Archives of Physical Medicine and Rehabilitation
journal homepage: www.archives-pmr.org
Archives of Physical Medicine and Rehabilitation 2014;95:1759-75

REVIEW ARTICLE (META-ANALYSIS)

Psychometric Properties of 2-Minute Walk Test: A Systematic Review

Tamis W. Pin, PhD

From the Department of Rehabilitation Science, Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong.

Abstract

Objective: To systematically review the psychometric evidence on the 2-minute walk test (2MWT).

Data Sources: Electronic searches of databases including MEDLINE, CINAHL, Academic Search Premier, SPORTDiscus, PsycINFO, EMBASE, the Cochrane Library, and DARE were done until February 2014 using a combination of subject headings and free texts.

Study Selection: Studies were included if psychometric properties of the 2MWT were (1) evaluated; (2) written as full reports; and (3) published in English language peer-reviewed journals.

Data Extraction: A modified consensus-based standard for the selection of health measurement instruments checklist was used to rate the methodological quality of the included studies. A quality assessment for statistical outcomes was used to assess the measurement properties of the 2MWT.

Data Synthesis: Best-evidence synthesis was collated from 25 studies of 14 patient groups. Only 1 study was found that examined the 2MWT in the pediatric population. The testing procedures of the 2MWT varied across the included studies. Reliability, validity (construct and criterion), and responsiveness of the 2MWT also varied across different patient groups. Moderate to strong evidence was found for reliability, convergent validity, discriminative validity, and responsiveness of the 2MWT in frail elderly patients. Moderate to strong evidence for reliability, convergent validity, and responsiveness was found in adults with lower limb amputations. Moderate to strong evidence for validity (convergent and discriminative) was found in adults who received rehabilitation after hip fractures or cardiac surgery. Limited evidence for the psychometric properties of the 2MWT was found in other population groups because of methodological flaws.

Conclusions: There is inadequate breadth and depth of psychometric evidence of the 2MWT for clinical and research purposes—specifically, minimal clinically important changes and responsiveness. More good-quality studies are needed, especially in the pediatric population. Consensus on standardized testing procedures of the 2MWT is also required.

Archives of Physical Medicine and Rehabilitation 2014;95:1759-75

Walk tests are simple, inexpensive, and safe performance-based tests that provide information on functional exercise capacity of individuals when compared with laboratory-based indexes of aerobic capacity such as cycle, treadmill, and step ergometry, which require expensive and cumbersome equipment.[1] Walk tests can be either time-based tests that measure the distance covered in a specific period, such as the 12-minute walk test (12MWT), the 6-minute walk test (6MWT), or the 2-minute walk test (2MWT), or distance-based tests that measure the time taken to complete a set distance, such as the 20-m shuttle test or the 1-mile walk test.[2]

Among the time-based walk tests, both the 2MWT and 6MWT are modifications of the 12MWT.[3] A high correlation has been demonstrated between the 2MWT and 6MWT and the 12MWT in a group of 30 patients with respiratory conditions.[3] Since the introduction of these shorter versions of walk tests in 1982, the 6MWT is the most commonly used and has been thoroughly investigated.[4,5] Standardized testing guidelines for the 6MWT were developed by the American Thoracic Society in 2002.[6] There is an argument that some individuals are unable to walk for 6 minutes because of muscle weakness, gait inefficiency, or poor endurance.[7,8] For individuals with acquired brain injury or cognitive impairment, it may be a challenge to concentrate to complete the 6MWT.[9,10] Hence, the 2MWT may be a more feasible walk test for these individuals in clinical situations.

The quality of the information provided by the outcome measures depends on the psychometric properties of the outcome measures.[11] The main psychometric properties of an outcome

Disclosures: none.

0003-9993/14/$36 - see front matter © 2014 by the American Congress of Rehabilitation Medicine
http://dx.doi.org/10.1016/j.apmr.2014.03.034

measure are reliability (provides consistent data in repeated measurements), validity (measures what it is intended to measure), and responsiveness (detects changes over time).[12] Reliability refers to the consistency of the measurement of the 2MWT in the absence of real changes within the study participants (test-retest reliability) and among the assessors (intra- and interrater reliability).[12] Validity generally comprises content, construct, and criterion validity.[13] For a single-item walking test such as the 2MWT, the construct and criterion validity are more relevant.[5] Construct validity refers to how consistent the measurement of the 2MWT is when compared with predefined hypothesis testing on the score of another instrument or on differences among relevant population groups.[14] Criterion validity demonstrates whether the 2MWT is an adequate reflection of a criterion standard of the same construct.[14] Studies of the psychometric properties of walking tests, such as the 6MWT or 2MWT, traditionally considered maximal exercise tests using the progressive cycle ergometry as the criterion standard, although the walk tests are measuring submaximal functional capacity.[15] Responsiveness refers to the ability of the 2MWT to detect changes over time.[12]

To the author's knowledge, the psychometric properties of the 2MWT have not yet been systematically examined. This lack of knowledge about the psychometric evidence of the 2MWT affects patient management decisions if clinicians are unable to accurately interpret their patients' 2MWT findings.[11] The objective of this study was to systematically evaluate the psychometric properties of the 2MWT so that clinicians can make an informed decision about when and for whom to use the 2MWT and about how to accurately interpret the test's findings in clinical practice. The existing research gaps on the 2MWT were also identified.

Methods

Study search and selection

A selection of databases, including MEDLINE (via OVID), EBSCOHost (Academic Search Premier, CINAHL, SPORTDiscus, MEDLINE), PsycINFO, EMBASE (via OVID), the Cochrane Library, and DARE, was used for the literature search. These databases were chosen because they cover a variety of disciplines and integrated information from the fields of biomedical clinical practice and health. A validated search filter for searching studies on measurement properties was used.[16] The phrases "2-minute walk test," "2-minute walk distance," and the common abbreviations (ie, 2MWT and 2MWD) were also entered as keyword searches. Appropriate Boolean symbols and linking terms were used (appendix 1). The bibliographies of key articles were hand searched to ensure that relevant articles were not missed.

The inclusion criteria for the search of this systematic review were studies (1) in which the psychometric properties of the 2MWT were evaluated; (2) written as full reports; and (3) published in English language peer-reviewed journals. Studies were excluded if (1) data were self-reported, and (2) the 2MWT was considered as the criterion standard for another outcome measure of interest.

The titles and abstracts of articles identified in the initial search were first screened against inclusion and exclusion criteria. No authors of included and excluded studies were approached to investigate whether relevant data could be extracted.

Quality evaluation of study methodology and measurement properties

The methodological quality of all the included studies of the 2MWT was assessed using the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) checklist (http://www.cosmin.nl/cosmin_1_0.html). The COSMIN checklist was originally designed and validated to assess the methodological quality of studies on the measurement properties of health-related, patient-reported outcome measures,[12] but since its development, it has been used to evaluate the methodological quality of studies on performance-based outcome measures.[5,17,18] The checklist evaluates 9 measurement properties—internal consistency, reliability, measurement error, content validity, structural validity, hypothesis testing (construct validity), cross-cultural validity, criterion validity, and responsiveness—and 2 subchecklists to determine the interpretability and generalizability of the studies.[13] Each property is scored using a 4-point scale with defined response options grading the study as excellent, good, fair, or poor. The overall quality score is the lowest rating of any item within that measurement property (ie, worst score counts).[13] In each measurement property, there is an item on sample sizes in which 30 participants are considered to be minimally acceptable sample size. The COSMIN checklist was first developed for evaluating psychometric evidence on health-related questionnaires, which would require a larger sample size. This criterion may not apply for psychometric studies of performance-based outcome measures[5]; otherwise, studies with small sample sizes may be rated poor regardless of the good quality of their other methodological aspects. Hence, this criterion on sample size was omitted for assessing the methodological quality of the included studies. The issue of sample sizes would be considered in the best-evidence synthesis for the 2MWT. The quality of statistical findings of the measurement properties of included studies was also rated as positive, indeterminate, or negative[19] (definitions of each rating criterion in appendix 2).

Synthesis of best evidence for measurement properties of 2MWT

The levels of evidence for the overall quality of each measurement property of the 2MWT were synthesized as strong, moderate, limited, conflicting, or unknown by pulling all the data from the included studies[20] (definitions for each rating criterion in appendix 3). Similar to the levels of evidence of clinical trials, the synthesis of best evidence depends on the methodological quality of the included studies (ie, the COSMIN score) and the quality and consistency of the statistical findings of each measurement property (ie, positive, indeterminate, or negative), as well as the number of studies examining each measurement property of the 2MWT.

Selection of the studies, data retraction of the characteristics and main findings of the included studies, grading of the included studies using the COSMIN checklist, and best-evidence synthesis

List of abbreviations:

COPD	**chronic obstructive pulmonary disease**
COSMIN	**COnsensus-based Standards for the selection of health Measurement INstruments**
GVHD	**graft-versus-host disease**
MCID	**minimal clinically important difference**
6MWT	**6-minute walk test**
12MWT	**12-minute walk test**
2MWT	**2-minute walk test**
$\dot{V}O_2max$	**maximum oxygen consumption**

for the measurement properties of the 2MWT were all performed by the author.

Results

Study selection

The selection procedures are summarized in figure 1. Among the 26 included studies, 2 studies[7,21] were performed by the same authors. The study published in 2007[7] was a subset of the study published in 2006.[21] Hence, these 2 studies were treated as 1 single entity,[21] in which detailed data on the measurement properties of the 2MWT were provided. Twenty-five studies examining the psychometric properties of the 2MWT were therefore included in this review.

The characteristics and demographics of the 25 included studies are summarized in table 1. All patient groups were adults, except one consisting of children with cystic fibrosis.[43] There were 5 studies with participants having respiratory conditions (n=3 with chronic obstructive pulmonary disease [COPD],[22,27,35] n=1 cystic fibrosis,[43] n=1 respiratory conditions[3]). Five studies[23,24,31,36,38] consisted of participants with lower limb amputation, 1 study[37] with chronic graft-versus-host disease (GVHD), and 10 studies with neurologic conditions (n=2 poliomyelitis,[33,41] n=2 stroke,[32,34] n=3 multiple sclerosis,[28-30] n=1 neurologic impairment,[39] n=1 Parkinson's disease,[40] n=1 intellectual disability[10]). The remaining 4 studies were of participants who were receiving rehabilitation (n=2 geriatric conditions,[21,26] n=1 hip fractures,[42] n=1 cardiac surgery[25]). The volume of literature in each of the 14 patient groups was fairly limited, except in respiratory conditions and lower limb amputation with each having 5 studies.

Testing protocol

The testing protocols of the 2MWT in the included studies are summarized in table 2. Many of the included studies did not state important information on the testing protocol. In studies that have explicitly stated the testing protocols of the 2MWT, the testing instructions were a modified version from the 6MWT.[10] The testing procedures were varied and inconsistent among the studies, including use of a distance-measuring wheel,[21,24,26] position of the assessor (walking in front[10,27] or behind the participant[21,23-26,31]), use of encouragement (strictly no encouragement,[21,23-26,30,31] standardized encouragement,[10,27,42] nonstandardized encouragement[43]), and strictly no practice run before the actual testing[31,36] (see table 2). Four studies measured the distance covered in the first 2 minutes as the results of the 2MWT in a 6MWT[30,38] or 12MWT.[22,34]

Reliability

For reliability, there were 12 studies of 8 patient groups (lower limb amputations, COPD, cystic fibrosis, frail elderly patients, neurologic impairment, poliomyelitis, respiratory conditions, stroke) (table 3). The methodological quality of the studies on participants with amputations was good.[24,38] The quality of studies on COPD,[27,35] respiratory conditions,[3] and cystic fibrosis[45] was poor. The 2 studies of frail elderly patients[26] and adults with neurologic impairment[39] had good and poor methodological quality, respectively. The methodological quality of the studies[33,41] on patients with poliomyelitis was fair. The studies of stroke patients scored good[32] and fair[34] in methodological quality. The main methodological flaws for those studies with poor quality were inadequate statistical analyses[3,27,43] and an unclear number of assessors involved in the study.[3,27,35,39,43] Those studies that reported the statistical results of the reliability all reached a high level of reliability (intraclass correlation coefficient $\geq$0.8; range, .83−.99), receiving a positive rating on the quality of this measurement property (see table 3).

Measurement errors

Four studies examined the measurement errors of the 2MWT for participants with lower limb amputation,[38] poliomyelitis,[33,41] and stroke[32] (see table 3). The quality of the studies on participants with poliomyelitis was fair, but the other studies had good quality. The minimal detectable change in participants with lower limb amputation was 34.4m,[38] but 22.9m[41] and 13.4m[32] in participants with poliomyelitis and stroke, respectively. No study reported the minimal clinical importance difference (MCID), so the quality score of this measurement property remains indeterminate (see table 3).

Validity

There were 16 studies of 12 patient groups (GVHD, lower limb amputations, cardiac surgery, COPD, respiratory conditions, frail elderly patients, hip fractures, intellectual disability, multiple sclerosis, neurologic impairment, poliomyelitis, stroke) evaluating the validity of the 2MWT (table 4). Most of the studies examined the construct validity (hypothesis testing), and 2 studies[22,35] examined the criterion validity. These 2 studies had fair[22] and good[35] methodological quality. Good or above methodological quality for the construct validity was demonstrated in studies of participants undergoing cardiac rehabilitation,[25] geriatric rehabilitation,[21] and the study of patients with hip fractures.[42] Two[23,36] of 3 studies of patients with lower limb amputation scored good or above in methodological quality, but the other study[31] was fair. The study[34] on stroke patients scored fair in methodological quality. One study[35] of participants with COPD scored good, but the other 2 studies of COPD[22] and respiratory conditions[3] scored poor in methodological quality. The rest of the studies of participants with GVHD,[37] intellectual disability,[10] multiple sclerosis,[29,30] neurologic impairment,[39] and poliomyelitis[41] all scored poor in methodological quality. The main flaws of the methodological quality for the studies in evaluating the validity of the 2MWT were an inadequate description of the comparator instrument including its measurement properties,[3,29,39] and unclear hypotheses or no hypotheses formulated a priori regarding the expected correlations.[3,10,29,30,37,39] The quality of the measurement properties of most of the studies was positive. Two studies[31,37] received an indeterminate rating on the quality of the measurement properties, since no actual data on the correlations with related measure outcomes were provided (see table 4). Both studies examining the criterion validity of the 2MWT[22,35] had a negative rating for the measurement property, as the correlation coefficient did not reach over .70 (see table 4).

Responsiveness

Nine studies investigated the responsiveness of the 2MWT on 8 patient groups (lower limb amputation, cardiac rehabilitation, COPD, cystic fibrosis, geriatrics, multiple sclerosis, Parkinson's disease, stroke) (see table 4). Six studies[21,27,28,34,35,43] examined the responsiveness of the 2MWT before and after

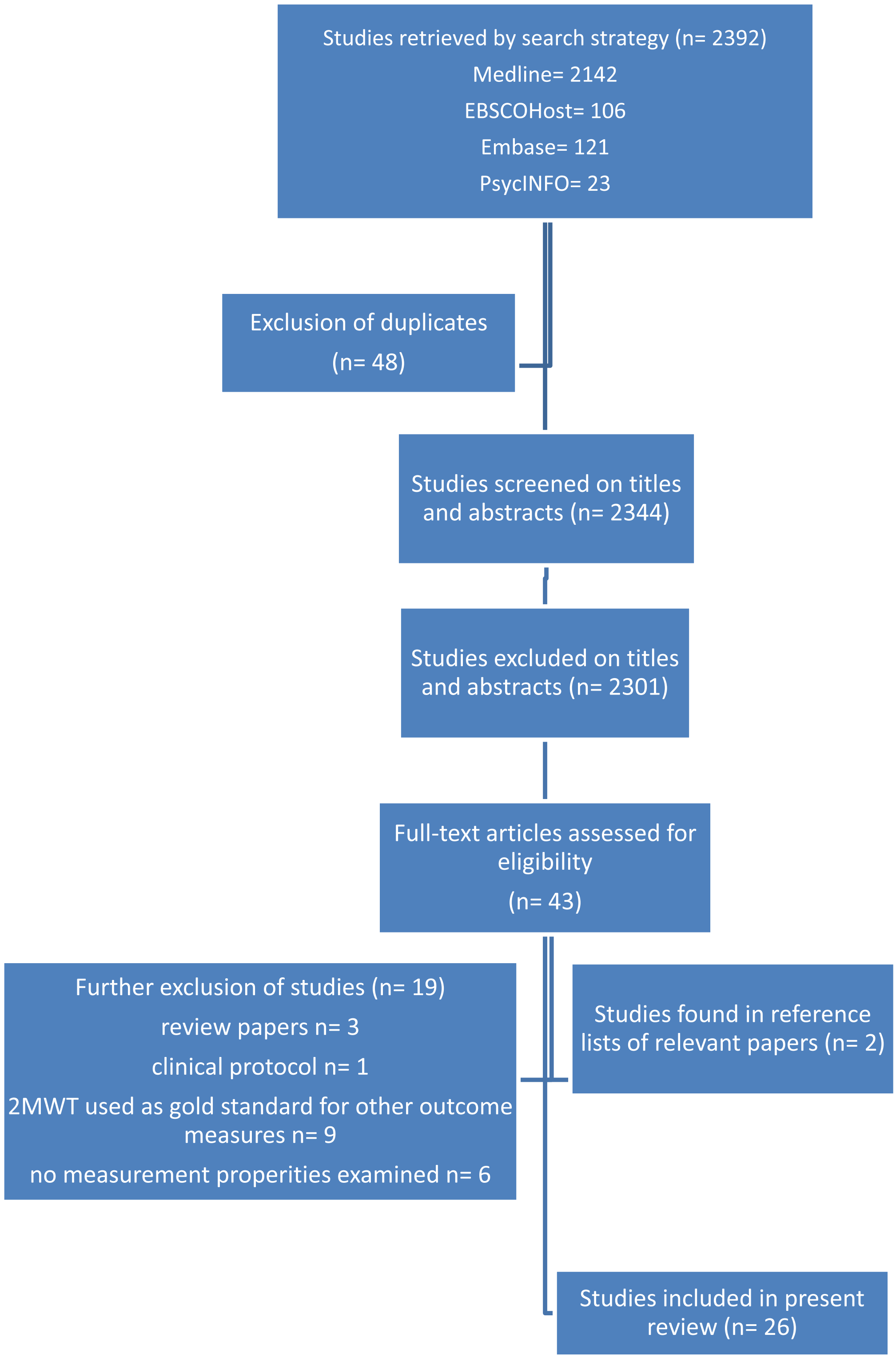

Fig 1 Flow diagram of search results and included studies.

intervention. Three studies[23,25,40] assessed whether the 2MWT could detect differences in participants with different severities of the condition. All studies scored poor in methodological quality, except the studies of patients with lower limb amputation (good),[23] frail elderly patients (good),[21] and patients with cardiac surgery (fair).[25] The main flaws in the methodological quality for this measurement property of the 2MWT included not using a longitudinal study design,[40] no or unclear hypothesis formulated a priori about the changed scores,[27,28,34,35,43] absence or inadequate information of the comparator instrument,[27,28,34,40,43] and inappropriate statistical analyses.[28,34,35,40] Hence those studies that had poor methodological quality also

Table 1 Characteristics of included studies

Study	N	Age (y)	Male/Female	Diagnosis	Testing Periods
Bernstein et al,[22] 1994	9	67±4 Range NA	9/0 (100/0)	COPD	Tested 6 times, 2–3wk apart
Brooks et al,[23] 2001	290	66.3±13.1 Range: 21–94	212/78 (73/17)	Lower limb amputation (n=179 transtibial, n=60 transfemoral, n=51 bilateral)	Baseline (fitting of prosthesis), within 48h before discharge, 3mo as FU
Brooks et al,[24] 2002	33	63.6±2.0 Range: 42–88	23/10 (70/30)	Transtibial amputation	2 consecutive days
Brooks et al,[25] 2004	122	63.4±8.6 Range NA	87/13 (87/13)	Postcardiac surgery	Preoperatively, 1d before discharge (postoperatively), 6–8wk FU
Brooks et al,[21] 2006	52	79.9±7.7 Range: 62–82	17/35 (33/67)	Frail elderly	On admission and last week of rehabilitation program before discharge
Butland et al,[3] 1982	30	61±12 Range NA	NA	Respiratory conditions (hypothesis testing)	One-off testing
Butland et al,[3] 1982	13	51±14 Range NA	NA	Respiratory conditions (reliability)	One-off testing
Connelly et al,[26] 1996	20	88.15±4.44 Range: 80–96	4/16 (20/80)	Frail healthy elderly	2 consecutive days
Eiser et al,[27] 2003	57	69±8 Range NA	30/27 (53/47)	COPD	3 consecutive weeks (reliability) Before and after use of bronchodilators in 2mo (responsiveness)
Filipovic et al,[28] 2011	49	35 (median, no SD given) Range: 18–56	10/39 (20.4/79.6)	Multiple sclerosis	Before and 31d after administration of IVMP
Gijbels et al,[29] 2011	40	48±10 Range NA	NA	Multiple sclerosis	One-off testing
Gijbels et al,[30] 2012	178	47±11 Range: 21–70	73/105 (41/59)	Multiple sclerosis	One-off testing
Gremeaux et al,[31] 2012	64	58±16 Range: 22–87	54/10 (84/16)	Lower limb amputation (n=47 transtibial, n=17 transfemoral)	One-off testing
Hiengkaew et al,[32] 2012	61	63.5±10 Range NA	43/18 (70/30)	Chronic stroke	2 testings, 5–10d apart
Horemans et al,[33] 2004	62	52±7 Range NA	23/40 (37/63)	Poliomyelitis	2 testings, 3wk apart
Kosak and Smith,[34] 2005	18	77±11 Range NA	6/12 (33/67)	Stroke	2 testings, 2d apart weekly for 4wk (intrarater reliability and responsiveness) 1 testing weekly for 4wk by another therapist (interrater reliability)
Leung et al,[35] 2006	45	71.8±8.3 Range NA	37/8 (82/18)	Stable COPD	One-off testing
Maring et al,[10] 2013	30	59.6±8.17 Range NA	NA	Elderly with intellectual disability	One-off testing
Parker et al,[36] 2010	52	55.2±15.8 Range: 20.1–88.7	41/11 (79/21)	Lower limb amputation (n=30 transtibial, n=16 transfemoral, n=6 bilateral)	One-off testing
Pidala et al,[37] 2012	584*	51.5 (median) Range: 2–79*	336/248 (58/42)*	GVHD	At enrollment and FUs (at least 3mo postenrollment)

(*continued on next page*)

Table 1 (*continued*)

Study	N	Age (y)	Male/Female	Diagnosis	Testing Periods
Resnik and Borgia,[38] 2011	44	66±13 Range: 31−85	42/2 (95/5)	Lower limb amputation (n=19 transtibial, n=2 through knee, n=23 transfemoral)	2 testings within 1wk
Rossier and Wade,[39] 2001	46	47±13 Range NA	28/18 (61/39)	Neurologic impairment	2 testings, 7d apart
Schenkman et al,[40] 2011	339	66.1±9.34 Range: 37−92	238/101 (70.2/29.8)	Parkinson's disease	One-off testing
Stolwijk-Swuste et al,[41] 2008	57	57.3±7.2 Range NA	21/36 (37/63)	Poliomyelitis	2 testings, 3wk apart
Unnanuntana et al,[42] 2012	162	66.69±9.8 Range NA	71/91 (44/56)	Waiting for total hip arthroplasty	One-off testing
Upton et al,[43] 1988	12	10 (NA) Range: 5−15	5/7 (42/58)	Stable cystic fibrosis (reliability)	In hospital and 3mo postdischarge (reliability)
Upton et al,[43] 1988	16	12.5 (NA) Range: 6−16	7/9 (44/56)	Cystic fibrosis with acute exacerbations (responsiveness)	Before, during, and after treatment (responsiveness)

NOTE. Values are mean ± SD, n (%), or as otherwise indicated.
Abbreviations: FU, follow-up; IVMP, intravenous methylprednisolone; NA, no information available.
* n=492 for 2MWT. No further breakdown on age, sex distribution of this subgroup. Data available for the whole cohort of the study (98% adults, 2% children).

had a negative rating on quality for this measurement property (see table 4).

Synthesis of best evidence

The best evidence for the measurement properties of the 2MWT is summarized in table 5. Findings from the studies[23,24,31,36,38] of adults with lower limb amputation were pooled together, since the clinical features of the patients were very similar and thus their performances in the 2MWT should be comparable. Similarly, the studies of adults with COPD[22,27,35] and respiratory conditions,[3] studies of frail elderly patients,[21,26] and studies of adults with multiple sclerosis[28-30] were pooled accordingly to synthesize evidence for the measurement properties of the 2MWT for the respective patient groups. However, the study of children with cystic fibrosis[43] was not pooled with the studies of adults with COPD or respiratory conditions, since children might show different performances in the 2MWT because of their developing cardiorespiratory system. There was moderate evidence for positive reliability (test-retest, intra- and interrater) of the 2MWT for adults with lower limb amputation, intra- and interrater reliability for frail elderly patients, and test-retest reliability for adults with poliomyelitis or stroke. The evidence was limited for positive intra- and interrater reliability of the 2MWT for adults with stroke. Unknown evidence was found for positive test-retest reliability for adults with COPD and respiratory conditions, adults with neurologic impairment, and children with cystic fibrosis. Some knowledge was available for the measurement errors of the 2MWT for adults with lower limb amputation, poliomyelitis, or stroke. Since no information is available on the MCID, according to the COSMIN checklist, the evidence for this measurement property of the 2MWT remains unknown (see tables 3 and 5).

There was strong evidence for the convergent validity of the 2MWT with disease-specific parameters in adults with lower limb amputation (see tables 4 and 5). Good evidence was demonstrated for positive convergent validity of the 2MWT with generic functional mobility measures for adults undergoing cardiac rehabilitation or geriatric rehabilitation, or adults after hip fractures. However, limited evidence was shown for the convergent validity of the 2MWT with the 6MWT and 12MWT for adults with stroke and with respiratory conditions, and so with general functional measures for adults with poliomyelitis (see tables 4 and 5). At present, there was unknown evidence to support positive convergent validity of the 2MWT with other walk tests for adults with multiple sclerosis, with disease-specific parameters for adults with GVHD, and with generic functional physical measures for adults with intellectual disability or neurologic impairment (see tables 4 and 5). There was moderate evidence in support of the ability of the 2MWT to discriminate frail elderly patients walking with and without aid, or the use of walking aids for patients with hip fractures (discriminate validity). Although the studies[22,35] on the criterion validity of the 2MWT had fair or above methodological quality, the positive correlation with maximum oxygen consumption ($\dot{V}O_2max$) for adults with COPD was low (r=.45−.55) (see tables 4 and 5), and 1 study[22] had a small sample size (n=9). Hence the evidence on this measurement property for this patient group remained limited.

Moderate evidence was demonstrated for positive responsiveness of the 2MWT across times (ie, before and after intervention for frail elderly patients and adults with lower limb amputation), but there was limited evidence for patients undergoing cardiac rehabilitation (see tables 4 and 5). However, some knowledge was available for the responsiveness of the 2MWT for adults with COPD or respiratory conditions, children with cystic fibrosis, adults with multiple sclerosis, adults with Parkinson's disease, and adults with stroke, but the evidence for this measurement property

Table 2 Testing protocols of the 2MWT in the included studies

	Environment	Instructions				Administration		
Study	Length of Walkway	Pretest Rest	Verbal Instructions	Practice Run	No. of Trials as Test Run	Accompanied by Assessor	Measurement of Distance Covered	Verbal Encouragement
Bernstein et al,[22] 1994	20ft	NA	Go as far as you can in 12min.*	3	1	No	NA	NA
Brooks et al,[23] 2001	NA	NA	Walk as far as you can in 2min.	2	1	Walked behind participants	NA	No
Brooks et al,[24] 2002	≥40m	NA	Walk as far as you can in 2min.	≥1 practice run, at least 1d before actual testing	1	Presumably walking behind participants	Calibrated wheel by assessor	No
Brooks et al,[25] 2004	NA	NA	Cover as much ground as possible in 2min.	2	1	Walked behind participants	NA	No
Brooks et al,[21] 2006	30m	30-min rest btw trials	Walk as far you can in 2min.	2	1	Presumably walking behind participants	Calibrated wheel by assessor	No
Butland et al,[3] 1982	NA	60-min rest btw trials (reliability)	NA	NA	4 (reliability)	NA	NA	NA
Connelly et al,[26] 1996	80m	NA	Walk as far as you can in 2min.	0	Mean of 3 trials used	Presumably walking behind participants	Precimeter by assessor	No
Eiser et al,[27] 2003	120m	30-min rest btw other tests	Walk as fast as possible.	NA	1	Walked in front of participants	NA	Yes, standardized encouragement
Filipovic et al,[28] 2011	NA	NA	NA	NA	NA	NA	NA	NA
Gijbels et al,[29] 2011	30m	NA	NA	NA	NA	NA	NA	NA
Gijbels et al,[30] 2012	30m	NA	Walk at fastest speed to cover as much distance as possible in 6min.†	NA	1	NA	NA	No
Gremeaux et al,[31] 2012	NA	NA	Cover greatest distance in 2min.	No	1	Walked behind participants	NA	No
Hiengkaew et al,[32] 2012	20m	3- to 5-min rest btw other tests	Walk as far as possible in 2min.	NA	1	NA	NA	NA
Horemans et al,[33] 2004	65m	5min	NA	NA	1	NA	NA	NA
Kosak and Smith,[34] 2005	122m rectangular hallway	NA. Therapist gave assistance if needed.	NA*	NA	NA	NA	NA	NA

(*continued on next page*)

Table 2 (*continued*)

Study	Environment	Instructions				Administration		
	Length of Walkway	Pretest Rest	Verbal Instructions	Practice Run	No. of Trials as Test Run	Accompanied by Assessor	Measurement of Distance Covered	Verbal Encouragement
Leung et al,[35] 2006	30m	20-min rest btw trials	Walk as far as possible in 2min.	1 practice run, 1d before actual testing	3	NA	NA	NA
Maring et al,[10] 2013	30m or 10–15m if outdoor corridor unavailable	≥10min	Walk as far as possible in 2min but do not run or jog.	1	1	Walked in front of participants Participants wearing gait belt for safety	NA	Yes, standardized encouragement given every 15–30s
Parker et al,[36] 2010	30m	NA	Walk as far as you can in 2min.	No	1	NA	NA	NA
Pidala et al,[37] 2012	25ft	NA	NA	NA	NA	NA	NA	NA
Resnik and Borgia,[38] 2011	≤30.5m	NA	Walk as far as you can in 6min.[†]	NA	1	NA	NA	NA
Rossier and Wade,[39] 2001	NA	NA	NA	NA	NA	NA	NA	NA
Schenkman et al,[40] 2011	NA	NA	Walk in 2min at fastest comfortable pace	2	1	NA	NA	NA
Stolwijk-Swuste et al,[41] 2008	NA	NA	NA	NA	NA	NA	NA	NA
Unnanuntana et al,[42] 2012	13.5m	NA	Walk at normal pace and turn around at ends of corridor without stopping.	NA	1	NA	NA	Yes, standardized encouragement
Upton et al,[43] 1988	35m	5- to 10-min rest btw trials	Walk as fast as they found comfortable, not to run.	NA	Greater distance of 2 trials used	NA	NA, taken nearest 5m	Yes

Abbreviations: btw, between; NA, information not available.

* Participants were tested with 2MWT, 6MWT, and 12MWT in one go; that is, distances covered at 2, 6, and 12 minutes were measured with 1 walk of 12 minutes.

† Participants were tested with 2MWT and 6MWT in one go; that is, distances covered at 2 and 6 minutes were measured with 1 walk of 6 minutes.

Table 3 Summary on measurement properties of the 2MWT (reliability and measurement errors)

Study	Reliability: Test-Retest	Reliability: Intrarater	Reliability: Interrater	COSMIN Score*/ Quality Score†	Measurement Error	COSMIN Score*/ Quality Score†
Amputation (lower limb)						
Brooks et al,[24] 2002		ICC=.90–.96	ICC=.98–.99	Good/+	NT	NA
Resnik and Borgia,[38] 2011	ICC=.83 (95% CI, .71–.90)	NT	NT	Good/+	MDC_{90}=34.4m	Good/?
COPD						
Eiser et al,[27] 2003	>95% between-subject variations, 5% intrasubject variations	NT	NT	Poor/−	NT	NA
Leung et al,[35] 2006	ICC>.99	NT	NT	Poor/+	NT	NA
Cystic fibrosis						
Upton et al,[43] 1988	Nonsignificant variations btw 1st and 2nd tests. No ICC or Pearson correlation given	NT	NT	Poor/−	NT	NA
Respiratory conditions						
Butland et al,[3] 1982	No statistical analysis results provided	NT	NT	Poor/−	NT	NA
Frail elderly						
Connelly et al,[26] 1996	NT	ICC=.93–.95	ICC=.82–.89	Good/+	NT	NA
Neurologic impairment						
Rossier and Wade,[39] 2001	ICC=.97	NT	NT	Poor/+	NT	NA
Poliomyelitis						
Horemans et al,[33] 2004	ICC=.94 (95% CI, .90–.96)	NT	NT	Fair/+	LoA=18.0m (15%)	Fair/?
Stolwijk-Swuste et al,[41] 2008	ICC=.93 (95% CI, .88–.96)	NT	NT	Fair/+	LoA=22.9m SDC=22.9m	Fair/?
Stroke						
Hiengkaew et al,[32] 2012	ICC=.98 (95% CI, .97–.99)	NT	NT	Good/+	MDC_{95}=13.4m or 23%	Good/?
Kosak and Smith,[34] 2005	NT	ICC=.85	ICC=.85	Fair/+	NT	NA

Abbreviations: btw, between; CI, confidence interval; ICC, intraclass correlation coefficient; LoA, limit of agreement; MDC_{90}, minimal detectable change at 90% CI; MDC_{95}, minimal detectable change at 95% CI; NA, not available; NT not tested; SDC, smallest detectable change.
* COSMIN score after removing the sample size item from the rating.
† Quality score of the measurement property: +, measurement property; −, no measurement property; ?, indeterminate.

of the 2MWT in these patient groups remained unknown because of the overall poor quality in methodology and inadequate statistical analyses for this measurement property in the relevant studies (see tables 4 and 5).

Discussion

Twenty-five studies were evaluated on both methodological quality and quality criteria for the measurement properties of the 2MWT using the COSMIN checklist, providing evidence for the psychometric properties of the 2MWT. The methodological quality of the included studies varied from poor to excellent (see tables 3 and 4). Based on the findings in this systematic review, it appears that the 2MWT is a reliable, valid, and responsive submaximal exercise test for adults with lower limb amputation and for frail elderly patients (see table 5).

Only 1 study[43] was found that examined the reliability and responsiveness of the 2MWT in children with cystic fibrosis, but it had poor methodological quality and a small sample size (n<30) (see table 1). The existence of only this single pediatric study indicates the paucity of research evidence on the psychometric properties of the 2MWT in children. The 2MWT, because of its short duration, has been proposed to be a more clinically feasible submaximal exercise test for individuals with poor concentration span or endurance,[9,10,25] such as children with acquired brain injury or moderately severe cerebral palsy. More research appears to be required before the 2MWT can be considered a reliable and valid outcome measure in pediatric patient groups in both clinical and research contexts.

The studies on hypothesis testing (construct validity) received low ratings on methodological quality, notably those studies of adults with COPD or respiratory conditions, adults with GVHD, and adults with various neurologic conditions (see table 4). The low ratings were mainly because few studies formulated clear hypotheses a priori on the expected correlation with other variables and mean differences between known groups. Thus, the construct validity of the 2MWT remains unclear for adults with respiratory conditions, GVHD, and neurologic conditions. Despite the shortcomings of the methodological quality of the studies, those studies examined the construct validity of the 2MWT with other walk tests such as the 6MWT, the 12MWT, or the 10-m walk test. The correlations were greater than .75, which is considered to

Table 4 Summary of measurement properties of the 2MWT (validity and responsiveness)

Study	Validity				Responsiveness	COSMIN Score*/ Quality Score†
	Criterion	COSMIN Score*/ Quality Score†	Construct (Hypotheses Testing)	COSMIN Score*/ Quality Score†		
Amputation (lower limb)						
Brooks et al,[23] 2001	NT	NA	r=.22–.48 with SF-36 PF r=.49 with Houghton Scale	Excellent/+	Significant increase in distances covered across 3 time points across 3 subgroups (P<.001)	Good/+
Gremeaux et al,[31] 2012	NT	NA	Significant correlation among the tests but no data given. Cutoff 130–150m for differentiating with/without functional limitations	Fair/?	NT	NA
Parker et al,[36] 2010	NT	NA	r=–.60 (n=52) with TAPES Activity Scale r=.45–.78 (n=27) with steps per day	Good/+	NT	NA
Cardiac surgery						
Brooks et al,[25] 2004	NT	NA	r=.44 (preoperative) and .48 (postoperative) with SF-36 PF r=.12 (preoperative) and –.03 (postoperative) with SF-36 SF r=.34 (preoperative) and .31 (postoperative) with NEADL Unable to show significant difference between adults with and without postoperative complications	Excellent/+	Significant changes for the distance covered over time (intake, discharge, and FU) (P<.0001)	Fair/+
COPD						
Bernstein et al,[22] 1994	r=.45 with $\dot{V}O_2$max r=.35 with $\dot{V}CO_2$max	Fair/–	r=.95 with 6MWT r=.94 with 12MWT	Poor/+	NT	NA
Eiser et al,[27] 2003	NT	NA	NT	NA	Significant increase in distance covered (P<.0001) after BD	Poor/–
Leung et al,[35] 2006	r=.454 with $\dot{V}O_2$max r=.555 with $\dot{V}O_2$max/kg	Good/–	r=.937 with 6MWT	Fair/+	Effect size=.61, SRM=1.25 postrehabilitation	Poor/–
Cystic fibrosis						
Upton et al,[43] 1988	NT	NA	NT	NA	Significant increase in distance covered posttreatment (P<.005)	Poor/–
Respiratory conditions						
Butland et al,[3] 1982	NT	NA	r=.96 with 6MWT r=.864 with 12MWT	Poor/+	NT	NA
Frail elderly						
Brooks et al,[21] 2006	NT	NA	r=.59 with FIM, .42 with MBI, and –.81 with TUG test at admission r=.42 with FIM, .35 with MBI, and	Good/+	Significant increase in distance covered postrehabilitation (P<.04)	Good/+

(*continued on next page*)

Table 4 (*continued*)

Study	Validity				Responsiveness	COSMIN Score*/ Quality Score†
	Criterion	COSMIN Score*/ Quality Score†	Construct (Hypotheses Testing)	COSMIN Score*/ Quality Score†		
			−.68 with TUG test at discharge Adults without walking aid walked further than with walker ($P\leq.04$).			
GVHD						
Pidala et al,[37] 2012	NT	NA	Significant association with Lee Chronic GVHD scale ($P<.001$), SF-36 RP/GH/VT ($P<.001$), FACT ($P<.001$), HAP score ($P<.001$), and NIH global score (patient-reported and clinician-reported chronic GVHD severity, $P\leq.006$). Inversely associated with hazard for death. Odds ratio of gastrointestinal (.997), liver, (.998) and lung (.998) involvement.	Poor/−	NT	NA
Hip fractures						
Unnanuntana et al,[42] 2012	NT	NA	$r=.41-.54$ with WOMAC $r=.35-.58$ with SF-36 PF $R^2=.452$ for use of walking aid and shorter distance covered	Good/+	NT	NA
Intellectual disability						
Maring et al,[10] 2013	NT	NA	$r=.748$ with MBI $r=.799$ with POMA I Adults with previous hospitalization covered less distance ($P=.002$).	Poor/+	NT	NA
Multiple sclerosis						
Filipovic et al,[28] 2011	NT	NA	NT	NA	Significant increase in distances covered after 1mo of IVMP ($P<.001$). Effect size=.54, SRM=.89, relative efficiency=95.1 (z=5.451) according to EDSS.	Poor/−
Gijbels et al,[29] 2011	NT	NA	Nonsignificant difference in distance covered in the 2MWT and the first 2min of the 6MWT. $R^2=.96$ with 6MWT	Poor/+	NT	NA
Gijbels et al,[30] 2012	NT	NA	Significant association with 6MWT (univariate regression coefficient=.97, mean absolute	Poor/+	NT	NA

(*continued on next page*)

Table 4 (*continued*)

Study	Validity: Criterion	COSMIN Score*/ Quality Score†	Validity: Construct (Hypotheses Testing)	COSMIN Score*/ Quality Score†	Responsiveness	COSMIN Score*/ Quality Score†
			estimation error=18 [95% CI, 15–21]), mean relative estimation error=5% Mildly affected adults walked further than moderately affected adults (P<.01).			
Neurologic impairment						
Rossier and Wade,[39] 2001	NT	NA	r=.75 with RMI r=.75 with 10-m timed walk Adults without mobility aid walked further than those with mobility aid (P<.001). Adults without leg sensory impairment walked further than those with sensory impairment (P<.001).	Poor/+	NT	NA
Parkinson's disease						
Schenkman et al,[40] 2011	NT	NA	NT	NA	Significant association with severity of the disease using H & Y Scale (F=24.79, P<.0001, Cohen F=.21) and to UPDRS motor score (F=4.94, P=.0282, Cohen F=.03)	Poor/−
Poliomyelitis						
Stolwijk-Swuste et al,[41] 2008	NT	NA	r=.69 with SF-36 PF r=.45 with WOMAC r=.61 with NHP	Poor/+	NT	NA
Stroke						
Kosak and Smith,[34] 2005	NT	NA	R=.997 with 6MWT R=.993 with 12MWT	Fair/+	SRM=1.34	Poor/−

Abbreviations: BD, bronchodilator; CI, confidence interval; EDSS, Expanded Disability Status Scale; FACT, Functional Assessment of Cancer Therapy; FU, follow-up; HAP, Human Activity Profile; H & Y, Hoehn and Yahr; IVMP, intravenous methylprednisolone; MBI, modified Barthel Index; NA, not available; NEADL, Nottingham Extended Activities of Daily Living; NHP, Nottingham Health Profile; NIH, National Institutes of Health; NT, not tested; POMA I, Performance-Oriented Mobility Assessment I; RMI, Rivermead Mobility Index; SF-36 PF, physical functioning subscale of the Medical Outcomes Study 36-Item Short-Form Health Survey; SF-36 RP/GH/VT, role physical, general health, and vitality subscale of the Medical Outcomes Study 36-Item Short-Form Health Survey; SF-36 SF, social functioning subscale of the Medical Outcomes Study 36-Item Short-Form Health Survey; SRM, standardized response mean; TAPES, Trinity Amputation and Prosthesis Experience Scales; TUG, Timed Up and Go; UPDRS, Unified Parkinson's Disease Rating Scale; $\dot{V}CO_2max$, maximum carbon dioxide consumption; $\dot{V}O_2max/kg$, maximum oxygen consumption per kilogram body weight; WOMAC, Western Ontario and McMaster Universities Osteoarthritis Index.

* COSMIN score after removing the sample size item from the rating.

† Quality score of the measurement property: +, measurement property; −, no measurement property; ?, indeterminate.

Table 5 Levels of evidence of the 2MWT

Patient Groups	Reliability			Measurement Error	Validity		Responsiveness
	Test-Retest	Intrarater	Interrater		Criterion	Construct	
Lower limb amputation	++	++	++	?	0	+++	++
Cardiac rehabilitation	0	0	0	0	0	++	+
COPD or respiratory conditions	?	0	0	0	+	++	?
Cystic fibrosis (children)	?	0	0	0	0	0	?
Frail elderly	0	++	++	0	0	++	++
GVHD	0	0	0	0	0	?	0
Hip fracture	0	0	0	0	0	++	0
Intellectual disability	0	0	0	0	0	?	0
Multiple sclerosis	0	0	0	0	0	?	?
Neurologic impairment	?	0	0	0	0	?	0
Parkinson's disease	0	0	0	0	0	0	?
Poliomyelitis	++	0	0	?	0	+	0
Stroke	++	+	+	?	0	+	?

Abbreviations: +++, strong evidence; ++, moderate evidence; +, limited evidence; ±, conflicting evidence; ?, unknown; 0, no information.

be a good correlation (see table 4).[44] The correlations with other generic mobility tests, such as the physical functioning of the Medical Outcomes Study 36-Item Short-Form Health Survey and modified Barthel Index, were moderate (see table 4).[44] These preliminary results are supportive of the 2MWT as a valid walk test for individuals with respiratory conditions, GVHD or neurologic conditions, but this speculation needs to be proven with more rigorous studies.

Both studies[22,35] used the progressive cycle ergometry as the criterion standard to examine the criterion validity of the 2MWT, resulting in fair correlation (r=.45−.55 with $\dot{V}O_2$max or $\dot{V}O_2$max/kg). It is questionable whether cycle ergometry, which assesses maximal exercise capacity (aerobic or endurance capacity to cycle using arms or legs), measures the same construct as the 2MWT, which assesses submaximal functional capacity (aerobic and anaerobic capacity during walking that requires brief and intense bouts of ambulation).[45,46] The difficulty of finding an appropriate criterion standard to examine the criterion validity in the construct of submaximal exercise capacity is also evident in walk tests of longer duration such as the 6MWT.[5] It is possible that like other functional performance tests, such as the timed Up and Go test (which assesses combined skills of mobility and balance),[4] examining the criterion validity of the 2MWT may not be possible. Based on the lack of evidence for this measurement property of the 2MWT, clinicians should use other outcome measures such as maximal cardiopulmonary exercise tests to assess the exercise fitness of their patients.

For clinicians, it is of paramount importance to use the 2MWT to assess the effectiveness of an intervention on their patients. Although a few studies calculated the minimal detectable change of the 2MWT in adults with lower limb amputation, poliomyelitis, and stroke (see table 3), the authors did not recommend an MCID for their specific population group. The evidence for this measurement property of the 2MWT remains unknown. An MCID needs to be established for the 2MWT so that clinicians are aware of the smallest change in distance covered in 2 minutes that signifies an important clinical change in the submaximal exercise capacity for individuals with different pathologic conditions.[47] More studies in this area are needed.

No standardized testing protocol is available for the 2MWT,[4] unlike the 6MWT.[6] Large variations in the testing protocols were found in the included studies (see table 2). Some factors such as use of verbal encouragement and practice runs before the actual testing may have affected the walking performance of the participants. A significant increase in walking distances has been shown in a group of adults with respiratory conditions, using the 6MWT, when verbal encouragement was given during the walk test.[48] It has been recommended that 2 practice walks be allowed before the walking performance is measured, to avoid a learning effect on the performance.[45] Based on the existing evidence on walk tests in general[45] and the standardized testing procedures for the 6MWT,[6] a testing protocol for the 2MWT is proposed in appendix 4. Furthermore, normative studies of the 2MWT for both adults and children are recommended to provide data for comparisons as well as to construct reference equations for this test, as is done for the 6MWT.[49,50]

Study limitations

There were limitations to this review. First, the selection and rating of the studies and data retraction were done by 1 person. Nevertheless, the studies were selected under clear and defined selection criteria, and the studies were rated using the stringent COSMIN checklist and quality criteria on measurement properties (see appendix 2), which would minimize the bias during the process. Second, only published studies in English language peer-reviewed journals were included in this review, raising the possibility of publication bias. Finally, the content validity of the COSMIN checklist is unknown when 1 of the criteria on sample sizes is removed during the rating process, as it was in this study.

Conclusions

This systematic review has provided evidence for reliability, validity (construct and criterion), and responsiveness of the 2MWT, which varied across different patient groups. The volume of literature on the psychometric properties of the 2MWT was fairly low except for patients with respiratory conditions and lower limb amputation. There is a significant paucity of psychometric evidence of the 2MWT in the pediatric population. There was moderate to strong evidence to support the 2MWT as a reliable, valid, and responsive outcome measures for adults with lower limb amputation and for frail elderly patients. Important psychometric information on the 2MWT such as minimal clinically important changes and normative data is still missing. At present, any changes in the 2MWT, whether across time points or after intervention, should be interpreted with caution. Studies to gain a consensus on the testing protocols of the 2MWT are also needed.

Keywords

Exercise test; Psychometrics; Rehabilitation; Reliability and validity; Reproducibility of results

Corresponding author

Tamis W. Pin, PhD, Department of Rehabilitation Sciences, Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong. *E-mail address:* tamis.pin@polyu.edu.hk.

Appendix 1 Search Strategies

MEDLINE/EMBASE via OvidSP

- *Construct search:* [Gait (MeSH) OR gait (mp) OR walk*(mp) OR Walking (MeSH) OR walking (mp) OR ambulat*(mp) OR mobility (mp)] AND [minute* (mp) OR metre*(mp) OR meter* (mp) OR two-min* (mp) OR 2-min* (mp) OR 2MWT (mp) OR 2MWD (mp)]
- *Measurement search:* test* (mp) OR instrument* (mp) OR performance-base* (mp) OR method* (mp) OR measur* (mp) OR objective (mp) OR assess* (mp) OR observat* (mp) OR scale* (mp) OR function* (mp) OR disabilit* (mp) OR "Outcome Assessment (Health Care)"(MeSH) OR outcome* (mp) OR investigat* (mp) OR examin* (mp) OR index* (mp) OR indice* (mp)
- *Measurement properties search:* instrumentation (sh) OR validation studies (pt) OR "Reproducibility of Results" (MeSH) OR reproducib* (mp) OR Psychometrics (MeSH) OR psychometr* (mp) OR clinimetr* (mp) OR clinometr* (mp) OR Observer Variation (MeSH) OR observer variation (mp) OR Discriminant Analysis (MeSH) OR reliab* (mp) OR valid* (mp) OR coefficient (mp) OR internal consistency (mp) OR (cronbach* (mp) AND [alpha (mp) OR alphas (mp)]) OR item correlation (mp) OR item correlations (mp) OR item selection (mp) OR item selections (mp) OR item reduction (mp) OR item reductions (mp) OR agreement (tw) OR precision (tw) OR imprecision (tw) OR precise values (tw) OR test-retest (mp) OR [test (mp) AND retest (mp)] OR (reliab* (mp) AND [test (mp) AND retest (mp)]) OR stability (mp) OR interrater (mp) OR inter-rater (mp) OR intrarater (mp) OR intra-rater (mp) OR intertester (mp) OR inter-tester (mp) OR intratester (mp) OR intra-tester (mp) OR interobserver (mp) OR inter-observer (mp) OR intraobserver (mp) OR intra-observer (mp) OR intertechnician (mp) OR inter-technician (mp) OR intratechnician (mp) OR intra-technician (mp) OR interexaminer (mp) OR inter-examiner (mp) OR intraexaminer (mp) OR intra-examiner (mp) OR interassay (mp) OR inter-essay (mp) OR intraassay (mp) OR intra-assay (mp) OR interindividual (mp) OR inter-individual (mp) OR intraindividual (mp) OR intra-individual (mp) OR inter-participant (mp) OR inter-participant(mp) OR intra-participant (mp) OR intraparticipant (mp) OR kappa (mp) OR kappa's (mp) OR kappas (mp) OR coefficient of variation (mp) OR "repeatab*" (tw) OR ["replicab*" (tw) OR repeated (tw)] AND [measure (tw) OR measures (tw) OR findings (tw) OR result (tw) OR results (tw) OR test (tw) OR tests (tw)] OR generaliza* (mp) OR generalisa* (mp) OR concordance (mp) OR intraclass (mp) OR correlation* (mp) OR discriminative (mp) OR known group (mp) OR factor analysis (mp) OR factor analyses (mp) OR factor structure (mp) OR factor structures (mp) OR dimensionality (mp) OR subscale* (mp) OR multitrait scaling analysis (mp) OR multitrait scaling analyses (mp) OR item discriminant (mp) OR interscale correlation (mp) OR interscale correlations (mp) OR [error (mp) OR errors (mp)] AND measure* (mp) OR correlat* (mp) OR evaluat* (mp) OR accuracy (mp) OR accurate (mp) OR precision (mp) OR mean (mp) OR individual variability (mp) OR interval variability (mp) OR rate variability (mp) OR variability analysis (mp) OR (uncertainty (mp) AND [measurement (mp) OR measuring (mp)]) OR standard error of measurement (mp) OR sensitiv* (mp) OR responsive* (mp) OR [limit (mp) AND detection (mp)] OR minimal detectable concentration (mp) OR interpretab* (mp) OR (small* (mp) AND [real (mp) OR detectable (mp)] AND [change (mp) OR difference (mp)] OR meaningful change (mp) OR minimal important change (mp) OR minimal important difference (mp) OR minimally important change (mp) OR minimally important difference (mp) OR minimal detectable change (mp) OR minimal detectable difference (mp) OR minimal real change (mp) OR minimal real difference (mp) OR minimally real change (mp) OR minimally real difference (mp) OR ceiling effect (mp) OR floor effect(mp) OR item response model (mp) OR IRT (mp) OR Rasch (mp) OR differential item functioning (mp) OR DIF (mp) OR computer adaptive testing (mp) OR item bank (mp) OR cross cultural equivalence (mp)
- *Exclusion filter:* addresses (pt) OR biography (pt) OR comment (pt) OR editorial (pt) OR festschrift (pt) OR interview (pt) OR lectures (pt) OR legal cases (pt) OR legislation (pt) OR letter (pt) OR case reports (pt) OR directory (pt) OR news (pt) OR newspaper article (pt) OR patient education handout (pt) OR popular works (pt) OR congresses (pt) OR consensus development conference (pt) OR consensus development conference, nih (pt) OR practice guideline (pt) NOT [Animals (MeSH) NOT Humans (MeSH)]

EBSCOHost (CINAHL, Academic Search Premier, MEDLINE, SPORTDiscus) and PsycInfo

- *Construct search:* two min* walk* test* (mp) OR 2 min* walk* test* (mp) OR 2-minute walk* test* (mp) OR two-minute walk* test* (mp) OR two-minute walk* distance* (mp) OR 2-minute walk* distance* (mp) OR 2 min* walk* distance* (mp) OR two min* walk* distance* (mp) OR 2MWT (mp) OR 2MWD (mp)
- *Measurement properties search:* Psychometrics (MeSH) OR reliability (MeSH) OR validity (MeSH) OR responsiveness (MeSH) OR repeatab* (mp) OR variab* (mp) OR reproduc* (mp)

Appendix 2 Quality Criteria for Rating Measurement Properties*

Measurement Property	Rating	Quality Criteria
Reliability		
Internal consistency	+	Cronbach α between .70 and .95 OR KR-20 between .70 and .90
	−	Cronbach α<.70 OR KR-20<.70
	?	Cronbach α not reported
Reliability	+	ICC>.70 OR weighted κ>.70 OR Pearson $r \geq .80$
	−	ICC≤.70 OR weighted $\kappa \leq .70$ OR Pearson r<.80
	?	Neither ICC, weighted κ, nor Pearson r determined
Measurement error	+	MIC>SDC OR MID>SDC OR MIC outside LoA
	−	MIC≤SDC OR MID≤SDC OR MIC equals or inside LoA
	?	MIC not defined
Validity		
Content validity	+	Target population considers all items in the questionnaire to be relevant AND considers the questionnaire to be complete.
	−	Target population considers items in the questionnaire to be irrelevant OR considers the questionnaire to be incomplete.
	?	No target population involved
Structural validity	+	Factors should explain ≥50% of the variance.
	−	Factors explain <50% of the variance.
	?	Explained variance not mentioned
Construct validity/ hypothesis testing	+	(Correlation with an instrument assessing the same construct ≥.50 OR ≥75% of the results were in accordance with the hypotheses) AND correlation with related constructs was higher than with unrelated constructs.
	−	Correlation with an instrument assessing the same construct <.50 OR <75% of the results were in accordance with the hypotheses OR correlation with related constructs was lower than with unrelated constructs.
	?	Sole correlations determined with unrelated constructs
Cross-cultural validity	+	(Original factor structure confirmed OR no important differential item functioning between language versions) AND the correlation between the translated or culturally adapted version and the original version was ≥.70.
	−	Original factor structure not confirmed OR important differential item functioning found between language versions OR the correlation between the translated or culturally adapted version and the original version was <.70.
	?	Confirmatory factor analysis not applied AND differential item functioning not assessed
Criterion validity (predictive or concurrent)	+	Correlation with standard was ≥.70 OR AUC≥.70 OR no statistically significant differences between the walking test and the criterion standard were found OR sensitivity and specificity ≥.70
	−	Correlation with standard was <.70 OR AUC<.70 OR no statistically significant differences between the walking test and the criterion standard were found OR sensitivity and specificity <.70
	?	No convincing arguments that criterion standard is actually the best standard OR doubtful design or method
Responsiveness	+	(Correlation with an instrument assessing the same construct ≥.50 OR ≥75% of the results were in accordance with the hypotheses OR AUC≥.70 OR sensitivity and specificity ≥.70) AND correlation with related constructs was higher than with unrelated constructs.
	−	Correlation with an instrument assessing the same construct <.50 OR <75% of the results were in accordance with the hypotheses OR AUC<.70 OR sensitivity and specificity <.70 OR correlation with related constructs was higher than with unrelated constructs.
	?	Sole correlations determined with unrelated constructs
Floor or ceiling effects	No	≤15% achieved the highest OR lowest possible scores.
	Yes	≥15% achieved the highest OR lowest possible scores.

Abbreviations: AUC, area under the receiver operating characteristics curve; ICC, intraclass correlation coefficient; KR-20, Kuder-Richardson formula(s); LoA, limit of agreement; MIC, minimal important change; MID, minimal important difference; SDC, significant detectable change; +, measurement property evident; −, no measurement property evident; ?, indeterminate.

* Adapted from Terwee et al,[19] 2007.

Appendix 3 Levels of Evidence for the Overall Quality of the Measurement Properties*

Level	Rating	Criteria
Strong	+++ or −−−	Consistent findings in multiple studies of good methodological quality OR in 1 study of excellent methodological quality
Moderate	++ or −−	Consistent findings in multiple studies of fair methodological quality OR in 1 study of good methodological quality
Limited	+ or −	One study of fair methodological quality
Conflicting	±	Conflicting findings
Unknown	?	Only studies of poor methodological quality

Abbreviations: +, positive result; −, negative result.
* Adapted from Schellingerhout et al[20] (2012) with kind permission from Springer Science+Business Media.

Appendix 4 Proposed 2MWT Testing Procedure

Location:

- The preferred walking course is a 30-m-long, flat, straight enclosed indoor corridor with a hard surface.
- The turnaround points should be clearly marked with a cone.
- A starting line should be visible on the floor with brightly colored tape.
- If the weather permits, and if a 30-m walkway is not available, the test may be performed outdoors.
- Length of the walkway and number of turns the subject must make should be recorded.

Preparation:

- If repeated testing is required, it should be performed about the same time of the day to minimize within-day variability.
- Patients should sit at rest in a chair for at least 10 minutes before the test. During this time, check for contraindications and measure blood pressure and pulse rate.
- Patients should wear appropriate shoes for walking.
- Patients should use their usual walking aids during the test (cane, walker, etc).
- Patients' usual medical regimen should be continued.
- A light meal is acceptable before early morning or early afternoon tests.
- Patients should not have exercised vigorously within 2 hours of beginning the test.

During testing:

- A pacer could be used during the 2MWT to ensure maximal effort from the patient. A different individual may act as the timer, or the pacer may act as the timer.
- The pacer, if used, should walk half a meter behind patients so as not to disturb their walking pace.
- No encouragement should be given to the patient, and the patient is not encouraged to talk during the test.
- Two trials are given as practice runs for the patient, and the performance of the third trial is taken as the actual measurement. Rest of at least 10 minutes is given to the patients between each trial to avoid fatigue.
- Distance traveled during the walk test can be recorded using markings on the wall or on the floor, or using a distance-measuring wheel by the pacer.

Instructions to the patient:

"The purpose of this test is to find out how far you can walk in 2 minutes. You will start from this point and follow the corridor/path to the cone. You should pivot briskly around the cone like this (demonstrate to the patient how to go around the cone briskly) and continue back the other way without stopping. You will walk back and forth between the 2 cones. Don't run or jog. When the 2 minutes are up, I will say 'STOP.' I want you to stop where you are. If you become too short of breath or tired during the test to continue, you can stop at any time. When you feel more comfortable, you may start walking again. I will walk behind you because I don't want to influence the pace at which you are walking. You should not talk during the test, but I do want you to tell me if you develop any chest pain or tightness or if you become dizzy or light-headed during the test. Do you have any questions? Are you ready? Please begin when I say 'GO.'"

At the completion of the 2MWT:

- Blood pressure, pulse rate, and distance walked are recorded.

References

1. Ohtake PJ. Field tests of aerobic capacity for children and older adults. Cardiopulm Phys Ther J 2005;16:5-11.
2. Noonan V, Dean E. Submaximal exercise testing: clinical application and interpretation. Phys Ther 2000;80:782-807.
3. Butland RJA, Pang J, Gross ER, Woodcock AA, Geddes DM. Two-, six- and 12-minutes walking tests in respiratory disease. BMJ 1982; 284:1607-8.
4. Finch E, Brooks D, Stratford PW, Mayo NE. Physical rehabilitation outcome measures. A guide to enhanced clinical decision making. 2nd ed. Ontario: BC Decker; 2002.
5. Bartels B, de Groot JF, Terwee CB. The six-minute walk test in chronic pediatric conditions: a systematic review of measurement properties. Phys Ther 2013;93:529-41.
6. ATS Committee on Proficiency Standards for Clinical Pulmonary Function Laboratories. ATS statement: guidelines for the six-minute walk test. Am J Respir Crit Care Med 2002;166:111-7.
7. Brooks D, Davis AM, Naglie G. The feasibility of six-minute and two-minute walk tests in in-patient geriatric rehabilitation. Can J Aging 2007;26:159-62.
8. Sibley KM, Tang A, Patterson KK, Brooks D, McIlroy WE. Changes in spatiotemporal gait variables over time during a test of functional capacity after stroke. J Neuroeng Rehabil 2009;6:27.
9. Katz-Leurer M, Rotem H, Keren O, Meyer S. The relationship between step variability, muscle strength and functional walking

performance in children with post-traumatic brain injury. Gait Posture 2009;29:154-7.
10. Maring JR, Costello E, Birkmeier MC, Richards M, Alexander LM. Validating functional measures of physical ability for aging people with intellectual development disability. Am J Intellect Dev Disabil 2013;118:124-40.
11. Roach KE. Measurement of health outcomes: reliability, validity and responsiveness. J Prosthet Orthot 2006;18:8-12.
12. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol 2010;63:737-45.
13. Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res 2012;21:651-7.
14. Mokkink LB, Terwee CB, Knol DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. BMC Med Res Methodol 2010;10:22.
15. Takken T, Engelbert R, van Bergen M, et al. Six-minute walking test in children with ESRD: discrimination validity and construct validity. Pediatr Nephrol 2009;24:2217-23.
16. Terwee CB, Jansma EP, Riphangen II, de Vet HC. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. Qual Life Res 2009;18:1115-23.
17. Dobson F, Hinman RS, Hall M, Terwee CB, Roos EM, Bennell KL. Measurement properties of performance-based measures to assess physical function in hip and knee osteoarthritis: a systematic review. Osteoarthritis Cartilage 2012;20:1548-62.
18. van Bloemendaal M, van de Water ATM, van de Port IGL. Walking tests for stroke survivors: a systematic review of their measurement properties. Disabil Rehabil 2012;34:2207-21.
19. Terwee CB, Bot SDM, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007;60:34-42.
20. Schellingerhout J, Verhagen A, Heymans M, Koes B, Vet H, Terwee C. Measurement properties of disease-specific questionnaires in patients with neck pain: a systematic review. Qual Life Res 2012; 21:659-70.
21. Brooks D, Davis AM, Naglie G. Validity of 3 physical performance measures in inpatient geriatric rehabilitation. Arch Phys Med Rehabil 2006;87:105-10.
22. Bernstein ML, Despars JA, Singh NP, Avalos K, Stansbury DW, Light RW. Reanalysis of the 12-minute walk in patients with chronic obstructive pulmonary disease. Chest 1994;105:163-7.
23. Brooks D, Parsons J, Hunter JP, Devlin M, Walker J. The 2-minute walk test as a measure of functional improvement in persons with lower limb amputation. Arch Phys Med Rehabil 2001;82:1478-83.
24. Brooks D, Hunter JP, Parsons J, Livsey E, Quirt J, Devlin M. Reliability of the two-minute walk test in individuals with transtibial amputation. Arch Phys Med Rehabil 2002;83:1562-5.
25. Brooks D, Parsons J, Tran D, et al. The two-minute walk test as a measure of functional capacity in cardiac surgery patients. Arch Phys Med Rehabil 2004;85:1525-30.
26. Connelly DM, Stevenson TJ, Vandervoort AA. Between- and within-rater reliability of walking tests in a frail elderly population. Physiother Can 1996;49:47-51.
27. Eiser N, Willsher D, Dore CJ. Reliability, repeatability and sensitivity to change of externally and self-paced walking tests in COPD patients. Respir Med 2003;97:407-14.
28. Filipovic Grcic P, Matijaca M, Lusic I, Capkun V. Responsiveness of walking-based outcome measures after multiple sclerosis relapses following steroid pulses. Med Sci Monit 2011;17:CR704-10.
29. Gijbels D, Eijnde BO, Feys P. Comparison of the 2- and 6-minute walk test in multiple sclerosis. Mult Scler 2011;17:1269-72.
30. Gijbels D, Dalgas U, Romberg A, et al. Which walking capacity tests to use in multiple sclerosis? A multicentre study providing the basis for a core set. Mult Scler 2012;18:364-71.
31. Gremeaux V, Damak S, Troisgros O, et al. Selecting a test for the clinical assessment of balance and walking capacity at the definitive fitting state after unilateral amputation: a comparative study. Prosthet Orthot Int 2012;36:415-22.
32. Hiengkaew V, Jitaree K, Chaiyawat P. Minimal detectable changes of the Berg Balance Scale, Fugl-Meyer Assessment Scale, Timed "Up & Go" Test, gait speeds, and 2-minute walk test in individuals with chronic stroke with different degrees of ankle plantarflexor tone. Arch Phys Med Rehabil 2012;93:1201-8.
33. Horemans HL, Beelen A, Nollet F, Lankhorst GJ. Reproducibility of walking at self-preferred and maximal speed in patients with post-poliomyelitis syndrome. Arch Phys Med Rehabil 2004;85:1929-32.
34. Kosak M, Smith T. Comparison of the 2-, 6-, and 12-minute walk tests in patients with stroke. J Rehabil Res Dev 2005;42:103-7.
35. Leung ASY, Chan KK, Sykes K, Chan KS. Reliability, validity, and responsiveness of a 2-min walk test to assess exercise capacity of COPD patients. Chest 2006;130:119-25.
36. Parker K, Kirby RL, Adderson J, Thompson K. Ambulation of people with lower-limb amputations: relationship between capacity and performance measures. Arch Phys Med Rehabil 2010;91:543-9.
37. Pidala J, Vogelsang G, Martin P, et al. Overlap subtype of chronic graft-versus-host disease is associated with an adverse prognosis, functional impairment, and inferior patient-reported outcomes: a Chronic Graft-versus-Host Disease Consortium study. Haematologica 2012;97:451-8.
38. Resnik L, Borgia M. Reliability of outcome measures for people with lower-limb amputations: distinguishing true change from statistical error. Phys Ther 2011;91:555-65.
39. Rossier P, Wade DT. Validity and reliability comparison of 4 mobility measures in patients presenting with neurologic impairment. Arch Phys Med Rehabil 2001;82:9-13.
40. Schenkman M, Ellis T, Christiansen C, et al. Profile of functional limitations and task performance among people with early- and middle-stage Parkinson disease. Phys Ther 2011;91:1339-54.
41. Stolwijk-Swuste JM, Beelen A, Lankhorst GJ, Nollet F; CARPA study group. SF36 physical functioning scale and 2-minute walk test advocated as core qualifiers to evaluate physical functioning in patients with late-onset sequelae of poliomyelitis. J Rehabil Med 2008; 40:387-94.
42. Unnanuntana A, Mait JE, Shaffer AD, Lane JM, Mancuso CA. Performance-based tests and self-reported questionnaires provide distinct information for the preoperative evaluation of total hip arthroplasty patients. J Arthroplasty 2012;27:770-775.e1.
43. Upton CJ, Tyrrell JC, Hiller EJ. Two minute walking distance in cystic fibrosis. Arch Dis Child 1988;63:1444-8.
44. Portney LG, Watkins MP. Foundations of clinical research. Applications to practice. 2nd ed. Englewood Cliffs: Pearson Prentice Hall; 2009.
45. Solway S, Brooks D, Lacasse Y, Thomas S. A qualitative systematic overview of the measurement properties of functional walk tests used in the cardiorespiratory domain. Chest 2001;119:256-70.
46. Motl RW, Suh Y, Balantrapu S, et al. Evidence for the different physiological significance of the 6- and 2-minute walk tests in multiple sclerosis. BMC Neurol 2012;12:6.
47. Terwee CB, Roorda LD, Knol DL, De Boer MR, de Vet HCW. Linking measurement error to minimal important change of patient-reported outcomes. J Clin Epidemiol 2009;62:1062-7.
48. Guyatt GH, Pugsley SO, Sullivan MJ, et al. Effect of encouragement on walking test performance. Thorax 1984;39:818-22.
49. Lammers AE, Hislop AA, Flynn Y, Haworth SG. The 6-minute walk test: normal values for children of 4-11 years of age. Arch Dis Child 2008;93:464-8.
50. Li AM, Yin J, Au JT, et al. Standard reference for the six-minute-walk test in healthy children aged 7 to 16 years. Am J Respir Crit Care Med 2007;176:174-80.