BMJ Open

Systematic review of the measurement properties of performance-based functional tests in patients with neck disorders

Steven McGee,[1] Taylor Sipos,[1] Thomas Allin,[1] Celia Chen,[1] Alexandra Greco,[1] Pavlos Bobos ,[1,2,3] Joy MacDermid,[1,2,4] CATWAD

To cite: McGee S, Sipos T, Allin T, *et al*. Systematic review of the measurement properties of performance-based functional tests in patients with neck disorders. *BMJ Open* 2019;**9**:e031242. doi:10.1136/bmjopen-2019-031242

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2019-031242).

Received 24 April 2019
Revised 08 October 2019
Accepted 04 November 2019

[1]School of Physical Therapy, Health and Rehabilitation Sciences, Western University, London, Ontario, Canada
[2]Western's Bone and Joint Intitute, Western University, London, Ontario, Canada
[3]Dalla Lana School of Public Health, Institute of Health Policy Management and Evaluation, Department of Clinical Epidemiology and Health Care Research, University of Toronto, Toronto, Ontario, Canada
[4]School of Rehabilitation Science, McMaster University, Hamilton, Ontario, Canada

Correspondence to
Dr Pavlos Bobos;
pbobos@uwo.ca

ABSTRACT

Objectives The purpose of this systematic review is to identify and synthesise studies evaluating performance-based functional outcome measures designed to evaluate the functional abilities of patients with neck pain.

Design Systematic review.

Data sources A literature search using PubMed, Scopus, CINAHL, EMBASE, COCHRANE, Google Scholar and a citation mapping strategy was conducted until July 2019.

Eligibility criteria More than half of the study's patient population had neck pain or a musculoskeletal neck disorder and completed a functional-based test. Clinimetric properties of at least one performance-based functional tests were reported. Both traumatic and non-traumatic origins of neck pain were considered.

Data extraction and synthesis Relevant data were then extracted from selected articles using an extraction guide. Selected articles were appraised using the Quality Appraisal for Clinical Measurement Research Reports Evaluation Form (QACMRR).

Results The search obtained 12 articles which reported on four outcome measures (functional capacity evaluations (FCE), Baltimore Therapeutic Equipment Work Simulator II (BTEWS II), Functional Impairment Test-Hand and Neck/Shoulder/Arm (FIT-HaNSA)) and a physiotherapy test package, to assess the functional abilities in patients with mechanical neck pain. Of the selected papers: one reports content validity, five construct validity, four reliability, one sensitivity to change and one both reliability and construct validity. QACMRR scores ranged from 68% to 95%.

Conclusions This review found very good quality evidence that the FIT-HaNSA has excellent inter and intra-rater reliability and very weak to weak convergent validity. Excellent quality evidence of fair test-retest reliability, weak convergent validity and very weak known groups validity for the BTEWS II test was found. Good to excellent quality evidence exists that an FCE battery has poor to excellent reliability and very weak to strong validity. Good to excellent quality of weak to strong validity and trivial to strong effect sizes were found for a physiotherapy test package.

Prospero registration number CRD42018112358

Strengths and limitations of this study

- ► The psychometric properties of performance outcome measures for neck pain were synthesised and critically appraised.
- ► This study assessed the risk of bias and the quality of measurements properties.
- ► The feasibility or usability of these tools was not assessed.

INTRODUCTION

Neck pain has been associated with high disability and is regarded as a substantial societal burden.[1] Approximately 70% of people experience neck pain within their lifetime and about 33% of adults experience neck pain every year.[2 3] Further concern is warranted as it has been suggested that the incidence of neck pain is increasing.[4–6] The economic burden due to neck disorders is high, including lost wages, costs of treatment and compensation expenditures to injured people.[7 8] Neck pain is second only to low back pain in annual workers' compensation costs in the USA and has been associated with many other comorbidities such as headaches, anxiety, depression, back pain and arthralgias.[6 9 10]

Outcome measures are a crucial component in monitoring patients with neck pain to determine the effects of treatment,[11 12] evaluation of interventions, guiding return to work and justifying treatment.[13 14] Several self-reported outcome measures currently exist to assess disability and function in those with neck pain (eg, the Neck Disability Index-NDI).[13] Evidence-based clinical practice guidelines suggest that measures assessing physical performance should also be used for people with neck pain.[15] Performance-based testing is where the assessment is based on actual performance of a task or activity. Physical performance can be assessed by testing a person's ability to execute a standardised activity in a standardised environment (ie, clinical setting).[16] Time to complete the activity, number of repetitions performed and weight lifted are frequently used to quantify the physical performance.[17] Conversely, self-report measures examine patients' perception and experience of their ability to perform functional tasks.[16] Previous research has demonstrated poor to fair relationships between physical performance and self-report measures of ability in patients with various musculoskeletal disorders suggesting that these measures assess different constructs of function.[17 18] Consequently, physical performance tests and self-report measures complement each other and may each contribute unique information about a patient's function.[19]

A fundamental component of monitoring outcomes is having reliable and valid tools with known measurement properties.[13 20] While recent research has investigated the psychometric properties of patient-reported outcomes in people with neck pain,[13 21] there is a gap in knowledge with respect to performance-based functional outcomes. The purpose of this systematic review was to identify and synthesise clinical measurement studies that evaluate measurement properties of performance-based functional tests in patients with neck disorders.

METHODS

Patient and public involvement

There was no patient or public involvement in the design or planning of this study.

Study design and protocol registration

We conducted a systematic review to evaluate the psychometric properties of performance-based functional tests for people with mechanical neck disorders. The protocol was registered in PROSPERO register with registration number CRD42018112358.

Search strategy

A database search using CINAHL, PubMed, Scopus and Google Scholar was performed to identify articles published until July 2019. The following search strategy was used to search all databases for eligible studies: (Reliability OR validity OR responsiveness OR calibration OR validation) OR (minimal detectable change) OR (clinically important difference) OR (psychometric properties) AND cervical OR neck OR c-spine AND (performance measure) OR (functional test) OR (functional outcome) OR (performance outcome). MeSH terms were searched in PubMed. A citation map of articles and systematic reviews selected for the full-text review was performed. This strategy was included to minimise the risk of publication bias. The full search strategy is summarised in online supplementary appendix 1. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) process[22] was followed to ensure all appropriate steps were taken in the selection process (figure 1).

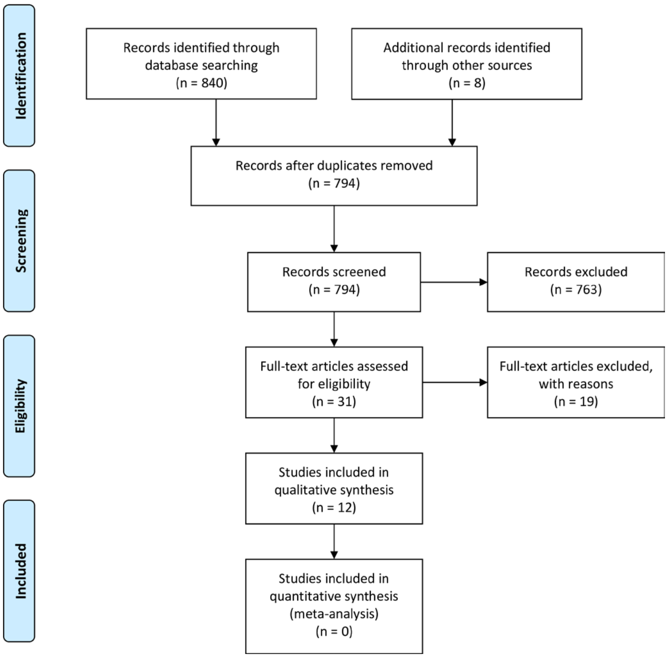
Identification
Screening
Eligibility
Included
Records identified through database searching (n = 840)
Additional records identified through other sources (n = 8)
Records after duplicates removed (n = 794)
Records screened (n = 794)
Records excluded (n = 763)
Full-text articles assessed for eligibility (n = 31)
Full-text articles excluded, with reasons (n = 19)
Studies included in qualitative synthesis (n = 12)
Studies included in quantitative synthesis (meta-analysis) (n = 0)

Figure 1 Selection of the studies for inclusion in the systematic review.

Inclusion criteria

Articles were included in the final review if all of the following criteria were met:

- >50% of the study's patient population had neck pain or a musculoskeletal neck disorder (eg, whiplash associated disorder (WAD II))
- Patients in the study completed a functional-based test
- Clinometric properties of at least one performance-based test were reported.

A test was considered functional-based if it met the following criteria:

- Assessment of a patient's ability to execute a standardised activity in a standardised environment
- Tests assessing muscular endurance (eg, cervical flexion test) or proprioception were not deemed functional-based as they are often not reflective of physical working conditions.

Both traumatic and non-traumatic origins of neck pain were considered. Definitions for the properties can be found in online supplementary appendix A.

Article selection

Titles and abstracts generated by the search strategy were screened by two authors (SM and PB) independently. Articles that met the inclusion criteria and selected for a full-text review were also reviewed in pairs of authors. Disagreements were resolved by the most experienced author (JCM)

Data extraction

Data extraction and critical appraisal were performed in pairs of two raters among the authors, after the completion of a calibration session in which the most experienced author (JM) reviewed the data extraction tools with the authors that performed the data extraction. When reviewers disagreed during data extraction and/or critical appraisal, and consensus could not be met, a third author arbitrated. A data extraction form[23] (see online supplementary appendix A and B), developed by one of the authors (JM), was used to ensure systematicity. Authors extracted sample size, patient population characteristics, functional tests performed and reported psychometric properties. The interpretation of ICC was as follows: ICC<0.50 indicating poor, 0.50≤ICC<0.75 indicating moderate, 0.75≤ICC<0.9 indicating good and ICC≥0.9 indicating excellent reliability were used as a common benchmark.[24] For validity estimates, correlation coefficient (Pearson's/Spearman) and the 95% CI were extracted if were available.[23 25] Evan's guidelines to interpret the strength of the correlation was used which included: 0.00–0.19 'very weak', 0.20–0.39 'weak', 0.40–0.59 'moderate', 0.60–0.79 'strong' and 0.80–1.00 'very strong'.[26] To assist clinical decision making, standard benchmark scores of trivial (<0.20), small (≥0.20 to<0.50), moderate (≥0.50 to<0.80) or large (≥0.80), as proposed by Cohen, were used.[27] For studies assessing construct validity specifically, results in accordance with predefined hypotheses were evaluated to interpret the findings.

Quality appraisal for clinical measurement research reports evaluation form

Pairs of authors critically appraised the quality of each study using a standardised 12-item evaluation tool (QACMRR) designed to assess the quality of studies determining measurement properties in outcome measures (see online supplementary appendix C). If disagreement was present, a third person (JM) assisted in resolving the discrepancy.[23] This tool has been found to have moderate to excellent preconsensus inter-rater reliability (ICC: 0.69–0.91, κ=0.62–1.00) across a number of systematic reviews.[23 25 28] The evaluation criteria of this tool included 12 items: (1) thorough literature review to define the research question; (2) specific inclusion/exclusion criteria; (3) specific hypotheses; (4) appropriate scope of psychometric properties; (5) sample size; (6) follow-up; (7) the authors referenced specific procedures for administration, scoring and interpretation of procedures; (8) measurement techniques were standardised; (9) data were presented for each hypothesis; (10) appropriate statistics-point estimates; (11) appropriate statistical error estimates; and (12) valid conclusions and recommendations.[23 25] Each item is scored from 0 to 2 with (score=2) is the best; (score=1) is acceptable but suboptimal; (score=0) is not done/documented, substantially inadequate or inappropriate. An article's total score, quality, was calculated by the sum of scores for each item, divided by the numbers of items and multiplied by 100%.[23 25] Overall, the quality summary of appraised articles ranges from (0%–30%) poor, (31%–50%) fair, (51%–70%) good, (71%–90%) very good and (>90%) excellent.

RESULTS

The search strategy resulted in 840 published articles. After duplications were removed, 31 articles were deemed relevant and were screened at full text. Overall, 12 articles met our inclusion criteria (figure 1). The excluded articles were removed due to inappropriate patient populations, investigations into self-report measures or tests assessing proprioception/muscular endurance rather than functional-based measures, or because the articles were found to be systematic reviews. The characteristics of the included studies and the summary of psychometric properties are presented in table 1. The quality assessment is summarised and presented in table 2. Percent agreement was calculated for quality scores between the two raters and it was 90%.

Participants

Participants in the selected articles had various types of neck pain including subacute, chronic and whiplash-associated disorder. The mean/median age of the samples of each study ranged from 30 to 48 years of age. The proportion of women in each article ranged from 34% to 78% of the study population. Two studies that had a mixed sample of subjects with various spinal pain did not report the demographics of the neck pain portion of their sample. One study did not contain any subjects and performed a review of epidemiological literature to establish content validity for work-related neck disorders (table 1).

Functional-based tests

The 12 articles that were included for review provided properties on the following functional based tests: functional capacity evaluations (FCE),[29–34] The Baltimore Therapeutic Equipment work simulator II (BTEWS II),[35] Functional Impairment Test- Hand and Neck/Shoulder/Arm (FIT-HaNSA),[36] as well as items off of a physiotherapy test package including a cervical and lumbar Progressive Isoinertial Lifting Evaluation (PILE-C, PILE-L) test[37–40] and 2×20 m with burden walking test (2×20M-WWB).[37–40] Descriptions of all functional-based tests and their relevant subtasks are provided in online supplementary appendix D.

Functional capacity evaluations

Six articles reported measurement properties for an FCE battery. We identified multiple versions of the FCE in the literature with one article reporting properties on the Workwell FCE,[30] two reporting on the Whiplash Associated Disorder (WAD) FCE[29 31] and three reporting on the neck-FCE.[32–34] These test batteries include various combinations of muscular strength, endurance and functional based tests. The measurement properties of the functional based tests used by the FCE are outlined in table 3.

Individuals with subacute to chronic WAD

Trippolini *et al* (2014)[30] evaluated the Workwell FCE test-retest reliability, measurement error, convergent validity

Table 1 Summary of studies reporting psychometric properties of functional-based tests in neck disorder patients

Study	Population	Sample size (n)	Functional tests	Intervention/test interval	Quality
Ljungquist *et al*[38]	Neck pain (55%), back pain, multiple pain sites	53	PILE-C, PILE-L	N/A	Good (68%)
Ljungquist *et al* [39]	Neck pain (50%), lumbar pain, thoracic pain, shoulder pain, multiple pain sites	68	PILE-C, PILE-L, 2×20 m WWB	8 days	Very good (79%)
Ljungquist *et al*[37]	Neck pain, lumbar pain, thoracic pain, shoulder pain, lower extremity pain, multiple pain sites	235	PILE-C, PILE-L, 2×20 m WWB	N/A	Very good (82%)
Ljungquist *et al*[40]	cervical pain (25%), lumbar pain, cervical (25%) and lumbar pain, multiple pain sites	186	PILE-C, PILE-L, 2×20 m WWB	6 months	Very good (79%)
Lomond and Cote[35]	Chronic neck and shoulder pain (100%)	32	BTEWS II	9.5 days	Very good (88%)
Pierrynowski *et al*[36]	Subacute and chronic WAD II	66	FIT-HaNSA	2–7 days	Very good (88%)
Reesink *et al*[34]	N/A	N/A	Neck-FCE	N/A	N/A
Reneman *et al*[32]	Chronic multifactorial neck pain	18	Neck-FCE	2 weeks	Good (67%)
Trippolini *et al*[31]	Sub acute and chronic WAD I and II	32	WAD FCE	7 days	Very good (75%)
Trippolini *et al*[30]	Sub acute and chronic WAD I and II	267	Workwell FCE	N/A	Excellent (92%)
Trippolini *et al*[29]	Sub acute and chronic WAD I and II	314	WAD FCE	N/A	Very good (86%)
Van der Meer *et al*[33]	Chronic WAD I and II	40	Neck FCE	N/A	Very good (86%)

CBT, cognitive-behavioural therapy; EXP, experimental; F, female; FCE, functional capacity evaluation; FIT-HaNSA, Functional Impairment Test-Hand and Neck/Shoulder/Arm; BTEWS II, Baltimore Therapeutic Equipment work simulator II; M, male; MVA, motor vehicle accident; N/A, not applicable; NRPS, Numeric Pain Rating Scale;PILE-C, Progressive Isoinertial Lifting Evaluation-Cervical; PILE-L, Progressive Isoinertial Lifting Evaluation; PT, physical therapy; WAD, Whiplash Associated Disorder.

and predictive criterion validity of future work capacity in workers diagnosed with WAD I or II. Interclass correlation coefficients (ICC) ranged from 0.66 to 0.96 (moderate to excellent). Limits of agreement relative to mean performance ranged from 21% to 57% for functional based subtests. Correlations between FCE sub scores and baseline work capacity were very weak to weak ranging between r=0.06 and r=0.39. FCE sub scores did not predict future work capacity at 1, 3, 6 and 12 months.

Trippolini *et al* (2015)[29] assessed the WAD FCE (31) and evaluated convergent validity and known-groups validity. FCE subscales showed very weak to strong correlations (0.15–0.68) with each of: pain, self-reported functional ability, self-reported disability, anxiety and depression. It was found that the FCE had known-group sex validity (males vs females) for 1 of 3 functional subtests (lifting waist-overhead) and reported significant performance differences between culture groups (German vs non-German language groups). To test construct validity, 29 a priori formulated hypotheses were tested, 4 related to gender differences, 20 related associations with other constructs, 5 related to cultural differences. In total 23 out of 29 hypotheses were confirmed (79 %).

Work-related neck disorders

Reesink et al. (2007)[34] developed an independent FCE for patients with musculoskeletal neck disorders (neck FCE). They performed a review of epidemiological literature and identified four physical risk factors for work-related neck disorders and used that information to develop an FCE consisting of eight functional-based tests. Content validity was established by following operational definitions of the risk factors when searching the literature and using current literature to provide a rationale to guide their development of the tasks comprising the FCE.

Chronic neck pain

Reneman et al. (2017)[32] measured test-retest reliability of the subscales of the neck FCE in patients with

Table 2 Quality of studies on psychometric properties of functional-based tests evaluated in neck disorder patients

	Item evaluation criteria												
Study	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**	**Q9**	**Q10**	**Q11**	**Q12**	**Total (%)**
Trippolini *et al*[30]	2	2	2	2	1	2	2	2	2	2	1	2	92%
Lomond and Cote[35]	2	2	1	2	0	2	2	2	2	2	2	2	88%
Pierrynowski *et al*[36]	2	2	1	2	0	2	2	2	2	2	2	2	88%
Trippolini et al[29]	2	2	2	0	1	N/A	2	2	2	2	2	2	86%
Van der Meer *et al*[33]	2	1	2	1	2	N/A	2	1	2	2	1	2	86%
Ljungquist *et al*[37] KGV†	2	2	2	0	0	N/A	2	2	2	2	2	2	82%
Ljungquist *et al*[38] Rel§	2	1	1	2	0	2	2	2	2	2	1	2	79%
Ljungquist *et al*[40] STC‡	1	1	1	2	1	1	2	2	2	2	2	2	79%
Trippolini *et al*[31]	2	2	1	1	0	0	2	2	2	2	2	2	75%
Ljungquist *et al*[39] KGV†	2	1	1	2	0	N/A	2	1	2	1	1	2	68%
Reneman *et al*[32]	1	2	1	1	1	0	1	2	2	2	2	1	67%
Reesink[34]*	–	–	–	–	–	–	–	–	–	–	–	–	N/A

12-item evaluation tool (QACMRR) designed to assess the quality of studies determining measurement properties in outcome measures. Questions 1–12 in the tool evaluate aspects of study question, study design, measurements, analyses, and study recommendations.
*Paper is not applicable for completion of study quality tool
†KGV, known-groups validity
‡STC, sensitivity-to-change
§Rel, reliability
KGV, known-groups validity; rel, reliability; STC, sensitivity-to-change.

multifactorial neck pain. Test-retest ICC's ranged from poor to excellent (0.39–0.96). Limits of agreement relative to mean performance range from 32.0% to 56.5% for functional based sub tests. Convergent validity was performed against the Neck Disability Index (NDI) items and total score.[33] The authors found weak to strong Pearson correlations (0.39–0.70) for the FCE sub scores to both NDI individual items and the NDI total score.

Table 3 Psychometric properties of the functional capacity evaluation

FCE battery	Type of properties	Statistical test	Value	Interpretation
Neck FCE	Test-retest	ICC	0.39–0.96	Poor-excellent
	Measurement Error	Ratio of LoA	32.0%–56.5%	
	Convergent Validity	Pearson or Spearman correlation	NDI total: 0.39–0.62 NDI items: 0.03–0.63	Weak to moderate very weak to strong
WAD FCE	Test-retest Reliability	ICC	0.66–0.96	moderate-excellent
	Convergent Validity	Pearson Correlation	Pain* 0.31–0.39 SFS: 0.42–0.61 NDI: 0.34–0.45 HADS-A: 0.27–0.36 HADS-D: 0.30–0.41	Weak Moderate-strong Weak-moderate weak Weak-moderate
	Discriminative Validity (German vs Non-German)	Linear Regression Analysis	p<0.001	Significant for All Tasks
	Discriminative Validity (sex)	t-test	p<0.001	Significant for Two Tasks
Workwell FCE	Convergent Validity	Pearson or Spearman Correlation	Work Capacity: 0.1–0.3	Very Weak – weak

	Predictive Validity	Pearson or Spearman Correlation
		Linear Mixed Model Regression of All Predictors
			0.06–0.39
			β=−0.04, 95% CI: −0.15–0.06
			p=0.428 (task 6)
				Very weak - Weak
				Not Significant

*Pain measured via Numeric Rating Scale.
FCE, Functional Capacity Evaluation; HADS-A, Hospital Anxiety and Depression Scale – Anxiety; HADS-D, Hospital Anxiety and Depression Scale – Depression; ICC, Intraclass correlation coefficient; LoA, Limits of Agreement; Mod, Moderate; NDI, Neck Disability Index; Neg, Negligible; SFS, Spinal Function Sort; Sig, Significant.

Table 4 Summary of Fit-HaNSA's psychometric properties in neck disorder patients

Test	Type of property	Statistical test	Value	Interpretation
Fit-HaNSA	Intra-rater Reliability	ICC	0.78	good
Fit-HaNSA	Inter-rater Reliability	ICC	0.84	good
Fit-HaNSA	Measurement Error	SEM LOA_{95} MDC_{90}	76 s 248 s 176 s	
Fit-HaNSA	Convergent Validity	Spearman Rank Correlation*	<0.4 ->0.75	Weak – Strong
Fit-HaNSA	Discriminative WAD II vs Control	F-test	62.6,<p,0.001	Significant
Fit-HaNSA Functional Sub-tasks	Intra-rater reliability	ICC	0.70–0.72	moderate
	Inter-reliability	ICC	0.54–0.80	–moderate – good
	Convergent Validity	Spearman Rank Correlation*	<0.4 ->0.75	Weak - Strong
	Discriminative Validity WAD II vs Control	F-test	42.0–53.3, p<0.001	Significant

*Correlations completed with Numeric Pain Rating Scale, Neck Disability Index, Disabilities of Arm, Shoulder, Hand and six cervical range of motion tests

Fit-HaNSA, Functional Impairment Test, Hand and Neck/Shoulder/Arm; ICC, Intraclass correlation coefficient; LOA_{95}, 95% Limits of Agreement; MDC_{90}, 90% Minimal Detectable Change; Mod, Moderate; SEM, SE of Measurement; WAD, Whiplash Associated Disorder.

The BTEWS II

Chronic neck pain

Lomond and Côté, (2011)[35] reported on the reliability, measurement error, minimum detectable change (MDC) and validity of the power output (PO) task during the BTEWS II test in patients with chronic neck and shoulder pain (table 4). Test-retest reliability, measured with Spearman Rank correlations and ICC's was moderate and measured at ρ=0.37 and $ICC_{2,1}$ = 0.54, respectively. The SE of measurement (SEM) and the minimal detectable change at 90% confidence (MDC_{90}) for the PO task were measured as 30.25 and 70.59, respectively. Weak Spearman Rank correlations between the PO task and the NDI, Shoulder Pain and Disability Index (SPADI) and Numeric Rating Scale (NRS) for pain tests were recorded. There were no significant performance differences between control and pain groups for the PO task.

Functional Impairment Test-Hand and Neck/Shoulder/Arm

Subacute to chronic WAD

Pierrynowski *et al* (2016)[36] reported on the reliability, measurement error, MDC and validity of the Functional Impairment Test-Hand and Neck/Shoulder/Arm (Fit-HaNSA) test in a sample of people with WAD II following motor vehicle collision (MVC) (table 5). Intra-rater reliability ICC's for patient subtask and total scores were moderate to good ranging between 0.70–0.78.[36] Inter-rater reliability ICC's for patient subtask and total scores were moderate to good and ranged between 0.54 and 0.84.[36] The Bland and Altman plot for the patient group showed a 26 s (s) bias in terms of improved performance on the second test (possible learning effect). The SD of difference was 124 s and 95% Limits of Agreement (LoA_{95}) was 248 s.[36] The SEM for people with WAD II was reported to be 76 s. The MDC_{90} was measured as 176 s.[36]

Table 5 Psychometric properties of Baltimore Therapeutic Equipment work simulator II–Power Output task

Test	Type of property	Statistical test	Value	Interpretation
BTEWS II	Test-retest reliability	ICC	0.53	moderate
		Spearman	0.37	Poor
BTEWS II	Measurement Error	SEM	30.25	
		MDC_{90}	70.59	
BTEWS II	Convergent Validity	Spearman	Not Reported	Weak
BTEWS II	Discriminative Validity (Pain vs Control)	Two-way Repeated Measures ANOVA	Not Reported	Non-significant

*Spearman correlations completed with Numeric Rating Scale, Neck Disability Index and Shoulder Pain and Disability Index.

ANOVA, analysis of variance; ICC, intraclass correlation coefficient; MDC_{90}, 90% minimal detectable change; SEM, SE of measurement.

Spearman rank correlations were also calculated between the Fit-HANSA, Numeric Pain Rating Scale (NPRS), NDI, the disabilities of arm, hand and shoulder (DASH) and six cervical range of motion measures. Most (59 of 78) of the correlations between performance and comparator measures were very weak to weak (r=<0.4).[36] All correlations between total Fit-HaNSA scores and subtask scores had good correlations (r=<0.75), except for Task 1-Task 3.[36] Significant performance differences between WAD II and control groups (known group validity) were recorded for the total Fit-HaNSA score and all three subtask scores.[36]

Physiotherapy test package subtests

Ljungquist *et al* published a series of articles[37–40] which evaluated the clinimetric properties of a physiotherapy test package for patients with spinal pain (table 6). This package included muscular strength & endurance tests, submaximal endurance tests, and three functional tests. These functional tests included the PILE-C, PILE-L, and 2×20M-WWB test. Ljungquist's series of articles reported on convergent validity, known-groups validity, reliability, measurement error and sensitivity to change for these tests.[37–40]

Undetermined duration of neck pain

In a 1999 article,[39] correlations between the tests of the package and pain (CR-10) and perceived exertion (Borg RPE) were determined. All correlations were very weak to moderate (0.10–0.48) except for moderate to strong correlations (0.55–0.65) between the PILE-C test and pain intensity and between 2×20M-WWB test and pain intensity.

In a 2003 article,[37] the PILE-C, PILE-L and 2×20M-WWB tests were tested to determine their ability to discriminate between known-groups (neck pain vs back pain). Subjects with spinal pain completed the CR-10, the University of Alabama Pain Behaviour Scale (UAB) and the Borg RPE test. Specific cut points were used to distinguish patients with high vs low pain intensity, high vs low pain behaviour, and high versus low perceived exertion in patients, respectively. Participants then completed the test package and it was determined if each subtest could discriminate between participants with high vs low pain intensity. The PILE-C and the 2×20M-WWB tests were hypothesised to be more difficult for persons with neck pain and the PILE-L was hypothesised to be more difficult for persons with back pain. Subjects with neck pain performed worse on the PILE-C test compared with those with back pain. Subjects with back pain did not perform worse than those with neck pain on the PILE-L test and subjects with back pain performed worse on the 2×20M-WWB test.

The functional tests were able to discriminate between all three subgroups with the exception of the PILE-C being unable to discriminate between participants with high versus low perceived exertion.

In a paper from 1999,[39] the PILE-C, PILE-L and 2×20M-WWB tests were found to have significant discriminative abilities in distinguishing healthy subjects from patients with spinal pain. The sensitivity and specificity for this known group discrimination for the PILE-C test, were reported to be 0.93 (very strong) and 0.69 (strong), respectively. The sensitivity and specificity for the PILE-L test were reported to be 0.85 (very strong) and 0.65 (strong), respectively.

The inter and intra rater reliability were tested on participants with spinal pain.[38] Limits of agreement were used to measure inter rater reliability and repeatability, defined as 2x the within-subject SD of each variable. Inter-rater agreement for two tests was deemed 'acceptable', while all three functional tests had 'clinically acceptable' intra-rater reliability.

Sensitivity-to-change was evaluated in the test package following 6 months of a physiotherapy intervention. Using ROC curves, Wilcoxon sign ranked tests and spearman correlation coefficients, only the 2×20m-WWB test and the PILE-C (women only) were deemed to be sensitive to change.[40] Additionally, moderate to large effect sizes were found for all test components.

DISCUSSION

This study synthesised 12 studies assessing clinometric properties of 4 different functional-based assessments. Given the limited number of studies, the substantial variation in the types of tests examined, the methods used to assess the clinical measurement properties, and the study populations, the current state of knowledge does not allow firm conclusions regarding recommendations for an optimal functional-based test at this time. Overall, the quality ranging from good to excellent (67%–92%,) as determined by the QACMRR, for a range of properties of the four different assessments in patients with acute or chronic neck pain that is musculoskeletal in origin. Studies obtaining higher percentages indicate research that has been consistent with best practice where studies with lower percentages are more likely to be inadequate or inappropriate

Functional capacity evaluation

The breadth of a functional-based test is variable and defined by the developers. An advantage of the functional assessment designed by Reesink *et al*[34] is that they mapped the eight subtests to risk factors identified in the literature for work-related neck disorders. The eight subtests consist of: material handling tasks, lifting floor to waist, overhead lift test, one-handed and two-handed carrying, overhead working, repetitive reaching, overhead lifting, and repetitive bending and overhead reaching. Given the systematic approach and rationale these authors used in developing the FCE and this approach being used in previous research,[41] we suggest that this test has strong content validity.

Six articles address the clinical measurement properties of this FCE ranging from good to excellent quality (67%–92%). There was evidence that the FCE was stable over

Table 6 Psychometric properties of performance-based tests included in physiotherapy test package

Test	Type of property	Statistical test	Value	Interpretation
PILE-C	Inter-rater reliability	Mean difference LoA	► 0.24 ► 2.46 and 1.82	

PILE-C	Inter-rater reliability	Repeatability (2X SD) % of range	M=3.93; F=1.19
			M=10.5%; F=6.1%	

PILE-C	Convergent validity	Spearman correlation	CR-10: 0.55–0.65* Borg RPE: 0.10–0.48	Moderate - Strong very weak - moderate
PILE-C	Discriminative: spinal pain vs control	Sensitivity and specificity	0.93, 0.69	Strong – Very Strong
PILE-C	Discriminative: spinal pain vs control	Wilcoxon sign ranked test	p=0.008	Significant
PILE-C	Discriminative: high vs low pain intensity	Mann-Whitney U	p=0.003	Significant
PILE-C	Discriminative: high vs low Pain behaviour	Mann-Whitney U	p=0.005	Significant
PILE-C	Discriminative: high vs low perceived exertion	Mann-Whitney U	p=0.154	Non-significant
PILE-C	Sensitivity to change	Effect Size	Subjects improving: 0.39–0.73 Subjects deteriorating: 0–0.4	Small – Moderate Trivial – Small
PILE-L	Inter-rater reliability	Mean difference LoA	► 0.11 ► 2.33 and 2.11	

PILE-L	Intra-rater reliability	Repeatability
		% of range
			M=4.0; F=3.59
			M=10.7%; F=18.5%	

PILE-L	Convergent validity	Spearman correlation	CR-10: 0.11–0.45 Borg RPE: 0.10–0.48	very weak – moderate very weak – moderate
PILE-L	Discriminative: spinal pain vs no spinal pain	Sensitivity and specificity	0.85, 0.65	Strong – Very Strong
PILE-L	Discriminative: spinal pain vs control	Wilcoxon sign ranked test	p=0.002	Significant
PILE-L	Discriminative: high vs low pain intensity	Mann-Whitney U	p=0.001	Significant
PILE-L	Discriminative: high vs low pain behaviour	Mann-Whitney U	p<0.001	Significant
PILE-L	Discriminative: high vs low perceived exertion	Mann-Whitney U	p<0.001	Significant
PILE-L	Sensitivity to change	Effect size	Subjects improving: 0.02–1.08 Subjects deteriorating 0.42–0.81	Trivial – Large Small – Large
2×20 m WWB	Inter-rater reliability	Mean difference LoA	0.05 −1.33 and 1.43	
2×20 m WWB	Intra-rater reliability	Repeatability % of range	3.2 10.7%	
2×20 m WWB	Convergent validity	Spearman correlation	CR-10: 0.55–0.65Borg RPE: 0.10–0.48	Moderate - Strong very weak – moderate
2×20 m WWB	Discriminative: spinal pain vs control	Wilcoxon sign ranked test	p=0.014	Significant
2×20 m WWB	Discriminative: high vs low pain intensity	Mann Whitney U	p<0.001	Significant
2×20 m WWB	Discriminative: high vs low pain behaviour	Mann Whitney U	p<0.001	Significant

Continued

Table 6 Continued

Test	Type of property	Statistical test	Value	Interpretation
2×20 m WWB	Discriminative: high vs low perceived exertion	Mann Whitney U	p<0.001	Significant
2×20 m WWB	Sensitivity to change	Effect size	Subjects improving: 0.38–0.78 Subjects deteriorating: 0.13–0.62	Small – Moderate Trivial – Moderate

*CR-10: Measurement of pain construct

F, Female; KGV, Known-groups Validity; LoA, Limits of Agreement; M, Male; Mod., Moderate; Neg., Negligible;PILE-C, Progressive Iso-intertial Lifting Evaluation – Cervical; PILE-L, Progressive Iso-intertial Lifting Evaluation – Lumbar; RPE, Rating of perceived exertion.

test-retest time of 7–14 days.[31 32] These measures demonstrate longer stability over time compared with self-report measures such as the Neck Disability Index (NDI) which has demonstrated test-retest reliability within only a short period of 0–3 days.[28] Whether this longer-term stability is a characteristic of functional-based tests or reflects differences in study populations in context requires further testing. These two studies had relatively lower quality scores on the QACMRR (67%–75%) compared with other studies in this review putting into question test-retest time. Although test-retest reliability has been assessed, inter-rater and intra-rater reliability has yet to be researched. Unlike self-report measures, we expect measurement error due to the evaluator and functional-based tests. Thus, future research should explore these aspects of reliability.

Convergent validity is often examined in clinical measurement studies. We suggest that this may be because these comparisons are easily performed by correlating different tests rather than providing strong confidence in the validity of the measurement. Often convenient comparisons are performed rather than those most relevant. Across many domains and measures it has become clear that the relationship between self-reported function and performance-based function or physical impairment is often very weak to moderate. Therefore, the value of assessment of these relationships as a form of validation has limited value. Several studies of very good to excellent quality have reported on the convergent validity of the FCE.[29 30 33] The highest quality article determined by the QACMRR (92%) found the relationship between the FCE and work capacity to be poorly associated with one another.[30] The same study found that the ability of the FCE to predict future work capacity was poor. This may be considered a more important comparison since ideally functional-based tests would relate to important outcomes like return to work. No studies to our knowledge report the responsiveness or sensitivity to change of the FCE. This is an important gap since the focus of rehabilitation is often to remediate limitations in goal impairments or work capacity, and assessment of these changes is critical to clinical decision-making and reporting outcomes. Thus, future research should evaluate the responsiveness of the FCE to provide insight in the measure's ability to detect change after an intervention.

Functional Impairment Test-Hand and Neck/Shoulder/Arm

One study of very good quality (88%) assessed the FIT-HaNSA, a test consisting of two reaching tasks (waist and eye-level) and sustained overhead task performance.[36] Overall, the FIT-HaNSA demonstrated excellent inter-rater reliability (0.84) and intra-rater reliability (0.78). The specific subtests included within the FIT-HaNSA similarly demonstrate fair to excellent (0.54–0.80) and good (0.70–0.72) inter-rater and intra-rater reliability respectively. The FIT-HaNSA also demonstrated a clear ability to distinguish between people with WAD two and healthy controls. Correlations between the FIT-HaNSA and other patient self-report disability and functional outcome measures (NPRS, NDI, DASH, CROM and FIT-HaNSA) were generally very weak to weak (ρ<0.4), consistent with other studies comparing performance and self-report.[17 18] The largest limitation in critically synthesising information for this test is that only a single study was found that reported the measurement properties for people with neck disorders. It should be noted however that it has been validated in other MSK disorders.[35 41] Although others have noted the lag in development of functional-based measures in comparison to self-report measures, FIT-HaNSA was recommended as a functional-based measure for people with shoulder disorders.[42] Further research is necessary to investigate the responsiveness of the FIT-HaNSA.

Baltimore Therapeutic Equipment work simulator II

Another study of very good quality (88%) assessed the efficacy of the BTEWS II where the participants performed a dynamic pushing and pulling task in which power output was recorded over a 10 s sample.[35] While the convergent validity aspect of this paper was assessed as consistent with best practice through the critical appraisal process, the relationship between the power output on the BTEWS and measures of pain and disability (NDI, SPADI, NRS) were poorly associated with each other. In addition, the power output component was not found to be significantly different between people with neck pain and healthy controls which suggests it might not be discriminative.

Discrimination between patients and healthy controls is a low standard for an outcome measure, and tests that cannot fulfil this benchmark should be viewed with caution. Because of the weak measurement properties demonstrated by the power output component of the BTEWS II, it does not appear to be a desirable functional-based measure to assess function in people with neck pain. However, we acknowledge for all of the functional-based tests the evidence pool is so shallow that there is high potential that future studies might lead to different conclusions. Future research should also investigate the reliability and responsiveness of the BTEWS II.

Physiotherapy test package subtests

Four studies ranging from good to very good quality (68%–82%) assessed relevant items from a physiotherapy test package, including a lift from floor-to-waist and a waist-to-shoulder task and a two-handed carrying task. The properties of these assessment items include weak to moderate correlations to pain, perceived exertion, and had "fair to good" reliability. The 2×20m-WWB and PILE-C tests were found to be sensitive-to-change which is valuable information as no other study has assessed this property in functional-based measures in patients with neck disorders. Thus, this measure may be of value in clinical settings when assessing functional capacity before and after a treatment intervention. All tests had discriminative ability for detecting participants with spinal pain vs healthy controls. Most of the three tests demonstrated poor construct validity in that they were poorly related to pain and perceived exertion and the results were not in accordance with pre-defined hypotheses. Thus, further research is necessary to investigate these constructs. Three of the four results from the studies assessing the physiotherapy test package had a mixed sample of patients with various pain sites including back pain. While the majority of each cohort in these studies had neck pain, careful consideration should be taken to apply these tests to a neck pain specific population.

Clinical implications

This study confirms that functional-based tests have had far less development and evaluation than self-report measures. Limitations include the number of tests and insufficient body of evidence to make confident recommendations with respect to functional-based testing. It is clear that self-report and functional-based measures provide different perspectives. Theoretically, functional-based tests are important to inform our understanding about the mechanisms of intervention and how interventions increase capacity. Future research may benefit by also comparing results from a functional-based measure to work capacity to when assessing construct validity. Overall more work is required to further establish the psychometric properties of functional-based tests in persons with neck disorders, including sensitivity-to-change, responsiveness, and predictive validity.

The FCE evaluated patients with neck pain of varying origin including WAD, work-related neck disorders, and chronic idiopathic neck pain. The BTEWs II evaluated functional capacity in patients with chronic neck pain, the FIT-HaNSA evaluated patients with WAD, and the physiotherapy test package did not specify the origin of musculoskeletal neck pain in their cohort. Thus, specific functional-based measures may be more applicable depending on the origin of the musculoskeletal neck pain being assessed.

The data presented suggest that the FIT-HaNSA has the strongest clinometric properties though this is based on a single higher quality paper specific to neck disorder.[36] Importantly, normative data have been published,[43] it has been validated in multiple studies in patients with shoulder conditions[44–46] and has been recommended when compared with other measures.[42] The FCE has a limited evidence base from which to draw, though it was developed with strong content validity and further evaluation may demonstrate its usefulness.

Limitations

A challenge in synthesising clinical measurement evidence is the wide range of properties and indicators that need to be considered. Unlike effectiveness studies where one can focus on the effect size of treatment there are many considerations that would affect the recommendations made about outcome measures. This is further complicated when the pool of evidence is shallow. Although the quality assessment tool (QACMRR) developed by one of the authors of this review which assess the quality of design of individual studies were useful for interpreting the evidentiary pool, there is no clear method to synthesise the extracted clinical measurement evidence. While some systematic reviews on treatment might only report findings from high-quality studies, it is important to see how outcome measures perform in different contexts. Further, the assessment of quality is complicated given that clinical measurement studies have so many dimensions. Therefore, exclusion of lower quality studies has questionable value. Thus, a more practical approach is to consider quality when interpreting the findings, rather than excluding studies.

The QACMRR focuses on whether the authors made appropriate decisions in selecting the scope and methods of their clinical measurement evaluations within a given study and provides descriptors of poor fair or good design options. Quality focuses on issues that might affect risk of bias or imprecision in estimates; whereas risk of bias assessments focusses on items that might result in a biassed estimate. For example, insufficient power is a precision (quality) issue, not a risk of bias. Although it is difficult to interpret the meaning of the percentage of the QACMRR as there are no established cut-offs for distinguishing good and poor-quality studies, it provides one way of ranking the articles in order of quality. We did not use COSMIN checklist since it was developed for

PROMS and some of the components/steps that involved are not applicable to performance-based tests.

Another limitation in this review was that the feasibility or usability of these tools was not assessed. While feasibility was not the focus of this review, information on the practical application of these functional-based measures provides valuable information to clinicians for determining whether these tests are appropriate to use in their given setting. Thus, future research should not only investigate further the psychometric properties of these tools, but also report the feasibility of using these tests so that they may be used in clinical settings and to identify limitations that restrict their application in practice.

CONCLUSION

This review found very good quality evidence that the FIT-HaNSA has excellent inter and intra-rater reliability and very weak to weak convergent validity. Excellent quality evidence of fair test-retest reliability, weak convergent validity, and very weak known groups validity for the BTEWS II test was found. Good to excellent quality evidence exists that an FCE battery has poor to excellent reliability and very weak to strong validity. Good to excellent quality of weak to strong validity and trivial to strong effect sizes were found for a physiotherapy test package. Functional-based evaluation in people with neck disorders is an area needing much research attention both to establish the measurement properties of existing measures, potentially to develop innovative new measures and to perform head-to-head comparisons of measures before an optimal functional-based test can be identified.

Twitter Pavlos Bobos @pavlosbob

Collaborators CATWAD Co-authors: Michele Sterling m.sterling@uq.edu.au, Anne Söderlund anne.soderlund@mdh.se, Michele Curatolo curatolo@uw.edu, James M Elliott j-elliott@northwestern.edu, David M Walton dwalton5@uwo.ca, Helge Kasch helgkasc@rm.dk, Linda Carroll linda.carroll@ualberta.ca, Hans Westergren Hans.Westergren@skane.se, Samuel A McLean Samuel_McLean@med.unc.edu, Gwendolen Jull g.jull@uq.edu.au, Genevieve Grant genevieve.grant@monash.edu, Luke Connelly l.connelly@uq.edu.au, Joy C MacDermid jmacderm@uwo.ca, Mandy Nielsen mandy.nielsen@griffith.edu.au, Pierre Côté pierre.cote@uoit.ca, Tonny Elmose Andersen tandersen@health.sdu.dk, Trudy Rebbeck trudy.rebbeck@sydney.edu.au, Annick Maujean a.maujean@uq.edu.au, Sarah Robins s.robins1@uq.edu.au, Kenneth Chen k.chen8@uq.edu.au, Julia Treleaven j.treleaven@uq.edu.au

Contributors SM contributed significantly to conception and design of the study, data extraction, critical appraisal, interpretation of data and drafting of the manuscript. TS, TA, PB and CC were involved in literature search, critical appraisal and interpretation of data and drafting. AG was involved in critical appraisal and drafting. JM was also involved in the conception and design of the study, drafting and revised the manuscript for important intellectual content. PB and CATWAD were involved in the drafting and review of the manuscript. All authors have given their final approval on the manuscript to be published.

Funding This work was supported by the Canadian Institutes of Health Research (CIHR) with funding reference number (FRN: SCA-145102).

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available.

ORCID iD

Pavlos Bobos http://orcid.org/0000-0002-5098-4840

REFERENCES

1 Carroll LJ, Hogg-Johnson S, van der Velde G, *et al*. Course and prognostic factors for neck pain in the general population: results of the bone and joint decade 2000-2010 Task force on neck pain and its associated disorders. *J Manipulative Physiol Ther* 2009;32:S87-96.

2 Croft PR, Lewis M, Papageorgiou AC, *et al*. Risk factors for neck pain: a longitudinal study in the general population. *Pain* 2001;93:317–25.

3 VosT, AllenC, AroraM, *et al*. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: a systematic analysis for the global burden of disease study 2015. *Lancet* 2016;388:1545–602.

4 Blanpied PR, Gross AR, Elliott JM, *et al*. Neck pain: revision 2017. *J Orthop Sports Phys Ther* 2017;47:A1–83.

5 Nygren A, Berglund A, von Koch M. Neck-and-shoulder pain, an increasing problem. strategies for using insurance material to follow trends. *Scand J Rehabil Med Suppl* 1995;32:107–12.

6 Wright A, Mayer TG, Gatchel RJ. Outcomes of disabling cervical spine disorders in compensation injuries. A prospective comparison to tertiary rehabilitation response for chronic lumbar spinal disorders. *Spine* 1999;24:178-83.

7 Rempel DM, Harrison RJ, Barnhart S. Work-Related cumulative trauma disorders of the upper extremity. *JAMA* 1992;267.

8 Borghouts JA, Koes BW, Vondeling H, *et al*. Cost-Of-Illness of neck pain in the Netherlands in 1996. *Pain* 1999;80:629–36.

9 Hogg-Johnson S, van der Velde G, Carroll LJ, *et al*. The burden and determinants of neck pain in the general population: results of the bone and joint decade 2000-2010 Task force on neck pain and its associated disorders. *J Manipulative Physiol Ther* : 2009;32:S46-60.

10 Bobos P, Nazari G, Palimeris S, *et al*. The contribution of health and psychological factors in patients with chronic neck pain and disability: a cross-sectional study. *JCDR* : 2018.

11 Bobos P, Billis E, Papanikolaou D-T, *et al*. Does deep cervical flexor muscle training affect pain pressure thresholds of myofascial trigger points in patients with chronic neck pain? A prospective randomized controlled trial. *Rehabil Res Pract* 2016;2016:1–8.

12 Nazari G, Bobos P, Billis E, *et al*. Cervical flexor muscle training reduces pain, anxiety, and depression levels in patients with chronic neck pain by a clinically important amount: a prospective cohort study. *Physiother Res Int* 2018;23.

13 Bobos P, MacDermid JC, Walton DM, *et al*. Patient-Reported outcome measures used for neck disorders: an overview of systematic reviews. *J Orthop Sports Phys Ther* 2018;48:775–88.

14 MacDermid JC, Walton DM, Bobos P, *et al*. A qualitative description of chronic neck pain has implications for outcome assessment and classification. *Open Orthop J* 2016;10:746–56.

15 Childs JD, Cleland JA, Elliott JM, *et al*. Neck pain: clinical practice guidelines linked to the International classification of functioning, disability, and health from the orthopedic section of the American physical therapy association. *J Orthop Sports Phys Ther* 2008;38:A1-A34.

16 Kay TM, Huijbregts M. *Physical rehabilitation outcome measures: a guide to enhanced clinical decision making*. Second Edition. Canada: Physiother, 2003.

17 Simmonds MJ, Olson SL, Jones S, *et al*. Psychometric characteristics and clinical usefulness of physical performance tests in patients with low back pain. *Spine* 1998;23:2412–21.

18 Stratford PW, Kennedy D, Pagura SMC, *et al*. The relationship between self-report and performance-related measures: Questioning the content validity of timed tests. *Arthritis Rheum* 2003;49:535–40.

19 Novy DM, Simmonds MJ, Lee CE. Physical performance tasks: what are the underlying constructs? *Arch Phys Med Rehabil* 2002;83:44–7.

20 MacDermid JC, Stratford P. Applying evidence on outcome measures to hand therapy practice. *J Hand Ther* 2004;17:165–73.

21 Alreni ASE, Harrop D, Lowe A, *et al*. Measures of upper limb function for people with neck pain. A systematic review of measurement and practical properties. *Musculoskelet Sci Pract* 2017;29:155–63.

22 Moher D, Shamseer L, Clarke M, *et al*. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev* 2015;4:1.

23 Law MC, MacDermid J. *Evidence-based rehabilitation : a guide to practice*. Thorofare, NJ: Slack Incorporated, 2014.

24 Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 2016;15:155–63.

25 Roy J-S, Desmeules F, MacDermid JC. Psychometric properties of presenteeism scales for musculoskeletal disorders: a systematic review. *J Rehabil Med* 2011;43:23–31.

26 Divaris K, Vann WF, Baker AD, *et al*. Examining the accuracy of caregivers' assessments of young children's oral health status. *J Am Dent Assoc* 2012;143:1237–47.

27 Cohen J. *Statistical power analysis for the behavioral sciences*, 1988.

28 MacDermid JC, Walton DM, Avery S, *et al*. Measurement properties of the neck disability index: a systematic review. *J Orthop Sports Phys Ther* 2009;39:400–12.

29 Trippolini MA, Dijkstra PU, Geertzen JHB, *et al*. Construct validity of functional capacity evaluation in patients with Whiplash-Associated disorders. *J Occup Rehabil* 2015;25:481–92.

30 Trippolini MA, Dijkstra PU, Côté P, *et al*. Can functional capacity tests predict future work capacity in patients with whiplash-associated disorders? *Arch Phys Med Rehabil* 2014;95:2357–66.

31 Trippolini MA, Reneman MF, Jansen B, *et al*. Reliability and safety of functional capacity evaluation in patients with whiplash associated disorders. *J Occup Rehabil* 2013;23:381–90.

32 Reneman MF, Roelofs M, Schiphorst Preuper HR. Reliability and agreement of neck functional capacity evaluation tests in patients with chronic multifactorial neck pain. *Arch Phys Med Rehabil* 2017;98:1476–9.

33 van der Meer S, Reneman MF, Verhoeven J, *et al*. Relationship between self-reported disability and functional capacity in patients with whiplash associated disorder. *J Occup Rehabil* 2014;24:419–24.

34 Reesink DD, Jorritsma W, Reneman MF. Basis for a functional capacity evaluation methodology for patients with work-related neck disorders. *J Occup Rehabil* 2007;17:436–49.

35 Lomond KV, Côté JN. Shoulder functional assessments in persons with chronic neck/shoulder pain and healthy subjects: reliability and effects of movement repetition. *Work* 2011;38:169–80.

36 Pierrynowski M, McPhee C, P Mehta S, *et al*. Intra and inter-rater reliability and convergent validity of FIT-HaNSA in individuals with grade Π whiplash associated disorder. *Open Orthop J* 2016;10:179–89.

37 Ljungquist T, Jensen IB, Nygren A, *et al*. Physical performance tests for people with long-term spinal pain: aspects of construct validity. *J Rehabil Med* 2003;35:69–75.

38 Ljungquist T, Harms-Ringdahl K, Nygren A, *et al*. Intra- and inter-rater reliability of an 11-test package for assessing dysfunction due to back or neck pain. *Physiother Res Int* 1999;4:214–32.

39 Ljungquist T, Fransson B, Harms-Ringdahl K, *et al*. A physiotherapy test package for assessing back and neck dysfunction--discriminative ability for patients versus healthy control subjects. *Physiother Res Int* : 1999;4:123–40.

40 Ljungquist T, Nygren A, Jensen I, *et al*. Physical performance tests for people with spinal pain--sensitivity to change. *Disabil Rehabil* 2003;25:856–66.

41 Reneman MF, Dijkstra PU, Westmaas M, *et al*. Test-Retest reliability of lifting and carrying in a 2-day functional capacity evaluation. *J Occup Rehabil* 2002;12:269–75.

42 Hegedus EJ, Vidt ME, Tarara DT. The best combination of physical performance and self-report measures to capture function in three patient groups. *Phys Ther Rev* 2014;19:196–203.

43 Roy J-S, Macdermid JC, Boyd KU, *et al*. Rotational strength, range of motion, and function in people with unaffected shoulders from various stages of life. *Sports Med Arthrosc Rehabil Ther Technol* 2009;1:4.

44 Kumta P, MacDermid JC, Mehta SP, *et al*. The FIT-HaNSA demonstrates reliability and convergent validity of functional performance in patients with shoulder disorders. *J Orthop Sports Phys Ther* 2012;42:455–64.

45 MacDermid JC, Ghobrial M, Quirion KB, *et al*. Validation of a new test that assesses functional performance of the upper extremity and neck (FIT-HaNSA) in patients with shoulder pathology. *BMC Musculoskelet Disord* 2007;8:42.

46 Hawkes DH, Alizadehkhaiyat O, Fisher AC, *et al*. Normal shoulder muscular activation and co-ordination during a shoulder elevation task based on activities of daily living: an electromyographic study. *J Orthop Res* 2012;30:53–60.