Age and Ageing 2012; **41:** 712–721
doi: 10.1093/ageing/afs099
Published electronically 10 August 2012

SYSTEMATIC REVIEW

Performance-based physical function in older community-dwelling persons: a systematic review of instruments

Ellen Freiberger[1], Paul de Vreede[2], Daniel Schoene[3], Elisabeth Rydwik[4], Volker Mueller[5], Kerstin Frändin[6], Marijke Hopman-Rock[7]

[1]Institute of Sport Science and Sports, Friedrich-Alexander-Universität, Gebbertstr. 123b, Erlangen-Nürnberg, Erlangen 91058, Germany
[2]Department of Public Health, Erasmus MC, University Medical Center, Office Ae-204, PO Box 2040, 3000 CA, Rotterdam, The Netherlands
[3]Neuroscience Research Australia, Falls and Balance Research Group, Sydney, Australia
[4]Research and Development Unit for the Elderly, Jakobsbergs Hospital, Karolinska Institute, Järfälla, Sweden
[5]Technical Library Friedrich-Alexander-Universität, Erlangen-Nürnberg, Erlangen, Germany
[6]Division of Physiotherapy, Department of Neurobiology, Care Sciences and Society, Karolinska Institute, Huddinge, Sweden
[7]Body@work, Research Center, TNO VU University Medical Center, Amsterdam, The Netherlands

Address correspondence to: E. Freiberger. Tel: (+49) 9131 8525464; Fax: (+49) 9131 8525002. Email: ellen.freiberger@sport.uni-erlangen.de/P. d. Vreede. Tel: (+31) 107043722; Fax: (+31) 107038460. Email: p.devreede@erasmusmc.nl

Abstract

Background: Identification of older persons at risk for the loss of independence, onset of (co)-morbidity or functional limitations through screening/assessment is of interest for the public health-care system. To date several different measurement instruments for overall physical function are frequently used in practice, but little information about their psychometric properties is available.
Objectives and Methods: Our aim was to assess instruments with an overall score related to functional status and/or physical performance on content and psychometric properties. Electronic databases (Medline, EMBASE, AMED, Cochrane Library and CINAHL) were searched, using MeSH terms and relevant keywords. Studies, published in English, were included if their primary or secondary purpose was to evaluate the measurement properties of measurement instruments for overall physical function in community-dwelling older persons aged 60 years and older. Reliability, validity, responsiveness and practicability were evaluated, adhering to a specified protocol.
Results: In total 78 articles describing 12 different functional assessment instruments were included and data extracted. Seven instruments, including their modified versions, were evaluated for reliability. Nine instruments, including their modified versions, were evaluated with regard to validity.
Conclusion: In conclusion, the Short Physical Performance Battery can be recommended most highly in terms of validity, reliability and responsiveness, followed by the Physical Performance Test and Continuous Scale Physical Functional Performance.

Keywords: *ageing, physical function, assessment, recommendation, older people*

Introduction

In community-dwelling older persons screening and assessment to detect early onset of functional decline or disability is a key factor [1–3]. This measurement of physical function gives guidance for geriatric treatment, and/or provides the base for the evaluation of the effectiveness of the treatment [1, 4]. Measures of functional status in older persons should

provide meaningful gradations on a continuum from vigorous to frail [1, 5]. Therefore, the classification and measurement of function in older persons affect clinical practice, as well as health-care systems, researchers and policy-makers [4, 6, 7]. The choice of the appropriate measurement instrument depends on the constructs being measured, ecological aspects of the instrument and their psychometric properties [8, 9].

Physical performance is commonly understood as the observable ability to perform tasks, e.g. chair rise [10–13]. Although there is no common definition of physical function, the term *functioning* is addressed by the WHO's International Classification of Functioning, Disability and Health (ICF) [14]. In the ICF model, the assessment of *physical function* seems crucial for evaluating aspects of health as well as the pathway to disability [7].

We use the term 'overall physical function' to address the measurement of different physiological domains which then generate an overall score. Because this score is easy to interpret in clinical practice, it can help identify people at risk, preferential at early stages of functional decline. Furthermore, this score is more robust than a one-item measurement and can be used to establish preventive strategies derived from different functional domains. The importance of a multi-dimensional measurement of physical function in older persons has been acknowledged in current primary care guidelines, thus making it mandatory to investigate the psychometric properties of these instruments [8, 15, 16].

To our knowledge at present no information on psychometric properties comparing different instruments on overall physical functioning in older adults is available. Therefore, the aim of this study was to conduct a systematic review of the psychometric properties of objective assessment instruments for performance-based overall physical function in older community-dwelling populations. A second aim was, to provide recommendations in practice for researchers, clinicians and health-care professionals. This systematic review is part of a series of systematic reviews initiated by the European Network for Action on Ageing and Physical Activity (EUNAAPA; http://www.eunaapa.org/) [16, 17, 18].

Methods

This systematic review adhered to a pre-specified protocol regarding search strategy and inclusion and exclusion criteria by the EUNAAPA review group based on a checklist for reliability, validity, responsiveness and practicability issues [9, 16, 17, 18, 19].

Search strategy

Electronic databases (Medline, EMBASE, AMED, Cochrane Library and CINAHL) were searched from their inception to March 2008, with an update in April 2010. Using MeSH terms and relevant keywords five semantic categories were entered: 'test battery', 'functional performance', 'reproducibility', 'age' (mean age ≥ 65 years) and 'setting' (community dwelling). Reference lists of review articles and included papers were scanned to identify further potential studies. The search was restricted to English language and peer-reviewed journal articles only.

Eligibility and exclusion criteria

To be included a study had to (i) investigate at least one of the mentioned psychometric properties of an overall index instrument; (ii) measure performance-based physical function or performance, providing an overall score; (iii) include a population 60 years of age or older or with a mean age above 65 years; (iv) address community-dwelling older persons; (v) have a sample size of at least 30 participants.

Studies that did not present an overall performance score, or used instruments that addressed only one physical dimension, e.g. balance, were excluded. Finally, studies were excluded if the instrument used was developed for populations with specific diseases or if the study was overall rated inadequate for reliability, validity and responsiveness.

Data extraction and evaluation of psychometric evidence

Two independent reviewers performed abstract scanning, selection of full-text articles and data extraction. Disagreements were judged by a third person. Full-text papers were obtained if abstracts fulfilled the inclusion criteria or eligibility could not be determined. In case further information was needed authors were contacted.

The quality of individual studies was rated as good (+), poor (−) or moderate (?) with the modified checklist as described in Table 1. Domains for grading the studies were 'study population', 'adequate description of test', 'adequate design for evaluating psychometric property' [16, 19]. Thus a clear description of the study population, measurement and design for evaluating psychometric properties was required to receive a positive rating (+). If a study failed to do so on one domain it was rated moderate (?) and rated poor (−) when it failed on two domains.

To evaluate the strength of evidence for the psychometric properties of instruments the domains 'Quality', 'Quantity' and 'Consistency' were used [101]. Quality of the evidence for individual studies was rated according to the checklist as presented in Table 1. Quantity was defined as the number of studies and magnitude of effect. Consistency was defined as the extent to which similar findings were reported [101]. Based on this, instruments were given an overall rating for each psychometric property: good (++), adequate (+), neutral (?) or inadequate (−).

Table 1. Quality criteria for clinimetric properties of physical function assessment

Property	Definition	Quality criteria[a,b]
Content validity	The extent to which the domain of interest is comprehensively sampled by the items in the instruments	+ Positive rating − Poor rating ? Moderate rating
Predictive validity	The extent to which the instrument had the ability to predict the onset of difficulties in functioning or negative health outcomes over time (e.g. mortality)	+ High scores with regard to methods, design and results (OR or AUC) ? Doubtful design or method (e.g. small sample size) − Inappropriate methods or lack of significant results
Construct validity	The ability to discriminate between subgroups e.g. age groups, gender	+ High scores with regard to methods, design and results (clear group definitions and significant results) ? Doubtful design or method (e.g. small sample size) − Inappropriate methods or lack of significant results
Concurrent validity	Established by simultaneously applying a previously validated tool or test, and comparing the results	+ Comparison with other instrument with significant results ($r > 0.80$) ? Doubtful design or method (e.g. small sample size); significant but small results $r > 0.60$–0.80 − Inappropriate methods or lack of significant results
Reliability	An indicator of the consistency of a measurement in terms of internal consistency with stability over time (reproducibility) and the degree of which the measurement is free of measurement error (internal consistency)	+ (good) intra-class correlation coefficient (ICC) or Kappa > 0.70 ? (moderate) ICC 0.70–0.60 or $r > 0.80$ − (poor) ICC or Kappa < 0.70, despite adequate design and method
Responsiveness	The instrument's ability to detect important change over time in the concept being measured, and may be defined as the extent to which a method detects minimal clinically relevant change over time	+ A power calculation for sample size presented, adequate design and sufficiently described ? Doubtful design or method (e.g. no hypotheses)
Floor and ceiling effects	The number of respondents who achieved the lowest or highest possible score	+ ≤15% of the respondents achieved the highest or lowest possible scores ? Doubtful design or method − >15% of the respondents achieved the highest or lowest possible scores, despite adequate design and methods
Overall quality of individual study	The degree to which one can assign qualitative meaning to quantitative scores	+ Clear description of study population, adequate description of instrument, adequate design for evaluating psychometric properties ? Doubtful description of either study population, or instrument but with reference given or method − Poor description of study population, OR instrument and no reference given, and poor method

AUC, area under the curve; ICC, intra-class correlation coefficient; OR, odds ratio.
[a]+, positive rating; −, poor rating; ?, moderate rating.
[b]Doubtful design or method = lacking of a clear description of the design or methods of the clinimetric study, sample size smaller than 30 subjects (e.g. subgroup analysis), or any important methodological weakness in the design or execution of the study.

Reliability

Reliability is an indicator of the consistency of a measurement in terms of internal consistency with stability over time (reproducibility) and defines the degree of which the measurement is free of measurement error (internal consistency) [9]. Reliability was rated as described in Table 1.

Validity

Predictive, construct and concurrent validity were considered because of the frequent use of overall physical function instruments for clinical use [9, 19]. Validity domains were rated as described in Table 1.

Responsiveness and practical issues

For applied research, it is of essential interest to know whether an instrument is suited to detect changes over time with respect to the construct being measured, and if there are floor or ceiling effects defined as the number of participants who achieved the lowest or highest possible score of the instrument [9, 19]. Practicability issues, such as the time needed to administer assessment or requirements of space, equipment, training for administration, were obtained.

Results

The literature search identified 2,383 abstracts. After screening the abstracts, 454 full papers were obtained and further screened for inclusion or exclusion. The flow of the study can be viewed in Figure 1.

In total 94 articles describing 12 different functional assessment instruments with their modified versions were included and data extracted. In the data extraction process 16 studies investigating 8 different instruments were excluded, due to addressing only one domain, or an overall inadequate rating. The included instruments were Continuous Scale Physical Performance (CS-PFP) [10] with two modifications [20, 21], MacArthur battery [22],

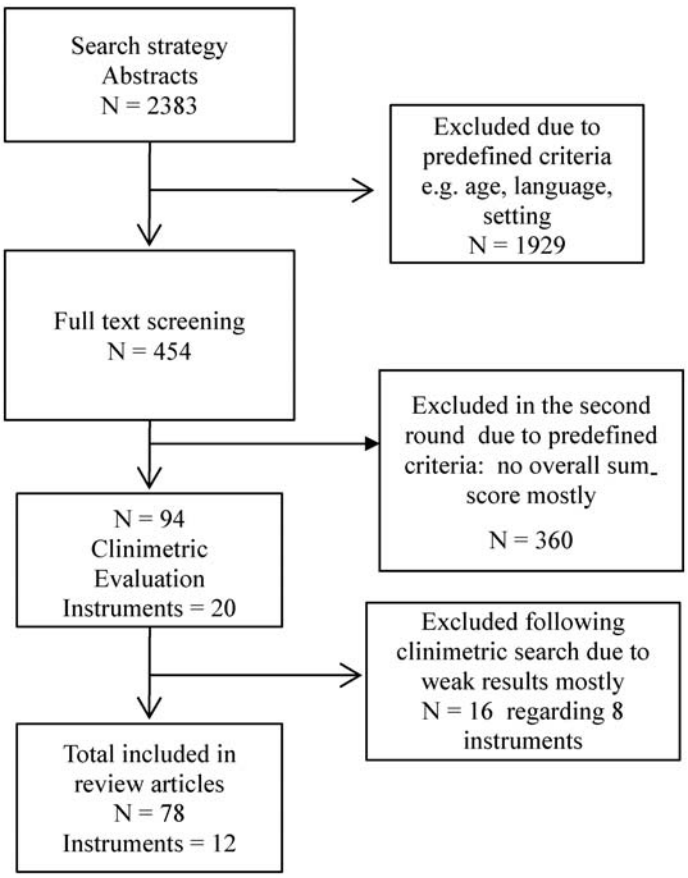

Figure 1. Flow of review.

Modified Timed Movement Battery (Mod TMB) [23], mobility-related limitation index (MOBLI Index) [24], Physical Capacity Evaluation (PCE) [25] with one modification [25], Performance-Oriented Mobility Assessment (POMA) [26] with one modification [27], Performance-based Physical Function Test (PPF) [28], Physical Performance Test (PPT) [29] with five modifications [30–34], Shinkai Summary Performance Score (SSPS) [35], Short Physical Performance Battery (SPPB) [36] with three modifications [37–39], Task Modification Scale (TMS) [40], Upper Extremity Summary Performance Score (UESPS) [37] with one modification [38].

Overall quality of included studies was moderate, mainly caused by a lack of a clear description of the measurement and/or of the population studied. Risk for bias was mainly due to instruments being tested in a selected population of community dwellers (e.g. often data on frail older adults are missing) and the occurrence of ceiling effects in some instruments.

Description of instruments

Relevant information about the included instruments is shown in Table 2. Six instruments (MOBLI Index, mod. TMB, PCE, PPF, PPT, SPPB) were tested in the general older community-dwelling population. Regarding the study population's ability to perform ADLs, six instruments (CS-PFP, MacArthur Battery, POMA, PPT, SSPS, SPPB) were tested in an older population with no ADL limitations, seven instruments (CS-PFP, MacArthur Battery, MOBLI Index, POMA, PPT, SPPB, TMS) were tested in an older population with some ADL limitations (difficulty in one or two domains) and four instruments (CS-PFP, PPT, SPPB, UESPS) were tested in an older population with more severe ADL limitations (difficulty in two or more domains). Psychometric properties of three instruments (CS-PFP, PPT, SPPB) were evaluated in all ADL categories.

Most instruments were used to measure physical function, physical performance and functional status, thus targeting different constructs. The CS-PFP, PCE and PPT typically combine the performance of a wide variety of daily activities, e.g. writing, eating and walking. The MacArthur Battery, Mobli Index, Mod TMB, POMA, PPF, SPPB, TMS combine items that focus on lower extremity performance, whereas the UESPS is directed to assess only the capability of the upper extremities. The MOBLI Index is the only included instrument taking lung function into account.

Measurement properties of instruments

Reliability

Seven instruments (58%) have been tested for reliability. No information regarding reliability could be obtained for

Table 2. Instrument characteristics

Instrument	Modified version	Population	Construct measured	Items
CS-PFP [10, 49, 72]	CS-PFP 10-item [11, 20, 46, 51, 73] ADAP [21, 74]	No (I)ADL limitations [10, 20, 21, 46, 49, 72–74]	Physical function [10, 20, 46, 49, 72, 73]; physical functional performance [21, 74]	16 daily tasks [10, 21, 49, 72, 74] 10 daily tasks [11, 20, 46, 73] 10 daily tasks, no stair climbing [51]
		Some ADL limitations (one or two domains) [10, 11, 20, 51, 72]	Physical function [10, 11, 20, 51, 72]	
		Limited ADL function (two or more domains) [10, 20, 72]	Physical function [10, 20, 72]	
MacArthur Battery [22, 75]		No (I)ADL limitations [22]	Physical performance [22]	Writing, standing balance, chair rise, Gait speed, foot tapping
		Some ADL limitations (one or two domains) [75]	Physical performance [75]	
MOBLI Index [24, 44]		General population (mixed) [24, 44]	Mobility-related limitations [24]	Gait speed, chair rise, peak expiratory flow rate
		Some ADL limitations (1 or 2 domains) [44]	Mobility-related limitations [44]	
Modified TMB [23]		General population (mixed) [23]	Functional mobility [23]	Nine mobility and transfer tasks
PCE [25, 76]	PCE short form [25]	General population (mixed) [25, 76]	Observed physical function [25, 76]	Three to five flexibility tasks, two to four manual dexterity tasks, foot tapping [76], get-up-and-go [25, 76], tandem stand [76], handgrip strength [25, 76]
POMA 40 items [77]	POMA 17 items [27, 78]	No (I)ADL limitations [27, 77, 78]	Functional balance [77] Balance and gait [78] [27]	24 Balance tasks (short version 9 tasks) 16 Gait tasks (short version 8 tasks)
		Some ADL limitations (one or two domains) [77]	Functional balance [77]	
PPF test [28, 79]		General population (mixed) [28, 79]	Physical function [79] Functional status [28]	Gait speed, standing balance, chair rise, grip strength
PPT-7 [3, 31, 48, 58–60, 80–85]	JPPT [34] PPT-8 [30, 75] PPT-9 [31, 50, 59, 84] PPT-9 Modified 1 [33] PPT-9 Modified 2 [32, 86]	General population (mixed) [31, 34, 58, 80, 86]	Physical function [31, 34, 58] Functional status [80] Physical frailty [86]	Writing, eating, jacket, lifting a book, picking up penny, turning 360°, walking 50 ft (25 ft [75]) Modifications: one flight of stairs [30–32, 50, 59, 75, 84, 86], stairs [31, 32, 50, 59, 84, 86]', chair rise [32, 33, 86], standing balance [32, 33, 86]
		No (I)ADL limitations [32, 48, 59, 81–84]	Physical function [81–83] Function [59] Functional limitations [48] Physical performance [32, 84]	
		Some ADL limitations (1 or 2 domains) [3, 30, 33, 60, 75, 81–83]	Physical function [3, 81–83] Functional status [60] Physical frailty [33] Physical performance [30, 75]	
		Limited ADL function (2 or more domains) [33, 50, 85]	Physical function [50, 85] Physical frailty [33]	
		No (I)ADL limitations [32]	Physical performance [32]	
SSPS [35]		No (I)ADL limitations [35]	Physical function [35]	Gait speed, handgrip strength, standing balance
SPPB [36–38, 41–43, 45, 47, 50–57, 60–**63**, 70, 71, 87–98]	Continuous SPPB score (CSPPS) [37, 57, 94]' SPPB extended balance [70, 99, 100] SPPB extended shoulder rotation [39]	General population (mixed) [36, 39, 43, 52, 56, 62, **63**, 87–92, 99]'	Physical function [43, 52] Functional status [39, 92, 99] Lower extremity function [62, **63**, 91, 92] Mobility function [89] Physical performance [36, 56, 87, 88, 90]	Gait speed, standing balance (three tasks or five tasks [99]), chair rise Modifications: narrow walk test [70], dynamic balance (two tasks [100]), shoulder rotation [39]

Continued

Table 2. Continued

Instrument	Modified version	Population	Construct measured	Items
		No (I)ADL limitations [42, 45, 61, 70, 71, 90, 93, 96]'	Physical function [70] Functional status [42] Lower extremity function [61, 71, 96] Mobility function [45] Lower extremity performance [93] Physical performance [90]	
		Some ADL limitations (1 or 2 domains) [47, 51, 53–55, 60, 88, 97, 98]'	Functional status [60, 98] Lower extremity function [47] Physical function [51, 55] Physical performance [53, 88, 97] Mobility function [54]	
		Limited ADL function (two or more domains) [37, 38, 41, 50, 57, 88, 94, 95, 100]	Physical function [97] Lower extremity function [41, 50, 88] Lower extremity mobility [95] Physical performance [94] Lower extremity performance [37, 38, 100]	
TMS [40]		Some ADL limitations (1 or 2 domains) [40]	Physical ability [40]	Chair rise, stairs, kneel rise, supine rise
UESPS [37]	Modified UESPS [38]	Limited ADL function (two or more domains) [37, 38]	Upper extremity performance [37, 38]	Putting on a blouse [37, 38] Lock and key test [37] Pegboard [37, 38] Handgrip strength [37, 38]

CS-PFP, Continuous Scale Physical functional Performance Test; ADAP, assessment of daily activity performance; MOBLI Index: mobility-related limitation index; Modified TMB, modified timed movement battery (TMB); PCE, physical capacity evaluation; POMA, performance-oriented mobility assessment; PPF, performance-based physical function test; PPT-7, 7-item physical performance test; JPPT, Japanese version physical performance test; PPT-9, 9-item physical performance test; PPT-8, 8-item physical performance test; SSPS, Shinkai summary performance score; SPPB, short physical performance battery; TMS, task modification scale; UESPS, upper extremities summary performance score; CSPPS, continuous summary physical performance score; ADL, activities of daily living.

the MOBLI Index, POMA, PPF, SSPS and UESPS. In most cases, reliability was obtained with a 1–2 week interval. One study investigated the change of reliability over time ranging from 6 to 36 months [34, 41]. Most studies determined reliability using ICCs and six instruments were found to have a good reliability rating: CS-PFP, MacArthur Battery, Mod TMB, PPT, SPPB, TMS, with ICCs ranging from 0.70 to 0.99. Two studies [20, 42] investigated whether a change in setting altered the reliability of their instrument (CS-PFP; SPPB) with both studies showing no change. One study showed that the experience of the tester has a significant influence on the reliability [21]. An in-depth table of reliability can be viewed in Supplementary data Appendix 2.

Validity

All but one (PPF) instruments have addressed concurrent and/or predictive validity. An in-depth table of validity can be viewed in Supplementary data Appendix 2.

We obtained predictive validity with regard to:

- mortality: SPPB [**36**, 43, 87, 88, 90, 93], MOBLI Index [44],
- dependency: CS-PFP [72] ; SSPS [**35**],
- difficulty (I)ADL: POMA [27], PPT [48], SPPB [38, 45, 61, 71], UESPS [38],
- falls: POMA [78], PPT [85],
- global health improvements: CS-PFP [51], SPPB [51],
- difficulty in walking: MOBLI Index [**24**], SPPB [38, 45, 47, 61, 62, 71], UESPS [38],
- disability in upper extremity performance: SPPB [38], UESPS [38].

The ability to distinguish between groups was found with regard to:

- dependency: CS-PFP [**10**, 72], PPT [81, 83],
- gender: CS-PFP [46], MacArthur Battery [**22**] SPPB [94, 97, 100],
- age groups: CS-PFP [46, 73], PCE [76], PPT [32, 33, 59, 60, 81, 83], SPPB [**37**, 57, 60],
- ethnic groups: MacArthur Battery [**22**], PPT [48],

- chronic conditions: MacArthur Battery [22], PPT [83], SPPB [97],
- ADL disability: MacArthur Battery [22], TMB [23], SPPB [90].

Good to moderate concurrent validity was found for following instruments:

- ADL disability: CS-PFP [46, 51], PCE [25, 76], PPT [3, 30, 32, 34, 58, 60] SPPB [36, 42, 51, 60, 91, 94, 95],
- balance: Mod. TMB [23], POMA [78], PPT [86], TMS [40],
- cognition: PPT [32],
- endurance: CS-PFP [10],
- flexibility: CS-PFP [10], PPT [81, 86],
- mobility: Mod TMB [23], PPT [3, 86], SPPB [51, 70],
- quality of life: CS-PFP [10, 11, 21, 46, 51], PPT [58], SPPB [51, 60], TMS [40],
- SPPB: CS-PFP [51], PPT [60],
- strength, power: CS-PFP [10, 21, 51], PPT [3, 81, 86], SPPB [51, 94], TMS [40],
- walking/gait speed: POMA [78], TMS [40].

Responsiveness

Information was available for six instruments (CS-PFP, MacArthur Battery, MOBLI Index, POMA, PPT, SPPB). Responsiveness after an intervention was demonstrated for the CS-PFP [20, 49, 51], PPT [30, 50] and SPPB [50–55] (effect sizes ranging from 0.48 to 1.25). Effect of falls on test performance was investigated for the POMA showing a higher functional decline in fallers compared with non-fallers [27]. MOBLI Index demonstrated responsiveness with Guyatt's responsive index ranging from 0.32 to 0.85 [24]. Significant changes over time have been determined for SPPB [41, 56, 57].

Practical issues

For the PPT, time to administer, varied between 4 min (high functioning [48, 58]) and 15 min (frail [85]). The mean administration time for PCE was 36 min [25] and for TMS was 15 min [40]. The time to perform the SPPB was 10–15 min [36].

Information on floor and ceiling effects was reported for PPT and SPPB. For PPT-7 and PPT-8, 0% of older persons with no limitations or some limitations achieved the lowest possible score and 0–4% achieved the highest possible score [58, 59]. For SPPB, 0–7% achieved the lowest score and 2–16% scored the highest score of 12 points [61, 62]. One study [62] found that in a group of non-disabled older persons; 77% had the highest possible score on the SPPB.

Summary of instrument properties

Table 3 summarises the evidence for reliability and validity of the instruments. The SPPB, PPT and CS-PFP were tested in older populations for all psychometric properties. The SPPB is the only test battery rated as good in reliability, whereas CS-PFP, PPT, PCE, Mod TMB and MacArthur Battery were rated as adequate in reliability. SPPB is the only instrument that shows good validity, whereas CS-PFP, POMA and PPT demonstrate adequate validity. SPPB is the only instrument that shows a good responsiveness, whereas CS-PFP, MOBLI Index and PPT-8 are rated as adequate.

Discussion

To our knowledge this is the first systematic review targeting performance-based, overall index instruments that measure physical function in community-dwelling older persons. From the 12 instruments that we evaluated, 3 met all our criteria: the SPPB [36], the PPT [29] and the CS-PFP [10]. Being able to screen or assess older persons at risk for functional decline or evaluate effects of an intervention targeted to older community-dwelling persons is of great interest to rehabilitation staff and geriatricians. It is therefore worrisome that many commonly used instruments are still not well validated and presumable less reliable.

The SPPB was the measurement with most positive ratings (e.g. highest score in reliability, validity and responsiveness) and has been extensively investigated in different populations ranging from vigorous to ADL limited or frail with 34 studies investigating at least one psychometric property. The focus of the SPPB is on lower extremity function, whereas the PPT and CS-PFP both evaluate lower and upper bodily function. It may be that the SPPB had an advantage in our review, because reliable measurement in fewer domains is usually relatively easier to achieve [64].

An increasing body of evidence suggests that in early stages of decline physical function can be stabilised or even reversed with a targeted intervention [65–68]. This demonstrates the dynamic process with older persons moving in and out of the status of disability or functional impairment. Based on this dynamic fluctuation the precise measurements of older persons being at risk for disability, demonstrating preclinical limitation in performance or functional limitation becomes a primary baseline to raise red flags or channel intervention. Sensitive measurement is also important for population-based research activities as well as to policy-makers [1, 36, 65, 69].

One limitation of this review is the focus on community-dwelling older people only, omitting articles about instruments developed for use in chronically diseased populations. Another limitation is that no attention was paid to the amount of missing data as a source of potential bias. Many instruments have missing data for frail persons and if these persons form a considerable part of the target population, it is recommended to reconsider using these instruments.

Table 3. Summary of reviewed outcome measure's properties

Instrument	Modified version	Number of studies	Reliability	Validity	Responsiveness
CS-PFP		3	+	++	+
	ADAP	2	+	+	?
	CS-PFP 10-item	4	+	++	+
	CS-PFP 10-item (without stairs)	1	0	0	0
MacArthur Battery		2	+	?	?
MOBLI Index		2	0	+	+
Mod TMB		1	+	+	0
PCE		2	+	+	0
	PCE short form	1	?	0	0
POMA		3	0	++	?
PPF		2	0	0	?
PPT-7		12	+	++	0
	JPPT	1	0	+	0
	PPT-8	2	+	?	+
	PPT-9	4	?	?	?
	PPT-9 (mod1)	1	0	?	0
	PPT-9 (mod2)	2	+	++	0
SSPS		1	0	+	0
SPPB		34	++	+++	++
	CSPPB	3	+	++	++
	SPPB extended balance	4	?	++	0
	SPPB extended shoulder rotation	1	−	?	0
TMS		1	+	+	0
UESPS		1	0	?	+
	Modified UESPS	1	0	+	0

0, no numerical results reported; −, not adequate; ?, weak; +, adequate; ++, good; +++, very good.
CS-PFP, continuous scale physical functional performance test; ADAP, assessment of daily activity performance; MOBLI Index, mobility-related limitation index; Modified TMB, modified timed movement battery; PCE, physical capacity evaluation; POMA, performance-oriented mobility assessment; PPF, performance-based physical function test; PPT-7, 7-item physical performance test; JPPT, Japanese version physical performance test; PPT-9, 9-item physical performance test; PPT-8, 8-item physical performance test; SSPS, Shinkai summary performance score; SPPB, short physical performance battery; TMS, task modification scale; UESPS, upper extremities summary performance score; CSPPS, continuous summary physical performance score; ADL, activities of daily living.

Although the importance of performance-based physical functioning instruments is widely acknowledged, the results of this review show that many instruments lack acceptable psychometric evidence and/or have several limitations. Few studies included frail people or those with ADL/IADL disabilities thus limiting the generalisability of the results. In addition we came across too many modified versions. It seems that the 'not invented here syndrome (NIH)' plays a major role in the modifications [64]. Usually, for the modified versions less data were found, further limiting generalisability (e.g. SPPB with continuous scoring was only tested in older adults with limited function). We found several different names for some instruments as well as differences in test items in the instruments themselves. For the SPPB, different distances for the walking tests can be found [39, 61, 70, 71]. Nevertheless, we could identify numerous modified versions of original instruments. It seems that researchers and practitioners either use measurements without a critical look, or change it without looking at the methodological consequences. Both aspects pose barriers and hinder further development in the field of physical function screening.

With regard to practical issues, we found little information about ecological aspects. The tester's expertise seems crucial for reliability [21]. Rarely reported, but important, is information on floor and ceiling effects of the instruments in a population ranging from high functioning to frail older persons. As stated in Table 2 most instruments are applied to only one target group, and not over the whole range. No information was found on the absolute (clinical) reliability of the instruments included in this review as measured by the magnitude of individual day-to-day variability. Researchers rarely report the absolute reliability of instruments although it may be an important determinant for use in clinical practice.

Several performance-based tests (including SPPB, CS-PFP and PPT) require significant space (e.g. to measure gait speed) or stairs, which may restrict usability by GPs or other allied health-care staff in primary care practices or on home visits. For the SPPB, a reduction of the distance to measure gait speed to 2.4 m has been shown to provide valid data [102].

For practitioners in the field of performance-based functional tests, it seems that the SPPB [36] is the best choice. But attention should be paid to different versions of the SPPB.

Conclusion

Only few valid and reliable overall measures for performance-based physical function exist. Some introspection in this field is necessary to avoid unnecessary modifications

and use instruments that have not been documented to be valid and reliable.

We recommend the original SPPB and the CS-PFP in geriatrician screening and assessment and for scientific purposes. For researchers, the use of the PPT and the CS-PFP is also recommended. These last instruments are more complex in use, but are reliable and valid for the assessment of physical function in community-dwelling older adults.

Key points

- Only a few valid and reliable overall measures for performance-based physical function exist.
- In the field of performance-based functional tests, SPPB is the best choice.
- Change in names or measurements of instruments pose a barrier for implication.

Acknowledgements

The collaboration for this manuscript was initiated during one of the meetings of the European Network for Action on Ageing and Physical Activity (EUNAAPA). The authors thank the European Commission, Directorate C—Public Health and Risk Assessment, for this network under the program of community action in the field of public health (2003–2008). The content of the article does not represent the opinion of the European Community, and the European Commission is not responsible for any use that might be made of the information presented in the text. We would like to thank Dr Priscilla MacRae for reviewing our paper as a native speaker.

Authors' contribution

E.F.: review concept and design, analysis and interpretation of data, manuscript preparation. P.d.V.: review concept and design, analysis and interpretation of data, manuscript preparation. D.S.: design of search strategy, manuscript preparation. E.R.: study protocol development, manuscript preparation. V.M.: design of search strategy, data analysis. K.F.: study protocol development. M.H.-R.: analysis and interpretation of data, manuscript preparation.

Funding

This works was supported by the Universität Erlangen-Nürnberg.

Supplementary data

Supplementary data mentioned in the text is available to subscribers in *Age and Ageing* online.

References

The very long list of references supporting this review has meant that only the most important are listed here and are represented by bold type throughout the text. The full list of references is available on Supplementary data in Age and Ageing online, Appendix 1.

1. Gill TM. Assessment of function and disability in longitudinal studies. J Am Geriatr Soc 2010; 58(Suppl. 2): S308–12.

4. Fried L, Ferrucci L, Darer J, Williamson JD, Anderson G. Untangling the concepts of disability, frailty, and comorbidity: Implications for improved targeting and care. J Gerontol Series A Biol Sci Med Sci 2004; 59: M255–63.

9. Terwee CB, Bot SD, de Boer MR *et al.* Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007; 60: 34–42.

10. Cress M, Buchner DM, Questad KA, Esselman PC, deLateur BJ, Schwartz RS. Continuous-scale physical functional performance in healthy older adults: a validation study. Arch Phys Med Rehabil 1996; 77: 1243–50.

14. WHO. International Classification of Functioning, Disability and Health. Geneva, 2001.

15. Beswick A, Rees K, Dieppe P *et al.* Complex interventions to improve physical function and maintain independent living in elderly people: a systematic review and meta-analysis. Lancet 2008; 371: 725–35.

16. Rydwik E, Bergland A, Forsén L, Frändin K. Psychometric properties of Timed Up and Go in Elderly People: a systematic review. Phys Occup Ther Geriatr 2011; 29: 102–25.

17. Forsen L, Loland NW, Vuillemin A *et al.* Self-administered physical activity questionnaires for the elderly: a systematic review of measurement properties. Sports Med 2010; 40: 601–23.

19. Terwee CB, Mokkink LB, van Poppel MNM, Chinapaw MJM, van Mechelen W, de Vet HCW. Qualitative attributes and measurement properties of physical activity questionnaires: a checklist. Sports Med 2010; 40: 525–37.

22. Seeman TE, Charpentier PA, Berkman LF *et al.* Predicting changes in physical performance in a high-functioning elderly cohort: MacArthur studies of successful aging. J Gerontol 1994; 49: M97–108.

23. Creel LG, Light KE, Thipgen MT. Concurrent and construct validity of scores on the Timed Movement Battery. Phys Ther 2001; 81: 789–98.

24. Lan TY, Deeg DJ, Guralnik JM, Melzer D. Responsiveness of the index of mobility limitation: comparison with gait speed alone in the longitudinal aging study amsterdam. J Gerontol A Biol Sci Med Sci 2003; 58: 721–7.

25. Daltroy LH, Phillips CB, Eaton HM *et al.* Objectively measuring physical ability in elderly persons: the Physical Capacity Evaluation. Am J Public Health 1995; 85: 558–60.

26. Tinetti ME. Performance-oriented assessment of mobility problems in elderly patients. J Am Geriatr Soc 1986; 34: 119–26.

28. Wang L, van Belle G, Kukull WB, Larson EB. Predictors of functional change: a longitudinal study of nondemented people aged 65 and older. J Am Geriatr Soc 2002; 50: 1525–34.

29. Reuben DB, Siu AL. An objective measure of physical function of elderly outpatients. The Physical Performance Test. J Am Geriatr Soc 1990; 38: 1105–12.

35. Shinkai S, Watanabe S, Kumagai S *et al*. Walking speed as a good predictor for the onset of functional dependence in a Japanese rural community population. Age Ageing 2000; 29: 441–6.

36. Guralnik JM, Simonsick EM, Ferrucci L *et al*. A short physical performance battery assessing lower extremity function: association with self-reported disability and prediction of mortality and nursing home admission. J Gerontol 1994; 49: M85–94.

37. Onder G, Penninx BW, Lapuerta P *et al*. Change in physical performance over time in older women: the Women's Health and Aging Study. J Gerontol A Biol Sci Med Sci 2002; 57: M289–93.

40. Manini T, Cook SB, VanArnam T, Marko M, Ploutz-Snyder L. Evaluating task modification as an objective measure of functional limitation: repeatability and comparability. J Gerontol A Biol Sci Med Sci 2006; 61: 718–25.

63. Sayers SP, Guralnik JM, Newman AB, Brach JS, Fielding RA. Concordance and discordance between two measures of lower extremity function: 400 meter self-paced walk and SPPB. Aging Clin Exp Res 2006; 18: 100–6.

64. Katz R, Allen TJ. Investigating the not invented here (nih) syndrome: a look at the performance, tenure, and communication patterns of 50 R & D project groups. R&D Manage 1982; 12: 7–20.

65. Daniels R, van Rossum E, de Witte L, Kempen G, van den Heuvel W. Interventions to prevent disability in frail community-dwelling elderly: a systematic review. BMC Health Serv Res 2008; 8: 278.

66. Haskell W, Lee I, Pate R *et al*. Physical activity and public health: updated recommendation for adults from the American College of Sports Medicine and the American Heart Association. Med Sci Sports Exerc 2007; 39: 1423–34.

67. Liu CK, Fielding RA. Exercise as an intervention for frailty. Clin Geriatr Med 2011; 27: 101–10.

68. Chodzko-Zajko W, Proctor DN, Fiatarone Singh MA *et al*. Exercise and physical activity for older adults. Med Sci Sports Exerc 2009; 41: 1510–30.

69. Hardy SE, Gill TM. Recovery from disability among community-dwelling older persons. JAMA 2004; 291: 1596–602.

Received 18 November 2011; accepted in revised form 19 April 2012