.[65] found a good correlation with Timed Up and Go test (*r* = 0.76) and Tinetti Balance subscale (*r* = 0.91) in a group of 31 participants. Moderate correlations exists between BBS and the Functional Independence Measure, according to Juneja et al.[66] (*r* = 0.57 at admission and *r* = 0.53 at discharge) and Wee et al.[67](*r* = 0.76).

Some studies[59,68,69] demonstrated that BBS has a correlation on the order of 0.50 or greater with various force platform indicators of postural control. BBS has an excellent correlation with other tests like the Postural Assessment Scale for Stroke Patients, Functional Reach Test, the Functional Independence Measure and the Rivermead Mobility Index.[70]

Construct validity of BBS was also found to be high in studies on younger subjects, aged above 45 years.[62,63]

If new and old studies are compared together, it seems that BBS has a good correlation especially with the TUG, but the level of evidence of this assumption is not known.

These data proof that BBS is a valid measurement instrument, and this review does not add new information about its construct validity.

Criterion Validity:

Moderate evidence shows that BBS is predictor of falls, with a cut-off score of 47.

The original validation suggested a cut-off score of 45 points. [64]

Concurrent validity of BBS in elderly persons was not assessed in the literature found.

Concurrent validity of BBS was evaluated extensively in other studies, especially about stroke patients. Juneja et al.[66] found moderate correlations between BBS and the Functional Independence Measure (*r* = 0.57 at admission and *r* = 0.53 at discharge) in stroke patients. According to the authors, this means that an assessment of balance at admission to inpatient rehabilitation unit, performed with the BBS, may enhance the ability to predict rehabilitation outcomes. Furthermore, it may assist to estimate approximate length of stay and predicting eventual discharge destination.[67] and it can be used to evaluate which patients have a tendency to fall.[71] Mao et al.[63] found excellent correlations with the balance subscale of the Fugl-Meyer at 14, 30, 90 and 180 days post stroke (*r*= 0.90 to 0.92) and excellent correlations with Postural Assessment Scale for Stroke patients (PASS) (*r* = 0.92 to 0.95). Liston et al.[59] reported excellent correlation of the BBS to the Barthel Index in chronic stroke patients (*r* = 0.76 – 0.81). Shumway-Cook et al.[72] Found an excellent correlation with Dynamic Gait Index in an elderly population, (*r* = 0.67).

Even if this review does not add any further infomation about the concurrent validity of the BBS, it can be stated that its concurrent validity is strong, based on the consistent result of the older literature.

Responsiveness

There is moderate evidence that BBS is not a responsive instrument for subjects with high baseline score.

Therefore, the BBS should not be used as a tool to detect performance changes in persons having high BBS baseline scores.

It is not clear if BBS is responsive in populations with lower baseline scores and what is the value of the MDC in this case. The MDC was revealed to be 7 points for responsiveness in stroke patients, but the evidence about it is limited and not confirmed by any other study. If the MDC was known, it would also be possible to determine the limit baseline score suitable to have a responsive measurement, free from ceiling effects.

Inability of the BBS to detect true change in balance at the high end of the scale has already been demonstrated in hemiplegic patients by Garland et al.[73]

On the other hand, Mao et al.[63] demonstrated that BBS has high responsiveness in patients with stroke at different phases. The participants of the study had an average baseline score of 22.3 points. This study suggest, once again, that it is possible that BBS is responsive in populations with low BBS baseline scores.

Interpretability

There is good evidence that BBS shows a ceiling effect, especially in healthy, independent people, or groups with mild disabilities, therefore a more challenging measurement should be used in this types of population.[3,54,50] An easier test is not needed, because no floor effect was reported. The most challenging items have been demonstrated to be item 13, “Standing with one foot in front” and item 14, “Standing on one foot”.[3] No information was found about the MIC.

Clinical implications

BBS is a suitable test, that can be used in the clinical practice to assess the balance of patients aged 60 or above. If the test is used to evaluate the improvements after a period of treatment, the clinician should be critical towards the results, because the responsiveness is not clear. It is advisable to add other tests or questionnaires to BBS, in order to evaluate properly such improvements. If the baseline score is very high, another measurement should be used, because BBS is not challenging enough for some people, especially if healthy or with mild disabilities, and it shows ceiling effects.

What was already known? What is new?

From older studies, it was already known that BBS is a reliable and valid measurement tool in some populations, for example stroke patients. Some sparse data about responsiveness were available.

This review adds the information that BBS is reliable also in elderly persons, healthy or not, aged 60 years and above. The MDC has been found to be higher than expected in the older literature. The cut-off score of the test may be higher than what expected in the previous literature. Some more information have been added about the responsiveness of the test, and the role of ceiling effect in the highest range of the baseline scores has been highlighted.

5.2 Clinimetric properties of the 10 Meters Walking Test

The results collected about the 10MWT showed that:

- Intrarater and interrater reliability are not known;
- There is moderate evidence that 10MWT has high test-retest reliability; with ICC = 0.95 for GCS and ICC = 0.97 for FGS;
- Construct validity: There is moderate evidence that there is an excellent correlation between 10MWT and 20MWT (*r* = 0.922); there is limited evidence that there is a low correlation between 10MWT and 4MWT(discrepancy of ± 0.15 to ± 0.17 m/s);
- Criterion validity: there is moderate evidence that there is a significant correlation between 10MWT and the Sit and Reach Test in healthy older adults; high correlation between 10MWT and Barthel Index score and between 10MWT and the Instrumental Activity of Daily Living; moderate correlation with maximum isometric strength of quadriceps;
- The responsiveness of 10MWT is not known.

These results are going to be compared to other studies, in order to check if something new has been discovered in this review.

Reliability

Internal consistency:

The 10MWT includes only one item, therefore no investigations are necessary to test its internal consistency.

Intrarater, interrater, test-retest reliability:

Intrarater reliability was not assessed in any of the studies found.

The interrater reliability of the 10MWT cannot be evaluated, based on the results found.

There is moderate evidence that both CGS and FGS have excellent test-retest reliability, with ICC = 0.95 or higher.

The results of the recent literature show, in general, slightly higher reliability of 10MWT, compared to some of the data available before 1999, but there is a general agreement that 10MWT has high ICCs. In two studies published before 1999, reliability of 10MWT was assessed. It was reported high test-retest reliability of the measurement in stroke patients, with ICC ranging from 0.72 to 0.98.[74,75] In more recent studies, the reliability of 10MWT was also assessed in groups in which subjects younger

than 60 years were present. Flansbjer et al.[76] reported ICC of 0.88 for intrarater reliability in persons with hemiparesis. Collen et al.[25] found excellent intrarater reliability of the10MWT in persons with stroke (ICC = 0.87 to 0.88) Scivoletto et al.[56] found excellent intra- and interrater reliability (ICCs between 0.95 and 0.99) in subjects with cronic spinal cord injury, aged 58.5 years in the average. Tyson et al.[21] recommend to use 10MWT, because it has excellent interrater reliability (ICC = 0.99) in subjects with traumatic brain injuries. Wolf et al.,[27] in a study on 28 adults, aged 57 years in the average, found excellent interrater reliability for comfortable gait speed on 10 meters, with ICC = 0.980 for people without impairments and ICC = 0.998 for patients with stroke.
Lim et al.[17] reported high test-retest reliability, with ICC = 0.81, in patients with Parkinson's disease.
Test-retest reliability in individuals with traumatic brain injuries was calculated to have ICC = 0.95 - 0.96 for comfortable and fast gait speed.[77]
All the studies found, old and new, agree that 10MWT has an excellent intrarater, interrater and test-retest reliability. On the base of this data, it could be confirmed that 10MWT is a reliable instrument.

Measurement error:
It is not possible to know exactly what is the value of the MDC, from the data collected. Comparing the results, it seems that the MDC is higher in patients with Parkinson's disease than on healthy subjects. This is reasonable, because the daily oscillations in the performances caused by the disease may increase the probability of a measurement error. There is a similarity between the results of Steffen et al.[54] on individuals with Parkinson's disease and another study, carried by Lim et al.,[17] on the same type of population. Steffen et al., indeed, reported a MDC of 0.18 m/s, while Lim et al. reported a MDC of 0.19 m/s.

Validity

Content Validity:
No information was found about content validity of 10MWT.

Construct Validity:
Moderate evidence demonstrated very high concordance between 10MWT and 20MWT. This information is very handy to know in the clinical setting, because it means that it is not necessary to have a room long 20 m to measure the gait speed of a patient.
As already mentioned in the section regarding the validity of the BBS, there is limited evidence that CGS and BBS have a moderate correlation (*r* = 0.46).
In the past, it was found that gait velocity is correlated to Timed Up and Go test and Functional Reach test in both healthy older people and geriatric patients.[78]
More recent articles investigated the construct validity of 10MWT, in groups of people including also individuals aged less than 60 years. For example, Flansbjer et al.[79] found an excellent correlation between comfortable gait speed and TUG (*r* = -0.84), FGS (*r* = 0.92), Stair climbing ascend (*r* =

-0.81), Stair climbing descend (*r* = -0.82), 6MWT (*r* = 0.89) and an excellent correlation between fast gait speed and TUG (*r* = -0.91), CGS (*r* = 0.88), Stair climbing ascend (*r* = -0.84), Stair climbing descend (*r* = -0.87) and 6MWT (ICC = 0.95).
If new and old studies are compared together, it seems that 10MWT, BBS and TUG have a good correlation, but the level of evidence of this assumption is not known.

Criterion validity:
Moderate evidence reported high correlation between 10MWT and Barthel Index (*r* =-0.78) and Instrumental Activities of Daily Living (*r* = -0.76).
No further studies were found about the criterion validity of 10MWT, therefore a comparison is not possible.

Responsiveness

The responsiveness of 10MWT is unknown, because only one article of poor quality was found.
No other studies were found on responsiveness of the 10MWT among elderly people.

Interpretability

In the literature included, 10MWT did not show any floor or ceiling effect.
A disadvantage of the 10MWT, compared to shorter distance walking tests, is that it has a higher chance of floor effect, due to the fact that certain patients can walk only shorter distances. For example, in an experiment performed by English et al. with patients with acute stroke, aged 65 in the average, gait speed showed a floor effect by 19% of the subjects at admission and 2 % at discharge. The patient who demonstrated the floor effect in gait speed required the assistance of more than one person to walk.[80]

Clinical implications

10MWT is a recommendable test for measuring the walking velocity of patients aged 60 or above, because of its high reliability, validity and the simplicity of interpretation, even among non-experienced clinicians. Furthermore, it is easy for the patients to understand the task, it does not require special instruments and it can be performed and repeated in a short time. It might not be suitable for all kinds of patients, especially for people with a disability that limits heavily the walking capacity.
The MDC and responsiveness of this test are not clear, it is therefore advisable to associate other tests to 10MWT in order to prove a true change over time.
If the goal of the clinician is to measure a change in the ADL, the BBS can be an example of test that can be administered with 10MWT, even if they measure different aspects of the activities of daily life, because their results are correlated to each other.

Another factor to be taken into account when performing the 10MWT is that the test is conducted in a controlled environment, for example a physiotherapy practice, and its results can not be directly translated to the external environment. For example, a person that can walk without problems in the practice may not be able to walk safely in the streets, where there are obstacles in the way or distractions that oblige the person to carry out a double task. For these reasons, it is advisable to associate the 10MWT with some other tests suitable to measure the ADL specifically for each individual.

Another point is the way to perform the 10MWT: there are, indeed, different procedures that can be used. It is not clear whether a procedure is better than others, because all of them are highly reliable. According to the literature, if the pathway does not offer some space for acceleration, the results might be disconnected from the reality.[57] This may happen especially in patients that have difficulties in starting movements, due to orthopedic or neurological issues, because the initial speed will lower the average speed. On the other hand, most of the clinics were the tests are performed do not provide long, straight pathways were the patients can walk without interruptions; in this case a path with long space for acceleration would not be available.

Furthermore, if one patient needs long starting time, he or she will probably need a similar acceleration space to reach the optimal speed when the test is performed again after some time. For these reasons, the advice is to choose one of the procedures and be consistent in the re-tests. The best compromise seems to be the version “2 + 6 + 2 meters”, because it leaves room for acceleration and deceleration, but it does not require a huge amount of space.

This hypothesis is confirmed by some studies.[56]

Another incongruity in the procedures was found in the speed chosen. Some authors investigated both comfortable and maximal gait velocity, while others chose to test only one modality. Officially, both speeds should be tested. The choice of testing two different walking speeds has probably something to do with the ADL. To have an adequate response to the needs of daily living, it is indeed necessary to be able to control the velocity of gait, accelerating or decelerating the walking pace, if necessary.[12] This does not have something to do only with the velocity of the gait itself, but also on the ability to control the body movements, the fear of walking faster and the general condition of a person. It is advisable to test comfortable and maximum walking speed every time, because it is more appropriate to record both values and it does not take a long time to do it.

What was already known? What is new?

From older studies, it was already known that 10MWT is a reliable and valid measurement tool, but not much was known about its responsiveness. This review confirms that 10MWT is a reliable instrument when used in adults aged 60 years or more. The element that was added about the validity of the test is that the 10MWT can replace the 20MWT in the assessment of walking speed and it probably cannot be replaced by the 4MWT. Some information was also added about the criterion validity of the test: 10MWT has high correlation with Barthel Index and the Instrumental Activity Daily Living. The responsiveness of 10MWT is still not known. It was discovered that there are different versions of the 10MWT, they are all reliable, but they should be performed with the same procedure all the times.

5.3 Limitations/strengths

There were some strengths and some limitations in this thesis.

One strong point is that recent studies were included, in order to offer the best evidence available. Older studies have been searched but they did not match the current required quality standards, or the whole article was not published online. This is why it was chosen to set a limitation to the year of publication in the inclusion/exclusion criteria.

Another strength was the high generalizability of the review, because of the wide range of participants' health conditions and the various countries were the studies were performed.

On the other hand, the diversity of samples gave vague results, especially regarding the minimum detectable change. Even if the total number of studies included was acceptable, it was not possible to compare studies concerning the same type of population.

An important limitation was the fact that the quality of the articles was assessed only by one person, this might have lead to some misjudgments.

6.Conclusion

The BBS, when performed by adults aged 60 years and over, has good internal consistency (Cronbach alpha = 0.77) and excellent intrarater, interrater and test-retest reliability, (ICCs ranging from 0.87 to 0.998). Its MDC is 5-7 points, depending on the baseline score.

Its validity has been confirmed though construct and criterion validity. Its responsiveness is limited, due to ceiling effects.

The 10MWT, when performed by adults aged 60 years and over, has excellent test-retest reliability (ICC = 0.95 for GCS and ICC = 0.97 for FGS). There is no clear agreement about MDC of 10MWT. Its validity has been confirmed though construct and criterion validity. The responsiveness of 10MWT is not known.

BBS and 10MWT are both suitable tests for measurements of the ADL in populations of elderly people, because they are highly reliable and valid, but their responsiveness is not clear.

7.Recommendations for future studies

Most of the studies found described reliability and validity of Berg Balance Scale and 10 Meter Walk Test. MDC was also assessed, but it is still not clear the amount of it, especially regarding the 10MWT. The responsiveness of these tests after a period of treatment is not clear. Because of this reason, different tests are often combined together to assess change in the elderly population, but the administration of multiple tests can be a considerable burden for the patients and clinicians.

It is therefore recommended to deepen the knowledge of responsiveness of BBS and 10MWT in different groups of elderly patients.

8.References

1. WHO. Definition of an older or elderly person. Proposed Working Definition of an Older Person in Africa for the MDS Project. [Internet] Available at:http://www.who.int/healthinfo/survey/ageingdefnolder/en/ [accessed 2013, Dec 27]

2. WHO. Active aging: a policy frame work. A contribution of the World Health Organization to the Second United Nations World Assembly on Ageing. [Internet] Available at: http://whqlibdoc.who.int/hq/2002/WHO_NMH_NPH_02.8.pdf [accessed 2014, Mar 28].

3. Wang C, Hsieh CL, Olson SL, Wang CH, Sheu CF, Lianget CC. Psychosometric properties of the Berg Balance Scale in a community-dwelling elderly resident population in Taiwan. J Formos Med Assoc. 2006 Jun;105(2):992-999

4. Rubenstein LZ. Falls in older people: epidemiology, risk factors and strategies for prevention. Age Ageing. 2006 Sep;35(2):37–41

5. Radosavljevic N, Nikolic D, Lazovic M, Petronic I, Milicevic V, Radosavljevic Z et al. Estimation of functional recovery in patients after hip fracture by Berg Balance Scale regarding the sex, age and comorbidity of partecipants. Geriatr Gerontol Int. 2013 Apr;13(2):365-371

6. Panel on Prevention of Falls in Older Persons. American Geriatrics Society, British Geriatrics Society. Summary of the updated American Geriatrics Society/British Geriatrics Society clinical practice guideline for prevention of falls in older persons. J Am Geriatr Soc. 2011 Jan;59(1):148–157

7. Hayes KW, Johnson ME. Measures of Adult General Performance. American College of Rheumatology. 2003 Oct;49(5):28–42

8. Berg KO, Wood-Dauphinee S, Williams JI, Gayton DG. Measuring balance in the elderly: preliminary development of an instrument. Physiother Can. 1989;41(6):304-311

9. Jogi P, Spaulding SJ, Zecevic AA, Overend TJ, Kramer JF. Comparison of the Original and Reduced Versions of the BBS and WOMAC Following Hip or Knee Arthroplasty. Physiotherapy Canada. 2010;63(1):107-114

10. Bennie S, Bruner K, Dizon A, Fritz H, Goodman B, Peterson S. Measurements of Balance: Comparison of the Timed "Up and Go" Test and Functional Reach Test with the Berg Balance Scale. Journal of Physical Therapy Science. 2003;15(3):93-97.

11. Matjacić Z, Bohinc K, Cikajlo I.Development of an objective balance assessment method for purposes of telemonitoring and telerehabilitation in elderly population.Disabil Rehabil.2010;32(3):259-66.

12. Rydwik E, Bergland A, Forsén L, Frändin K. Investigation into the reliability and validity of the measurement of elderly people’s clinical walking speed: A systematic review. Physiotherapy Theory and Practice. 2012 Apr;28(3):238–256

13. Fritz S, Lusardi M. Walking speed: the sixth vital sign. Journal of Geriatric Physical Therapy. 2009;32(2):2-5

14. Von Heideken Wågert P, Gustafson Y, Lundin-Olsson L. Large variation in walking, standing up from a chair and balance in women and men over 85 years: an observational study. Australian Journal of Physiotherapy. 2009;55(1):39-45

15. Espy DD, Yang F, Bhatt T, Pay YC. Independent influence of gait speed and step length on stability and fall risk. Gait & Posture. 2010 Jul;32(3):378–382

16. Purser JL, Weinberger M, Cohen HJ, Pieper CF, Morey MC, Li T, et al. Walking speed predicts health status and hospital costs for frail elderly male veterans. J Rehabil Res Dev. 2005 Jul-Aug;42(4):535– 46

17. Lim L, van Wegen E, de Goede C, Jones D, Rochester L, Hetherington V, et al. Measuring gait and gait-related activities in Parkinson's patients' own home environment: a reliability, responsiveness and feasibility study. Parkinsonism Relat Disord.2005;11(3):19-24

18. Bohannon RW, Williams Andrews A. Normal walking speed: a descriptive meta-analysis. Physiotherapy. 2011 Sep;97(3):182–189

19. Worsfold C, Simpson JM. Standardisation of a three-metre walking test for elderly people. Physiotherapy. 2001 Mar;87(3):125-133

20. Brouwer B, Musselman K, Culham E. Physical function and health status among seniors with and without a fear of falling. Gerontology. 2004 May-Jun;50(3):135–141

21. Tyson S, Connell L. The psychometric properties and clinical utility of measures of walking and mobility in neurological conditions: a systematic review. Clin Rehabil. 2009;23(11):1018-1033.

22. Hollman JH, Beckman BA, Brandt RA, Merriwether EN, Williams RT, Nordrum JT. Minimum detectable change in gait velocity during acute rehabilitation following hip fracture. J Geriatr Phys Ther. 2008;31(2):53-56.

23. Paltamaa J, Sarasoja T, Leskinen E, Wikström J, Mälkiä E. Measures of physical functioning predict self-reported performance in self-care, mobility, and domestic life in ambulatory persons with multiple sclerosis. Archives of physical medicine and rehabilitation. 2007 Dec;88(12):1649-1657.

24. Van Hedel HJ, Dietz V, Curt A. Assessment of walking speed and distance in subjects with an incomplete spinal cord injury. Neurorehabil Neural Repair. 2007 Jul-Aug;21(4):295-301.

25. Collen F, Wade D, Bradshaw CM. Mobility after stroke: reliability of measures of impairment and disability. Disability & Rehabilitation. 1990 Jan-Mar;12(1):6-9.

26. Watson, M. Refining the ten-metre walking test for use with neurologically impaired people. Physiotherapy. 2002 Jul;88(7):386-397

27. Wolf SL, Catlin PA, Gage K, Gurucharri K, Robertson R, Stephen K. Establishing the reliability and validity of measurements of walking time using the Emory Functional Ambulation Profile. Phys Ther. 1999 Dec;79(12):1122-1133.

28. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al.The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes:results of the COSMIN study. Journal of Clinical Epidemiology. 2010 Jul;63(7):737-745

29. De Vet H, Terwee CB, Mokkink LB, Knol DL. Measurement in Medicine. 3rd ed. New York Cambridge; 2014, p.83.

30. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Quality of Life Research. 2010 May;19(4):539-549

31. Conradsson M, Lundin-Olsson L, Lindelöf N, Littbrand H, Malmqvist L, Gustafson Y, et al. Berg Balance Scale: intrarater test-retest reliability among older people dependent in activities of daily living and living in residential care facilities. Phys Ther. 2007 Sep;87(9):1155-1163

32. Holmefur M, Aarts P, Hoare B, Krumlinde-Sundholm L. Test-retest and alternate forms reliability of the assisting hand assessment. J Rehabil Med. 2009 Nov;41(11):886–891

33. De Vet H, Terwee CB, Mokkink LB, Knol DL. Measurement in Medicine. 3rd ed. New York Cambridge; 2014. pp.154-185

34. De Vet H, Terwee CB, Mokkink LB, Knol DL. Measurement in Medicine. 3rd ed. New York Cambridge; 2014. p.228

35. Terwee CB, Mokkink LB, Knol DL, Ostelo RWJG, Bouter LM, De Vet HCW. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res. 2012 May;21(4):651-657

36. Van Tulder M, Furlan A, Bombardier C, Bouter L, the Editorial Board of the Cochrane Collaboration Back Review Group. Updated Method Guidelines for Systematic Reviews in the Cochrane Collaboration Back Review Group. SPINE. 2003;28(12):1290–1299

37. Terwee CB, Bot SDM, Boer MR, Windt DAWM, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. Journal of Clinical Epidemiology. 2007 Jan;60(1):34–42.

38. Boulgarides LK, McGinty SM, Willett JA, Barnes CW. Use of Clinical and Impairment-Based Tests to Predict Falls by Community-Dwelling Older Adults. Phys Ther 2003 Apr;838(4):328-339

39. Brusse KJ, Zimdars S, Zalwski KR, Steffen TM. Testing functional performance in people with Parkinson disease. Phys Ther. 2005 Feb;85(2):134-141

40. Chiu AYY, Au-Yeung SSY, Lo SK. A Comparison of Four functional tests in discriminating fallers from non-fallers in older people. Disability and Rehabilitation 2003 Jan;25(1):45-50

41. Donoghue D, Stokes EK. How much change is true change? The minimum detectable change of the Berg Balance Scale in elderly people. J Rehabil Med 2009 Apr;41(5):343-346

42. De Figuiredo KMOB, de Lima KC, Maciel ACC, Guerra RO. Interobserver reproducibility of the Berg Balance Scale by novice and experienced physiotherapists. Physiotherapy Theory and Practice. 2009 Jan-Feb;25(1):30-36

43. Halsaa KE, Brovold T, Graver V, Sandvik L, Bergland A. Assessment of interrater reliability and internal consistency of the norvegian version of the Berg Balance Scale. Arch Phys Med Rehabil 2007 Jan;88(1):94-98

44. Holbein-Jenny MA, Billek-Sawhney B, Beckman E, Smith T. "Balance in personal care home residents: a comparison of the Berg Balance Scale, the Multi-Directional Reach Test, and the Activities-Specific Balance Confidence Scale." J Geriatr Phys Ther 2005;28(2):48-53

45. Lajoie Y, Gallager SP. Predicting falls within the elderly community: comparison of postural sway, reaction time, the Berg Balance Scale and the Activities-specific Balance Confidence (ABC) scale for comparing fallers and non-fallers. Arch Gerontol Geriatr.2004 Jan-Feb;38(1):11-26

46. Leerar PJ, Miller EW. Concurrent validity of distance-walks and timed-walks in the well-elderly. Journal of Geriatric Physical Therapy. 2002;25(2):3-7

47. Maeda A, Yuasa T, Nakamura K, Higuki S, Motohashi Y. Physical performance tests after stroke: reliability and validity. Am J Phys Med Rehabil 2000 Nov-Dec;79(6):519-525

48. Miyamoto ST, Lombardi Junior I, Berg KO, Ramos LR, Natour J. Brazilian version of the Berg Balance Scale. Brazilian Journal of Medical and Biological Research. 2004 Sep;37(9):1411-1421

49. Montero-Odasso M, Shapira M, Soriano ER, Varela M, Kaplan R, Camera LA, et al. Gait velocity as a single predictor of adverse events in healthy seniors aged 75 years or older. Journal of Gerontology. 2005;60(10):1304-1309

50. Pardasaney PK, Latham NK, Jette AM, Wagenaar RC, Ni P, Slavin MD et al. Sensitivity of change and responsiveness of four balance measures for community-dwelling older adults. Physical Therapy. 2012 Mar ;92(3):388-397

51. Perera S, Mody SH, Woodman RC, Studenski SA. Meaningful change and responsiveness in common physical performance measures in older adults. Geriatr Sc. 2006 May;54(5):743-749

52. Peters DM, Fritz SL, Krotish DE. Assessing the reliability and validity of a shorter walk test compared with the 10-Meter Wallk Test for measurements of gait speed in healthy, older adults. J Geriatr Phys Ther 2013 Jan-Mar;36(1):24-30

53. Steffen TM, Hacker TA, Mollinger L. Age- and gender- related test performance in community-dwelling elderly people: six-minute walk test, Berg Balance Scale, Timed Up and Go Test, and gait speeds. Phys Ther. 2002 Feb ;82(2):128-137

54. Steffen T, Seney M. Test-retest reliability and minimal detectable change on balance and ambulation testsl the 36-Item Short Form Health Survey, and the Unified Parkinson Disease Rating Scale in people with parkinsonism. Phys Ther 2008 Jun;88(6):733-746

55. Stevenson TJ. Detecting change in patients with stroke using the Berg Balance Scale. Australian Journal of Physiotherapy. 2001;47(1):29-38

56. Scivoletto G, Tamburella F, Laurenza L, Foti C, Ditunno JF, Molinari M. Validity and reliability of the 10-m walk test and the 6-min walk test in spinal cord injury patients. Spinal Cord. 2011 Jan;49(6):736-740.

57. Lindemann U, Najafi B, Zijlstra W, Hauer K, Muche R, Becker C, et al. Distance to achieve steady state walking speed in frail elderly persons. 2008 Jan;27(1):91-96

58. Berg K, Wood-Dauphinee S, Williams JI. The Balance Scale: reliability assessment with elderly residents and patients with acute stroke. Scand J of Rehabil Med. 1995 Mar;27(1):27-36

59. Liston RAL, Brouwer BJ. Reliability and validity of measures obtained from stroke patients using the Balance Master. Archives of Physical Medicine and Rehabilitation. 1996 May;77(5):425-430

60. Thorbahn LDB, Newton RA. Use of the Berg balance test to predict falls in elderly persons. Physical Therapy. 1996;76(6):576-585

61. Newstead A, Hinman M, Tomberlin J. Reliability of the Berg Balance Scale and Balance Master limits of stability tests for individuals with brain injury. Journal of Neurologic Physical Therapy. 2005 Mar;29(1):18-23

62. Usuda S, Araya K, Umehara K, Endo M, Shumizu T, Endo F. Construct validity of functional balance scale in stroke inpatients. J Phys Ther Sci. 1998 Dec;10(1):53-56

63. Mao HF, Hsueh IP, Tang PF, Sheu CF, Hsieh CL. Analysis and comparison of the psychometric properties of three balance measures for stroke patients. Stroke. 2002 May;33:1022-1027

64. Berg KO, Wood-Dauphinee S, Williams JI, Maki BE. Measuring balance in the elderly: validation of an instrument. Canadian Journal of Public Health. 1992 Jul-Aug;83(2):S7-S11

65. Berg KO, Maki BE, Williams JL, Holliday PJ, Wood-Dauphinee S. Clinical and laboratory measures of postural balance in an elderly population. Arch Phys Med Rehabil. 1992 Nov;73(11):1073-1080

66. Juneja G, Czyrny JJ, Linn RT. Admission balance and outcomes of patients admitted for acute inpatient rehabilitation. American Journal of Physical Medicine and Rehabilitation. 1998 Sep-Oct;77(5):388-393

67. Wee JYM, Bagg SD, Palepu A. The Berg Balance Scale as a predictor of length of stay and discharge destination in an acute stroke rehabilitation setting. Archives of Physical Medicine and Rehabilitation. 1999 Apr;80(4):448-452

68. Niam S, Cheung W, Sullivan PE, Kent S, Gu X. Balance and physical impairment after stroke. Archives of Physical Medicine and Rehabilitation. 1999 Oct;80(10):1227-1233

69. Stevenson TJ, Garland SJ. Standing balance during internally produced perturbations in subjects with hemiplegia: validation of the balance scale. Archives of Physical Medicine and Rehabilitation. 1996 Jul;77(7):656-672

70. Blum L, Korner-Bitensky N. Usefulness of the Berg Balance Scale in a Stroke Rehabilitation: a Systematic Review. Phys Ther. 2008 May;88(5):559-566

71. Andersson AG, Kamwendo K, Seiger A, Appelros P. How to identify potential fallers in a stroke unit: validity indexes of 4 test methods. J Rehabil Med. 2006 May;38(3):186–191.

72. Shumway-Cook A, Baldwin M, Polissar NL, Gruber W. Predicting the probability for falls in community-dwelling older adults. Physical Therapy. 1997 Aug;77(8):812-819.

73. Garland SJ, Stevenson TJ, Ivanova T.Postural responses to unilateral arm perturbation in young, elderly, and hemiplegic subjects. Arch Phys Med Rehabil.1997 Oct;78(10):1072-1077.

74. Evans MD, Goldie PA, Hill KD. Systematic and random error in repeated measurements of temporal and distance parameters of gait after stroke. Arch Phys Med Rehabil. 1997 Jul;78(7):725-729

75. Hill KD, Goldie PA, Baker PA, Greenwood KM. Retest reliability of the temporal and distance characteristics of hemiplegic gait using a footswitch system. Arch Phys Med Rehabil.1994 May;75(5):577-83

76. Flansbjer UB, Blom J, Brogårdh C. The reproducibility of Berg Balance Scale and the Single-Leg Stance in chronic stroke and the relationship between the two tests. American Academy of Physical Medicine and Rehabilitation. 2012 Mar;4(3):165-170

77. Van Loo M, Moseley A, Bosman J, de Bie RA, Hassett L. Test-retest reliability of walking speed, step length, and step width measurements after traumatic brain injury: a pilot study. Brain Inj. 2004 Oct;18(10):1041-1048

78. Langhammer B, Lindmark B. Performance-related values for gait velocity, Timed Up-and-Go and functional reach in healthy older people and institutionalized geriatric patients. Phys Occup Ther Geriatr. 2007;25(3):55-69

79. Flansbjer UB, Holmbäck AM, Downham D, Patten C, Lexell J. Reliability of gait performance tests in men and women with hemiparesis after stroke. J Rehabil Med. 2005 Mar;37(2):75-82.

80. English CK, Hillier SL, Stiller K, Warden-Flood A. The sensitivity of three commonly used outcome measures to detect change amongst patients receiveing inpatient rehabilitation following stroke. Clinical Rehabilitation 2006 Jan;20(1):52-55

Appendix. 1 Berg Balance Scale

Name: ________________________________ Date: __________________

Location: ______________________________ Rater: __________________

ITEM DESCRIPTION SCORE (0-4)

1. Sitting to standing ________
2. Standing unsupported ________
3. Sitting unsupported ________
4. Standing to sitting ________
5. Transfers ________
6. Standing with eyes closed ________
7. Standing with feet together ________
8. Reaching forward with outstretched arm ________
9. Retrieving object from floor ________
10. Turning to look behind ________
11. Turning 360 degrees ________
12. Placing alternate foot on stool ________
13. Standing with one foot in front ________
14. Standing on one foot ________

Total ________

GENERAL INSTRUCTIONS
Please document each task and/or give instructions as written. When scoring, please record the lowest response category that applies for each item.

In most items, the subject is asked to maintain a given position for a specific time. Progressively more points are deducted if:

- the time or distance requirements are not met
- the subject's performance warrants supervision
- the subject touches an external support or receives assistance from the examiner

Subject should understand that they must maintain their balance while attempting the tasks. The choices of which leg to stand on or how far to reach are left to the subject. Poor judgment will adversely influence the performance and the scoring.

Equipment required for testing is a stopwatch or watch with a second hand, and a ruler or other indicator of 2, 5, and 10 inches. Chairs used during testing should be a reasonable height. Either a step or a stool of average step height may be used for item # 12.

Berg Balance Scale

1. SITTING TO STANDING
INSTRUCTIONS: Please stand up. Try not to use your hand for support.
(   ) 4   able to stand without using hands and stabilize independently
(   ) 3   able to stand independently using hands
(   ) 2   able to stand using hands after several tries
(   ) 1   needs minimal aid to stand or stabilize
(   ) 0   needs moderate or maximal assist to stand

2. STANDING UNSUPPORTED
INSTRUCTIONS: Please stand for two minutes without holding on.
(   ) 4   able to stand safely for 2 minutes
(   ) 3   able to stand 2 minutes with supervision
(   ) 2   able to stand 30 seconds unsupported
(   ) 1   needs several tries to stand 30 seconds unsupported
(   ) 0   unable to stand 30 seconds unsupported

If a subject is able to stand 2 minutes unsupported, score full points for sitting unsupported. Proceed to item #4.

3. SITTING WITH BACK UNSUPPORTED BUT FEET SUPPORTED ON FLOOR OR ON A STOOL
INSTRUCTIONS: Please sit with arms folded for 2 minutes.
(   ) 4   able to sit safely and securely for 2 minutes
(   ) 3   able to sit 2 minutes under supervision
(   ) 2   able to able to sit 30 seconds
(   ) 1   able to sit 10 seconds
(   ) 0   unable to sit without support 10 seconds

4. STANDING TO SITTING
INSTRUCTIONS: Please sit down.
(   ) 4   sits safely with minimal use of hands
(   ) 3   controls descent by using hands
(   ) 2   uses back of legs against chair to control descent
(   ) 1   sits independently but has uncontrolled descent
(   ) 0   needs assist to sit

5. TRANSFERS
INSTRUCTIONS: Arrange chair(s) for pivot transfer. Ask subject to transfer one way toward a seat with armrests and one way toward a seat without armrests. You may use two chairs (one with and one without armrests) or a bed and a chair.
(   ) 4   able to transfer safely with minor use of hands
(   ) 3   able to transfer safely definite need of hands
(   ) 2   able to transfer with verbal cuing and/or supervision
(   ) 1   needs one person to assist
(   ) 0   needs two people to assist or supervise to be safe

6. STANDING UNSUPPORTED WITH EYES CLOSED
INSTRUCTIONS: Please close your eyes and stand still for 10 seconds.
(   ) 4   able to stand 10 seconds safely
(   ) 3   able to stand 10 seconds with supervision
(   ) 2   able to stand 3 seconds
(   ) 1   unable to keep eyes closed 3 seconds but stays safely
(   ) 0   needs help to keep from falling

7. STANDING UNSUPPORTED WITH FEET TOGETHER
INSTRUCTIONS: Place your feet together and stand without holding on.
( ) 4 able to place feet together independently and stand 1 minute safely
( ) 3 able to place feet together independently and stand 1 minute with supervision
( ) 2 able to place feet together independently but unable to hold for 30 seconds
( ) 1 needs help to attain position but able to stand 15 seconds feet together
( ) 0 needs help to attain position and unable to hold for 15 seconds

8. REACHING FORWARD WITH OUTSTRETCHED ARM WHILE STANDING
INSTRUCTIONS: Lift arm to 90 degrees. Stretch out your fingers and reach forward as far as you can. (Examiner places a ruler at the end of fingertips when arm is at 90 degrees. Fingers should not touch the ruler while reaching forward. The recorded measure is the distance forward that the fingers reach while the subject is in the most forward lean position. When possible, ask subject to use both arms when reaching to avoid rotation of the trunk.)
( ) 4 can reach forward confidently 25 cm (10 inches)
( ) 3 can reach forward 12 cm (5 inches)
( ) 2 can reach forward 5 cm (2 inches)
( ) 1 reaches forward but needs supervision
( ) 0 loses balance while trying/requires external support

9. PICK UP OBJECT FROM THE FLOOR FROM A STANDING POSITION
INSTRUCTIONS: Pick up the shoe/slipper, which is place in front of your feet.
( ) 4 able to pick up slipper safely and easily
( ) 3 able to pick up slipper but needs supervision
( ) 2 unable to pick up but reaches 2-5 cm(1-2 inches) from slipper and keeps balance independently
( ) 1 unable to pick up and needs supervision while trying
( ) 0 unable to try/needs assist to keep from losing balance or falling

10. TURNING TO LOOK BEHIND OVER LEFT AND RIGHT SHOULDERS WHILE STANDING
INSTRUCTIONS: Turn to look directly behind you over toward the left shoulder. Repeat to the right. Examiner may pick an object to look at directly behind the subject to encourage a better twist turn.
( ) 4 looks behind from both sides and weight shifts well
( ) 3 looks behind one side only other side shows less weight shift
( ) 2 turns sideways only but maintains balance
( ) 1 needs supervision when turning
( ) 0 needs assist to keep from losing balance or falling

11. TURN 360 DEGREES
INSTRUCTIONS: Turn completely around in a full circle. Pause. Then turn a full circle in the other direction.
( ) 4 able to turn 360 degrees safely in 4 seconds or less
( ) 3 able to turn 360 degrees safely one side only 4 seconds or less
( ) 2 able to turn 360 degrees safely but slowly
( ) 1 needs close supervision or verbal cuing
( ) 0 needs assistance while turning

12. PLACE ALTERNATE FOOT ON STEP OR STOOL WHILE STANDING UNSUPPORTED
INSTRUCTIONS: Place each foot alternately on the step/stool. Continue until each foot has touch the step/stool four times.
( ) 4 able to stand independently and safely and complete 8 steps in 20 seconds
( ) 3 able to stand independently and complete 8 steps in > 20 seconds
( ) 2 able to complete 4 steps without aid with supervision
( ) 1 able to complete > 2 steps needs minimal assist
( ) 0 needs assistance to keep from falling/unable to try

13. STANDING UNSUPPORTED ONE FOOT IN FRONT
INSTRUCTIONS: (DEMONSTRATE TO SUBJECT) Place one foot directly in front of the other. If you feel that you cannot place your foot directly in front, try to step far enough ahead that the heel of your forward foot is ahead of the toes of the other foot. (To score 3 points, the length of the step should exceed the length of the other foot and the width of the stance should approximate the subject's normal stride width.)

(   ) 4   able to place foot tandem independently and hold 30 seconds
(   ) 3   able to place foot ahead independently and hold 30 seconds
(   ) 2   able to take small step independently and hold 30 seconds
(   ) 1   needs help to step but can hold 15 seconds
(   ) 0   loses balance while stepping or standing

14. STANDING ON ONE LEG
INSTRUCTIONS: Stand on one leg as long as you can without holding on.

(   ) 4   able to lift leg independently and hold > 10 seconds
(   ) 3   able to lift leg independently and hold 5-10 seconds
(   ) 2   able to lift leg independently and hold ≥ 3 seconds
(   ) 1   tries to lift leg unable to hold 3 seconds but remains standing independently.
(   ) 0   unable to try of needs assist to prevent fall

(   )   TOTAL SCORE (Maximum = 56)

Appendix 2. COSMIN boxes

Box A. Internal consistency

1 Does the scale consist of effect indicators, i.e. is it based on a reflective model?

Design requirements

2 Was the percentage of missing items given?

3 Was there a description of how missing items were handled?

4 Was the sample size included in the internal consistency analysis adequate?

5 Was the unidimensionality of the scale checked? i.e. was factor analysis or IRT model applied?

6 Was the sample size included in the unidimensionality analysis adequate?

7 Was an internal consistency statistic calculated for each (unidimensional) (sub)scale separately?

8 Were there any important flaws in the design or methods of the study?

Statistical methods

9 for Classical Test Theory (CTT): Was Cronbach's alpha calculated?

10 for dichotomous scores: Was Cronbach's alpha or KR-20 calculated?

11 for IRT: Was a goodness of fit statistic at a global level calculated? e.g. $\chi 2$, reliability coefficient of estimated latent trait value (index of (subject or item) separation)

Box B. Reliability: relative measures (including test-retest reliability, inter-rater reliability and intra-rater reliability)

Design requirements

1 Was the percentage of missing items given?

2 Was there a description of how missing items were handled?

3 Was the sample size included in the analysis adequate?

4 Were at least two measurements available?

5 Were the administrations independent?

6 Was the time interval stated?

7 Were patients stable in the interim period on the construct to be measured?

8 Was the time interval appropriate?

9 Were the test conditions similar for both measurements? e.g. type of administration, environment, instructions

10 Were there any important flaws in the design or methods of the study?

Statistical methods

11 for continuous scores: Was an intraclass correlation coefficient (ICC) calculated?

12 for dichotomous/nominal/ordinal scores: Was kappa calculated?

13 for ordinal scores: Was a weighted kappa calculated?

14 for ordinal scores: Was the weighting scheme described? e.g. linear, quadratic

Box C. Measurement error: absolute measures

Design requirements

1 Was the percentage of missing items given?

2 Was there a description of how missing items were handled?

3 Was the sample size included in the analysis adequate?

4 Were at least two measurements available?

5 Were the administrations independent?

6 Was the time interval stated?

7 Were patients stable in the interim period on the construct to be measured?

8 Was the time interval appropriate?

9 Were the test conditions similar for both measurements? e.g. type of administration, environment, instructions

10 Were there any important flaws in the design or methods of the study?

Statistical methods

11 for CTT: Was the Standard Error of Measurement (SEM), **S**mallest Detectable Change (SDC) or Limits of Agreement (LoA) calculated?

Box D. Content validity (including face validity)

General requirements

1 Was there an assessment of whether all items refer to relevant aspects of the construct to be measured?

2 Was there an assessment of whether all items are relevant for the study population? (e.g. age, gender, disease characteristics, country, setting)

3 Was there an assessment of whether all items are relevant for the purpose of the measurement instrument? (discriminative, evaluative, and/or predictive)

4 Was there an assessment of whether all items together comprehensively reflect the construct to be measured?

5 Were there any important flaws in the design or methods of the study?

Box E. Structural validity

1 Does the scale consist of effect indicators, i.e. is it based on a reflective model?

Design requirements

2 Was the percentage of missing items given?

3 Was there a description of how missing items were handled?

4 Was the sample size included in the analysis adequate?

5 Were there any important flaws in the design or methods of the study?

Statistical methods

6 for CTT: Was exploratory or confirmatory factor analysis performed?

7 for IRT: Were IRT tests for determining the (uni-) dimensionality of the items performed?

Box F. Hypotheses testing

Design requirements

1 Was the percentage of missing items given?

2 Was there a description of how missing items were handled?

3 Was the sample size included in the analysis adequate?

4 Were hypotheses regarding correlations or mean differences formulated a priori (i.e. before data collection)?

5 Was the expected direction of correlations or mean differences included in the hypotheses?

6 Was the expected absolute or relative magnitude of correlations or mean differences included in the hypotheses?

7 for convergent validity: Was an adequate description provided of the comparator instrument(s)?

8 for convergent validity: Were the measurement properties of the comparator instrument(s) adequately described?

9 Were there any important flaws in the design or methods of the study?

Statistical methods

10 Were design and statistical methods adequate for the hypotheses to be tested?

Box G. Cross-cultural validity

Design requirements

1 Was the percentage of missing items given?

2 Was there a description of how missing items were handled?

3 Was the sample size included in the analysis adequate?

4 Were both the original language in which the HR-PRO instrument was developed, and the language in which the HR-PRO instrument was translated described?

5 Was the expertise of the people involved in the translation process adequately described? e.g. expertise in the disease(s) involved, expertise in the construct to be measured, expertise in both languages

6 Did the translators work independently from each other?

7 Were items translated forward and backward?

8 Was there an adequate description of how differences between the original and translated versions were resolved?

9 Was the translation reviewed by a committee (e.g. original developers)?

10 Was the HR-PRO instrument pre-tested (e.g. cognitive interviews) to check interpretation, cultural relevance of the translation, and ease of comprehension?

11 Was the sample used in the pre-test adequately described?

12 Were the samples similar for all characteristics except language and/or cultural background?

13 Were there any important flaws in the design or methods of the study?

Statistical methods

14 for CTT: Was confirmatory factor analysis performed?

15 for IRT: Was differential item function (DIF) between language groups assessed?

Box H. Criterion validity

Design requirements

1 Was the percentage of missing items given?

2 Was there a description of how missing items were handled?

3 Was the sample size included in the analysis adequate?

4 Can the criterion used or employed be considered as a reasonable 'gold standard'?

5 Were there any important flaws in the design or methods of the study?

Statistical methods

6 for continuous scores: Were correlations, or the area under the receiver operating curve calculated?

7 for dichotomous scores: Were sensitivity and specificity determined?

Box I. Responsiveness

Design requirements

1 Was the percentage of missing items given?

2 Was there a description of how missing items were handled?

3 Was the sample size included in the analysis adequate?

4 Was a longitudinal design with at least two measurement used?

5 Was the time interval stated?

6 If anything occurred in the interim period (e.g. intervention, other elevant events), was it adequately described?

7 Was a proportion of the patients changed (i.e. improvement or deterioration)?

Design requirements for hypotheses testing

For constructs for which a gold standard was not available:

8 Were hypotheses about changes in scores formulated a priori (i.e. before data collection)?

9 Was the expected direction of correlations or mean differences of the change scores of HR-PRO instruments included in these hypotheses?

10 Were the expected absolute or relative magnitude of correlations or mean differences of the change scores of HR-PRO instruments included in these hypotheses?

11 Was an adequate description provided of the comparator instrument(s)?

12 Were the measurement properties of the comparator instrument(s) adequately described?

13 Were there any important flaws in the design or methods of the study?

Statistical methods

14 Were design and statistical methods adequate for the hypotheses to be tested?

Design requirement for comparison to a gold standard

For constructs for which a gold standard was available:

15 Can the criterion for change be considered as a reasonable gold standard?

16 Were there any important flaws in the design or methods of the study?

Statistical methods

17 for continuous scores: Were correlations between change scores, or the area under the Receiver Operator Curve (ROC) curve calculated?

18 for dichotomous scales: Were sensitivity and specificity (changed versus not changed) determined?

Box Interpretability

Percentage of missing items	
Description of how missing items were handled	
Distribution of the (total) scores	
Percentage of the respondents who had the lowest possible (total) score	
Percentage of the respondents who had the highest possible (total) score	
Scores and change scores (i.e. means and SD) for relevant (sub) groups, e.g. for normative groups, subgroups of patients, or the general population	
Minimal Important Change (MIC) or Minimal Important Difference (MID)	

Box Generalizability

Median or mean age (with standard deviation or range)	
Distribution of sex	
Important disease characteristics (e.g. severity, status, duration) and description of treatment	
Setting(s) in which the study was conducted (e.g. general population, primary care or hospital/rehabilitation care)	
Countries in which the study was conducted	
Language in which the HR-PRO instrument was evaluated	
Method used to select patients (e.g. convenience, consecutive, or random)	
Percentage of missing responses (response rate)	

Appendix 3. Literature research

A full list of the literature researches, including detailed research strings, has been reported in this section.

Database: MEDLINE: (Total: 495)

1.
(Berg Balance Scale OR BBS) AND (reliability OR validity OR responsiveness)
Aged: 65+
Year: from 1999 to 2014
Full Text Available

Results: 308

2.
(Berg Balance Scale OR BBS) AND (reliability OR validity OR responsiveness)
Aged: 80 and over
Year: from 1999 to 2014
Full Text Available

Results: 157

3.
(Ten Meter Walk Test OR 10MWT) AND (reliability OR validity OR responsiveness)
Aged: 65+
Year: from 1999 to 2014
Full Text Available

Results: 24

4.
(Ten Meter Walk Test OR 10MWT) AND (reliability OR validity OR responsiveness)
Aged: 80 and over
Year: from 1999 to 2014
Full Text Available

Results: 6

Database: CINAHL: (Total: 510)

1.
(Berg Balance Scale OR BBS) AND (reliability OR validity OR responsiveness)
Aged: 65+
Year: from 1999 to 2014
Full Text Available

Results: 306

2.
(Berg Balance Scale OR BBS) AND (reliability OR validity OR responsiveness)
Aged: 80 and over
Year: from 1999 to 2014
Full Text Available

Results: 175

3.
(Ten Meter Walk Test OR 10MWT) AND (reliability OR validity OR responsiveness)
Aged: 65+
Year: from 1999 to 2014
Full Text Available

Results: 22

4.
(Ten Meter Walk Test OR 10MWT) AND (reliability OR validity OR responsiveness)
Aged: 80 and over
Year: from 1999 to 2014
Full Text Available

Results: 7

Database: PubMED: (Total: 89)

1.
(((Berg Balance Scale) OR (BBS) AND (reliability OR validity OR responsiveness)) AND aged [MeSH Terms]
Year: from 1999 to 2014
Full Text Available

Results: 82

2.
((Ten Meter Walk Test) OR 10MWT) AND (reliability OR validity OR responsiveness) AND aged [MeSH Terms]
Year: from 1999 to 2014
Full Text Available

Results: 7

Database: SPORTDiscus: (Total: 62)

1.
(Berg Balance Scale OR BBS) AND (reliability OR validity OR responsiveness)
Subject: Thesaurus Term: Older people
Year: from 1999 to 2014
Full Text Available

Results: 56

2.
(Ten Meter Walk Test OR 10MWT) AND (reliability OR validity OR responsiveness)
Subject: Thesaurus Term: Older people
Year: from 1999 to 2014
Full Text Available

Results: 8

Further reference screening

The references listed in the studies included in the result section were checked to find suitable articles. Furthermore, the references of the following literature was screened:

- Graham JE, Ostir GV, Fisher SR, Ottenbacher KJ. Assessing walking speed in clinical research: a systematic review. Journal of Evaluation in Clinical Practice. 2008;14:552–562
- Perera S, Samuel S, Schmid WA, Duncan PW, Studenski S, Lai SM, Richards L. Improvements in Speed-Based Gait Classifications Are Meaningful. Stroke.2007;38:2096-2100

- Howe TE, Rochester L, Neil F, Skelton DA, Ballinger C. Exercise for improving balance in older people (Review). The Cochrane Library. 2012:5.

- Godi M, Franchignoni F, Caligari M, Giordano A, Turcato AM, Nardone A. Comparison of Reliability, Validity and Responsiveness of the Mini-BESTest and Berg Balance Scale in Patients With Balance Disorders. Pys Ther. 2013;93:158-167

- Blum L. Korner-Bitensky N. Usefulness of the Berg Balance Scale in stroke rehabilitation: a systematic review. Phys Ther. 2008;88:559-566

- Noohu MM, Dey AB, Hussain ME. Relevance of balance measurement tools and balance training for fall prevention in older adults. Journal of Clinical Gerontology & Geriatrics. 2013;30:1-5

- Tyson S, Connell L. The psychometric prooperties and clinical utility of measures of walking and mobility in neurological conditions: a systematic review. 2009; 23:1018-1033

- Storey AST, Myrah AM, Bauck RA, Brinkman DM, Friess SN, Webber SC. Indoor and outdoor mobility following total knee arthroplasty. 2013;65(3);279-288

- Rydwik E, Bergland A, Forsén L, Frändin K. Investigation into the reliability and validity of the measurement of elderly people’s clinical walking speed: A systematic review. Physiotherapy Theory and Practice. 2012;28(3):238–256

- Hayes KW, Johnson ME. Measures of Adult General Performance. American College of Rheumatology. 2003;49(5):28–42

Appendix 4. Studies grading

All the studies included were evaluated following the COSMIN standards. Every box contains a number of items that were graded; the following tables show how each item was graded.

Not all the items have the same grading system, therefore the grading system is reported under the table.

When grading an item, abbreviations have been used:

++ = Excellent

+ = Good

+/- = Fair

- = Poor

The final grade is the lowest grade present in the box.

BBS

Box A. Internal Consistency

Author	**Item number**											**Final grade**
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	
Halsaa et al.	**Yes**	**++**	**++**	**+**	**++**	**+**	**++**	**++**	**++**	**NA**	**NA**	**Good**
Comments:	Good study, some minor flaws.											
Steffen et al. (2008)	**Yes**	**++**	**++**	**+/-**	**+**	**+/-**	**-**	**++**	**++**	**NA**	**NA**	**Poor**
Comments:	The unidimensionality of the scale was not assessed. Internal consistency of BBS found in other studies was reported, but it is not clear whether the factor analysis was performed in these studies.											
Wang et al.	**Yes**	**++**	**++**	**++**	**++**	**++**	**++**	**++**	**++**	**NA**	**NA**	**Excellent**
Comments:												

The items could be graded as:
1 = Yes / No
2 =Excellent / Good
3 =Excellent / Good / Fair
4 =Excellent / Good / Fair / Poor
5 =Excellent / Good / Fair / Poor
6 =Excellent / Poor
7 =Excellent / Fair / Poor
8 =Excellent / Fair / Poor
9 = Excellent / Fair / Poor / NA
10 =Excellent / Fair / Poor / NA
11 = Excellent / Poor **/** NA

Box B. Reliability (intrarater, interrater, test-retest)

Author	Item number														Final grade
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	
Conradsson et al.	++	++	+/-	++	++	++	++	++	++	++	++	NA	NA	NA	**Fair**
Comments:	The sample size (45 participants) was the only reason for the lower grade.														
De Figueiredo et al.	++	++	-	++	++	++	++	++	++	++	++	++	NA	NA	**Poor**
Comments:	Very small sample size: 12 participants.														
Halsaa et al.	++	++	+	++	++	++	++	++	++	++	+	++	NA	NA	**Good**
Comments:	Translation procedure to Norwegian described, but not in details.														
Holbein-Jenny et al.	+	+	-	++	++	++	+	++	++	++	+	NA	NA	NA	**Poor**
Comments:	Small sample size: 26 participants.														
Miyamoto et al.	++	++	+/-	++	++	++	++	++	++	++	++	NA	NA	NA	**Fair**
Comments:	The sample size (36 participants) was the only reason for the lower grade.														
Steffen et al. (2008)	++	++	+/-	++	++	++	+	++	++	++	++	NA	NA	NA	**Fair**
Comments:	The sample size (37 participants) was the main reason for the lower grade.														
Wang et al.	++	++	+	++	++	++	++	++	++	++	++	NA	NA	NA	**Good**
Comments:	Sample size for reliability assessment: 68 participants, out of 268 in total.														

The items could be graded as:
1 = Excellent / Good
2 =Excellent / Good / Fair
3 =Excellent / Good / Fair / Poor
4 =Excellent / Poor
5 =Excellent / Good / Fair / Poor
6 =Excellent / Fair
7 =Excellent / Good / Fair / Poor
8 =Excellent / Fair / Poor
9 =Excellent / Good / Fair / Poor
10 =Excellent / Fair / Poor
11 =Excellent / Good / Fair / Poor / NA
12 =Excellent / Poor / NA
13=Excellent / Fair / Poor / NA
14=Excellent / Good / NA

Box C. Measurement Error

Author	Item number											Final grade
	1	2	3	4	5	6	7	8	9	10	11	
Conradsson et al.	++	++	+/-	++	++	++	++	++	++	++	++	**Fair**
Comments:	The sample size (45 participants) was the only reason for the lower grade.											
Donoghue et al.	++	++	++	++	++	++	++	++	++	++	++	**Excellent**
Comments:												
Steffen et al. (2008)	++	++	+/-	++	++	++	+	++	++	++	++	**Fair**
Comments:	The sample size (37 participants) was the main reason for the lower grade.											
Stevenson	++	++	+	++	++	++	++	++	++	++	++	**Fair**
Comments:	Sample size = 48 participants.											

The items could be graded as:
1 = Excellent / Good
2 =Excellent / Good / Fair
3 =Excellent / Good / Fair / Poor
4 =Excellent / Poor
5 =Excellent / Good / Fair / Poor
6 =Excellent / Fair
7 =Excellent / Good / Fair / Poor
8 =Excellent / Fair / Poor
9 =Excellent / Good / Fair / Poor
10 =Excellent / Fair / Poor
11 =Excellent / Good / Poor

Box F. Hypothesis Testing (Construct Validity)

Author	Item number										Final grade
	1	2	3	4	5	6	7	8	9	10	
Brusse et al.	++	++	-	+/-	+	+	++	++	++	++	**Poor**
Comments:	Small sample size (25 participants).										
Holbein-Jenny et al.	+	+	-	-	NA	NA	+/-	+/-	++	++	**Poor**
Comments:	Small sample size (26 participants), measurement properties briefly described, no hypothesis present.										
Wang et al.	++	++	++	+/-	+	+	++	++	++	++	**Fair**
Comments:	Hypothesis not formulated.										

The items could be graded as:
1 = Excellent / Good
2 =Excellent / Good / Fair
3 =Excellent / Good / Fair / Poor
4 =Excellent / Good / Fair / Poor
5 =Excellent / Good / NA
6 =Excellent / Good / NA
7 =Excellent / Good / Fair / Poor
8 =Excellent / Good / Fair / Poor
9 =Excellent / Fair / Poor
10 =Excellent / Good / Fair / Poor / NA

Box G. Cross-cultural Validity

Author	Item number															Final grade
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
Miyamoto et al.	++	++	+/-	++	+	++	++	++	++	++	++	++	++	++	NA	Fair
Comments:	The sample size (36 participants) was the reason for the lower grade.															

The items could be graded as:
1 = Excellent / Good
2 =Excellent / Good / Fair
3 =Excellent / Good / Fair / Poor
4 =Excellent / Poor
5 =Excellent / Good / Fair
6 =Excellent / Good / Fair / Poor
7 =Excellent / Good / Fair / Poor
8 =Excellent / Good
9 =Excellent / Good
10 =Excellent / Good / Fair / Poor
11 =Excellent / Fair / Poor
12 =Excellent / Good / Fair / Poor
13=Excellent / Fair / Poor
14=Excellent / Poor / NA
15=Excellent / Poor / NA

Box H. Criterion validity

Author	Item number							Final grade
	1	2	3	4	5	6	7	
Boulgarides et al.	++	++	+	++	++	-	NA	Poor
Comments:	AUC not calculated							
Chiu et al.	++	++	+	++	++	++	++	Good
Comments:	Participants number: 78							
Lajoie et al.	++	++	++	++	++	-	NA	Poor
Comments:	AUC not calculated							

The item could be graded as:
1 = Excellent / Good
2 =Excellent / Good / Fair
3 =Excellent / Good / Fair / Poor
4 =Excellent / Good / Fair / Poor
5 =Excellent / Fair / Poor
6 =Excellent / Poor / NA
7 =Excellent / Poor/ NA

Box I. Responsiveness

Author	Item number																		Final grade
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
Pardasaney et al.	++	++	++	++	++	++	++	++	++	+	++	++	++	++	NA	NA	NA	NA	Good
Comments:	The expected magnitude of correlations or differences of the change scores were not stated in the hypotheses.																		
Stevenson	++	++	+/-	++	++	++	+	++	+	+	++	++	+/-	++	NA	NA	NA	NA	Fair
Comments:	It is doubtful whether the treatment was effective: it was only two weeks long and it was not standardadized.																		

The items could be graded as:
1 = Excellent / Good
2 =Excellent / Good / Fair
3 =Excellent / Good / Fair / Poor
4=Excellent / Poor
5 =Excellent / Poor
6 =Excellent / Good / Fair
7 =Excellent / Good / Fair / Poor
8 =Excellent / Fair / Poor / NA
9 =Excellent / Good / NA
10 =Excellent / Good / NA
11 =Excellent / Fair / Poor / NA
12 =Excellent / Good / Fair / Poor / NA
13=Excellent / Fair / Poor / NA
14=Excellent / Fair / Poor / NA
15=Excellent / Good / Fair / Poor / NA
16 = Excellent / Fair / Poor
17 =Excellent / Poor / NA
18 =Excellent / Poor / NA

10MWT

Box B. Reliability (intrarater, interrater, test-retest)

Author	Item number														Final grade
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	
Brusse et al.	++	++	-	++	++	++	++	++	++	++	+	NA	NA	NA	Poor
Comments:	Small sample size (25 participants)														
Maeda et al.	+	+	+	++	++	++	+	++	+	++	+/-	NA	NA	NA	Fair
Comments:	Test-retest reliability was calculate after one minute and after one year: only the values of one minute interval was considered, because the participants were stable between measurements.														
Peters et al.	++	++	+/-	++	++	++	++	++	++	++	++	NA	NA	NA	Fair
Comments:	The sample size (43 participants) was the only reason for the lower grade.														
Steffen et al. (2002)	++	++	+	++	++	++	++	++	++	++	+	NA	NA	NA	Good
Comments:	ICC calculated, but formula not described														
Steffen et al. (2008)	++	++	+/-	++	++	++	+	++	++	++	++	NA	NA	NA	Fair
Comments:	The sample size (37 participants) was the reason for the lower grade.														

The items could be graded as:
1 = Excellent / Good
2 =Excellent / Good / Fair
3 =Excellent / Good / Fair / Poor
4 =Excellent / Poor
5 =Excellent / Good / Fair / Poor
6 =Excellent / Fair
7 =Excellent / Good / Fair / Poor
8 =Excellent / Fair / Poor
9 =Excellent / Good / Fair / Poor
10 =Excellent / Fair / Poor
11 =Excellent / Good / Fair / Poor / NA
12 =Excellent / Poor / NA
13=Excellent / Fair / Poor / NA
14=Excellent / Good / NA

Box C. Measurement Error

Author	**Item number**											**Final grade**
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	
Perera et al.	++	++	++	++	++	++	-	++	+	++	++	**Poor**
Comments:	Patients not stable between the measurements											
Peters et al.	++	++	+/-	++	++	++	++	++	++	++	++	**Fair**
Comments:	The sample size (43 participants) was the only reason for the lower grade.											
Steffen et al. (2008)	++	++	+/-	++	++	++	+	++	++	++	++	**Fair**
Comments:	Small sample size (37 participants)											

The items could be graded as:
1 = Excellent / Good
2 =Excellent / Good / Fair
3 =Excellent / Good / Fair / Poor
4 =Excellent / Poor
5 =Excellent / Good / Fair / Poor
6 =Excellent / Fair
7 =Excellent / Good / Fair / Poor
8 =Excellent / Fair / Poor
9 =Excellent / Good / Fair / Poor
10 =Excellent / Fair / Poor
11 =Excellent / Good / Poor

Box F. Hypothesis Testing (Construct Validity)

Author	**Item number**										**Final grade**
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
Brusse et al.	++	++	-	+/-	+	+	++	++	++	++	**Poor**
Comments:	Small sample size (25 participants).										
Leerar et al.	++	++	+	+	++	+	+	+	++	++	**Good**
Comments:	Good study, some minor flaws.										
Peters et al.	++	++	+/-	+/-	+	+	++	+/-	++	++	**Fair**
Comments:	Hypotheses vague, measurement proper not adequately described.										

The items could be graded as:
1 = Excellent / Good
2 =Excellent / Good / Fair

3 =Excellent / Good / Fair / Poor
4 =Excellent / Good / Fair / Poor
5 =Excellent / Good / NA
6 =Excellent / Good / NA
7 =Excellent / Good / Fair / Poor
8 =Excellent / Good / Fair / Poor
9 =Excellent / Fair / Poor
10 =Excellent / Good / Fair / Poor / NA

Box H. Criterion Validity

Author	**Item number**							**Final grade**
	1	**2**	**3**	**4**	**5**	**6**	**7**	
Maeda et al.	**+**	**+**	**+**	**+**	**++**	**++**	**NA**	**Good**
Comments:	No evidence provided about the clinimetric properties of the criteria used, but assumable that they can be considered adequate “gold standards”.							
Montero-Odasso et al.	**++**	**++**	**++**	**++**	**++**	**-**	**NA**	**Poor**
Comments:	AUC not calculated							

The item could be graded as:
1 = Excellent / Good
2 =Excellent / Good / Fair
3 =Excellent / Good / Fair / Poor
4 =Excellent / Good / Fair / Poor
5 =Excellent / Fair / Poor
6 =Excellent / Poor / NA
7 =Excellent / Poor/ NA

Box I. Responsiveness

Author	**Item number**																		**Final grade**
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	
Perera et al.	**++**	**++**	**++**	**++**	**++**	**++**	**++**	**NA**	**NA**	**NA**	**NA**	**NA**	**NA**	**NA**	**+**	**++**	**-**	**NA**	**Poor**
Comments:	Guyatt's responsiveness ratio was used, which is not an optimal statistical method for responsiveness. Correlations or Area under the ROC Curve (AUC) were not calculated.																		

The items could be graded as:
1 = Excellent / Good
2 =Excellent / Good / Fair
3 =Excellent / Good / Fair / Poor
4=Excellent / Poor
5 =Excellent / Poor
6 =Excellent / Good / Fair
7 =Excellent / Good / Fair / Poor
8 =Excellent / Fair / Poor / NA
9 =Excellent / Good / NA
10 =Excellent / Good / NA
11 =Excellent / Fair / Poor / NA
12 =Excellent / Good / Fair / Poor / NA
13=Excellent / Fair / Poor / NA
14=Excellent / Fair / Poor / NA
15=Excellent / Good / Fair / Poor / NA
16 = Excellent / Fair / Poor
17 =Excellent / Poor / NA
18 =Excellent / Poor / NA

Study: Pardasaney	
Box Interpretability	
Percentage of missing items	0
Description of how missing items were handled	na
Distribution of the (total) scores	Ceiling
Percentage of the respondents who had the lowest possible (total) score	0
Percentage of the respondents who had the highest possible (total) score	10% at the beginning, 23% after the treatment period
Scores and change scores (i.e. means and SD) for relevant (sub) groups, e.g. for normative groups, subgroups of patients, or the general population	Baseline: 50.1, after treament: mean difference: 1.4 points, SD = 3.1
Minimal Important Change (MIC) or Minimal Important Difference (MID)	MID= 2.50

Study: Wang	
Box Interpretability	
Percentage of missing items	0
Description of how missing items were handled	na
Distribution of the (total) scores	ceiling
Percentage of the respondents who had the lowest possible (total) score	0
Percentage of the respondents who had the highest possible (total) score	33.2%
Scores and change scores (i.e. means and SD) for relevant (sub) groups, e.g. for normative groups, subgroups of patients, or the general population	53.3 ± 4.1
Minimal Important Change (MIC) or Minimal Important Difference (MID)	Not reported

Study: Boulgarides	
Box Generalisability	
Median or mean age (with standard deviation or range)	Mean =74.02; SD =5.64; Range 65-90
Distribution of sex	60F, 39M
Important disease characteristics (e.g. severity, status, duration) and description of treatment	Mix. Exclusion criteria: heart or pulmonary problems; not able to stand for 5 minutes
Setting(s) in which the study was conducted (e.g. general population, primary care or hospital/rehabilitation care)	General population
Countries in which the study was conducted	USA
Language in which the HR-PRO instrument was evaluated	English
Method used to select patients (e.g. convenience, consecutive, or random)	Random
Percentage of missing responses (response rate)	7/106

Study: Brusse	
Box Generalisability	
Median or mean age (with standard deviation or range)	71-86 (mean:76)
Distribution of sex	11F, 14M
Important disease characteristics (e.g. severity, status, duration) and description of treatment	Parkinson's disease, able to ambulate with or wihout assistive device
Setting(s) in which the study was conducted (e.g. general population, primary care or hospital/rehabilitation care)	Community support group PD and neurologist's group office
Countries in which the study was conducted	USA
Language in which the HR-PRO instrument was evaluated	English
Method used to select patients (e.g. convenience, consecutive, or random)	Random
Percentage of missing responses (response rate)	0

Study: Chiu	
Box Generalisability	
Median or mean age (with standard deviation or range)	Fallers: Mean = 82.12; SD =8.19; Non fallers: Mean = 81.59; SD = 8.31
Distribution of sex	26M 52F
Important disease characteristics (e.g. severity, status, duration) and description of treatment	No pathologies requiring medical interventions
Setting(s) in which the study was conducted (e.g. general population, primary care or hospital/rehabilitation care)	Fallers: Falls clinic; Non fallers:general population
Countries in which the study was conducted	China
Language in which the HR-PRO instrument was evaluated	Not reported
Method used to select patients (e.g. convenience, consecutive, or random)	Convenience
Percentage of missing responses (response rate)	0

Study: Conradsson	
Box Generalisability	
Median or mean age (with standard deviation or range)	Mean: 82.3 (SD= 6.6; Range: 68-96)
Distribution of sex	36F 9M
Important disease characteristics (e.g. severity, status, duration) and description of treatment	Mini Mental State range 4-30
Setting(s) in which the study was conducted (e.g. general population, primary care or hospital/rehabilitation care)	Residential care facilities (patients dependant in ADL)
Countries in which the study was conducted	Sweden
Language in which the HR-PRO instrument was evaluated	n.a.
Method used to select patients (e.g. convenience, consecutive, or random)	Convenience
Percentage of missing responses (response rate)	0

Study: Donoghue	
Box Generalisability	
Median or mean age (with standard deviation or range)	Mean 80.5, range 65-95
Distribution of sex	Male: 34.7 %
Important disease characteristics (e.g. severity, status, duration) and description of treatment	Completely independent (15.3%) to requiring physical assistance (1.1%)
Setting(s) in which the study was conducted (e.g. general population, primary care or hospital/rehabilitation care)	Various physio rehabilitation centres
Countries in which the study was conducted	Ireland
Language in which the HR-PRO instrument was evaluated	English
Method used to select patients (e.g. convenience, consecutive, or random)	Random
Percentage of missing responses (response rate)	0

Study: De Figueiredo	
Box Generalisability	
Median or mean age (with standard deviation or range)	75.8 ± 8.4 (range = 63-87)
Distribution of sex	10F, 2M
Important disease characteristics (e.g. severity, status, duration) and description of treatment	Able to perform the test, osteoporosis, osteoarthritis, hpertension, diabetes.
Setting(s) in which the study was conducted (e.g. general population, primary care or hospital/rehabilitation care)	Private rehabilitation clinic
Countries in which the study was conducted	Brazil
Language in which the HR-PRO instrument was evaluated	Portuguese
Method used to select patients (e.g. convenience, consecutive, or random)	Convenience
Percentage of missing responses (response rate)	0

Study: Halsaa	
Box Generalisability	
Median or mean age (with standard deviation or range)	69-95 (82 ± 5.5)
Distribution of sex	58F 25M
Important disease characteristics (e.g. severity, status, duration) and description of treatment	25 inpatients geriatric rehabilitation 58 geriatric day hospital
Setting(s) in which the study was conducted (e.g. general population, primary care or hospital/rehabilitation care)	Patients of geriatric rehabilitation unit or geriatric day hospital
Countries in which the study was conducted	Norway
Language in which the HR-PRO instrument was evaluated	Norvegian
Method used to select patients (e.g. convenience, consecutive, or random)	Consecutive
Percentage of missing responses (response rate)	0

Study: Holbein-Jenny	
Box Generalisability	
Median or mean age (with standard deviation or range)	Mean = 83.5, SD = 4.9
Distribution of sex	21F, 5M
Important disease characteristics (e.g. severity, status, duration) and description of treatment	Independent but in need for some assistance
Setting(s) in which the study was conducted (e.g. general population, primary care or hospital/rehabilitation care)	Personal care homes
Countries in which the study was conducted	USA
Language in which the HR-PRO instrument was evaluated	English
Method used to select patients (e.g. convenience, consecutive, or random)	Convenience
Percentage of missing responses (response rate)	Not reported

Study: Lajoie	
Box Generalisability	
Median or mean age (with standard deviation or range)	Fallers: mean =75.5; SD = 3.14 Non fallers: mean 73.80; SD = 2.75
Distribution of sex	3M, 86F
Important disease characteristics (e.g. severity, status, duration) and description of treatment	Exclusion: not able to stand 1 minute for 4 times
Setting(s) in which the study was conducted (e.g. general population, primary care or hospital/rehabilitation care)	Nursing homes, senior residences
Countries in which the study was conducted	Canada
Language in which the HR-PRO instrument was evaluated	Not reported
Method used to select patients	Convenience
Percentage of missing responses (response rate)	0

Study: Leerar	
Box Generalisability	
Median or mean age (with standard deviation or range)	60-92
Distribution of sex	Not reported
Important disease characteristics (e.g. severity, status, duration) and description of treatment	Healthy
Setting(s) in which the study was conducted (e.g. general population, primary care or hospital/rehabilitation care)	General elderly population
Countries in which the study was conducted	USA
Language in which the HR-PRO instrument was evaluated	English
Method used to select patients	Random
Percentage of missing responses (response rate)	6/17 didn't answer one question (reason for falling)

Study:Maeda	
Box Generalisability	
Median or mean age (with standard deviation or range)	Males: Stroke: 69.6 ± 8.3 Healthy: 72.1 ± 7.4 Females: Stroke: 70.6 ± 9.1 Healthy: 72.5 ± 7.4
Distribution of sex	Stroke: 21M, 19Fw Healthy: 17M, 23F
Important disease characteristics (e.g. severity, status, duration) and description of treatment	Stroke survivors/healthy independent
Setting(s) in which the study was conducted (e.g. general population, primary care or hospital/rehabilitation care)	Community/ general population
Countries in which the study was conducted	Japan
Language in which the HR-PRO instrument was evaluated	Not reported
Method used to select patients	Convenience
Percentage of missing responses (response rate)	Not reported

Study: Miyamoto	
Box Generalisability	
Median or mean age (with standard deviation or range)	65+
Distribution of sex	Not reported
Important disease characteristics (e.g. severity, status, duration) and description of treatment	Excluded: unable to stand without help, with lower limb amputations
Setting(s) in which the study was conducted (e.g. general population, primary care or hospital/rehabilitation care)	Rheumatology Outpatient Clinic
Countries in which the study was conducted	Brazil
Language in which the HR-PRO instrument was evaluated	Portuguese
Method used to select patients (e.g. convenience, consecutive, or random)	Consecutive
Percentage of missing responses (response rate)	0

Study: Montero-Odasso	
Box Generalisability	
Median or mean age (with standard deviation or range)	Mean = 78.9; SD = 3
Distribution of sex	71.3% female
Important disease characteristics (e.g. severity, status, duration) and description of treatment	Healthy
Setting(s) in which the study was conducted (e.g. general population, primary care or hospital/rehabilitation care)	University hospital
Countries in which the study was conducted	Argentina
Language in which the HR-PRO instrument was evaluated	Not reported
Method used to select patients (e.g. convenience, consecutive, or random)	Random
Percentage of missing responses (response rate)	0

Study: Pardasaney	
Box Generalisability	
Median or mean age (with standard deviation or range)	65 or more average 75.9 (7), 28% ≥ 80
Distribution of sex	F: 68.5 %
Important disease characteristics (e.g. severity, status, duration) and description of treatment	Functional limitations, community-dwelling
Setting(s) in which the study was conducted (e.g. general population, primary care or hospital/rehabilitation care)	General population, in general high functionaal level and motivated to to exercise, but study generalisable because heterogeneous population
Countries in which the study was conducted	USA
Language in which the HR-PRO instrument was evaluated	English
Method used to select patients (e.g. convenience, consecutive, or random)	Random (volunteer)
Percentage of missing responses (response rate)	0

Study: Perera	
Box Generalisability	
Median or mean age (with standard deviation or range)	77.6 ± 7.7; 74.1 ± 5.7; 69.8 ± 10.3
Distribution of sex	Female: 50%, 43.7%, 44%
Important disease characteristics (e.g. severity, status, duration) and description of treatment	1. mobility limitation; 2 community-dwelling; 3.poststroke.
Setting(s) in which the study was conducted (e.g. general population, primary care or hospital/rehabilitation care)	General population
Countries in which the study was conducted	USA
Language in which the HR-PRO instrument was evaluated	English
Method used to select patients	Random
Percentage of missing responses (response rate)	25 %(group3, reason: floor or ceiling at the beginning, no opportunity to report decline)

Study: Peters	
Box Generalisability	
Median or mean age (with standard deviation or range)	84.3 ± 6.9
Distribution of sex	32w 11m
Important disease characteristics (e.g. severity, status, duration) and description of treatment	Healthy
Setting(s) in which the study was conducted (e.g. general population, primary care or hospital/rehabilitation care)	Local retirement community
Countries in which the study was conducted	USA
Language in which the HR-PRO instrument was evaluated	English
Method used to select patients	Convenience
Percentage of missing responses (response rate)	1/129

Study: Steffen (2002)	
Box Generalizability	
Median or mean age (with standard deviation or range)	61-89
Distribution of sex	37M, 59F
Important disease characteristics (e.g. severity, status, duration) and description of treatment	Healthy
Setting(s) in which the study was conducted (e.g. general population, primary care or hospital/rehabilitation care)	General population.
Countries in which the study was conducted	Wisconsin, USA
Language in which the HR-PRO instrument was evaluated	English
Method used to select patients	Random (advertisement)
Percentage of missing responses (response rate)	2/97

Study: Steffen 2008	
Box Generalisability	
Median or mean age (with standard deviation or range)	Mean = 71
Distribution of sex	26M, 11F
Important disease characteristics (e.g. severity, status, duration) and description of treatment	Parkinsonism H&Y 1-4, average 2
Setting(s) in which the study was conducted (e.g. general population, primary care or hospital/rehabilitation care)	Local fitness centres
Countries in which the study was conducted	USA
Language in which the HR-PRO instrument was evaluated	English
Method used to select patients (e.g. convenience, consecutive, or random)	Random
Percentage of missing responses (response rate)	0 BBS; 1 Gait speed

Study: Stevenson	
Box Generalisability	
Median or mean age (with standard deviation or range)	73.5 (7.0)
Distribution of sex	24:24
Important disease characteristics (e.g. severity, status, duration) and description of treatment	Stroke
Setting(s) in which the study was conducted (e.g. general population, primary care or hospital/rehabilitation care)	Rehabilitation unit
Countries in which the study was conducted	Canada
Language in which the HR-PRO instrument was evaluated	English
Method used to select patients (e.g. convenience, consecutive, or random)	Convenience
Percentage of missing responses (response rate)	0

Study: Wang	
Box Generalisability	
Median or mean age (with standard deviation or range)	65-90, mean 73.8 ± 5.18
Distribution of sex	149(55.6%) males, 119 (44.4%) females For reliability testing: 32m 36w
Important disease characteristics (e.g. severity, status, duration) and description of treatment	Healthy, living independently in the community
Setting(s) in which the study was conducted (e.g. general population, primary care or hospital/rehabilitation care)	General population
Countries in which the study was conducted	Taiwan
Language in which the HR-PRO instrument was evaluated	Not reported
Method used to select patients (e.g. convenience, consecutive, or random)	Random
Percentage of missing responses (response rate)	0