

BMJ Open

BMJ Open is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers' comments and the authors' responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Open is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (<http://bmjopen.bmj.com>).

If you have any questions on BMJ Open's open peer review process please email info.bmjopen@bmj.com

BMJ Open

Systematic Review of the Measurement Properties of Performance-based Functional Tests in Patients with Neck Disorders

Journal:	<i>BMJ Open</i>
Manuscript ID	bmjopen-2019-031242
Article Type:	Research
Date Submitted by the Author:	24-Apr-2019
Complete List of Authors:	<p>McGee, Steven; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Sipos, Taylor; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Alin, Thomas; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Chen, Celia; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Greco, Alexandra; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Bobos, Pavlos; Western University, Health and Rehabilitation Sciences; University of Toronto, Dalla Lana School of Public Health, Institute of Health Policy Management and Evaluation MacDermid, Joy ; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Group, CATWAD; Michele Sterling m.sterling@uq.edu.au, Anne Söderlund anne.soderlund@mdh.se, Michele Curatolo, curatolo@uw.edu, Jim Elliott j-elliott@northwestern.edu, David Walton dwalton5@uwo.ca, Helge Kasch helgkasc@rm.dk, Carroll, Linda linda.carroll@ualberta.ca Hans.Westergren@skane.se, McLean, Samuel A 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, MacDermid, Joy jmacderm@uwo.ca, Mandy Nielsen mandy.nielsen@gri</p>
Keywords:	functional, psychometric properties, neck pain, cervical, outcome measures

SCHOLARONE™
Manuscripts

1
2
3 1 **Title:** Systematic Review of the Measurement Properties of Performance-based Functional
4 Tests in Patients with Neck Disorders

5
6
7 3 ¹Steven McGee, PT

8
9 4 ²Taylor Sipos, PT

10
11 5 ³Thomas Allin, PT

12
13 6 ⁴Celia Chen, PT

14
15 7 ⁵Alexandra Greco, PT

16
17 8 ⁶Pavlos Bobos, PT, PhD(c) (corresponding author)

18
19 9 ⁷Joy MacDermid, PT, PhD

20
21 10 ⁸CATWAD

22
23 11

12 **Authors' information**

24
25 13 ¹Steven McGee PT, School of Physical Therapy, Department of Health and Rehabilitation
26 Sciences, Western University, London, Ontario, Canada, (smcgee7@uwo.ca)

27
28
29 15

30
31 16 ²Taylor Sipos PT, School of Physical Therapy, Department of Health and Rehabilitation Sciences,
32 Western University, London, Ontario, Canada, (jsipos@uwo.ca)

33
34
35 18

36
37 19 ³Thomas Allin PT, School of Physical Therapy, Department of Health and Rehabilitation Sciences,
38 Western University, London, Ontario, Canada, (tallin@uwo.ca)

39
40
41 21

42
43 22 ⁴Celia Chen PT, School of Physical Therapy, Department of Health and Rehabilitation Sciences,
44 Western University, London, Ontario, Canada, (qchen224@uwo.ca)

45
46
47 24

48
49 25 ⁵Alexandra Greco PT, School of Physical Therapy, Department of Health and Rehabilitation
50 Sciences, Western University, London, Ontario, Canada, (agreco33@uwo.ca)

51
52
53 27

54
55 28 ⁶Pavlos Bobos PT, PhD(c), (corresponding author) Doctoral Candidate, Western's Bone and Joint
56 Institute, Department of Health and Rehabilitation Sciences, Western University, Elborn College,
57 1201 Western Road, N6G 1H1, London, Ontario, Dalla Lana School of Public Health, Institute of

31 Health Policy Management and Evaluation, Department of Clinical Epidemiology, University of
32 Toronto, Canada, (pbobos@uwo.ca), tel: +1 519 661 2111 x88912

33
34 ⁷Joy C MacDermid BScPT, PhD, Professor, Physical Therapy and Surgery, Western University,
35 London, ON and Co-director Clinical Research Lab, Hand and Upper Limb Centre, St. Joseph's
36 Health Centre, London, Ontario; Professor Rehabilitation Science McMaster University,
37 Hamilton, ON, Canada (jmacderm@uwo.ca)

38
39 ⁵CATWAD Coauthors: Michele Sterling m.sterling@uq.edu.au, Anne Söderlund
40 anne.soderlund@mdh.se, Michele Curatolo, curatolo@uw.edu, Jim Elliott [j-](mailto:j-elliott@northwestern.edu)
41 elliott@northwestern.edu, David Walton dwalton5@uwo.ca, Helge Kasch helgkasc@rm.dk,
42 Carroll, Linda linda.carroll@ualberta.ca Hans.Westergren@skane.se, McLean, Samuel A
43 Samuel_McLean@med.unc.edu, Gwendolen Jull g.jull@uq.edu.au, Genevieve Grant
44 genevieve.grant@monash.edu, Luke Connelly l.connelly@uq.edu.au, MacDermid, Joy
45 jmacderm@uwo.ca, Mandy Nielsen mandy.nielsen@griffith.edu.au, Pierre Cote
46 pierre.cote@uoit.ca, Tonny Elmoose Andersen tandersen@health.sdu.dk, Trudy Rebbeck
47 trudy.rebbeck@sydney.edu.au, Annick Maujean a.maujean@uq.edu.au, Sarah Robins
48 s.robins1@uq.edu.au, Kenneth Chen k.chen8@uq.edu.au, Julia Treleaven
49 j.treleaven@uq.edu.au

50
51 **Key Words:** functional, psychometric properties, neck, cervical, outcome measures

52
53 **Word Count:** 4239

54
55
56
57
58
59
60

61 Abstract

62 **Objective:** The purpose of this systematic review is to identify and synthesize studies evaluating
63 performance-based outcome measures designed to evaluate the functional abilities of patients
64 with mechanical neck pain.

65 **Setting:** Not applicable

66 **Participants:** Participants with neck disorders

67 **Methods:** A literature search using PubMed, Scopus, CINAHL, Embase, COCHRANE, Google
68 Scholar, and a citation mapping strategy was conducted through June 2018. Selected articles
69 were appraised using the COSMIN risk of bias checklist tool and the Quality Appraisal for
70 Clinical Measurement Research Reports Evaluation Form (QACMRR). Relevant data were then
71 extracted from selected articles using an extraction guide.

72 **Results:** The search obtained 12 articles which reported on 4 outcome measures reporting to
73 assess the functional abilities in patients with mechanical neck pathology. Of the selected papers:
74 1 reports content validity, 5 construct validity, 4 reliability, 1 sensitivity to change, and 1 both
75 reliability and construct validity. COSMIN sub-scores ranged from “inadequate” to “very good”
76 and QACMRR scores ranged from 68% to 95%.

77 **Conclusions:** A limited number of performance-based tests have been developed or validated
78 for assessing neck function. The pool of research in this area is sparse and insufficient to make
79 conclusive recommendations.

80 **Prospero registration:** CRD42018112358

81

82

83 **Strengths and limitations of this study**

- 84 • The psychometric properties of performance outcome measures for neck pain were
85 synthesized and critically appraised
- 86 • This study assessed the risk of bias and the quality of measurements properties
- 87 • The feasibility or usability of these tools was not assessed

89 Introduction

90 Neck pain has been associated with high disability and is regarded as a substantial
91 societal burden. (1,2) Approximately 70% of people experience neck pain within their lifetime
92 and about 33% of adults experience neck pain every year. (3,4) Further concern is warranted as it
93 has been suggested that the incidence of neck pain is increasing. (5,6,7) The economic burden
94 due to neck disorders is high, including lost wages, costs of treatment, and compensation
95 expenditures to injured people. (8,9) Neck pain is second only to low back pain in annual
96 workers' compensation costs in the United States.(7)

97 Outcome measures are a crucial component in monitoring patients with neck pain to
98 determine the effects of treatment, evaluation of interventions, guiding return to work, and
99 justifying treatment. Several self-reported outcome measures currently exist to assess disability
100 and function in those with neck pain (e.g. the Neck Disability Index (NDI) or the numeric pain
101 rating scale (NPRS). (10) Evidence-based clinical practice guidelines suggest that measures
102 assessing physical performance should also be used for people with neck pain. (11)
103 Performance-based testing is where the assessment is based on actual performance of a task or
104 activity. Physical performance can be assessed by testing a person's ability to execute a
105 standardized activity in a standardized environment (i.e. clinical setting). (12) Time to complete
106 the activity, number of repetitions performed, and weight lifted are frequently used to quantify

1
2
3 107 the physical performance. (13) Conversely, self-report measures examine patients' perception
4
5 108 and experience of their ability to perform functional tasks. (12) Previous research has
6
7
8 109 demonstrated poor to fair relationships between physical performance and self-report measures
9
10 110 of ability in patients with various musculoskeletal disorders suggesting that these measures
11
12 111 assess different constructs of function. (13,14) Consequently, physical performance tests and
13
14 112 self-report measures complement each other and may each contribute unique information about a
15
16
17 113 patient's function. (15)

18
19 114 A fundamental component of monitoring outcomes is having reliable and valid tools
20
21 115 with known measurement properties. (16,17) While recent research has investigated the
22
23 116 psychometric properties of patient-reported outcomes in people with neck pain (1,10, 18,19,20)
24
25
26 117 there is a gap in knowledge with respect to performance-based functional outcomes. The purpose
27
28 118 of this systematic review was to identify and synthesize clinical measurement studies that
29
30 119 evaluate psychometric properties of performance-based functional tests in patients with neck
31
32
33 120 disorders.

34
35 121

36 37 122 **METHODS**

38 39 123 **Patient and Public Involvement**

40
41
42 124 No patient involved
43
44
45 125

46 47 126 **Study Design and Protocol Registration**

48
49 127 We conducted a systematic review to evaluate the psychometric properties of
50
51 128 performance-based functional tests for people with mechanical neck disorders. The protocol was
52
53
54 129 registered in PROSPERO register with registration number CRD42018112358.
55
56
57
58
59
60

130

131 **Search Strategy**

132 A database search using CINAHL, PubMed, Scopus and Google Scholar was performed
133 to identify articles published before July 2018. The following search strategy was used to search
134 all databases for eligible studies: (Reliability OR validity OR responsiveness OR calibration OR
135 validation OR (minimal detectable change) OR (clinically important difference) OR
136 (psychometric properties) AND cervical OR neck OR c-spine AND (performance measure) OR
137 (functional test) OR (functional outcome) OR (performance outcome)). A citation map of articles
138 and systematic reviews selected for the full-text review was performed. This strategy was
139 included to minimize the risk of publication bias. The Preferred Reporting Items for Systematic
140 Reviews and Meta-Analyses (PRISMA) process (21) was followed to ensure all appropriate
141 steps were taken in the selection process (**FIGURE 1**).

142

143 **Inclusion and Exclusion Criteria**

144 Articles were included in the final review if all of the following criteria were met: 1)
145 >50% of the study's patient population had neck pain or a musculoskeletal neck disorder 2)
146 Patients in the study completed a functional-based test 3) Clinometric properties of at least one
147 performance-based test were reported. Definitions for the properties can be found in

148 **APPENDIX A.**

149

150 **Article Selection**

151 Titles and abstracts generated by the search strategy were screened by two authors
152 independently. Articles that met the inclusion criteria and selected for a full text review were also

1
2
3 153 reviewed in pairs of authors. Disagreements were resolved by the most experienced author
4
5 154 (JCM)
6
7
8 155

9 10 156 **Data Extraction**

11
12 157 Data extraction and critical appraisal was performed in pairs of two raters among the
13
14 158 authors, after the completion of a calibration session. When reviewers disagreed during data
15
16 159 extraction and/or critical appraisal, and consensus could not be met, a third author arbitrated. A
17
18 160 data extraction form (17) (**APPENDIX A and APPENDIX B**), developed by one of the authors
19
20 161 (JCM.), was used to ensure systematicity. Authors extracted sample size, patient population
21
22 162 characteristics, functional tests performed and reported psychometric properties.
23
24
25
26 163

27 28 164 **Risk of Bias and Quality Assessment**

29
30 165 Two authors used the Consensus-based Standards for the selection of health
31
32 166 Measurement Instruments (COSMIN) (22) checklist to assess risk of bias in the articles selected
33
34 167 for publication. The COSMIN checklist was recently adapted to evaluate risk of bias in studies
35
36 168 on measurement properties of patient reported outcome measures (PROMs). (22) After
37
38 169 completing a calibration session, each article was scored on the 4-point scale as “very good”,
39
40 170 “adequate”, “doubtful” or “inadequate” for each of the checklist criteria for relevant
41
42 171 measurement properties (e.g. reliability, responsiveness, etc.). To determine the overall score for
43
44 172 each measurement property, the worst score counts method was used wherein the lowest score
45
46 173 for the checklist criteria of the relevant property was taken as the overall score. (23) Pairs of
47
48 174 authors critically appraised the quality of each study using a standardized 12-item evaluation tool
49
50 175 (QACMRR) designed to assess the quality of studies determining measurement properties in
51
52
53
54
55
56
57
58
59
60

1
2
3 176 outcome measures (**APPENDIX C**). (24) Total scores on the tool can range from 0 to 24, with a
4
5 177 higher score indicating higher quality. Scores can be normalized to range between 0-100%. This
6
7
8 178 tool has been found to have good to excellent pre-consensus inter-rater reliability (ICC: 0.69-
9
10 179 0.91) across a number of systematic reviews. (17,24-28) Raw scores were converted to
11
12 180 standardized percentage scores and ranked based on percentage values. There were no formal
13
14
15 181 mechanisms developed to weight the studies based on quality scores.
16
17 182

19 183 **RESULTS**

21
22 184 The search strategy resulted in 840 published articles. After duplications were removed,
23
24 185 31 articles were deemed relevant and were screened at full text. Overall, 12 articles met our
25
26 186 inclusion criteria (**FIGURE 1**). The characteristics of the included studies and the summary of
27
28 187 psychometric properties are presented in **TABLE 1**. The risk of bias and the quality assessment
29
30 188 is summarized and presented in **TABLE 2-3**. The 12 articles that were included for review
31
32 189 provided properties on the following performance based tests: Functional Capacity Evaluations
33
34 190 (FCE) (29,30,31,32,33,34), The Baltimore Therapeutic Equipment Work Simulator II (BTEWS
35
36 191 II) (35), Functional Impairment Test- Hand and Neck/Shoulder/Arm (FIT-HaNSA) (36), as well
37
38 192 as items off of a physiotherapy test package including a cervical and lumbar Progressive
39
40
41
42 193 Isoinertial Lifting Evaluation (PILE-C, PILE-L) test (37,38,39,40) and 2 x 20 m with burden
43
44 194 walking test (2x20M-WWB) (37,38,39,40). Descriptions of all performance-based tests and their
45
46 195 relevant subtasks are provided in **APPENDIX D**.

48
49 196

51 197 **FCE**

1
2
3 198 Six articles reported measurement properties for a FCE battery. We identified multiple
4
5 199 versions of the FCE in the literature with one article reporting properties on the Workwell FCE
6
7
8 200 (30), two reporting on the Whiplash Associated Disorder (WAD) FCE (29,31) and three
9
10 201 reporting on the neck-FCE. (32,33,34) These test batteries include various combinations of
11
12 202 muscular strength, endurance and functional based tests. The measurement properties of the
13
14 203 functional based tests used by the FCE are outlined in **TABLE 4**.

15
16
17 204 An article evaluating the Workwell FCE (30) reported convergent validity and predictive
18
19 205 criterion validity of future work capacity in workers diagnosed with WAD I or II. Correlations
20
21 206 between FCE sub scores and baseline work capacity ranged between $r=0.06$ and $r=0.39$. FCE
22
23 207 subscores did not predict future work capacity at 1, 3, 6 and 12 months.

24
25
26 208 An article evaluating the WAD FCE (29) evaluated test-retest reliability and
27
28 209 measurement error in sick listed workers diagnosed with WAD grade 1 or 2. Interclass
29
30 210 Correlation Coefficients (ICC) ranged from 0.66 to 0.96 (moderate to excellent). Limits of
31
32 211 agreement relative to mean performance ranged from 21 to 57% for functional based sub-tests.
33
34 212 Another WAD FCE article (31) evaluated convergent validity and known-groups validity. FCE
35
36 213 subscales showed small to moderate correlations with each of: pain, self-reported functional
37
38 214 ability, self-reported disability, anxiety and depression. It was found that the FCE had known-
39
40 215 group sex validity (males vs females) for 1 of 3 functional subtests (lifting waist-overhead) and
41
42 216 reported significant performance differences between culture groups (german vs non-german
43
44 217 language groups).

45
46
47 218 Reesink et al. developed an independent FCE for patients with musculoskeletal neck
48
49 219 disorders (neck FCE). (34) They performed a review of epidemiological literature and identified
50
51 220 four physical risk factors for work-related neck disorders and used that information to develop an
52
53
54
55
56
57
58
59
60

221 FCE consisting of eight performance-based tests. Content validity was established by following
222 operational definitions of the risk factors when searching the literature and using current
223 literature to provide a rationale to guide their development of the tasks comprising the FCE.
224 Because of the unconventional methods used by this study to establish content validity, the
225 authors of this review determined that the tools used to critically appraise other articles would be
226 inappropriate and were given scores of N/A for the COSMIN and QACMRR. An additional
227 article measured test-retest reliability of the subscales of the neck FCE in patients with
228 multifactorial neck pain. (32) Test retest ICC's ranged from poor to excellent. Limits of
229 agreement relative to mean performance range from 32.0% to 56.5% for functional based sub
230 tests. Convergent validity was performed against the Neck Disability Index (NDI) items and total
231 score. (33) The authors found weak to moderate Pearson correlations for the FCE sub scores to
232 both NDI individual items and the NDI total score.

234 **BTEWS II**

235 Lomond and Cote reported on the reliability, measurement error, minimum detectable change
236 (MDC) and validity of the power output (PO) task during the BTEWS II test in patients with
237 chronic neck and shoulder pain (**TABLE 5**). (35) Test-retest reliability, measured with Spearman
238 Rank correlations and ICC's was measured at $\rho=0.37$ and $ICC_{2,1} = 0.54$, respectively. The
239 standard error of measurement (SEM) and the minimal detectable change at 90% confidence
240 (MDC_{90}) for the PO task were measured as 30.25 and 70.59, respectively.

241 Weak Spearman Rank correlations between the PO task and the NDI, Shoulder Pain and
242 Disability Index (SPADI) and Numeric Rating Scale (NRS) for pain tests were recorded. There
243 were no significant performance differences between control and pain groups for the PO task.

244

Fit-HaNSA

246 Pierrynowski and colleagues reported on the reliability, measurement error, MDC and
247 validity of the Fit-HaNSA test in a sample of people with WAD II following motor vehicle
248 collision (MVC) (**TABLE 6**). (36) Intra-rater reliability ICC's for patient subtask and total
249 scores ranged between 0.70-0.78. (36) Inter-rater reliability ICC's for patient subtask and total
250 scores ranged between 0.54-0.84. (36) The Bland and Altman plot for the patient group showed a
251 26 s bias in terms of improved performance on the second test (possible learning effect). The
252 standard deviation of difference was 124 s and 95% Limits of Agreement (LoA₉₅) was 248 s.
253 (36) The SEM for people with WAD II was reported to be 76 s. (36) The MDC₉₀ was measured
254 as 176 s. (36)

255 Spearman rank correlations were also calculated between the Fit-HANSA, Numeric Pain
256 Rating Scale (NPRS), NDI, the disabilities of arm, hand and shoulder (DASH) and 6 cervical
257 range of motion measures. Most (59 of 78) of the correlations between performance and
258 comparator measures were poor ($r < 0.4$). (36) All correlations between total Fit-HaNSA scores
259 and subtask scores had good correlations ($r < 0.75$), except for Task 1-Task 3. (36) Significant
260 performance differences between WAD II and control groups (known group validity) were
261 recorded for the total Fit-HaNSA score and all 3 subtask scores. (36)

262

Physiotherapy Test Package Subtests

264 Ljungquist et al published a series of articles which evaluated the clinometric properties
265 of a physiotherapy test package for patients with spinal pain (**TABLE 7**). (37,38,39,40) This
266 package included muscular strength & endurance tests, submaximal endurance tests, and three

1
2
3 267 functional tests. These functional tests included the PILE-C, PILE-L, and 2x20M-WWB test.

4
5 268 Ljungquist's series of articles reported on convergent validity, known-groups validity, reliability,
6
7
8 269 measurement error and sensitivity to change for these tests. (37,38,39,40)
9

10 270 In a 1999 article (38), correlations between the tests of the package and pain (CR-10) and
11
12 271 perceived exertion (Borg RPE) were determined. All correlations were weak, except for a
13
14 272 moderate correlation between the PILE-C test and pain intensity and a moderate correlation
15
16 273 between 2x20M-WWB test and pain intensity.
17
18

19 274 In a paper from 1999, the PILE-C, PILE-L and 2x20M-WWB tests were found to have
20
21 275 significant discriminative abilities in distinguishing healthy subjects from patients with spinal
22
23 276 pain. (37) The sensitivity and specificity for this known group discrimination for the PILE-C test,
24
25 277 were reported to be 0.93 and 0.69, respectively. (37) The sensitivity and specificity for the PILE-
26
27 278 L test were reported to be 0.85 and 0.65, respectively. In a 2003 article, the PILE-C, PILE-L and
28
29 279 2x20M-WWB tests were tested to determine their ability to discriminate between known-groups
30
31 280 (neck pain vs back pain). (40) Subjects with spinal pain completed the CR-10, the University of
32
33 281 Alabama Pain Behavior scale (UAB) and the Borg RPE test. Specific cut points were used to
34
35 282 distinguish patients with high vs. low pain intensity, high vs. low pain behavior, and high vs. low
36
37 283 perceived exertion in patients, respectively. Participants then completed the test package and it
38
39 284 was determined if each subtest could discriminate between participants with high vs. low pain
40
41 285 intensity. The functional tests were able to discriminate between all 3 subgroups with the
42
43 286 exception of the PILE-C being unable to discriminate between participants with high vs. low
44
45 287 perceived exertion.
46
47
48
49

50
51 288 The inter and intra rater reliability were tested on participants with spinal pain. (38)
52
53 289 Limits of agreement were used to measure inter rater reliability and repeatability, defined as 2x
54
55
56
57
58
59
60

1
2
3 290 the within-subject standard deviation of each variable. Interrater agreement for 2 tests was
4
5 291 deemed “acceptable”, while all 3 functional tests had “clinically acceptable” intrarater reliability.
6
7
8 292 (38) Sensitivity-to-change was evaluated in the test package following 6 months of a
9
10 293 physiotherapy intervention. Using ROC curves, Wilcoxon sign ranked tests and spearman
11
12 294 correlation coefficients, only the 2x20m-WWB test and the PILE-C (women only) were deemed
13
14
15 295 to be sensitive to change. (39) Additionally, moderate to high effect sizes were found for all test
16
17 296 components.
18
19 297

21 298 **DISCUSSION**

22
23
24 299 This study synthesized 12 studies assessing clinometric properties of 4 different
25
26 300 performance-based functional assessments. Given the limited number of studies, the substantial
27
28 301 variation in the types of tests examined, the methods used to assess the clinical measurement
29
30 302 properties, and the study populations, the current state of knowledge does not allow firm
31
32 303 conclusions regarding recommendations for an optimal performance-based test at this time.
33
34
35 304 Overall, there is weak to strong evidence for a range of properties of the 4 different assessments
36
37 305 in patients with acute or chronic neck pain that is musculoskeletal in origin.
38
39

40 306 **FCE**

41
42 307 The breadth of a performance-based test is variable and defined by the developers. An
43
44 308 advantage of the functional assessment designed by Reesink et al. (34) is that they mapped the
45
46 309 eight subtests to risk factors identified in the literature for work-related neck disorders. The eight
47
48 310 subtests consist of: material handling tasks, lifting floor to waist, overhead lift test, one-handed
49
50 311 and two-handed carrying, overhead working, repetitive reaching, overhead lifting, and repetitive
51
52 312 bending and overhead reaching. Given the systematic approach and rationale these authors used
53
54
55
56
57
58
59
60

1
2
3 313 in developing the FCE and this approach being used in previous research (41), we suggest that
4
5 314 this test has strong content validity. However, the nature of the reporting of content validity
6
7
8 315 made it difficult to formally assess this paper using the COSMIN tool.
9

10 316 Six articles address the clinical measurement properties of this FCE. There is adequate
11
12 317 evidence that the FCE is stable over test-retest time of 7-14 days. (29,32) These measures
13
14 318 demonstrate longer stability over time compared to self-report measures such as the Neck
15
16 319 Disability Index (NDI) which has demonstrated test-retest reliability within only a short period
17
18 320 of 0-3 days. (17) Whether this longer-term stability is a characteristic of performance-based tests
19
20 321 or reflects differences in study populations in context requires further testing. Although test-
21
22 322 retest reliability has been assessed, inter-rater and intra-rater reliability has yet to be researched.
23
24 323 Unlike self-report measures, we expect measurement error due to the evaluator and performance-
25
26 324 based tests. Thus, future research should explore these aspects of reliability.
27
28
29

30 325 Convergent validity is often examined in clinical measurement studies. We suggest that
31
32 326 this may be because these comparisons are easily performed by correlating different tests rather
33
34 327 than providing strong confidence in the validity of the measurement. Often convenient
35
36 328 comparisons are performed rather than those most relevant. Across many domains and measures
37
38 329 it has become clear that the relationship between self-reported function and performance-based
39
40 330 function or physical impairment is often low to moderate. Therefore the value of assessment of
41
42 331 these relationships as a form of validation has limited value. Several studies of varying quality
43
44 332 have reported on the convergent validity of the FCE. (30,31,33) One article of adequate quality
45
46 333 found the relationship between the FCE and work capacity to be poorly associated with one
47
48 334 another. (30) The same study found that the ability of the FCE to predict future work capacity
49
50
51 335 was poor. This may be considered a more important comparison since ideally performance-based
52
53
54
55
56
57
58
59

1
2
3 336 tests would relate to important outcomes like return to work. No studies to our knowledge report
4
5 337 the responsiveness or sensitivity to change of the FCE. This is an important gap since the focus
6
7
8 338 of rehabilitation is often to remediate limitations in goal impairments or work capacity, and
9
10 339 assessment of these changes is critical to clinical decision-making and reporting outcomes. Thus,
11
12 340 future research should evaluate the responsiveness of the FCE to provide insight in the measure's
13
14
15 341 ability to detect change after an intervention.

16 17 342 **FIT-HaNSA**

18
19 343 One very good quality study assessed the FIT-HaNSA, a test consisting of two reaching tasks
20
21 344 (waist and eye-level) and sustained overhead task performance. (36) Overall, the FIT-HaNSA
22
23
24 345 demonstrates excellent inter-rater reliability and strong intra-rater reliability. The specific
25
26 346 subtests included within the FIT-HaNSA similarly demonstrate moderate to strong inter-rater
27
28 347 and intra-rater reliability. The FIT-HaNSA also demonstrated a clear ability to distinguish
29
30 348 between people with WAD 2 and healthy controls. Correlations between the FIT-HaNSA and
31
32 349 other patient self-report disability and functional outcome measures (NPRS, NDI, DASH,
33
34 350 CROM and FIT-HaNSA) were generally poor ($\rho < 0.4$), consistent with other studies comparing
35
36 351 performance and self-report. (13,14) The largest limitation in critically synthesizing information
37
38 352 for this test is that only a single study was found that reported the measurement properties for
39
40 353 people with neck disorders. It should be noted however that it has been validated in other MSK
41
42 354 disorders. (1–6) Although others have noted the lag in development of performance-based
43
44 355 measures in comparison to self-report measures, FIT-HaNSA was recommended as a
45
46
47 356 performance-based measure for people with shoulder disorders. (2)

48 49 357 **BTEWS II**

50
51
52
53
54
55
56
57
58
59
60

1
2
3 358 One study of doubtful to adequate quality according to the COSMIN risk of bias tool
4
5 359 assessed the efficacy of the BTEWS II where the participants performed a dynamic pushing and
6
7 360 pulling task in which power output was recorded over a 10 second sample. (35) While the
8
9 361 convergent validity aspect of this paper was assessed as adequate through the critical appraisal
10
11 362 process, the relationship between the power output on the BTEWS and measures of pain and
12
13 363 disability (NDI, SPADI, NRS) were poorly associated with each other. In addition, the power
14
15 364 output component was not found to be significantly different between people with neck pain and
16
17 365 healthy controls which suggests it might not be discriminative. Discrimination between patients
18
19 366 and those without any symptoms is a low benchmark, and tests that cannot fulfil this benchmark
20
21 367 should be viewed with caution. Because of the weak measurement properties demonstrated by
22
23 368 the power output component of the BTEWS II, it does not appear to be a desirable performance-
24
25 369 based measure to assess function in people with neck pain. However, we acknowledge for all of
26
27 370 the performance-based tests the evidence pool is so shallow that there is high potential that future
28
29 371 studies might lead to different conclusions.
30
31
32
33
34

35 372 **Physiotherapy Test Package Subtests**

36
37 373 Four studies assessing relevant items from a physiotherapy test package, including a lift
38
39 374 from floor-to-waist and a waist-to-shoulder task and a two-handed carrying task, ranged in
40
41 375 quality from “inadequate” to “very good”. The properties of these assessment items include weak
42
43 376 to moderate correlations to pain, perceived exertion, and had “adequate” reliability. The 2x20m-
44
45 377 WWB and PILE-C tests were found to be sensitive-to-change which is valuable information as
46
47 378 no other study has assessed this property in performance-based measures in patients with neck
48
49 379 disorders. Thus, this measure may be of value in clinical settings when assessing functional
50
51 380 capacity before and after a treatment intervention. All tests had discriminative ability for
52
53
54
55
56
57
58
59
60

1
2
3 381 detecting participants with spinal pain vs healthy controls. Most of the three tests demonstrated
4
5 382 poor construct validity in that they were poorly related to pain and perceived exertion, although
6
7
8 383 this was observed in a study of “doubtful” quality. Thus, further research of better quality is
9
10 384 necessary to investigate these constructs.

11 385 **Limitations**

12
13
14 386 A challenge in synthesizing clinical measurement evidence is the wide range of
15
16 387 properties and indicators that need to be considered. Unlike effectiveness studies where one can
17
18 388 focus on the effect size of treatment there are many considerations that would affect the
19
20 389 recommendations made about outcome measures. This is further complicated when the pool of
21
22 390 evidence is shallow. Although the COSMIN and the quality assessment tool (QACMRR)
23
24 391 developed by one of the authors of this review which assess risk of bias and the quality of design
25
26 392 of individual studies respectively, were useful for interpreting the evidentiary pool, there is no
27
28 393 clear method to synthesize the extracted clinical measurement evidence. While some systematic
29
30 394 reviews on treatment might only report findings from high-quality studies, it is important to see
31
32 395 how outcome measures perform in different contexts. Further, the assessment of risk of bias and
33
34 396 quality are complicated given that clinical measurement studies have so many dimensions.
35
36 397 Therefore, exclusion of lower quality studies has questionable value. Thus, a more practical
37
38 398 approach is to consider quality when interpreting the findings, rather than excluding studies.

39
40 399 The COSMIN and the QACMRR provide different perspectives since one focuses on the
41
42 400 risk of bias and the other the quality of the research design. For example, the article by Van de
43
44 401 Meer et al. was determined to be doubtful according to the COSMIN which is the lowest score
45
46 402 attainable on the tool whereas the QACMRR yielded a score of 86%. Additionally, the COSMIN
47
48 403 score for the Reneman 2017 paper in this review was found to be adequate, a much better result
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 404 than many other articles in this review but yielded the lowest score on the QACMRR of 67%.
4
5 405 This difference is likely attributed to the QACMRRs focus on different design issues. For
6
7 406 example, it provides lower scores where there are problems with small sample size or poor
8
9 407 subject retention, whereas the COSMIN did not ask any specific questions that captured these
10
11 408 qualities. The QACMRR focuses on whether the authors made appropriate decisions in selecting
12
13 409 the scope and methods of their clinical measurement evaluations within a given study and
14
15 410 provides descriptors of poor fair or good design options. Quality focuses on issues that might
16
17 411 affect risk of bias or imprecision in estimates; whereas risk of bias assessments focusses on items
18
19 412 that might result in a biased estimate. For example, insufficient power is a precision (quality)
20
21 413 issue, not a risk of bias. Although it is difficult to interpret the meaning of the percentage of the
22
23 414 QACMRR as there are no established cut-offs for distinguishing good and poor-quality studies, it
24
25 415 provides one way of ranking the articles in order of quality. Since the COSMIN rates bias
26
27 416 according to specific measurement properties whereas the the QACMRR evaluates the overall
28
29 417 study design, we found that these tools provide complementary perspectives on the studies.
30
31 418 Therefore, agreement on the scores was not expected.
32
33
34
35
36
37

38 419 Another limitation in this review was that the feasibility or usability of these tools was
39
40 420 not assessed. While feasibility was not the focus of this review, information on the practical
41
42 421 application of these performance-based measures provides valuable information to clinicians for
43
44 422 determining whether these tests are appropriate to use in their given setting. Thus, future research
45
46 423 should not only investigate further the psychometric properties of these tools, but also report the
47
48 424 feasibility of using these tests so that they may be used in clinical settings and to identify
49
50 425 limitations that restrict their application in practice.
51
52
53

54 426
55
56
57
58
59
60

427 CONCLUSION

428 This study confirms that performance-based tests have had far less development and
429 evaluation than self-report measures. Limitations include the number of tests and insufficient
430 body of evidence to make confident recommendations with respect to performance-based testing.
431 It is clear that self-report and performance-based measures provide different perspectives.
432 Theoretically, performance-based tests are important to inform our understanding about the
433 mechanisms of intervention and how interventions increase capacity. Overall more work is
434 required to further establish the psychometric properties of performance-based tests in persons
435 with neck disorders, including sensitivity-to-change, responsiveness, and predictive validity. The
436 data presented suggest that the FIT-HaNSA has the strongest clinometric properties though this
437 is based on a single high-quality paper specific to neck disorder. (36, 5) Importantly, normative
438 data have been published (6), it has been validated in multiple studies in patients with shoulder
439 conditions (1,3,4) and has been recommended when compared to other measures (2). The FCE
440 has a limited evidence base from which to draw, though it was developed with strong content
441 validity and further evaluation may demonstrate its usefulness. Performance-based evaluation in
442 people with neck disorders is an area needing much research attention both to establish the
443 measurement properties of existing measures, potentially to develop innovative new measures
444 and to perform head-to-head comparisons of measures before an optimal performance-based
445 tests can be identified.

446

447 Authors' contributions

448 SM contributed significantly to conception and design of the study, data extraction, critical
449 appraisal, interpretation of data and drafting of the manuscript. TS, TA, PB, and CC were involved
450 in literature search, critical appraisal and interpretation of data and drafting. AG was involved in

1
2
3 451 critical appraisal and drafting. JM was also involved in the conception and design of the study,
4 452 drafting, and revised the manuscript for important intellectual content. PB and CATWAD were
5 453 involved in the drafting and review of the manuscript. All authors have given their final approval
6 454 on the manuscript to be published
7
8
9

10 455

11 456 **Declarations**

12 457 **Ethics approval and consent to participate**

13 458 Not applicable
14
15
16
17 459

18 460 **Consent for publication**

19 461 Not applicable
20
21
22 462

23 463 **Availability of data and material**

24 464 Data sharing is not applicable to this article as no datasets were generated or analyzed during the
25 465 current study
26
27
28 466

29 467 **Funding Statement**

30 468 This work was supported by the Canadian Institutes of Health Research (CIHR) with funding
31 469 reference number (FRN: SCA-145102).
32
33
34
35 470

36 471 **Competing Interest Statement**

37 472 None to report.
38
39
40
41 473

42 474

43 475 **REFERENCES**

44 476

- 45 477 1. Bobos P, MacDermid JC, Walton DM, Gross A, Santaguida PL. Patient-Reported
46 478 Outcome Measures Used on Neck Disorders. An Overview of Systematic Reviews. *J*
47 479 *Orthop Sport Phys Ther.* June 2018;1-76. doi:10.2519/jospt.2018.8131.
48
49
50
51
52
53
54
55
56
57
58
59
60

- 1
2
3 480 2. Carroll LJ, Hogg-Johnson S, Côté P, et al. Course and Prognostic Factors for Neck Pain
4
5 481 in Workers. *Spine (Phila Pa 1976)*. 2008;33(Supplement):S93-S100.
6
7 482 doi:10.1097/BRS.0b013e31816445d4.
8
9
- 10 483 3. Croft PR, Lewis M, Papageorgiou AC, et al. Risk factors for neck pain: a longitudinal
11
12 484 study in the general population. *Pain*. 2001;93(3):317-325.
13
14 485 <http://www.ncbi.nlm.nih.gov/pubmed/11514090>. Accessed July 11, 2018.
15
16
- 17 486 4. Vos T, Allen C, Arora M, et al. Global, regional, and national incidence, prevalence, and
18
19 487 years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis
20
21 488 for the Global Burden of Disease Study 2015. *Lancet*. 2016;388(10053):1545-1602.
22
23 489 doi:10.1016/S0140-6736(16)31678-6.
24
25
- 26 490 5. Blanpied PR, Gross AR, Elliott JM, et al. Neck Pain: Revision 2017. *J Orthop Sport Phys*
27
28 491 *Ther*. 2017;47(7):A1-A83. doi:10.2519/jospt.2017.0302.
29
30
- 31 492 6. Nygren A, Berglund A, von Koch M. Neck-and-shoulder pain, an increasing problem.
32
33 493 Strategies for using insurance material to follow trends. *Scand J Rehabil Med Suppl*.
34
35 494 1995;32:107-112.
36
37
- 38 495 7. Wright A, Mayer TG, Gatchel RJ. Outcomes of disabling cervical spine disorders in
39
40 496 compensation injuries. A prospective comparison to tertiary rehabilitation response for
41
42 497 chronic lumbar spinal disorders. *Spine (Phila Pa 1976)*. 1999;24:178-183.
43
44
- 45 498 8. Rempel DM, Harrison RJ, Barnhart S. Work-related cumulative trauma disorders of the
46
47 499 upper extremity. *JAMA*. 1992;267:838-842.
48
49 500 <https://doi.org/10.1001/jama.1992.03480060084035>
50
51
52
53
54
55
56
57
58
59
60

- 1
2
3 501 9. Borghouts JA, Koes BW, Vondeling H, Bouter LM. Cost-of-illness of neck pain in The
4
5 502 Netherlands in 1996. *Pain*. 1999;80:629-636. <https://doi.org/10.1016/S0304->
6
7 503 [3959\(98\)00268-1](https://doi.org/10.1016/S0304-3959(98)00268-1)
8
9
10 504 10. Alreni ASE, Harrop D, Lowe A, Tanzila Potia, Kilner K, McLean SM. Measures of
11
12 505 upper limb function for people with neck pain. A systematic review of measurement and
13
14 506 practical properties. *Musculoskelet Sci Pract*. 2017;29:155-163.
15
16 507 doi:10.1016/j.msksp.2017.02.004.
17
18
19 508 11. Childs JD, Cleland JA, Elliott JM, et al. Neck pain: Clinical practice guidelines linked to
20
21 509 the International Classification of Functioning, Disability, and Health from the
22
23 510 Orthopedic Section of the American Physical Therapy Association. *J Orthop Sports Phys*
24
25 511 *Ther*. 2008;38(9):A1-A34. doi:10.2519/jospt.2008.0303.
26
27
28 512 12. Finch E, Canadian Physiotherapy Association. *Physical Rehabilitation Outcome*
29
30 513 *Measures : A Guide to Enhanced Clinical Decision Making*. BC Decker; 2002.
31
32 514 <https://www.ncbi.nlm.nih.gov/nlmcatalog/101175542>. Accessed July 19, 2018.
33
34
35 515 13. Simmonds MJ, Olson SL, Jones S, et al. Psychometric characteristics and clinical
36
37 516 usefulness of physical performance tests in patients with low back pain. *Spine (Phila Pa*
38
39 517 *1976)*. 1998;23(22):2412-2421. <http://www.ncbi.nlm.nih.gov/pubmed/9836355>.
40
41 518 Accessed July 19, 2018.
42
43
44 519 14. Stratford PW, Kennedy D, Pagura SMC, Gollish JD. The relationship between self-report
45
46 520 and performance-related measures: questioning the content validity of timed tests.
47
48 521 *Arthritis Rheum*. 2003;49(4):535-540. doi:10.1002/art.11196.
49
50
51
52
53
54
55
56
57
58
59
60

- 1
2
3 522 15. Novy DM, Simmonds MJ, Lee CE. Physical performance tasks: what are the underlying
4
5 523 constructs? *Arch Phys Med Rehabil.* 2002;83(1):44-47.
6
7
8 524 <http://www.ncbi.nlm.nih.gov/pubmed/11782832>. Accessed July 19, 2018.
9
10 525 16. MacDermid JC, Stratford P. Applying evidence on outcome measures to hand therapy
11
12 526 practice. *Journal of Hand Therapy.* 2004;17(2):165-173. doi: 10.1197/j.jht.2004.02.005.
13
14 527 17. Macdermid, J. C., Walton, D. M., Avery, S., Blanchard, A., Etruw, E., Mcalpine, C., &
15
16 528 Goldsmith, C. H. (2009). Measurement properties of the neck disability index: a
17
18 529 systematic review. *Journal of orthopaedic & sports physical therapy*, 39(5), 400-C12.
19
20 530 18. Misailidou V, Malliou P, Beneka A, Karagiannidis A, Godolias G. Assessment of
21
22 531 patients with neck pain: a review of definitions, selection criteria, and measurement tools.
23
24 532 *J Chiropr Med.* 2010;9(2):49-59. doi:10.1016/j.jcm.2010.03.002.
25
26 533 19. Holly LT, Matz PG, Anderson PA, et al. Functional outcomes assessment for cervical
27
28 534 degenerative disease. *J Neurosurg Spine.* 2009;11(2):238-244.
29
30 535 doi:10.3171/2009.2.SPINE08715.
31
32 536 20. Pietrobon R, Coeytaux RR, Carey TS, Richardson WJ, DeVellis RF. Standard scales for
33
34 537 measurement of functional outcome for cervical pain or dysfunction: a systematic review.
35
36 538 *Spine (Phila Pa 1976).* 2002;27(5):515-522.
37
38 539 <http://www.ncbi.nlm.nih.gov/pubmed/11880837>. Accessed July 19, 2018.
39
40 540 21. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred Reporting Items
41
42 541 for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *J Clin Epidemiol.*
43
44 542 2009;62(10):1006-1012. doi:10.1016/j.jclinepi.2009.06.005.
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

- 1
2
3 543 22. Mokkink, L. B., De Vet, H. C., Prinsen, C. A., Patrick, D. L., Alonso, J., Bouter, L. M.,
4
5 544 & Terwee, C. B. (2018). COSMIN risk of Bias checklist for systematic reviews of
6
7 545 patient-reported outcome measures. *Quality of Life Research*, 27(5), 1171-1179.
- 8
9
10 546 23. Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. Rating the
11
12 547 methodological quality in systematic reviews of studies on measurement properties: a
13
14 548 scoring system for the COSMIN checklist. *Qual Life Res* 2012; 21: 651–7
- 15
16
17 549 24. MacDermid J. Appendix G: critical appraisal of study quality for psychometric articles,
18
19 550 evaluation form. In: Law M, MacDermid J, eds. *Evidence-Based Rehabilitation: A Guide*
20
21 551 *to Practice*. 2nd ed. Thorofare, NJ: SLACK Incorporated; 2008:387-388.
- 22
23
24 552 25. Roy J-S, Desmeules F, MacDermid JC. Psychometric properties of presenteeism scales
25
26 553 for musculoskeletal disorders: a systematic review. *J Rehabil Med*. 2011;43(1):23-31.
27
28 554 doi:10.2340/16501977-0643.
- 29
30
31 555 26. Vessey J, Strout TD, DiFazio RL, Walker A. Measuring the youth bullying experience: a
32
33 556 systematic review of the psychometric properties of available instruments. *J Sch Health*.
34
35 557 2014;84(12):819-843. doi:10.1111/josh.12210.
- 36
37
38 558 27. Roy J-S, MacDermid JC, Woodhouse LJ. A systematic review of the psychometric
39
40 559 properties of the Constant-Murley score. *J shoulder Elb Surg*. 2010;19(1):157-164.
41
42 560 doi:10.1016/j.jse.2009.04.008.
- 43
44
45 561 28. Roy J-S, MacDermid JC, Woodhouse LJ. Measuring shoulder function: A systematic
46
47 562 review of four questionnaires. *Arthritis Rheum*. 2009;61(5):623-632.
48
49 563 doi:10.1002/art.24396.
- 50
51
52
53
54
55
56
57
58
59
60

- 1
2
3 564 29. Trippolini MA, Reneman MF, Jansen B, Dijkstra PU, Geertzen JHB. Reliability and
4
5 565 safety of functional capacity evaluation in patients with whiplash associated disorders. *J*
6
7 566 *Occup Rehabil.* 2013;23(3):381-390. doi:10.1007/s10926-012-9403-z.
8
9
10 567 30. Trippolini MA, Dijkstra PU, Côté P, Scholz-Odermatt SM, Geertzen JH, Reneman MF.
11
12 568 Can functional capacity tests predict future work capacity in patients with whiplash-
13
14 569 associated disorders? *Arch Phys Med Rehabil.* 2014;95(12):2357-2366.
15
16 570 doi:10.1016/j.apmr.2014.07.406.
17
18
19 571 31. Trippolini MA, Dijkstra PU, Geertzen JHB, Reneman MF. Construct Validity of
20
21 572 Functional Capacity Evaluation in Patients with Whiplash-Associated Disorders. *J Occup*
22
23 573 *Rehabil.* 2015;25(3):481-492. doi:10.1007/s10926-014-9555-0.
24
25
26 574 32. Reneman MF, Roelofs M, Schiphorst Preuper HR. Reliability and Agreement of Neck
27
28 575 Functional Capacity Evaluation Tests in Patients With Chronic Multifactorial Neck Pain.
29
30 576 *Arch Phys Med Rehabil.* 2017;98(7):1476-1479. doi:10.1016/j.apmr.2016.12.005.
31
32
33 577 33. Van der Meer S, Reneman MF, Verhoeven J, van der Palen J. Relationship between self-
34
35 578 reported disability and functional capacity in patients with whiplash associated disorder. *J*
36
37 579 *Occup Rehabil.* 2014;24(3):419-424. doi:10.1007/s10926-013-9473-6
38
39
40 580 34. Reesink DD, Jorritsma W, Reneman MF. Basis for a functional capacity evaluation
41
42 581 methodology for patients with work-related neck disorders. *J Occup Rehabil.*
43
44 582 35. Lomond K V, Côté JN. Shoulder functional assessments in persons with chronic
45
46 583 neck/shoulder pain and healthy subjects: Reliability and effects of movement repetition.
47
48 584 *Work.* 2011;38(2):169-180. doi:10.3233/WOR-2011-1119.
49
50
51 585 36. Pierrynowski M, McPhee C, P Mehta S, C MacDermid J, Gross A. Intra and Inter-Rater
52
53 586 Reliability and Convergent Validity of FIT-HaNSA in Individuals with Grade II
54
55
56
57
58
59
60

- 1
2
3 587 Whiplash Associated Disorder. *Open Orthop J.* 2016;10(1):179-189.
4
5 588 doi:10.2174/1874325001610010179.
6
7
8 589 37. Ljungquist T, Fransson B, Harms-Ringdahl K, Björnham A, Nygren A. A physiotherapy
9
10 590 test package for assessing back and neck dysfunction--discriminative ability for patients
11
12 591 versus healthy control subjects. *Physiother Res Int.* 1999;4(2):123-140.
13
14 592 <http://www.ncbi.nlm.nih.gov/pubmed/10444762>. Accessed July 11, 2018.
15
16
17 593 38. Ljungquist T, Harms-Ringdahl K, Nygren A, Jensen I. Intra- and inter-rater reliability of
18
19 594 an 11-test package for assessing dysfunction due to back or neck pain. *Physiother Res Int.*
20
21 595 1999;4(3):214-232. <http://www.ncbi.nlm.nih.gov/pubmed/10581627>. Accessed July 11,
22
23 596 2018.
24
25
26 597 39. Ljungquist T, Nygren A, Jensen I, Harms-Ringdahl K. Physical performance tests for
27
28 598 people with spinal pain--sensitivity to change. *Disabil Rehabil.* 2003;25(15):856-866.
29
30 599 doi:10.1080/0963828031000090579.
31
32
33 600 40. Ljungquist T, Jensen IB, Nygren A, Harms-Ringdahl K. Physical performance tests for
34
35 601 people with long-term spinal pain: aspects of construct validity. *J Rehabil Med.*
36
37 602 2003;35(2):69-75. <http://www.ncbi.nlm.nih.gov/pubmed/12691336>. Accessed July 11,
38
39 603 2018.
40
41
42 604 41. Reneman MF, Dijkstra PU, Westmaas M, Göeken LNH. Test-retest reliability of lifting
43
44 605 and carrying in a 2-day functional capacity evaluation. *J Occup Rehabil.* 2002;12(4):269-
45
46 606 275. <http://www.ncbi.nlm.nih.gov/pubmed/12389478>. Accessed July 19, 2018.
47
48
49 607
50 608
51 609
52 610
53 611
54 612

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

613
614
615
616

For peer review only

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
Ljungquist et al. 1999	Neck pain, back pain, multiple pain sites, chronic pain	53	PILE-C, PILE-L	N/A
Ljungquist et al. 1999	Neck pain, lumbar pain, thoracic pain, shoulder pain, multiple pain sites, chronic pain	68	PILE-C, PILE-L, 2 WWB	20m 8 days
Ljungquist et al. 2003	Neck pain, lumbar pain, thoracic pain, shoulder pain, lower extremity pain, multiple pain sites, chronic pain	235	PILE-C, PILE-L, 2 WWB	20m N/A
Ljungquist et al. 2003	cervical pain, lumbar pain, cervical and lumbar pain, multiple pain sites, chronic pain	186	PILE-C, PILE-L, 2 WWB	20m 6 months
Lomond and Cote. 2011	Chronic neck and shoulder pain	32	BTEWS II	9.5 days
Pierrynowski et al. 2016	Sub-acute and chronic WAD II	66	FIT-HaNSA	2-7 days
Reesink et al. 2007	N/A	N/A	Neck-FCE	N/A
Reneman et al. 2017	Chronic multifactorial neck pain	18	Neck-FCE	2 weeks
Trippolini et al. 2013	Sub acute and chronic WAD I and II	32	WAD FCE	7 days
Trippolini et al. 2014	Sub acute and chronic WAD I and II	267	Workwell FCE	N/A
Trippolini et al. 2015	Sub acute and chronic WAD I and II	314	WAD FCE	N/A
Van der Meer et al. 2013	Chronic WAD I and II	40	Neck FCE	N/A

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

PILE-C, Progressive Isoinertial Lifting Evaluation-Cervical; PILE-L, Progressive Isoinertial Lifting Evaluation; CBT, Cognitive-Behavioural Therapy; PT, Physical Therapy; NRPS, Numeric Pain Rating Scale; BTEWS II, Baltimore Therapeutic Equipment Work Simulator II; WAD, Whiplash Associated Disorder; MVA, Motor Vehicle Accident; FIT-HaNSA, Functional Impairment Test-Hand and Neck/Shoulder/Arm; FCE, Functional Capacity Evaluation; EXP, Experimental; M, Male; F, Female

For peer review only

TABLE 2. Summary of Psychometric Properties Reported in Studies and COSMIN risk of bias checklist scores

Study	Psychometric Properties Reported	COSMIN Score
Ljungquist et al. 1999	Known-groups Validity	Adequate
	Convergent Validity	Very Good
Ljungquist et al. 1999	Reliability	Inadequate
	Measurement Error	Adequate
Ljungquist et al. 2003	Known-groups Validity	Very Good
Ljungquist et al. 2003	Sensitivity to Change	Doubtful
Lomond and Cote. 2011	Reliability	Doubtful
	Measurement Error	Doubtful
	Known-groups Validity	Doubtful
	Convergent Validity	Adequate
Pierrynowski et al. 2016	Reliability	Very Good
	Measurement Error	Adequate
	Known-groups Validity	Very Good
	Convergent Validity	Very Good
Reesink et al. 2007	Content Validity	N/A*
Reneman et al. 2017	Reliability	Adequate
	Measurement Error	Adequate
Trippolini et al. 2013	Reliability	Adequate
	Measurement Error	Adequate
Trippolini et al. 2014	Convergent Validity	Very Good
	Predictive Criterion Validity	Very Good
Trippolini et al. 2015	Known-groups Validity	Very Good
	Convergent Validity	Inadequate
Van der Meer et al. 2013	Convergent Validity	Doubtful

COSMIN, Consensus-based Standards for the Selection of health Measurement Instruments

*Paper is not applicable for completion of COSMIN checklist

TABLE 3. Quality of Studies on Psychometric Properties of Functional-based Tests Evaluated in Neck Disorder Patients

Study	Item Evaluation Criteria												Total (%)
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	
Trippolini et al, 2014	2	2	2	2	1	2	2	2	2	2	1	2	92%
Lomond and Cote, 2011	2	2	1	2	0	2	2	2	2	2	2	2	88%
Pierrynowski et al, 2016	2	2	1	2	0	2	2	2	2	2	2	2	88%
Trippolini et al, 2015	2	2	2	0	1	N/A	2	2	2	2	2	2	86%
Van der Meer et al, 2013	2	1	2	1	2	N/A	2	1	2	2	1	2	86%
Ljungquist et al 2003 KGV	2	2	2	0	0	N/A	2	2	2	2	2	2	82%
Ljungquist et al 1999 Rel	2	1	1	2	0	2	2	2	2	2	1	2	79%
Ljungquist et al 2003 STC	1	1	1	2	1	1	2	2	2	2	2	2	79%
Trippolini et al, 2013	2	2	1	1	0	0	2	2	2	2	2	2	75%
Ljungquist et al 1999 KGV	2	1	1	2	0	N/A	2	1	2	2	1	2	68%
Reneman et al, 2017	1	2	1	1	1	0	1	2	2	2	2	1	67%
Reesink, 2007*	-	-	-	-	-	-	-	-	-	-	-	-	N/A

*Paper is not applicable for completion of study quality tool

TABLE 4. Psychometric Properties of the Functional Capacity Evaluation

FCE Battery	Type of Properties	Statistical Test	Value	Quality
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 Negligible to moderate
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 Weak Negligible-weak Weak
	Known-groups Validity (German vs Non-German)	Linear Regression Analysis	p<0.001	Significant for All tasks
	Known-groups 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	Weak
	Predictive Validity	Pearson or Spearman Correlation	0.06-0.39	Weak
		Linear Mixed Model Regression of All Predictors	$\beta=-0.04$, 95% CI: -0.15 – 0.06 p=0.428 (task 6)	Not Significant

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

*Pain measured via Numeric Rating Scale

TABLE 5. Summary of Fit-HaNSA's psychometric properties in neck disorder patients

Test	Type of Property	Statistical Test	Value	Quality
Fit-HaNSA	Intra-rater Reliability	ICC	0.78	Strong
Fit-HaNSA	Inter-rater Reliability	ICC	0.84	Strong
Fit-HaNSA	Measurement Error	SEM	76 s	
		LOA ₉₅	248 s	
		MDC ₉₀	176 s	
Fit-HaNSA	Convergent Validity	Spearman Rank Correlation	<0.4 - >0.75	Moderate - Strong
Fit-HaNSA	Known-groups Validity 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	Strong
	Inter-reliability	ICC	0.54-0.80	Moderate
	Convergent Validity	Spearman Rank Correlation	<0.4 - >0.75	Moderate - Strong
	Known-groups Validity WAD II vs Control	F-test	42.0-53.3, p<0.001	Significant

Fit-HaNSA, Functional Impairment Test, Hand and Neck/Shoulder/Arm; ICC, Intraclass correlation coefficient; SEM, Standard Error of Measurement; LOA₉₅, 95% Limits of Agreement; MDC₉₀, 90% Minimal Detectable Change; WAD, Whiplash Associated Disorder; Mod, Moderate

*Correlations completed with Numeric Pain Rating Scale, Neck Disability Index, Disabilities of Arm, Shoulder, Hand and 6 cervical range of motion tests

TABLE 6. Psychometric Properties of Baltimore Therapeutic Equipment Work Simulator II – Power Output Task

Test	Type of Property	Statistical Test	Value	Quality
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	Known-groups Validity (Pain vs Control)	Two-way Repeated Measures ANOVA	Not Reported	Non-significant

ICC, Intraclass correlation coefficient; SEM, Standard Error of Measurement; MDC₉₀, 90% Minimal Detectable Change; ANOVA, Analysis of Variance

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

TABLE 7. Psychometric Properties of performance-based tests included in physiotherapy test package

Test	Type of Property	Statistical Test	Value	Quality
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: Unreported* Borg RPE: Unreported	Moderate Low
PILE-C	KGV: spinal pain vs. control	Sensitivity and Specificity	0.93, 0.69	
PILE-C	KGV: spinal pain vs. control	Wilcoxon Sign Ranked Test	p=0.008	Significant
PILE-C	KGV: High vs. low pain intensity	Mann-Whitney U	p=0.003	Significant
PILE-C	KGV: High vs. low Pain behavior	Mann-Whitney U	p=0.005	Significant
PILE-C	KGV: 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	Low – Moderate Negligible – Low
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: Unreported Borg RPE: Unreported	Low Low
PILE-L	KGV: spinal pain vs no spinal pain	Sensitivity and Specificity	0.85, 0.65	
PILE-L	KGV: spinal pain vs control	Wilcoxon Sign Ranked Test	p=0.002	Significant

PILE-L	KGV: High vs. low pain intensity	Mann-Whitney U	p=0.001	Significant
PILE-L	KGV: High vs. low pain behaviour	Mann-Whitney U	p<0.001	Significant
PILE-L	KGV: 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	Negligible – Strong Weak – Strong
2 x 20m WWB	Inter-rater Reliability	Mean Difference LoA	0.05 -1.33 and 1.43	
2 x 20m WWB	Intra-rater Reliability	Repeatability % of Range	3.2 10.7%	
2 x 20m WWB	Convergent Validity	Spearman Correlation	CR-10: Unreported Borg RPE: Unreported	Moderate Low
2 x 20m WWB	KGV: spinal pain vs control	Wilcoxon Sign Ranked Test	p=0.014	Significant
2 x 20m WWB	KGV: High vs. low pain intensity	Mann Whitney U	p<0.001	Significant
2 x 20m WWB	KGV: High vs. low pain behaviour	Mann Whitney U	p<0.001	Significant
2 x 20m WWB	KGV: High vs. low perceived exertion	Mann Whitney U	p<0.001	Significant
2 x 20m WWB	Sensitivity to change	Effect Size	Subjects improving: 0.38-0.78 Subjects deteriorating: 0.13-0.62	Weak – Moderate Negligible – Moderate

PILE-C, Progressive Iso-inertial Lifting Evaluation – Cervical; PILE-L, Progressive Iso-inertial Lifting Evaluation – Lumbar; LoA, Limits of Agreement; SD, Standard Deviation; M, Male; F, Female; RPE, Rating of perceived exertion; KGV, Known-groups Validity; Neg., Negligible; Mod., Moderate, *CR-10: Measurement of pain construct

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2019-031242 on 24 November 2019. Downloaded from <http://bmjopen.bmj.com/> on April 19, 2024 by guest. Protected by copyright.

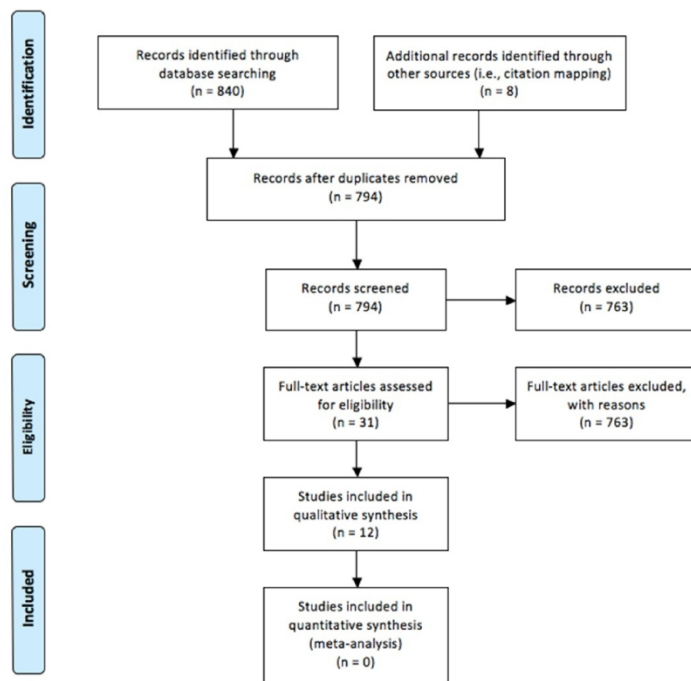


Figure 1. Selection of the studies for inclusion in the systematic review

215x279mm (300 x 300 DPI)

APPENDICES

APPENDIX A. Data extraction guide for studies evaluating the quality of studies evaluating the clinical measurement properties of outcome measures

Instructions

Clinical measurement studies may evaluate a wide spectrum of measurement properties; or evaluate aspects that relate to the implementability or interpretation of outcome measures. Individual clinical measurement studies cannot address every aspect of the measurement properties of an instrument. Ideally systematic reviews will synthesize the quality and content of research evidence addressing the clinical measurement properties of individual outcome measures. The summative knowledge about the measurement properties, cultural transferability, and utility across different contexts provides the scope of information needed to select an outcome measure for a specific patient (population), purpose and context.

This guide should facilitate extraction of data from individual clinical measurement studies. An explanation of the measurement property addressed in each item and how it might be measured within a given study is listed to facilitate finding and extracting that information. The accompanying extraction form can then be used to collect the specific information on these measurements or utility properties from specific studies.

The purpose of data extraction is to extract the specific information reported by authors within a study, not to evaluate the validity or value of that piece of information. Evaluation of the quality of the published version of the clinical measurement study (also called critical appraisal) is performed in a separate step. See the accompanying critical appraisal tool and guide. It is advisable to extract detailed specific information from the study; recognizing that this information may later be synthesized or subject to meta-analysis.

There is no standardized process for synthesizing clinical measurement information. Based on the findings of extraction you may elect to present the synthesize data in a descriptive way by creating a summary table of the data extracted in each category. If you find some studies with similar designs, you may be able to conduct a meta-analysis of some properties like clinically important difference (CID) or minimal detectable change (MDC); if appropriate given the sample and technique - this can be valuable as it may provide more stable estimates of these important properties.

366bmjopen-2019-031242 on 24 November 2019. Downloaded from <https://bmjopen.bmj.com/>. Protected by copyright.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

<u>Population studied</u>		
Population	A description of the study population	Sample size, pathology/disorder, demographics, setting, acute vs. chronic, where subjects were chosen from. Report meaningful demographics and indicators of the population studied.
Intervention	Interventions (if applicable) applied during longitudinal studies	Description of the nature, frequency, intensity of the intervention and the follow-up interval.
<u>Reliability</u>		
Reliability Description	The extent to which the same results are obtained on repeated administrations of the same measure when no change in status has occurred (reliability) or the precision of the scores on repeated measurements (agreement).	Test procedures or measures are typically reapplied on repeated occasions in individuals considered to have a stable condition during that time frame which repeated testing occurs. Repeated testing may be performed on different occasions (test-retest) for self-report measures, OR by the same rater (intra-rater) or different raters (inter-rater) if it is an observer-based scale. In some cases different test instruments (inter-instrument) are evaluated. The most common statistic used is the intraclass correlation coefficient for quantitative data (Shrout & Fleiss, 1979) and kappa (Landis & Koch, 1977) for nominal data. Standard error of measurement is used to present a quantitative estimate of the reliability—in the original units of measure. Report the type of reliability evaluated and coefficients obtained.
Reliability (relative)	The relationship (ratio) between variability in test scores when repeating the test on the same person in comparison to the overall variability (including variation between people)—typically indicated by a reliability coefficient	ICCs (Shrout & Fleiss, 1979) or another reliability coefficient and their associated confidence intervals are extracted.
Reliability (absolute)	Absolute reliability is portrayed as the quantity of error that could be anticipated upon repeated testing - reported in the original units of measure.	This may be reported as 1. Standard error of measurement (in older articles you may see coefficient of variation),

		<p>2. Altman and Bland graphical technique (Bland & Altman, 1990; Bland & Altman, 1987; Bland & Altman, 1986) where the difference on repeated tests for each individual (limits of agreement) is plotted versus their mean score. The mean difference and the boundaries of 2SD are shown to define the limits of agreement.</p>
<p>Minimum Detectable Change</p>	<p>Calculated from the reliability coefficient and the level of confidence specified for error margins. This indicator reflects the amount of change required before you can be confident that change exceeds the random error that occurs in stable patients.</p>	<p>Extract the number and level of confidence.</p>
<p><u>Content/structural validity</u></p>		
<p>Internal consistency</p>	<p>The extent to which items on a test or subscale are related (an indication of the consistency of the concept measured).</p>	<p>Cronbach's alpha is the inter-item correlation usually reported. Report alpha and whether it relates to the entire instrument or specific subscales.</p>
<p>Content Validity</p>	<p>The extent to which the conceptual domain or construct that a test is designed to measure is adequately reflected by the items in the measure. In assessing content validity, it is important to consider the population to whom the measure applies, the completeness of the content, the relevancy and emphasis of the content assessed.</p>	<p>A variety of techniques can be used to assess the extent to which items on a given measure reflected the necessary content to capture the concept of interest. Some of the techniques you will find are listed. Extract what was done to determine content validity and what was found.</p> <ol style="list-style-type: none"> 1) Patients and experts were involved during item selection/reduction - report how they were used and key decisions 2) Patients were consulted for reading and comprehension - report key findings 3) Cognitive interviews (Cibelli, 1994; Ojanen & Gogates, 2006) were done with patients to determine how items were interpreted by respondents; their perceptions of the items - report key findings 4) Expert panels or Delphi procedures were used to select items or evaluate the validity of the instrument - report key findings and decisions

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

36/bmjopen-2019-022120.n.pdf Downloaded from <http://bmjopen.bmj.com/> on 08 November 2019 by guest. Protected by copyright.

		<p>5) During translation specific study, the meaning of the questions to another cultural or language group was studied - report key findings and decisions</p> <p>6) ICF linking (Cieza et al., 2002), or other coding of content was performed - report the results which may include the distribution of content across ICF domains, or the distribution of specific codes</p>
<p>Floor-Ceiling Effects</p>	<p>The measure is unable to indicate a worsening score in patients who have clinically deteriorated and/or an improved score in patients who have clinically improved</p>	<p>There are a variety of potential methods; so the method and conclusion should be reported. Descriptive statistics of the distribution of scores that may be presented graphically or numerically may be used to indicate this. Other studies report the percentage of patients sustained a floor or ceiling effect defined by the number of people who fall in the extremes ranges. Note different studies may define the extreme ranges for floor/ceiling differently, so extract how it was defined and % of patients who obtained floor or ceiling category scores.</p>
<p>Factorial validity</p>	<p>The extent to which factor analysis supports assumptions surrounding constructs measured as defined by the measure or as indicated by subscale structure</p>	<p>Factor analysis may be reported as raw results; or compared to the inherent structure of the instrument or factor analysis upon which its construction was based. Report the type of factor analysis performed (exploratory or confirmatory), rotations used and the number of factors derived; specify whether this confirms the expected instrument structure or original factor structure.</p>
<p>Item response /Rasch Analyses</p>	<p>The extent to which items cross a range of difficulty, or a spectrum of the concept measured. The measurement scaling of the items.</p>	<p>Using item response theory or Rasch analysis, items are fit to a model to demonstrate interval scaling and determine item difficulty (Pallant & Tennant, 2007). Analyses might address item difficulty, person's ability curves, and comparison of ability estimation. Most commonly, the item difficulty and the composition of the test that fulfills interval scaling are defined. Data to be extracted include information on the scaling of the items, whether the interval scaling has been established; and the presence or absence of differential item functioning</p>

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

366bmjopen-2019-09-242 on 24 November 2019. Protected by copyright.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

		(DIF), where items perform differently on different types of respondents.
Construct Validity		
Construct Validity - correlational	<p>Constructs are artificial frameworks that are not directly observable. Construct validity assesses the extent to which measures perform according to a priori defined constructs. Construct validity can be cross-sectional or longitudinal (predictive).</p> <p>Constructed hypotheses can assess convergent validity where measures are thought to represent similar constructs or divergent validity where it is assumed they measure different constructs.</p> <p>For cross-cultural validation, the expected relationships are those that have been reported in validation of the instrument in its original language/format.</p>	<p>When extracting data about correlational validity, the pre-constructed hypothesis and whether it is supported should be documented. For correlational construct validity, this will be the nature and strength of the prespecified relationship and the correlations that support that. Relation to other indices/constructs that are similar (convergent) or different (divergent) can be reported. Ideally, hypotheses are formulated/reported and supported by correlations that are in accordance with the hypotheses. Note that there is no consistent agreement on what subjective term should be applied to validity correlations.</p> <p>Note that there is no consistent agreement on what subjective term should be applied to validity correlations. Some authors use subjective terminology defined for reliability such as: strong (>0.70) and moderate (0.40-0.70) correlations; others use the correlations like effect size benchmarks that 0.4 indicates a moderate effect and 0.6 a large effect. For validity assessment is more important than correlations prespecified constructed hypotheses, although not all papers are written clearly with respect to this.</p>
Convergent	The Relationship between similar scales/tests. Correlations are generally expected to be moderate to strong if the relationship is one where there is confidence that they measure a similar construct.	Extract test names, prespecified expected relationship and correlations observed.
Divergent	Divergent validity assesses the extent to which different scales/tests that are designed to	Extract test names, prespecified expected relationship and correlations observed.

	measure different constructs demonstrate that they are different by a lack of correlation between them.	
Construct validity - known groups	Known groups analysis supports the validity of a measure by demonstrating that the measurement is able to differentiate between groups that are prespecified and <u>known</u> to be different on the construct being assessed.	Data extraction should include the nature of the subgroups and the size of the difference observed between them (and its statistical significance). Typically, statistical tests of difference are performed. Since known groups analysis can provide data that is useful in clinical practice as benchmarks for comparing these known groups, it is a more practical form of construct validity than correlation. Data extraction/presentation should reflect this by presenting the group central tendency, their margins and statistical significance in an accessible manner.
Longitudinal Validity	This form of validity supports the validity of a measure by demonstrating that the change that occurs over time onto similar instruments is correlated in a manner consistent with the nature of the relationship between the scales. It is measured over a retest interval when clinically relevant change could be expected.	Extract test names and correlations Note: since longitudinal validity is based on four measures (pre-and post-test on two different measures), and since error tends to mitigate the strength of correlations, strong longitudinal correlations can be difficult to obtain.
Criterion validity Description	Criterion validation is determined by comparing a given outcome measure to an accepted standard of measure. For subjective constructs like pain and disability, it can be argued that there is no criterion since there is no external gold standard. Therefore, for self-report measures, validation focuses on construct validity. For performance measures, it is common to have a criterion measure that is considered to be highly precise and rigorous as the criterion comparator.	Authors will state that their measure is being compared against a specific instrument and report the correlation or agreement between the measures. Extract the test names and results: correlations if other as reported.
Concurrent criterion	Concurrent validity is assessed by comparing a scale and its criterion at a single point in time	Extract the test names and correlations.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

Predictive criterion	Predictive validity is evaluated by determining the extent to which the results of administering an outcome measure at one point in time can accurately predict a future status or outcome.	Extract the test names and correlations and time interval. (and important cutoffs if those were established/reported), if diagnostic test methodology was used to examine prediction, and sensitivity specificity and other diagnostic criteria were reported, they should be extracted.
<u>Responsiveness/Clinical Change</u>		
Responsiveness	Does the instrument detect changes over time that matters to patients?	Extract indicators of responsiveness include: effect size, standard response mean and the method for assessing whether patients were improved, stable or worse. (Beaton, 2000)
Clinically Important Difference (CID)	CID is the difference in scores that patients find to be observable and clinically important. It is assessed by comparing scores to an external benchmark of clinical relevance such as a global rating of change or some other method. The terminology used to rate the nature of this difference will affect the estimation process. Differences in methods include how clinically importance is framed and the metrics/process by which that is determined.	Extract the MID or CID and note the method/cut-off used to define importance. Extract how the clinically important differences were framed to respondents; or determined. For example, minimal, moderate, extreme improvement or better/not better, etc.

36/bmjopen-2019-031242 on 24 November 2019. Downloaded from <http://bmjopen.bmj.com/> on April 19, 2024 by guest. Protected by copyright.

APPENDIX B. Data extraction form for studies evaluating the clinical measurement properties of outcome measures

Authors: _____ Year: _____ Rater: _____

Instructions

When using the data extraction form, it is important to realize that the purpose of data extraction is to remove or extract the specific information reported by authors within a study, not to evaluate the validity or value of that piece of information. To make data extraction as useful as possible, and to avoid the need for repeated data extractions, it is advisable to read the accompanying guide and then be as specific as possible when extracting information.

DATA EXTRACTED	
Population studied	
Population	
Intervention	
Reliability	
Reliability (relative)	
Reliability (absolute)	
Minimum Detectable Change	
Content/structural validity	

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

Internal consistency	
Content Validity	
Floor-Ceiling Effects	
Factorial validity	
Item response /Rasch Analyses	
Construct/Criterion Validity	
Known groups	
Convergent	
Divergent	
Longitudinal Validity	
Concurrent criterion	
Predictive criterion	
Responsiveness/Clinical Change	

Responsiveness	
Minimally Clinical Important Difference	

For peer review only

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

APPENDIX C. Quality Appraisal for Clinical Measurement Research Reports Evaluation Form

Rater (Group) _____

Author(s) (Study Author(s)) _____

Year (Year of publication) _____

1. Was the relevant background work cited to define what is currently known about the measurement properties of measures under study, and the potential contributions of the current research question to informing that knowledge base?

2

1

0

2. Were appropriate inclusion/exclusion criteria defined? *

2

1

0

3. Were specific clinical measurement questions/hypotheses identified?

2

1

0

4. Was an appropriate scope of measurement properties considered?

2

1

0

5. Was an appropriate sample size used?

2

1

0

6. Was appropriate retention/follow-up obtained? (for studies involving retesting; otherwise n/a)

- 1
2
3 2
4 1
5 0
6
7 7. Were specific descriptions provided of the measure under study and the method(s) used to administer
8 it?
9 2
10 1
11 0
12
13 8. Were standardized procedures used to administer all study measures in a manner that minimized
14 potential sources of error/bias (including the study measure and its comparators)?
15 2
16 1
17 0
18
19 9. Were analyses conducted for each specific hypothesis or purpose?
20 2
21 1
22 0
23
24 10. Were appropriate statistical tests performed to obtain point estimates of the measurement
25 properties?
26 2
27 1
28 0
29
30 11. Were appropriate ancillary analyses done to quantify the confidence in the estimates of the clinical
31 measurement property (Precision/Confidence intervals; benchmark comparisons/ROC curves, alternate forms of
32 analysis like SEM/MID, etc.)?
33 2
34 1
35 0
36
37 12. Were clear, specific and accurate conclusions made about the clinical measurement properties; that
38 were associated with appropriate clinical measurement recommendations and supported by the study objectives,
39 analysis and results?
40 2
41
42
43
44
45
46
47

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

Subtotals (of column 1 and 2)

Total Score (sum of subtotals/24*100)

APPENDIX D. Description of each performance battery from selected articles

Battery	Description of Tasks
Relevant FCE Subtasks ^{25,26,27,28,29,30}	<p>Material Handling Tasks: All lifting tests were executed with a wooden crate (40 × 30 × 26 cm) of 2.5 kg, and four to five weight increments of 2.5 kg or 5 kg each were used until the maximum amount of weight was reached. Maximum performance was recorded in kg.</p> <p>Lifting floor to waist: Measured after five lifts of crate from floor to table and vice versa (time limit < 90 s): hands remained on the crate during the test. Increase weight in 4-5 steps until maximum is reached</p> <p>Overhead lift test: Five lifts from waist to crown height and vice versa within 90 s in standing position. Increase weight in 4–5 steps until maximum is reached</p> <p>Two-handed carrying: Carrying of a crate for a short distance measured after five carries of 1.5 m distance at waist height. Hands remain on the crate during the test.</p> <p>One-handed carrying: Carrying wooden crate for 15 m within 90 s beginning with the right hand and thereafter the left hand.</p> <p>Overhead working: Standing with hands at crown height for manipulation of nuts and bolts. The time that the position was held is recorded (sec).</p> <p>Repetitive reaching: fast horizontal movements of the upper extremity in a sitting position. Marbles are removed from bowls at arm length distance at table height from left to right and vice versa, with right and then left arm. The time taken to remove 30 marbles is recorded (sec).</p> <p>Overhead lift test: Five lifts from waist to crown height and vice versa within 90 s in standing position. Increase weight in 4–5 steps until maximum is reached</p>

	<p>Repetitive bending and overhead reaching: 20 marbles in 2 bowls at table height and crown height. Standing in front of bowl of marbles and moving the marbles as fast as possible from table height to crown height.</p>
<p>A Physiotherapy Test Package^{33,34,35,36}</p>	<p>PILE Tests: “The lifting tests were performed standing in front of bookshelves with shelves at 0.76m and 1.37 m from the floor. Subjects were asked to lift weights in a plastic box from floor to waist level (0–0.76 m) for the lumbar PILE test, or from waist to shoulder height (0.76–1.37 m) for the cervical PILE test. The initial weight was 3.6 kg for women and 6.9 kg for men. A ‘lifting movement’ involved a single transfer from one level to the next and back again. After every four such lifting movements (= 20 s), the weight was increased by 2.5 kg for women and 4.5 kg for men. The weight managed during the last lifting movement was recorded and used as a test result, as well as this maximum weight divided by the ‘adjusted weight’”.</p> <p>2x20m WWB: “Subjects were asked to walk 20 m at a comfortable speed along a corridor, to turn around where 20 m was marked and then to walk 20 m back to the starting point. In the first walking test they carried no extra weight, but in the second they carried one carrier bag in each hand, containing 4 kg each for the women, 8 kg each for the men. The time taken was recorded to get the walking speed. The tests were discontinued after 50 s”.</p>
<p>BTEWS II³¹</p>	<p>“The protocol consisted of performing a series of shoulder functional tasks before and after a fatiguing activity. Functional tasks consisted of active shoulder range of motion (ROM) in both flexion and abduction and cumulative power output (PO) accumulated over 10s during a repetitive pushing/pulling task in a horizontal plane at shoulder level”.</p>
<p>FIT - HaNSA³²</p>	<p>“The FIT-HaNSA protocol consists of three timed tasks and each task is performed for a maximum of 300 seconds (s) with approximately 30 s pause between them (set-up time for next task). Task 1 (waist-up) requires the patient to alternately “grab, lift, move and place” three 1000 g containers located on waist level and 25 cm above waist level shelves, using their affected arm, at a metronome pace of 60 beats per minute for 300 s or until they felt unable to continue. The time to complete Task 1 is measured using a stopwatch. Task 2 (eye down) is identical to Task 1 except that the two shelves are placed at eye-level and 25 cm below. Task 3 (overhead work) requires a patient to repeatedly screw and unscrew bolts in a sagittal plane oriented plate</p>

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

positioned at eye-level using both arms". More complete description at https://srs-mcmaster.ca/wp-content/uploads/2015/04/FIT-HaNSAProtocol_April2007.pdf

For peer review only



PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	1
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	2
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	3
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	3
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	4
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	4
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	3-4
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	3-4
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	4
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	4
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	5
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	NA
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I ²) for each meta-analysis.	NA



PRISMA 2009 Checklist

Page 1 of 2

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	NA
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	NA
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	6-7
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICCO, follow-up period) and provide the citations.	6-7
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	6-10
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	6-10
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	6-10
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	6-10
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	NA
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	11-13
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	14-16
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	16
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	18

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit: www.prisma-statement.org.

Page 2 of 2

For peer review only - <http://bmjopen.bmj.com/site/about/guidelines.xhtml>

BMJ Open

Systematic Review of the Measurement Properties of Performance-based Functional Tests in Patients with Neck Disorders

Journal:	<i>BMJ Open</i>
Manuscript ID	bmjopen-2019-031242.R1
Article Type:	Original research
Date Submitted by the Author:	22-Aug-2019
Complete List of Authors:	McGee, Steven; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Sipos, Taylor; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Allin, Thomas; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Chen, Celia; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Greco, Alexandra; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Bobos, Pavlos; Western University, Health and Rehabilitation Sciences; University of Toronto, Dalla Lana School of Public Health, Institute of Health Policy Management and Evaluation MacDermid, Joy ; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Group, CATWAD; Michele Sterling, Anne Söderlund, Michele Curatolo, Jim Elliott, David M Walton, Helge Kasch, Linda Carroll, Hans Westergren, Samuel McLean, Gwendolen Jull, Genevieve Grant Luke Connelly, Joy C MacDermid, Mandy Nielsen, Pierre Côté, Tonny Elmoose Andersen, Trudy Rebbeck Annick Maujean, Sarah Robins, Kenneth Chen, Julia Treleaven
Primary Subject Heading:	Rehabilitation medicine
Secondary Subject Heading:	Rehabilitation medicine
Keywords:	functional, psychometric properties, neck pain, cervical, outcome measures

SCHOLARONE™
Manuscripts

1
2
3 1 **Title:** Systematic Review of the Measurement Properties of Performance-based Functional
4 Tests in Patients with Neck Disorders

5
6
7 3 ¹Steven McGee, PT

8
9 4 ²Taylor Sipos, PT

10
11 5 ³Thomas Allin, PT

12
13 6 ⁴Celia Chen, PT

14
15 7 ⁵Alexandra Greco, PT

16
17 8 ⁶Pavlos Bobos, PT, PhD(c) (corresponding author)

18
19 9 ⁷Joy MacDermid, PT, PhD

20
21 10 ⁸CATWAD

22
23 11

12 **Authors' information**

24
25 13 ¹Steven McGee PT, School of Physical Therapy, Department of Health and Rehabilitation
26 Sciences, Western University, London, Ontario, Canada, (smcgee7@uwo.ca)

27
28
29 15

30
31 16 ²Taylor Sipos PT, School of Physical Therapy, Department of Health and Rehabilitation Sciences,
32 Western University, London, Ontario, Canada, (jsipos@uwo.ca)

33
34
35 18

36
37 19 ³Thomas Allin PT, School of Physical Therapy, Department of Health and Rehabilitation Sciences,
38 Western University, London, Ontario, Canada, (tallin@uwo.ca)

39
40
41 21

42
43 22 ⁴Celia Chen PT, School of Physical Therapy, Department of Health and Rehabilitation Sciences,
44 Western University, London, Ontario, Canada, (qchen224@uwo.ca)

45
46
47 24

48
49 25 ⁵Alexandra Greco PT, School of Physical Therapy, Department of Health and Rehabilitation
50 Sciences, Western University, London, Ontario, Canada, (agreco33@uwo.ca)

51
52
53 27

54
55 28 ⁶Pavlos Bobos PT, PhD(c), (corresponding author) Doctoral Candidate, Western's Bone and Joint
56 Institute, Department of Health and Rehabilitation Sciences, Western University, Elborn College,
57 1201 Western Road, N6G 1H1, London, Ontario, Dalla Lana School of Public Health, Institute of

31 Health Policy Management and Evaluation, Department of Clinical Epidemiology and Health Care
32 Research, University of Toronto, Canada, (pbobos@uwo.ca), tel: +1 519 661 2111 x88912

33
34 ⁷Joy C MacDermid BScPT, PhD, Professor, Physical Therapy and Surgery, Western University,
35 London, ON and Co-director Clinical Research Lab, Hand and Upper Limb Centre, St. Joseph's
36 Health Centre, London, Ontario; Professor Rehabilitation Science McMaster University,
37 Hamilton, ON, Canada (jmacderm@uwo.ca)

38
39 ⁵CATWAD Coauthors: Michele Sterling m.sterling@uq.edu.au, Anne Söderlund
40 anne.soderlund@mdh.se, Michele Curatolo curatolo@uw.edu, Jim Elliott [j-](mailto:j-elliott@northwestern.edu)
41 elliott@northwestern.edu, David M Walton dwalton5@uwo.ca, Helge Kasch helgkasc@rm.dk,
42 Linda Carroll linda.carroll@ualberta.ca, Hans Westergren Hans.Westergren@skane.se, Samuel A
43 McLean, Samuel_McLean@med.unc.edu, Gwendolen Jull g.jull@uq.edu.au, Genevieve Grant
44 genevieve.grant@monash.edu, Luke Connelly l.connelly@uq.edu.au, Joy C MacDermid,
45 jmacderm@uwo.ca, Mandy Nielsen mandy.nielsen@griffith.edu.au, Pierre Cote
46 pierre.cote@uoit.ca, Tonny Elmoose Andersen tandersen@health.sdu.dk, Trudy Rebbeck
47 trudy.rebbeck@sydney.edu.au, Annick Maujean a.maujean@uq.edu.au, Sarah Robins
48 s.robins1@uq.edu.au, Kenneth Chen k.chen8@uq.edu.au, Julia Treleaven
49 j.treleaven@uq.edu.au

50
51 **Key Words:** functional, psychometric properties, neck, cervical, outcome measures

52
53 **Word Count:** 4509

1
2
3 **61 Abstract**

4
5 **62 Objectives:** The purpose of this systematic review is to identify and synthesize studies evaluating
6
7
8 **63** performance-based outcome measures designed to evaluate the functional abilities of patients with
9
10 **64** neck pain.

11
12 **65 Design:** Systematic review

13
14
15 **66 Data Sources:** A literature search using PubMed, Scopus, CINAHL, EMBASE, COCHRANE,
16
17 **67** Google Scholar, and a citation mapping strategy was conducted till July 2019

18
19 **68 Eligibility criteria:** More than half of the study's patient population had neck pain or a
20
21
22 **69** musculoskeletal neck disorder and completed a functional-based test. Clinimetric properties of at
23
24 **70** least one performance-based functional tests were reported. Both traumatic and non-traumatic
25
26 **71** origins of neck pain were considered.

27
28 **72 Data extraction and synthesis:** Relevant data were then extracted from selected articles using an
29
30
31 **73** extraction guide. Selected articles were appraised the Quality Appraisal for Clinical Measurement
32
33 **74** Research Reports Evaluation Form (QACMRR).

34
35 **75 Results:** The search obtained 12 articles which reported on 4 outcome measures (Functional
36
37
38 **76** Capacity Evaluations (FCE), Baltimore Therapeutic Equipment Work Simulator II (BTEWS II),
39
40 **77** Functional Impairment Test- Hand and Neck/Shoulder/Arm (FIT-HaNSA)) reporting to assess the
41
42 **78** functional abilities in patients with mechanical neck pathology. Of the selected papers: 1 reports
43
44 **79** content validity, 5 construct validity, 4 reliability, 1 sensitivity to change, and 1 both reliability
45
46 **80** and construct validity. QACMRR scores ranged from 68% to 95%.

47
48
49 **81 Conclusions:** This review found very good quality evidence that the FIT-HaNSA has
50
51 **82** excellent inter and intra-rater reliability and very weak to weak convergent validity. Excellent
52
53
54 **83** quality evidence of fair test-retest reliability, weak convergent validity, and very weak known

84 groups validity for the BTEWS II test was found. Good to excellent quality evidence exists that an
85 FCE battery has poor to excellent reliability and very weak to strong validity. Good to excellent
86 quality of weak to strong validity and trivial to strong effect sizes were found for a physiotherapy
87 test package.

88 **Prospero registration:** CRD42018112358

91 **Strengths and limitations of this study**

- 92 • The psychometric properties of performance outcome measures for neck pain were
93 synthesized and critically appraised
- 94 • This study assessed the risk of bias and the quality of measurements properties
- 95 • The feasibility or usability of these tools was not assessed

97 **Introduction**

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

1
2
3 106 Outcome measures are a crucial component in monitoring patients with neck pain to
4
5 107 determine the effects of treatment[11,12], evaluation of interventions, guiding return to work, and
6
7
8 108 justifying treatment.[13,14] Several self-reported outcome measures currently exist to assess
9
10 109 disability and function in those with neck pain (e.g. the Neck Disability Index - NDI). [13]
11
12 110 Evidence-based clinical practice guidelines suggest that measures assessing physical performance
13
14 111 should also be used for people with neck pain.[15] Performance-based testing is where the
15
16 112 assessment is based on actual performance of a task or activity. Physical performance can be
17
18 113 assessed by testing a person's ability to execute a standardized activity in a standardized
19
20 114 environment (i.e. clinical setting).[16] Time to complete the activity, number of repetitions
21
22 115 performed, and weight lifted are frequently used to quantify the physical performance.[17]
23
24 116 Conversely, self-report measures examine patients' perception and experience of their ability to
25
26 117 perform functional tasks. [16] Previous research has demonstrated poor to fair relationships
27
28 118 between physical performance and self-report measures of ability in patients with various
29
30 119 musculoskeletal disorders suggesting that these measures assess different constructs of function.
31
32 120 [17,18] Consequently, physical performance tests and self-report measures complement each other
33
34 121 and may each contribute unique information about a patient's function. [19]
35
36
37
38
39

40 122 A fundamental component of monitoring outcomes is having reliable and valid tools with
41
42 123 known measurement properties.[20,21] While recent research has investigated the psychometric
43
44 124 properties of patient-reported outcomes in people with neck pain [21,22] there is a gap in
45
46 125 knowledge with respect to performance-based functional outcomes. The purpose of this systematic
47
48 126 review was to identify and synthesize clinical measurement studies that evaluate measurement
49
50 127 properties of performance-based functional tests in patients with neck disorders.
51
52
53

54 128
55
56
57
58
59
60

129 **METHODS**

130 **Patient and Public Involvement**

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

132

133 **Study Design and Protocol Registration**

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

137

138 **Search Strategy**

139 A database search using CINAHL, PubMed, Scopus and Google Scholar was performed
140 to identify articles published till July 2019. The following search strategy was used to search all
141 databases for eligible studies: (Reliability OR validity OR responsiveness OR calibration OR
142 validation) OR (minimal detectable change) OR (clinically important difference) OR
143 (psychometric properties) AND cervical OR neck OR c-spine AND (performance measure) OR
144 (functional test) OR (functional outcome) OR (performance outcome). MeSH terms were searched
145 in PubMed. A citation map of articles and systematic reviews selected for the full-text review was
146 performed. This strategy was included to minimize the risk of publication bias. The full search
147 strategy is summarized in **APPENDIX 1**. The Preferred Reporting Items for Systematic Reviews
148 and Meta-Analyses (PRISMA) process[23] was followed to ensure all appropriate steps were taken
149 in the selection process (**FIGURE 1**).

150

151 **Inclusion Criteria**

1
2
3 152 Articles were included in the final review if all of the following criteria were met:

- 4
5 153
- 6 • >50% of the study's patient population had neck pain or a musculoskeletal neck disorder
 - 7 (e.g. whiplash associated disorder (WAD II))
 - 8 154
 - 9 155 • Patients in the study completed a functional-based test
 - 10 156 • Clinometric properties of at least one performance-based test were reported.

11
12
13
14
15 157 A test was considered functional-based if it met the following criteria:

- 16
17 158
- 18 • assessment of a patient's ability to execute a standardized activity in a standardized
 - 19 159 environment
 - 20 160 • tests assessing muscular endurance (e.g. cervical flexion test) or proprioception were not
 - 21 161 deemed functional-based as they are often not reflective of physical working conditions.

22
23
24 162 Both traumatic and non-traumatic origins of neck pain were considered. Definitions for the
25 163 properties can be found in **APPENDIX A**.

26
27
28
29
30
31 164

32 33 165 **Article Selection**

34
35 166 Titles and abstracts generated by the search strategy were screened by two authors (SM
36 167 and PB) independently. Articles that met the inclusion criteria and selected for a full text review
37 168 were also reviewed in pairs of authors. Disagreements were resolved by the most experienced
38 169 author (JCM)

39
40
41
42
43
44
45 170

46 47 171 **Data Extraction**

48
49 172 Data extraction and critical appraisal was performed in pairs of two raters among the authors, after
50 173 the completion of a calibration session in which the most experienced author (JCM) reviewed the
51 174 data extraction tools with the authors that performed the data extraction. When reviewers disagreed

1
2
3 175 during data extraction and/or critical appraisal, and consensus could not be met, a third author
4
5 176 arbitrated. A data extraction form [24] (**APPENDIX A and APPENDIX B**), developed by one of
6
7
8 177 the authors (JCM.), was used to ensure systematicity. Authors extracted sample size, patient
9
10 178 population characteristics, functional tests performed and reported psychometric properties. The
11
12 179 ICC interpretation of $ICC < 0.40$ indicating poor, $0.40 \leq ICC < 0.75$ indicating fair-to-good and
13
14 180 $ICC \geq 0.75$ indicating excellent reliability were used as a common benchmark. For validity
15
16 181 estimates, correlation coefficient (Pearson's/Spearman) and the 95% confidence intervals were
17
18 182 extracted if were available. [24,25] Evan's guidelines to interpret the strength of the correlation
19
20 183 was used which included: 0.00–0.19 “very weak”, 0.20–0.39 “weak”, 0.40–0.59 “moderate”,
21
22 184 0.60–0.79 “strong”, and 0.80–1.00 “very strong”. To assist clinical decision making, standard
23
24 185 benchmark scores of trivial (< 0.20), small (≥ 0.20 to < 0.50), moderate (≥ 0.50 to < 0.80) or large
25
26 186 (≥ 0.80), as proposed by Cohen, were used. [26]
27
28
29
30
31 187
32
33 188

189 **Quality Appraisal for Clinical Measurement Research Reports Evaluation Form**

34
35
36 190 Pairs of authors critically appraised the quality of each study using a standardized 12-item
37
38 191 evaluation tool (QACMRR) designed to assess the quality of studies determining measurement
39
40 192 properties in outcome measures (**APPENDIX C**). If disagreement was present a third person (JM)
41
42 193 assist in resolving the discrepancy. [24] This tool has been found to have good to excellent pre-
43
44 194 consensus inter-rater reliability (ICC: 0.69-0.91) across a number of systematic reviews.[24,25,27]
45
46 195 The evaluation criteria of this tool included twelve items: 1) Thorough literature review to define
47
48 196 the research question; 2) Specific inclusion/exclusion criteria; 3) Specific hypotheses; 4)
49
50 197 Appropriate scope of psychometric properties; 5) Sample size; 6) Follow-up; 7) The authors
51
52
53
54
55
56
57
58
59
60

1
2
3 198 referenced specific procedures for administration, scoring, and interpretation of procedures; 8)
4
5 199 Measurement techniques were standardized; 9) Data were presented for each hypothesis; 10)
6
7
8 200 Appropriate statistics-point estimates; 11) Appropriate statistical error estimates; and 12) Valid
9
10 201 conclusions and recommendations. [24,25] Each item is scored from 0 to 2 with (score=2) is the
11
12 202 best; (score=1) is acceptable but suboptimal; (score=0) is not done/documentated, substantially
13
14 203 inadequate or inappropriate. An article's total score – quality - was calculated by the sum of scores
15
16 204 for each item, divided by the numbers of items and multiplied by 100%. [24,25] Overall, the quality
17
18 205 summary of appraised articles ranges from (0%-30%) Poor, (31%-50%) Fair, (51%-70%) Good,
19
20 206 (71%-90%) Very Good, and (>90%) Excellent
21
22
23
24 207
25
26 208
27

28 209 **RESULTS**

30
31 210 The search strategy resulted in 840 published articles. After duplications were removed, 31
32
33 211 articles were deemed relevant and were screened at full text. Overall, 12 articles met our inclusion
34
35 212 criteria (**FIGURE 1**). The excluded articles were removed due to inappropriate patient
36
37 213 populations, investigations into self-report measures or tests assessing proprioception/muscular
38
39 214 endurance rather than functional-based measures, or because the articles were found to be
40
41 215 systematic reviews. The characteristics of the included studies and the summary of psychometric
42
43 216 properties are presented in **TABLE 1**. The quality assessment is summarized and presented in
44
45 217 **TABLE 2**. Percent agreement was calculated for quality scores between the 2 raters and it was
46
47 218 90%.
48
49
50
51

52 219

54 220 **Participants**

55
56
57
58
59
60

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

228

229 **Functional-Based Tests**

230 The 12 articles that were included for review provided properties on the following
231 functional based tests: Functional Capacity Evaluations (FCE)[28–33], The Baltimore Therapeutic
232 Equipment Work Simulator II (BTEWS II) [34], Functional Impairment Test- Hand and
233 Neck/Shoulder/Arm (FIT-HaNSA) [35], as well as items off of a physiotherapy test package
234 including a cervical and lumbar Progressive Isoinertial Lifting Evaluation (PILE-C, PILE-L) test
235 [36–39] and 2 x 20 m with burden walking test (2x20M-WWB) [36–39]. Descriptions of all
236 functional-based tests and their relevant subtasks are provided in **APPENDIX D**.

237

238 **Functional Capacity Evaluations (FCE)**

239 Six articles reported measurement properties for an FCE battery. We identified multiple
240 versions of the FCE in the literature with one article reporting properties on the Workwell FCE
241 [29], two reporting on the Whiplash Associated Disorder (WAD) FCE [28,30] and three reporting
242 on the neck-FCE.[31–33] These test batteries include various combinations of muscular strength,

243 endurance and functional based tests. The measurement properties of the functional based tests
244 used by the FCE are outlined in **TABLE 3**.

245

246 *Individuals with Sub-acute to chronic WAD*

247 Trippolini et al. (2014)[29] evaluated the Workwell FCE test-retest reliability,
248 measurement error, convergent validity and predictive criterion validity of future work capacity in
249 workers diagnosed with WAD I or II. Interclass Correlation Coefficients (ICC) ranged from 0.66
250 to 0.96 (good to excellent). Limits of agreement relative to mean performance ranged from 21 to
251 57% for functional based sub-tests. Correlations between FCE sub scores and baseline work
252 capacity were very weak to weak ranging between $r=0.06$ and $r=0.39$. FCE sub scores did not
253 predict future work capacity at 1, 3, 6 and 12 months.

254 Trippolini et al. (2015)[28] assessed the WAD FCE (31) and evaluated convergent validity
255 and known-groups validity. FCE subscales showed very weak to strong correlations (0.15-0.68)
256 with each of: pain, self-reported functional ability, self-reported disability, anxiety and depression.
257 It was found that the FCE had known-group sex validity (males vs females) for 1 of 3 functional
258 subtests (lifting waist-overhead) and reported significant performance differences between culture
259 groups (German vs non-German language groups).

260

261 *Work-Related Neck Disorders*

262 Reesink et al. (2007)[33] developed an independent FCE for patients with musculoskeletal
263 neck disorders (neck FCE). They performed a review of epidemiological literature and identified
264 four physical risk factors for work-related neck disorders and used that information to develop an
265 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)[31] measured test-retest reliability of the subscales of the neck FCE in patients with 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.[32] 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.

The Baltimore Therapeutic Equipment Work Simulator II (BTEWS II)

Chronic Neck Pain

Lomond and Côté, (2011)[34] 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 of fair and measured at $\rho=0.37$ and $ICC_{2,1} = 0.54$, respectively. The standard error 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.

289 **Functional Impairment Test- Hand and Neck/Shoulder/Arm (Fit-HaNSA)**

290 *Sub-acute to chronic WAD*

291 Pierrynowski et al. (2016)[35] reported on the reliability, measurement error, MDC and
292 validity of the Fit-HaNSA test in a sample of people with WAD II following motor vehicle
293 collision (MVC) (**TABLE 5**). Intra-rater reliability ICC's for patient subtask and total scores were
294 good to excellent ranging between 0.70-0.78. [35] Inter-rater reliability ICC's for patient subtask
295 and total scores were fair to excellent and ranged between 0.54-0.84. [35] The Bland and Altman
296 plot for the patient group showed a 26 seconds (s) bias in terms of improved performance on the
297 second test (possible learning effect). The standard deviation of difference was 124 s and 95%
298 Limits of Agreement (LoA₉₅) was 248 seconds. [35] The SEM for people with WAD II was
299 reported to be 76 s. The MDC₉₀ was measured as 176 s. [35]

300 Spearman rank correlations were also calculated between the Fit-HANSA, Numeric Pain
301 Rating Scale (NPRS), NDI, the disabilities of arm, hand and shoulder (DASH) and 6 cervical range
302 of motion measures. Most (59 of 78) of the correlations between performance and comparator
303 measures were very weak to weak ($r < 0.4$). [35] All correlations between total Fit-HaNSA scores
304 and subtask scores had good correlations ($r < 0.75$), except for Task 1-Task 3. [35] Significant
305 performance differences between WAD II and control groups (known group validity) were
306 recorded for the total Fit-HaNSA score and all 3 subtask scores. [35]

307

308 **Physiotherapy Test Package Subtests**

309 Ljungquist et al. published a series of articles[36–39] which evaluated the clinimetric
310 properties of a physiotherapy test package for patients with spinal pain (**TABLE 6**). This
311 package included muscular strength & endurance tests, submaximal endurance tests, and three

1
2
3 312 functional tests. These functional tests included the PILE-C, PILE-L, and 2x20M-WWB test.

4
5 313 Ljungquist's series of articles reported on convergent validity, known-groups validity, reliability,

6
7 314 measurement error and sensitivity to change for these tests. [36–39]

8
9 315

10 11 12 316 *Undetermined duration of neck pain*

13
14 317 In a 1999 article [38], correlations between the tests of the package and pain (CR-10) and
15
16 318 perceived exertion (Borg RPE) were determined. All correlations were very weak to moderate
17
18 319 (0.10-0.48) except for moderate to strong correlations (0.55-0.65) between the PILE-C test and
19
20 320 pain intensity and between 2x20M-WWB test and pain intensity.

21
22 321 In a 2003 article[36], the PILE-C, PILE-L and 2x20M-WWB tests were tested to determine
23
24 322 their ability to discriminate between known-groups (neck pain vs back pain). Subjects with spinal
25
26 323 pain completed the CR-10, the University of Alabama Pain Behavior scale (UAB) and the Borg
27
28 324 RPE test. Specific cut points were used to distinguish patients with high vs. low pain intensity,
29
30 325 high vs. low pain behavior, and high vs. low perceived exertion in patients, respectively.
31
32 326 Participants then completed the test package and it was determined if each subtest could
33
34 327 discriminate between participants with high vs. low pain intensity. The functional tests were able
35
36 328 to discriminate between all 3 subgroups with the exception of the PILE-C being unable to
37
38 329 discriminate between participants with high vs. low perceived exertion.

39
40 330 In a paper from 1999[38], the PILE-C, PILE-L and 2x20M-WWB tests were found to have
41
42 331 significant discriminative abilities in distinguishing healthy subjects from patients with spinal pain.
43
44 332 The sensitivity and specificity for this known group discrimination for the PILE-C test, were
45
46 333 reported to be 0.93 (very strong) and 0.69 (strong), respectively. The sensitivity and specificity for
47
48 334 the PILE-L test were reported to be 0.85 (very strong) and 0.65 (strong), respectively.

1
2
3 335 The inter and intra rater reliability were tested on participants with spinal pain.[37] Limits
4
5 336 of agreement were used to measure inter rater reliability and repeatability, defined as 2x the within-
6
7
8 337 subject standard deviation of each variable. Interrater agreement for 2 tests was deemed
9
10 338 “acceptable”, while all 3 functional tests had “clinically acceptable” intra-rater reliability.

11
12 339 Sensitivity-to-change was evaluated in the test package following 6 months of a
13
14 340 physiotherapy intervention. Using ROC curves, Wilcoxon sign ranked tests and spearman
15
16 341 correlation coefficients, only the 2x20m-WWB test and the PILE-C (women only) were deemed
17
18 342 to be sensitive to change. [39] Additionally, moderate to large effect sizes were found for all test
19
20 343 components.
21
22
23

24 344

25 26 345 **DISCUSSION**

27
28 346 This study synthesized 12 studies assessing clinometric properties of 4 different functional-
29
30 347 based assessments. Given the limited number of studies, the substantial variation in the types of
31
32 348 tests examined, the methods used to assess the clinical measurement properties, and the study
33
34 349 populations, the current state of knowledge does not allow firm conclusions regarding
35
36 350 recommendations for an optimal functional-based test at this time. Overall, the quality ranging
37
38 351 from good to excellent (67-92%,) as determined by the QACMRR, for a range of properties of the
39
40 352 4 different assessments in patients with acute or chronic neck pain that is musculoskeletal in origin.
41
42 353 Studies obtaining higher percentages indicate research that has been consistent with best practice
43
44 354 where studies with lower percentages are more likely to be inadequate or inappropriate
45
46
47
48

49 355 **FCE**

50
51 356 The breadth of a functional-based test is variable and defined by the developers. An
52
53 357 advantage of the functional assessment designed by Reesink et al.[33] is that they mapped the
54
55
56
57
58
59
60

1
2
3 358 eight subtests to risk factors identified in the literature for work-related neck disorders. The eight
4
5 359 subtests consist of: material handling tasks, lifting floor to waist, overhead lift test, one-handed
6
7
8 360 and two-handed carrying, overhead working, repetitive reaching, overhead lifting, and repetitive
9
10 361 bending and overhead reaching. Given the systematic approach and rationale these authors used
11
12 362 in developing the FCE and this approach being used in previous research [40], we suggest that
13
14 363 this test has strong content validity.

15
16
17 364 Six articles address the clinical measurement properties of this FCE ranging from good to
18
19 365 excellent quality (67-92%). There was evidence that the FCE was stable over test-retest time of
20
21 366 7-14 days. [30,31] These measures demonstrate longer stability over time compared to self-report
22
23 367 measures such as the Neck Disability Index (NDI) which has demonstrated test-retest reliability
24
25 368 within only a short period of 0-3 days. [27] Whether this longer-term stability is a characteristic of
26
27 369 functional-based tests or reflects differences in study populations in context requires further
28
29 370 testing. These two studies had relatively lower quality scores on the QACMRR (67-75%)
30
31 371 compared to other studies in this review putting into question test-retest time. Although test-retest
32
33 372 reliability has been assessed, inter-rater and intra-rater reliability has yet to be researched. Unlike
34
35 373 self-report measures, we expect measurement error due to the evaluator and functional-based tests.
36
37 374 Thus, future research should explore these aspects of reliability.

38
39
40 375 Convergent validity is often examined in clinical measurement studies. We suggest that
41
42 376 this may be because these comparisons are easily performed by correlating different tests rather
43
44 377 than providing strong confidence in the validity of the measurement. Often convenient
45
46 378 comparisons are performed rather than those most relevant. Across many domains and measures
47
48 379 it has become clear that the relationship between self-reported function and performance-based
49
50 380 function or physical impairment is often very weak to moderate. Therefore, the value of assessment
51
52
53
54
55
56
57
58
59
60

1
2
3 381 of these relationships as a form of validation has limited value. Several studies of very good to
4
5 382 excellent quality have reported on the convergent validity of the FCE. [28,29,32] The highest
6
7
8 383 quality article determined by the QACMRR (92%) found the relationship between the FCE and
9
10 384 work capacity to be poorly associated with one another. [29] The same study found that the ability
11
12 385 of the FCE to predict future work capacity was poor. This may be considered a more important
13
14 386 comparison since ideally functional-based tests would relate to important outcomes like return to
15
16 387 work. No studies to our knowledge report the responsiveness or sensitivity to change of the FCE.
17
18
19 388 This is an important gap since the focus of rehabilitation is often to remediate limitations in goal
20
21 389 impairments or work capacity, and assessment of these changes is critical to clinical decision-
22
23 390 making and reporting outcomes. Thus, future research should evaluate the responsiveness of the
24
25
26 391 FCE to provide insight in the measure's ability to detect change after an intervention.
27

28 392 **FIT-HaNSA**

29
30
31 393 One study of very good quality (88%) assessed the FIT-HaNSA, a test consisting of two
32
33 394 reaching tasks (waist and eye-level) and sustained overhead task performance. [35] Overall, the
34
35 395 FIT-HaNSA demonstrated excellent inter-rater reliability (0.84) and intra-rater reliability (0.78).
36
37 396 The specific subtests included within the FIT-HaNSA similarly demonstrate fair to excellent (0.54-
38
39 397 0.80) and good (0.70-0.72) inter-rater and intra-rater reliability respectively. The FIT-HaNSA also
40
41
42 398 demonstrated a clear ability to distinguish between people with WAD 2 and healthy controls.
43
44 399 Correlations between the FIT-HaNSA and other patient self-report disability and functional
45
46 400 outcome measures (NPRS, NDI, DASH, CROM and FIT-HaNSA) were generally very weak to
47
48 401 weak ($p < 0.4$), consistent with other studies comparing performance and self-report. [17,18] The
49
50 402 largest limitation in critically synthesizing information for this test is that only a single study was
51
52
53 403 found that reported the measurement properties for people with neck disorders. It should be noted
54
55
56
57
58
59
60

1
2
3 404 however that it has been validated in other MSK disorders. [34,40] Although others have noted
4
5 405 the lag in development of functional-based measures in comparison to self-report measures, FIT-
6
7
8 406 HaNSA was recommended as a functional-based measure for people with shoulder disorders. [41]
9

10 407 **BTEWS II**

11
12 408 Another study of very good quality (88%) assessed the efficacy of the BTEWS II where
13
14 409 the participants performed a dynamic pushing and pulling task in which power output was recorded
15
16
17 410 over a 10 second sample.[34] While the convergent validity aspect of this paper was assessed as
18
19 411 consistent with best practice through the critical appraisal process, the relationship between the
20
21 412 power output on the BTEWS and measures of pain and disability (NDI, SPADI, NRS) were poorly
22
23 413 associated with each other. In addition, the power output component was not found to be
24
25 414 significantly different between people with neck pain and healthy controls which suggests it might
26
27 415 not be discriminative. Discrimination between patients and those without any symptoms is a low
28
29 416 benchmark, and tests that cannot fulfil this benchmark should be viewed with caution. Because of
30
31 417 the weak measurement properties demonstrated by the power output component of the BTEWS II,
32
33 418 it does not appear to be a desirable functional-based measure to assess function in people with
34
35 419 neck pain. However, we acknowledge for all of the functional-based tests the evidence pool is so
36
37 420 shallow that there is high potential that future studies might lead to different conclusions.
38
39
40

41 421 **Physiotherapy Test Package Subtests**

42
43 422 Four studies ranging from good to very good quality (68-82%) assessed relevant items
44
45 423 from a physiotherapy test package, including a lift from floor-to-waist and a waist-to-shoulder task
46
47 424 and a two-handed carrying task. The properties of these assessment items include weak to
48
49 425 moderate correlations to pain, perceived exertion, and had “fair to good” reliability. The 2x20m-
50
51 426 WWB and PILE-C tests were found to be sensitive-to-change which is valuable information as no
52
53
54
55
56
57
58
59
60

1
2
3 427 other study has assessed this property in functional-based measures in patients with neck disorders.
4
5 428 Thus, this measure may be of value in clinical settings when assessing functional capacity before
6
7 429 and after a treatment intervention. All tests had discriminative ability for detecting participants
8
9 430 with spinal pain vs healthy controls. Most of the three tests demonstrated poor construct validity
10
11 431 in that they were poorly related to pain and perceived exertion. Thus, further research is necessary
12
13 432 to investigate these constructs.
14

15 433 **Clinical Implications**

16
17 434 This study confirms that functional-based tests have had far less development and
18
19 435 evaluation than self-report measures. Limitations include the number of tests and insufficient body
20
21 436 of evidence to make confident recommendations with respect to functional-based testing. It is clear
22
23 437 that self-report and functional-based measures provide different perspectives. Theoretically,
24
25 438 functional-based tests are important to inform our understanding about the mechanisms of
26
27 439 intervention and how interventions increase capacity. Overall more work is required to further
28
29 440 establish the psychometric properties of functional-based tests in persons with neck disorders,
30
31 441 including sensitivity-to-change, responsiveness, and predictive validity.
32
33

34
35 442 The data presented suggest that the FIT-HaNSA has the strongest clinometric properties
36
37 443 though this is based on a single higher quality paper specific to neck disorder. [35] Importantly,
38
39 444 normative data have been published [42], it has been validated in multiple studies in patients with
40
41 445 shoulder conditions [43–45] and has been recommended when compared to other measures [41].
42
43 446 The FCE has a limited evidence base from which to draw, though it was developed with strong
44
45 447 content validity and further evaluation may demonstrate its usefulness.
46
47

48 448 **Limitations**

1
2
3 449 A challenge in synthesizing clinical measurement evidence is the wide range of properties
4
5 450 and indicators that need to be considered. Unlike effectiveness studies where one can focus on the
6
7 451 effect size of treatment there are many considerations that would affect the recommendations made
8
9 452 about outcome measures. This is further complicated when the pool of evidence is shallow.
10
11 453 Although the quality assessment tool (QACMRR) developed by one of the authors of this review
12
13 454 which assess the quality of design of individual studies were useful for interpreting the evidentiary
14
15 455 pool, there is no clear method to synthesize the extracted clinical measurement evidence. While
16
17 456 some systematic reviews on treatment might only report findings from high-quality studies, it is
18
19 457 important to see how outcome measures perform in different contexts. Further, the assessment of
20
21 458 quality is complicated given that clinical measurement studies have so many dimensions.
22
23 459 Therefore, exclusion of lower quality studies has questionable value. Thus, a more practical
24
25 460 approach is to consider quality when interpreting the findings, rather than excluding studies.

26
27 461 The QACMRR focuses on whether the authors made appropriate decisions in selecting the
28
29 462 scope and methods of their clinical measurement evaluations within a given study and provides
30
31 463 descriptors of poor fair or good design options. Quality focuses on issues that might affect risk of
32
33 464 bias or imprecision in estimates; whereas risk of bias assessments focusses on items that might
34
35 465 result in a biased estimate. For example, insufficient power is a precision (quality) issue, not a risk
36
37 466 of bias. Although it is difficult to interpret the meaning of the percentage of the QACMRR as there
38
39 467 are no established cut-offs for distinguishing good and poor-quality studies, it provides one way
40
41 468 of ranking the articles in order of quality. We did not use COSMIN checklist since it was developed
42
43 469 for PROMS and some of the components/steps that involved are not applicable to performance-
44
45 470 based tests.

1
2
3 471 Another limitation in this review was that the feasibility or usability of these tools was not
4
5 472 assessed. While feasibility was not the focus of this review, information on the practical
6
7 473 application of these functional-based measures provides valuable information to clinicians for
8
9 474 determining whether these tests are appropriate to use in their given setting. Thus, future research
10
11 475 should not only investigate further the psychometric properties of these tools, but also report the
12
13 476 feasibility of using these tests so that they may be used in clinical settings and to identify
14
15 477 limitations that restrict their application in practice.
16
17
18
19 478

21 479 **CONCLUSION**

24 480 This review found very good quality evidence that the FIT-HaNSA has excellent inter and
25
26 481 intra-rater reliability and very weak to weak convergent validity. Excellent quality evidence of fair
27
28 482 test-retest reliability, weak convergent validity, and very weak known groups validity for the
29
30 483 BTEWS II test was found. Good to excellent quality evidence exists that an FCE battery has poor
31
32 484 to excellent reliability and very weak to strong validity. Good to excellent quality of weak to strong
33
34 485 validity and trivial to strong effect sizes were found for a physiotherapy test package. Functional-
35
36 486 based evaluation in people with neck disorders is an area needing much research attention both to
37
38 487 establish the measurement properties of existing measures, potentially to develop innovative new
39
40 488 measures and to perform head-to-head comparisons of measures before an optimal functional-
41
42 489 based test can be identified.
43
44
45
46
47 490

49 491 **Authors' contributions**

51 492 SM contributed significantly to conception and design of the study, data extraction, critical
52
53 493 appraisal, interpretation of data and drafting of the manuscript. TS, TA, PB, and CC were involved
54
55 494 in literature search, critical appraisal and interpretation of data and drafting. AG was involved in
56
57
58
59

1
2
3 495 critical appraisal and drafting. JM was also involved in the conception and design of the study,
4 496 drafting, and revised the manuscript for important intellectual content. PB and CATWAD were
5 497 involved in the drafting and review of the manuscript. All authors have given their final approval
6 498 on the manuscript to be published
7
8
9

10 499

11 500 **Declarations**

12 501 **Ethics approval and consent to participate**

13 502 Not applicable
14
15
16
17 503

18 504 **Consent for publication**

19 505 Not applicable
20
21
22 506

23 507 **Availability of data and material**

24 508 Data sharing is not applicable to this article as no datasets were generated or analyzed during the
25 509 current study
26
27
28
29 510

30 511 **Funding Statement**

31 512 This work was supported by the Canadian Institutes of Health Research (CIHR) with funding
32 513 reference number (FRN: SCA-145102).
33
34
35
36 514

37 515 **Competing Interest Statement**

38 516 None to report.
39
40
41
42
43 517

44 518 **References**

- 45 519 1 Carroll LJ, Hogg-Johnson S, van der Velde G, *et al.* Course and Prognostic Factors for
46
47 520 Neck Pain in the General Population. Results of the Bone and Joint Decade 2000-2010
48
49 521 Task Force on Neck Pain and Its Associated Disorders. *J Manipulative Physiol Ther*
50
51
52 522 Published Online First: 2009. doi:10.1016/j.jmpt.2008.11.013
53
54
55
56
57
58
59
60

- 1
2
3 523 2 Croft PR, Lewis M, Papageorgiou AC, *et al.* Risk factors for neck pain: A longitudinal
4
5 524 study in the general population. *Pain* Published Online First: 2001. doi:10.1016/S0304-
6
7 525 3959(01)00334-7
8
9
10 526 3 Vos T, Allen C, Arora M, *et al.* Global, regional, and national incidence, prevalence, and
11
12 527 years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis
13
14 528 for the Global Burden of Disease Study 2015. *Lancet* Published Online First: 2016.
15
16 529 doi:10.1016/S0140-6736(16)31678-6
17
18
19 530 4 Blanpied PR, Gross AR, Elliott JM, *et al.* Neck Pain: Revision 2017. *J Orthop Sport Phys*
20
21 531 *Ther* Published Online First: 2017. doi:10.2519/jospt.2017.0302
22
23
24 532 5 Nygren A, Berglund A, Von Koch M. Neck-and-shoulder pain, an increasing problem.
25
26 533 Strategies for using insurance material to follow trends. In: *Scandinavian Journal of*
27
28 534 *Rehabilitation Medicine, Supplement*. 1995.
29
30
31 535 6 Wright A, Mayer TG, Gatchel RJ. Outcomes of disabling cervical spine disorders in
32
33 536 compensation injuries: A prospective comparison to tertiary rehabilitation response for
34
35 537 chronic lumbar spinal disorders. *Spine (Phila Pa 1976)* Published Online First: 1999.
36
37 538 doi:10.1097/00007632-199901150-00020
38
39
40 539 7 Rempel DM, Harrison RJ, Barnhart S. Work-Related Cumulative Trauma Disorders of the
41
42 540 Upper Extremity. *JAMA J Am Med Assoc* Published Online First: 1992.
43
44 541 doi:10.1001/jama.1992.03480060084035
45
46
47 542 8 Borghouts JAJ, Koes BW, Vondeling H, *et al.* Cost-of-illness of neck pain in The
48
49 543 Netherlands in 1996. *Pain* Published Online First: 1999. doi:10.1016/S0304-
50
51 544 3959(98)00268-1
52
53
54 545 9 Hogg-Johnson S, van der Velde G, Carroll LJ, *et al.* The Burden and Determinants of

- 1
2
3 546 Neck Pain in the General Population. Results of the Bone and Joint Decade 2000-2010
4
5 547 Task Force on Neck Pain and Its Associated Disorders. *J Manipulative Physiol Ther*
6
7 548 Published Online First: 2009. doi:10.1016/j.jmpt.2008.11.010
8
9
10 549 10 Bobos P, Nazari G, Palimeris S, *et al.* The contribution of health and psychological factors
11
12 550 in patients with chronic neck pain and disability: A cross-sectional study. *J Clin*
13
14 551 *Diagnostic Res* Published Online First: 2018. doi:10.7860/JCDR/2018/31284.11203
15
16 552 11 Bobos P, Billis E, Papanikolaou D-T, *et al.* Does Deep Cervical Flexor Muscle Training
17
18 553 Affect Pain Pressure Thresholds of Myofascial Trigger Points in Patients with Chronic
19
20 554 Neck Pain? A Prospective Randomized Controlled Trial. *Rehabil Res Pract* Published
21
22 555 Online First: 2016. doi:10.1155/2016/6480826
23
24
25 556 12 Nazari G, Bobos P, Billis E, *et al.* Cervical flexor muscle training reduces pain, anxiety,
26
27 557 and depression levels in patients with chronic neck pain by a clinically important amount:
28
29 558 A prospective cohort study. *Physiother Res Int* 2018;**23**. doi:10.1002/pri.1712
30
31
32 559 13 Bobos P, MacDermid JC, Walton DM, *et al.* Patient-Reported Outcome Measures Used
33
34 560 for Neck Disorders: An Overview of Systematic Reviews. *J Orthop Sport Phys Ther*
35
36 561 2018;**48**:1–76. doi:10.2519/jospt.2018.8131
37
38
39 562 14 Macdermid JC, Walton DM, Bobos P, *et al.* The Open Orthopaedics Journal A Qualitative
40
41 563 Description of Chronic Neck Pain has Implications for Outcome Assessment and
42
43 564 Classification. *Open Orthop J* 2016;**10**:746–56. doi:10.2174/1874325001610010746
44
45
46 565 15 Childs JD, Cleland JA, Elliott JM, *et al.* Neck pain: Clinical practice guidelines linked to
47
48 566 the international classification of functioning, disability, and health from the orthopaedic
49
50 567 section of the american physical therapy association. *J. Orthop. Sports Phys. Ther.* 2008.
51
52 568 doi:10.2519/jospt.2008.0303
53
54
55
56
57
58
59
60

- 1
2
3 569 16 Kay TM, Huijbregts M. Physical Rehabilitation Outcome Measures: A Guide to Enhanced
4
5 570 Clinical Decision Making, Second Edition. *Physiother Canada* Published Online First:
6
7 571 2003. doi:10.2310/6640.2003.35271
8
9
10 572 17 Simmonds MJ, Olson SL, Jones S, *et al.* Psychometric characteristics and clinical
11
12 573 usefulness of physical performance tests in patients with low back pain. *Spine (Phila Pa*
13
14 574 *1976)* Published Online First: 1998. doi:10.1097/00007632-199811150-00011
15
16
17 575 18 Stratford PW, Kennedy D, Pagura SMC, *et al.* The relationship between self-report and
18
19 576 performance-related measures: Questioning the content validity of timed tests. *Arthritis*
20
21 577 *Rheum* Published Online First: 2003. doi:10.1002/art.11196
22
23
24 578 19 Novy DM, Simmonds MJ, Lee CE. Physical performance tasks: What are the underlying
25
26 579 constructs? *Arch Phys Med Rehabil* Published Online First: 2002.
27
28 580 doi:10.1053/apmr.2002.27397
29
30
31 581 20 MacDermid JC, Stratford P. Applying evidence on outcome measures to hand therapy
32
33 582 practice. *J Hand Ther* Published Online First: 2004. doi:10.1197/j.jht.2004.02.005
34
35
36 583 21 Bobos P, MacDermid JC, Walton DM, *et al.* Patient-Reported Outcome Measures Used
37
38 584 for Neck Disorders: An Overview of Systematic Reviews. *J Orthop Sport Phys Ther*
39
40 585 2018;**48**:775–88. doi:10.2519/jospt.2018.8131
41
42
43 586 22 Alreni ASE, Harrop D, Lowe A, *et al.* Measures of upper limb function for people with
44
45 587 neck pain. A systematic review of measurement and practical properties. *Musculoskelet*
46
47 588 *Sci Pract* 2017;**29**:155–63. doi:10.1016/J.MSKSP.2017.02.004
48
49
50 589 23 Moher D, Shamseer L, Clarke M, *et al.* Preferred reporting items for systematic review
51
52 590 and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev* Published Online
53
54 591 First: 2015. doi:10.1186/2046-4053-4-1
55
56
57
58
59
60

- 1
2
3 592 24 Law MC, MacDermid J. *Evidence-based rehabilitation : a guide to practice*. Thorofare,
4
5 593 NJ: : Slack Incorporated 2014.
6
7
8 594 25 Roy JS, Desmeules F, MacDermid JC. Psychometric properties of presenteeism scales for
9
10 595 musculoskeletal disorders: A systematic review. *J Rehabil Med* Published Online First:
11
12 596 2011. doi:10.2340/16501977-0643
13
14
15 597 26 Cohen J. Statistical power analysis for the behavioral sciences. *Stat. Power Anal. Behav.*
16
17 598 *Sci.* 1988. doi:10.1234/12345678
18
19 599 27 Macdermid JC, Walton DM, Avery S, *et al.* Measurement properties of the neck disability
20
21 600 index: A systematic review. In: *Journal of Orthopaedic and Sports Physical Therapy*.
22
23 601 2009. 400–16. doi:10.2519/jospt.2009.2930
24
25
26 602 28 Trippolini MA, Dijkstra PU, Geertzen JHB, *et al.* Construct Validity of Functional
27
28 603 Capacity Evaluation in Patients with Whiplash-Associated Disorders. *J Occup Rehabil*
29
30 604 Published Online First: 2015. doi:10.1007/s10926-014-9555-0
31
32
33 605 29 Trippolini MA, Dijkstra PU, Côté P, *et al.* Can functional capacity tests predict future
34
35 606 work capacity in patients with whiplash-associated disorders? *Arch Phys Med Rehabil*
36
37 607 Published Online First: 2014. doi:10.1016/j.apmr.2014.07.406
38
39
40 608 30 Trippolini MA, Reneman MF, Jansen B, *et al.* Reliability and safety of functional capacity
41
42 609 evaluation in patients with whiplash associated disorders. *J Occup Rehabil* Published
43
44 610 Online First: 2013. doi:10.1007/s10926-012-9403-z
45
46
47 611 31 Reneman MF, Roelofs M, Schiphorst Preuper HR. Reliability and Agreement of Neck
48
49 612 Functional Capacity Evaluation Tests in Patients With Chronic Multifactorial Neck Pain.
50
51 613 *Arch Phys Med Rehabil* Published Online First: 2017. doi:10.1016/j.apmr.2016.12.005
52
53
54 614 32 Van Der Meer S, Reneman MF, Verhoeven J, *et al.* Relationship between self-reported
55
56
57
58
59
60

- 1
2
3 615 disability and functional capacity in patients with Whiplash Associated Disorder. *J Occup*
4
5 616 *Rehabil* Published Online First: 2014. doi:10.1007/s10926-013-9473-6
6
7
8 617 33 Reesink DD, Jorritsma W, Reneman MF. Basis for a functional capacity evaluation
9
10 618 methodology for patients with work-related neck disorders. *J Occup Rehabil* Published
11
12 619 Online First: 2007. doi:10.1007/s10926-007-9086-z
13
14
15 620 34 Lomond K V., Côté JN. Shoulder functional assessments in persons with chronic
16
17 621 neck/shoulder pain and healthy subjects: Reliability and effects of movement repetition.
18
19 622 In: *Work*. 2011. doi:10.3233/WOR-2011-1119
20
21
22 623 35 Pierrynowski M, McPhee C, P. Mehta S, *et al*. Intra and Inter-Rater Reliability and
23
24 624 Convergent Validity of FIT-HaNSA in Individuals with Grade II Whiplash Associated
25
26 625 Disorder. *Open Orthop J* 2016;**10**:179–89. doi:10.2174/1874325001610010179
27
28
29 626 36 Ljungquist T, Jensen IB, Nygren Å, *et al*. Physical performance tests for people with long-
30
31 627 term spinal pain: Aspects of construct validity. *J Rehabil Med* Published Online First:
32
33 628 2003. doi:10.1080/16501970306117
34
35
36 629 37 Ljungquist T, Harms-Ringdahl K, Nygren A, *et al*. Intra- and inter-rater reliability of an
37
38 630 11-test package for assessing dysfunction due to back or neck pain. *Physiother Res Int*
39
40 631 Published Online First: 1999. doi:10.1002/pri.167
41
42
43 632 38 Ljungquist T, Fransson B, Harms-Ringdahl K, *et al*. A physiotherapy test package for
44
45 633 assessing back and neck dysfunction--discriminative ability for patients versus healthy
46
47 634 control subjects. *Physiother Res Int* Published Online First: 1999. doi:10.1002/pri.158
48
49 635 39 Ljungquist T, Nygren Å, Jensen I, *et al*. Physical performance tests for people with spinal
50
51 636 pain - Sensitivity to change. *Disabil Rehabil* Published Online First: 2003.
52
53
54 637 doi:10.1080/0963828031000090579
55
56
57
58
59
60

- 1
2
3 638 40 Reneman MF, Dijkstra PU, Westmaas M, *et al.* Test-retest reliability of lifting and
4
5 639 carrying in a 2-day functional capacity evaluation. *J Occup Rehabil* Published Online
6
7 640 First: 2002. doi:10.1023/A:1020274624791
8
9
10 641 41 Hegedus EJ, Vidt ME, Tarara DT. The best combination of physical performance and self-
11
12 642 report measures to capture function in three patient groups. *Phys Ther Rev* 2014;**19**:196–
13
14 643 203. doi:10.1179/1743288X13Y.0000000121
15
16
17 644 42 Roy J-S, Macdermid JC, Boyd KU, *et al.* Rotational strength, range of motion, and
18
19 645 function in people with unaffected shoulders from various stages of life. *Sports Med*
20
21 646 *Arthrosc Rehabil Ther Technol* 2009;**1**:4. doi:10.1186/1758-2555-1-4
22
23
24 647 43 Kumta P, MacDermid JC, Mehta SP, *et al.* The FIT-HaNSA demonstrates reliability and
25
26 648 convergent validity of functional performance in patients with shoulder disorders. *J*
27
28 649 *Orthop Sports Phys Ther* 2012;**42**:455–64. doi:10.2519/jospt.2012.3796
29
30
31 650 44 Macdermid JCJC, Ghobrial M, Badra Quirion K, *et al.* Validation of a new test that
32
33 651 assesses functional performance of the upper extremity and neck (FIT-HaNSA) in patients
34
35 652 with shoulder pathology. *BMC Musculoskelet Disord* 2007;**8**:42. doi:10.1186/1471-2474-
36
37 653 8-42
38
39
40 654 45 Hawkes DH, Alizadehkhayat O, Fisher AC, *et al.* Normal shoulder muscular activation
41
42 655 and co-ordination during a shoulder elevation task based on activities of daily living: An
43
44 656 electromyographic study. *J Orthop Res* 2012;**30**:53–60. doi:10.1002/jor.21482
45
46
47 657
48
49 658
50 659
51 660
52 661
53 662
54 663
55
56
57
58
59
60

1
2
3 664
4 665
5 666
6 667
7 668
8 669
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For peer review only

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. 1999	Neck pain (55%), back pain, multiple pain sites,	53	PILE-C, PILE-L	N/A	Good (68%)
Ljungquist et al. 1999	Neck pain (50%), lumbar pain, thoracic pain, shoulder pain, multiple pain sites,	68	PILE-C, PILE-L, 2 x 20m WWB	8 days	Very Good (79%)
Ljungquist et al. 2003	Neck pain, lumbar pain, thoracic pain, shoulder pain, lower extremity pain, multiple pain sites,	235	PILE-C, PILE-L, 2 x 20m WWB	N/A	Very Good (82%)
Ljungquist et al. 2003	cervical pain (25%), lumbar pain, cervical (25%) and lumbar pain, multiple pain sites,	186	PILE-C, PILE-L, 2 x 20m WWB	6 months	Very Good (79%)
Lomond and Cote. 2011	Chronic neck and shoulder pain (100%)	32	BTEWS II	9.5 days	Very Good (88%)
Pierrynowski et al. 2016	Sub-acute and chronic WAD II	66	FIT-HaNSA	2-7 days	Very Good (88%)
Reesink et al. 2007	N/A	N/A	Neck-FCE	N/A	N/A
Reneman et al. 2017	Chronic multifactorial neck pain	18	Neck-FCE	2 weeks	Good (67%)
Trippolini et al. 2013	Sub acute and chronic WAD I and II	32	WAD FCE	7 days	Very Good (75%)
Trippolini et al. 2014	Sub acute and chronic WAD I and II	267	Workwell FCE	N/A	Excellent (92%)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

Trippolini et al. 2015	Sub acute and chronic WAD I and II	314	WAD FCE	N/A	Very Good (86%)
Van der Meer et al. 2013	Chronic WAD I and II	40	Neck FCE	N/A	Very Good (86%)

PILE-C, Progressive Isoinertial Lifting Evaluation-Cervical; PILE-L, Progressive Isoinertial Lifting Evaluation; CBT, Cognitive-Behavioural Therapy; PT, Physical Therapy; NRPS, Numeric Pain Rating Scale; BTEWS II, Baltimore Therapeutic Equipment Work Simulator II; WAD, Whiplash Associated Disorder; MVA, Motor Vehicle Accident; FIT-HANSA, Functional Impairment Test-Hand and Neck/Shoulder/Arm; FCE, Functional Capacity Evaluation; EXP, Experimental; M, Male; F, Female

For peer review only

TABLE 2. Quality of Studies on Psychometric Properties of Functional-based Tests Evaluated in Neck Disorder Patients

Study	Item Evaluation Criteria												Total (%)
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	
Trippolini et al, 2014	2	2	2	2	1	2	2	2	2	2	1	2	92%
Lomond and Cote, 2011	2	2	1	2	0	2	2	2	2	2	2	2	88%
Pierrynowski et al, 2016	2	2	1	2	0	2	2	2	2	2	2	2	88%
Trippolini et al, 2015	2	2	2	0	1	N/A	2	2	2	2	2	2	86%
Van der Meer et al, 2013	2	1	2	1	2	N/A	2	1	2	2	1	2	86%
Ljungquist et al 2003 KGV	2	2	2	0	0	N/A	2	2	2	2	2	2	82%
Ljungquist et al 1999 Rel	2	1	1	2	0	2	2	2	2	2	1	2	79%
Ljungquist et al 2003 STC	1	1	1	2	1	1	2	2	2	2	2	2	79%
Trippolini et al, 2013	2	2	1	1	0	0	2	2	2	2	2	2	75%
Ljungquist et al 1999 KGV	2	1	1	2	0	N/A	2	1	2	2	1	2	68%
Reneman et al, 2017	1	2	1	1	1	0	1	2	2	2	2	1	67%

Reesink, 2007*	-	-	-	-	-	-	-	-	-	-	-	-	N/A
----------------	---	---	---	---	---	---	---	---	---	---	---	---	-----

*Paper is not applicable for completion of study quality tool

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	Good-excellent
	Convergent Validity	Pearson Correlation	Pain* 0.31-0.39	Weak
			SFS: 0.42-0.61	Moderate-strong
			NDI: 0.34-0.45	Weak-moderate
			HADS-A: 0.27-0.36 HADS-D: 0.30-0.41	Weak Weak-moderate
Known-groups Validity (German vs Non-German)	Linear Regression Analysis	p<0.001	Significant for All Tasks	
Known-groups 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	0.06-0.39	Very weak - Weak
		Linear Mixed Model Regression of All Predictors	β =-0.04, 95% CI: -0.15 – 0.06 p=0.428 (task 6)	Not Significant

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

*Pain measured via Numeric Rating Scale

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	Excellent
Fit-HaNSA	Inter-rater Reliability	ICC	0.84	Excellent
Fit-HaNSA	Measurement Error	SEM	76 s	
		LOA ₉₅	248 s	
		MDC ₉₀	176 s	
Fit-HaNSA	Convergent Validity	Spearman Rank Correlation	<0.4 - >0.75	Weak – Strong
Fit-HaNSA	Known-groups Validity 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	Good
	Inter-reliability	ICC	0.54-0.80	Fair - Excellent
	Convergent Validity	Spearman Rank Correlation	<0.4 - >0.75	Weak - Strong
	Known-groups Validity WAD II vs Control	F-test	42.0-53.3, p<0.001	Significant

Fit-HaNSA, Functional Impairment Test, Hand and Neck/Shoulder/Arm; ICC, Intraclass correlation coefficient; SEM, Standard Error of Measurement; LOA₉₅, 95% Limits of Agreement; MDC₉₀, 90% Minimal Detectable Change; WAD, Whiplash Associated Disorder; Mod, Moderate

*Correlations completed with Numeric Pain Rating Scale, Neck Disability Index, Disabilities of Arm, Shoulder, Hand and 6 cervical range of motion tests

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	Fair
		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	Known-groups Validity (Pain vs Control)	Two-way Repeated Measures ANOVA	Not Reported	Non-significant

ICC, Intraclass correlation coefficient; SEM, Standard Error of Measurement; MDC₉₀, 90% Minimal Detectable Change; ANOVA, Analysis of Variance

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

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	KGV: spinal pain vs. control	Sensitivity and Specificity	0.93, 0.69	Strong – Very Strong
PILE-C	KGV: spinal pain vs. control	Wilcoxon Sign Ranked Test	p=0.008	Significant
PILE-C	KGV: High vs. low pain intensity	Mann-Whitney U	p=0.003	Significant
PILE-C	KGV: High vs. low Pain behavior	Mann-Whitney U	p=0.005	Significant
PILE-C	KGV: 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	Very weak – moderate Very weak – moderate

				Borg RPE: 0.10 - 0.48	
PILE-L	KGV: spinal pain vs no spinal pain	Sensitivity and Specificity	0.85, 0.65		Strong – Very Strong
PILE-L	KGV: spinal pain vs control	Wilcoxon Sign Ranked Test	p=0.002		Significant
PILE-L	KGV: High vs. low pain intensity	Mann-Whitney U	p=0.001		Significant
PILE-L	KGV: High vs. low pain behaviour	Mann-Whitney U	p<0.001		Significant
PILE-L	KGV: 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 x 20m WWB	Inter-rater Reliability	Mean Difference LoA	0.05 -1.33 and 1.43		
2 x 20m WWB	Intra-rater Reliability	Repeatability % of Range	3.2 10.7%		
2 x 20m WWB	Convergent Validity	Spearman Correlation	CR-10: 0.55 - 0.65 Borg RPE: 0.10 - 0.48		Moderate - Strong very weak – moderate
2 x 20m WWB	KGV: spinal pain vs control	Wilcoxon Sign Ranked Test	p=0.014		Significant
2 x 20m WWB	KGV: High vs. low pain intensity	Mann Whitney U	p<0.001		Significant
2 x 20m WWB	KGV: High vs. low pain behaviour	Mann Whitney U	p<0.001		Significant
2 x 20m WWB	KGV: High vs. low perceived exertion	Mann Whitney U	p<0.001		Significant
2 x 20m WWB	Sensitivity to change	Effect Size	Subjects improving: 0.38-0.78		Small – Moderate Trivial – Moderate

Subjects deteriorating:
0.13-0.62

PILE-C, Progressive Iso-inertial Lifting Evaluation – Cervical; PILE-L, Progressive Iso-inertial Lifting Evaluation – Lumbar; LoA, Limits of Agreement; SD, Standard Deviation; M, Male; F, Female; RPE, Rating of perceived exertion; KGV, Known-groups Validity; Neg., Negligible; Mod., Moderate, *CR-10: Measurement of pain construct

For peer review only

36bmjopen-2019-031242 on 24 November 2019. Downloaded from <http://bmjopen.bmj.com/> on April 19, 2024 by guest. Protected by copyright.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

For peer review only

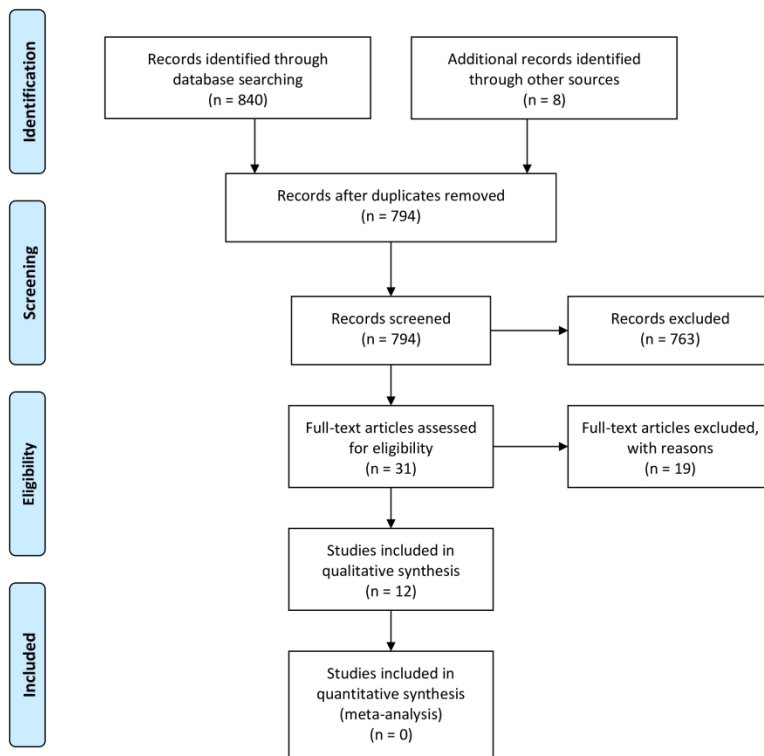


Figure 1

215x279mm (300 x 300 DPI)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Appendix 1: Search terms

EMBASE-OVID

1. exp "outcome and process assessment (health care)"/ or "outcome assessment (health care)"/ or treatment outcome/
2. outcome?.ti.
3. exp "Range of Motion, Articular"/
4. Pain Measurement/
5. exp disability evaluation/
6. "Recovery of Function"/
7. Questionnaires/
8. self-report.tw.
9. ((impairment or disability or function) adj2 (measure? or scale? or evaluation?)).tw.
10. range of motion.tw.
11. (strength adj2 (measure? or scale? or evaluation?)).tw.
12. (outcome? adj2 (measure* or scale? or indicator?)).tw.
13. or/1-12
14. "reproducibility of results"/
15. exp "Sensitivity and Specificity"/
16. reliability.mp.
17. validity.mp.
18. responsiveness.mp.
19. Psychometrics/
20. rasch.mp.
21. factor analysis, statistical/
22. factor analysis.tw.
23. differential functioning.mp.
24. (validity or validation).mp. [mp=title, original title, abstract, name of substance word, subject heading word, unique identifier]
25. (validity or validation).mp.
26. item difficulty.mp.
27. translation.tw.
28. or/14-27
29. 13 and 28
30. Neck Pain/
31. exp Brachial Plexus Neuropathies/
32. exp neck injuries/ or exp whiplash injuries/
33. cervical pain.mp.
34. neckache.mp.
35. whiplash.mp.
36. cervicodynia.mp.
37. cervicgia.mp.
38. brachialgia.mp.
39. brachial neuritis.mp.
40. brachial neuralgia.mp.
41. neck pain.mp.

- 1
- 2
- 3 42. neck injur*.mp.
- 4 43. brachial plexus neuropath*.mp.
- 5 44. brachial plexus neuritis.mp.
- 6 45. thoracic outlet syndrome/ or cervical rib syndrome/
- 7 46. Torticollis/
- 8 47. exp brachial plexus neuropathies/ or exp brachial plexus neuritis/
- 9 48. cervico brachial neuralgia.ti,ab.
- 10 49. cervicobrachial neuralgia.ti,ab.
- 11 50. (monoradicul* or monoradicl*).tw.
- 12 51. or/30-50
- 13 52. exp headache/ and cervic*.tw.
- 14 53. exp genital diseases, female/
- 15 54. genital disease*.mp.
- 16 55. or/53-54
- 17 56. 52 not 55
- 18 57. 51 or 56
- 19 58. neck/
- 20 59. neck muscles/
- 21 60. exp cervical plexus/
- 22 61. exp cervical vertebrae/
- 23 62. atlanto-axial joint/
- 24 63. atlanto-occipital joint/
- 25 64. Cervical Atlas/
- 26 65. spinal nerve roots/
- 27 66. exp brachial plexus/
- 28 67. (odontoid* or cervical or occip* or atlant*).tw.
- 29 68. axis/ or odontoid process/
- 30 69. Thoracic Vertebrae/
- 31 70. cervical vertebrae.mp.
- 32 71. cervical plexus.mp.
- 33 72. cervical spine.mp.
- 34 73. (neck adj3 muscles).mp.
- 35 74. (brachial adj3 plexus).mp.
- 36 75. (thoracic adj3 vertebrae).mp.
- 37 76. neck.mp.
- 38 77. (thoracic adj3 spine).mp.
- 39 78. (thoracic adj3 outlet).mp.
- 40 79. trapezius.mp.
- 41 80. cervical.mp.
- 42 81. cervico*.mp.
- 43 82. 80 or 81
- 44 83. exp genital diseases, female/
- 45 84. genital disease*.mp.
- 46 85. exp *Uterus/
- 47 86. 83 or 84 or 85
- 48 87. 82 not 86
- 49
- 50
- 51
- 52
- 53
- 54
- 55
- 56
- 57
- 58
- 59
- 60

- 1
- 2
- 3
- 4 88. 58 or 59 or 60 or 61 or 62 or 63 or 64 or 65 or 66 or 67 or 68 or 69 or 70 or 71 or 72 or 73 or
- 5 74 or 75 or 76 or 77 or 78 or 79 or 87
- 6 89. exp pain/
- 7 90. exp injuries/
- 8 91. pain.mp.
- 9 92. ache.mp.
- 10 93. sore.mp.
- 11 94. stiff.mp.
- 12 95. discomfort.mp.
- 13 96. injur*.mp.
- 14 97. neuropath*.mp.
- 15 98. or/89-97
- 16 99. 88 and 98
- 17 100. Radiculopathy/
- 18 101. exp temporomandibular joint disorders/ or exp temporomandibular joint dysfunction
- 19 syndrome/
- 20 102. myofascial pain syndromes/
- 21 103. exp "Sprains and Strains"/
- 22 104. exp Spinal Osteophytosis/
- 23 105. exp Neuritis/
- 24 106. Polyradiculopathy/
- 25 107. exp Arthritis/
- 26 108. Fibromyalgia/
- 27 109. spondylitis/ or discitis/
- 28 110. spondylosis/ or spondylolysis/ or spondylolisthesis/
- 29 111. radiculopathy.mp.
- 30 112. radiculitis.mp.
- 31 113. temporomandibular.mp.
- 32 114. myofascial pain syndrome*.mp.
- 33 115. thoracic outlet syndrome*.mp.
- 34 116. spinal osteophytosis.mp.
- 35 117. neuritis.mp.
- 36 118. spondylosis.mp.
- 37 119. spondylitis.mp.
- 38 120. spondylolisthesis.mp.
- 39 121. or/100-120
- 40 122. 88 and 121
- 41 123. exp neck/
- 42 124. exp cervical vertebrae/
- 43 125. Thoracic Vertebrae/
- 44 126. neck.mp.
- 45 127. (thoracic adj3 vertebrae).mp.
- 46 128. cervical.mp.
- 47 129. cervico*.mp.
- 48 130. 128 or 129
- 49 131. exp genital diseases, female/
- 50
- 51
- 52
- 53
- 54
- 55
- 56
- 57
- 58
- 59
- 60

132. genital disease*.mp.
133. exp *Uterus/
134. or/131-133
135. 130 not 134
136. (thoracic adj3 spine).mp.
137. cervical spine.mp.
138. 123 or 124 or 125 or 126 or 127 or 135 or 136 or 137
139. Intervertebral Disk/
140. (disc or discs).mp.
141. (disk or disks).mp.
142. 139 or 140 or 141
143. 138 and 142
144. herniat*.mp.
145. slipped.mp.
146. prolapse*.mp.
147. displace*.mp.
148. degenerat*.mp.
149. (bulge or bulged or bulging).mp.
150. 144 or 145 or 146 or 147 or 148 or 149
151. 143 and 150
152. intervertebral disk degeneration/ or intervertebral disk displacement/
153. intervertebral disk displacement.mp.
154. intervertebral disc displacement.mp.
155. intervertebral disk degeneration.mp.
156. intervertebral disc degeneration.mp.
157. 152 or 153 or 154 or 155 or 156
158. 138 and 157
159. 57 or 99 or 122 or 151 or 158
160. animals/ not (animals/ and humans/)
161. 159 not 160
162. exp *neoplasms/
163. exp *wounds, penetrating/
164. 162 or 163
165. 161 not 164
166. 29 and 165
167. guidelines as topic/
168. practice guidelines as topic/
169. guideline.pt.
170. practice guideline.pt.
171. (guideline? or guidance or recommendations).ti.
172. consensus.ti.
173. or/167-172
174. meta-analysis/
175. exp meta-analysis as topic/
176. (meta analy* or metaanaly* or met analy* or metanaly*).tw.
177. review literature as topic/

- 1
- 2
- 3 178. (collaborative research or collaborative review* or collaborative overview*).tw.
- 4 179. (integrative research or integrative review* or intergrative overview*).tw.
- 5 180. (quantitative adj3 (research or review* or overview*)).tw.
- 6 181. (research integration or research overview*).tw.
- 7 182. (systematic* adj3 (review* or overview*)).tw.
- 8 183. (methodologic* adj3 (review* or overview*)).tw.
- 9 184. exp technology assessment biomedical/
- 10 185. (hta or thas or technology assessment*).tw.
- 11 186. ((hand adj2 search*) or (manual* adj search*)).tw.
- 12 187. ((electronic adj database*) or (bibliographic* adj database*)).tw.
- 13 188. ((data adj2 abstract*) or (data adj2 extract*)).tw.
- 14 189. (analys* adj3 (pool or pooled or pooling)).tw.
- 15 190. mantel haenszel.tw.
- 16 191. (cohrane or pubmed or pub med or medline or embase or psycinfo or psychlit or psychinfo or
- 17 psychlit or cinahl or science citation indes).ab.
- 18 192. or/174-191
- 19 193. 173 or 192
- 20 194. 166 and 193
- 21
- 22
- 23
- 24
- 25
- 26
- 27
- 28
- 29
- 30
- 31
- 32
- 33
- 34
- 35
- 36
- 37
- 38
- 39
- 40
- 41
- 42
- 43
- 44
- 45
- 46
- 47
- 48
- 49
- 50
- 51
- 52
- 53
- 54
- 55
- 56
- 57
- 58
- 59
- 60

APPENDICES

APPENDIX A. Data extraction guide for studies evaluating the quality of studies evaluating the clinical measurement properties of outcome measures

Instructions

Clinical measurement studies may evaluate a wide spectrum of measurement properties; or evaluate aspects that relate to the implementability or interpretation of outcome measures. Individual clinical measurement studies cannot address every aspect of the measurement properties of an instrument. Ideally systematic reviews will synthesize the quality and content of research evidence addressing the clinical measurement properties of individual outcome measures. The summative knowledge about the measurement properties, cultural transferability, and utility across different contexts provides the scope of information needed to select an outcome measure for a specific patient (population), purpose and context.

This guide should facilitate extraction of data from individual clinical measurement studies. An explanation of the measurement property addressed in each item and how it might be measured within a given study is listed to facilitate finding and extracting that information. The accompanying extraction form can then be used to collect the specific information on these measurements or utility properties from specific studies.

The purpose of data extraction is to extract the specific information reported by authors within a study, not to evaluate the validity or value of that piece of information. Evaluation of the quality of the published version of the clinical measurement study (also called critical appraisal) is performed in a separate step. See the accompanying critical appraisal tool and guide. It is advisable to extract detailed specific information from the study; recognizing that this information may later be synthesized or subject to meta-analysis.

There is no standardized process for synthesizing clinical measurement information. Based on the findings of extraction you may elect to present the synthesize data in a descriptive way by creating a summary table of the data extracted in each category. If you find some studies with similar designs, you may be able to conduct a meta-analysis of some properties like clinically important difference (CID) or minimal detectable change (MDC); if appropriate given the sample and technique - this can be valuable as it may provide more stable estimates of these important properties.

366bmjopen-2019-031242 on 24 November 2019. Downloaded from <https://bmjopen.bmj.com/>. Protected by copyright.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

<u>Population studied</u>		
Population	A description of the study population	Sample size, pathology/disorder, demographics, setting, acute vs. chronic, where subjects were chosen from. Report meaningful demographics and indicators of the population studied.
Intervention	Interventions (if applicable) applied during longitudinal studies	Description of the nature, frequency, intensity of the intervention and the follow-up interval.
<u>Reliability</u>		
Reliability Description	The extent to which the same results are obtained on repeated administrations of the same measure when no change in status has occurred (reliability) or the precision of the scores on repeated measurements (agreement).	Test procedures or measures are typically reapplied on repeated occasions in individuals considered to have a stable condition during that time frame which repeated testing occurs. Repeated testing may be performed on different occasions (test-retest) for self-report measures, OR by the same rater (intra-rater) or different raters (inter-rater) if it is an observer-based scale. In some cases different test instruments (inter-instrument) are evaluated. The most common statistic used is the intraclass correlation coefficient for quantitative data (Shrout & Fleiss, 1979) and kappa (Landis & Koch, 1977) for nominal data. Standard error of measurement is used to present a quantitative estimate of the reliability—in the original units of measure. Report the type of reliability evaluated and coefficients obtained.
Reliability (relative)	The relationship (ratio) between variability in test scores when repeating the test on the same person in comparison to the overall variability (including variation between people)—typically indicated by a reliability coefficient	ICCs (Shrout & Fleiss, 1979) or another reliability coefficient and their associated confidence intervals are extracted.
Reliability (absolute)	Absolute reliability is portrayed as the quantity of error that could be anticipated upon repeated testing - reported in the original units of measure.	This may be reported as 1. Standard error of measurement (in older articles you may see coefficient of variation),

		2. Altman and Bland graphical technique (Bland & Altman, 1990; Bland & Altman, 1987; Bland & Altman, 1986) where the difference on repeated tests for each individual (limits of agreement) is plotted versus their mean score. The mean difference and the boundaries of 2SD are shown to define the limits of agreement.
Minimum Detectable Change	Calculated from the reliability coefficient and the level of confidence specified for error margins. This indicator reflects the amount of change required before you can be confident that change exceeds the random error that occurs in stable patients.	Extract the number and level of confidence.
<u>Content/structural validity</u>		
Internal consistency	The extent to which items on a test or subscale are related (an indication of the consistency of the concept measured).	Cronbach's alpha is the inter-item correlation usually reported. Report alpha and whether it relates to the entire instrument or specific subscales.
Content Validity	The extent to which the conceptual domain or construct that a test is designed to measure is adequately reflected by the items in the measure. In assessing content validity, it is important to consider the population to whom the measure applies, the completeness of the content, the relevancy and emphasis of the content assessed.	<p>A variety of techniques can be used to assess the extent to which items on a given measure reflected the necessary content to capture the concept of interest. Some of the techniques you will find are listed. Extract what was done to determine content validity and what was found.</p> <ol style="list-style-type: none"> 1) Patients and experts were involved during item selection/reduction - report how they were used and key decisions 2) Patients were consulted for reading and comprehension - report key findings 3) Cognitive interviews (Cibelli, 1994; Ojanen & Gogates, 2006) were done with patients to determine how items were interpreted by respondents, their perceptions of the items - report key findings 4) Expert panels or Delphi procedures were used to select items or evaluate the validity of the instrument - report key findings and decisions

		<p>5) During translation specific study, the meaning of the questions to another cultural or language group was studied - report key findings and decisions</p> <p>6) ICF linking (Cieza et al., 2002) or other coding of content was performed - report the results which may include the distribution of content across ICF domains, or the distribution of specific codes</p>
Floor-Ceiling Effects	The measure is unable to indicate a worsening score in patients who have clinically deteriorated and/or an improved score in patients who have clinically improved	There are a variety of potential methods; so the method and conclusion should be reported. Descriptive statistics of the distribution of scores that may be presented graphically or numerically may be used to indicate this. Other studies report the percentage of patients sustained a floor or ceiling effect defined by the number of people who fall in the extremes ranges. Note different studies may define the extreme ranges for floor/ceiling differently, so extract how it was defined and % of patients who obtained floor or ceiling category scores.
Factorial validity	The extent to which factor analysis supports assumptions surrounding constructs measured as defined by the measure or as indicated by subscale structure	Factor analysis may be reported as raw results; or compared to the inherent structure of the instrument or factor analysis upon which its construction was based. Report the type of factor analysis performed (exploratory or confirmatory), rotations used and the number of factors derived; specify whether this confirms the expected instrument structure or original factor structure.
Item response /Rasch Analyses	The extent to which items cross a range of difficulty, or a spectrum of the concept measured. The measurement scaling of the items.	Using item response theory or Rasch analysis, items are fit to a model to demonstrate interval scaling and determine item difficulty (Pallant & Tennant, 2007). Analyses might address item difficulty, person's ability curves, and comparison of ability estimation. Most commonly, the item difficulty and the composition of the test that fulfills interval scaling are defined. Data to be extracted include information on the scaling of the items, whether the interval scaling has been established; and the presence or absence of differential item functioning

366bmjopen-2019-0242 on 24 November 2019. Downloaded from <http://bmjopen.bmj.com/> on 01/11/2024 by guest. Protected by copyright.

		(DIF), where items perform differently on different types of respondents.
Construct Validity		
Construct Validity - correlational	<p>Constructs are artificial frameworks that are not directly observable. Construct validity assesses the extent to which measures perform according to a priori defined constructs. Construct validity can be cross-sectional or longitudinal (predictive).</p> <p>Constructed hypotheses can assess convergent validity where measures are thought to represent similar constructs or divergent validity where it is assumed they measure different constructs.</p> <p>For cross-cultural validation, the expected relationships are those that have been reported in validation of the instrument in its original language/format.</p>	<p>When extracting data about correlational validity, the pre-constructed hypothesis and whether it is supported should be documented. For correlational construct validity, this will be the nature and strength of the prespecified relationship and the correlations that support that. Relation to other indices/constructs that are similar (convergent) or different (divergent) can be reported. Ideally, hypotheses are formulated/reported and supported by correlations that are in accordance with the hypotheses. Note that there is no consistent agreement on what subjective term should be applied to validity correlations.</p> <p>Note that there is no consistent agreement on what subjective term should be applied to validity correlations. Some authors use subjective terminology defined for reliability such as: strong (>0.70) and moderate (0.40-0.70) correlations; others use the correlations like effect size benchmarks that 0.4 indicates a moderate effect and 0.6 a large effect. For validity assessment is more important than correlations prespecified constructed hypotheses, although not all papers are written clearly with respect to this.</p>
Convergent	The Relationship between similar scales/tests. Correlations are generally expected to be moderate to strong if the relationship is one where there is confidence that they measure a similar construct.	Extract test names, prespecified expected relationship and correlations observed.
Divergent	Divergent validity assesses the extent to which different scales/tests that are designed to	Extract test names, prespecified expected relationship and correlations observed.

	measure different constructs demonstrate that they are different by a lack of correlation between them.	
Construct validity - known groups	Known groups analysis supports the validity of a measure by demonstrating that the measurement is able to differentiate between groups that are prespecified and <u>known</u> to be different on the construct being assessed.	Data extraction should include the nature of the subgroups and the size of the difference observed between them (and its statistical significance). Typically, statistical tests of difference are performed. Since known groups analysis can provide data that is useful in clinical practice as benchmarks for comparing these known groups, it is a more practical form of construct validity than correlational. Data extraction/presentation should reflect this by presenting the group central tendency, their margins and statistical significance in an accessible manner.
Longitudinal Validity	This form of validity supports the validity of a measure by demonstrating that the change that occurs over time onto similar instruments is correlated in a manner consistent with the nature of the relationship between the scales. It is measured over a retest interval when clinically relevant change could be expected.	Extract test names and correlations Note: since longitudinal validity is based on four measures (pre-and post-test on two different measures), and since error tends to mitigate the strength of correlations, strong longitudinal correlations can be difficult to obtain.
Criterion validity Description	Criterion validation is determined by comparing a given outcome measure to an accepted standard of measure. For subjective constructs like pain and disability, it can be argued that there is no criterion since there is no external gold standard. Therefore, for self-report measures, validation focuses on construct validity. For performance measures, it is common to have a criterion measure that is considered to be highly precise and rigorous as the criterion comparator.	Authors will state that their measure is being compared against a specific instrument and report the correlation or agreement between the measures. Extract the test names and results: correlations if other as reported.
Concurrent criterion	Concurrent validity is assessed by comparing a scale and its criterion at a single point in time	Extract the test names and correlations.

<p>Predictive criterion</p>	<p>Predictive validity is evaluated by determining the extent to which the results of administering an outcome measure at one point in time can accurately predict a future status or outcome.</p>	<p>Extract the test names and correlations and time interval. (and important cutoffs if those were established/reported), if diagnostic test methodology was used to examine prediction, and sensitivity specificity and other diagnostic criteria were reported, they should be extracted.</p>
<p><u>Responsiveness/Clinical Change</u></p>		
<p>Responsiveness</p>	<p>Does the instrument detect changes over time that matters to patients?</p>	<p>Extract indicators of responsiveness include: effect size, standard response mean and the method for assessing whether patients were improved, stable or worse. (Beaton, 2000)</p>
<p>Clinically Important Difference (CID)</p>	<p>CID is the difference in scores that patients find to be observable and clinically important. It is assessed by comparing scores to an external benchmark of clinical relevance such as a global rating of change or some other method. The terminology used to rate the nature of this difference will affect the estimation process. Differences in methods include how clinically importance is framed and the metrics/process by which that is determined.</p>	<p>Extract the MID or CID and note the method/cut-off used to define importance. Extract how the clinically important differences were framed to respondents; or determined. For example, minimal, moderate, extreme improvement or better/not better, etc.</p>

36/bmjopen-2019-021425 on November 24, 2019. Protected by copyright. http://bmjopen.bmj.com/ on April 19, 2024 by guest. Protected by copyright.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

APPENDIX B. Data extraction form for studies evaluating the clinical measurement properties of outcome measures

Authors: _____ Year: _____ Rater: _____

Instructions

When using the data extraction form, it is important to realize that the purpose of data extraction is to remove or extract the specific information reported by authors within a study, not to evaluate the validity or value of that piece of information. To make data extraction as useful as possible, and to avoid the need for repeated data extractions, it is advisable to read the accompanying guide and then be as specific as possible when extracting information.

	DATA EXTRACTED
	Population studied
Population	
Intervention	
	Reliability
Reliability (relative)	
Reliability (absolute)	
Minimum Detectable Change	
	Content/structural validity

Internal consistency	
Content Validity	
Floor-Ceiling Effects	
Factorial validity	
Item response /Rasch Analyses	
Construct/Criterion Validity	
Known groups	
Convergent	
Divergent	
Longitudinal Validity	
Concurrent criterion	
Predictive criterion	
Responsiveness/Clinical Change	

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

Responsiveness	
Minimally Clinical Important Difference	

For peer review only

APPENDIX C. Quality Appraisal for Clinical Measurement Research Reports Evaluation Form

Rater (Group) _____

Author(s) (Study Author(s)) _____

Year (Year of publication) _____

1. Was the relevant background work cited to define what is currently known about the measurement properties of measures under study, and the potential contributions of the current research question to informing that knowledge base?

2

1

0

2. Were appropriate inclusion/exclusion criteria defined? *

2

1

0

3. Were specific clinical measurement questions/hypotheses identified?

2

1

0

4. Was an appropriate scope of measurement properties considered?

2

1

0

5. Was an appropriate sample size used?

2

1

0

6. Was appropriate retention/follow-up obtained? (for studies involving retesting; otherwise n/a)

- 1
2
3 2
4 1
5 0
6
7 7. Were specific descriptions provided of the measure under study and the method(s) used to administer
8 it?
9 2
10 1
11 0
12
13 8. Were standardized procedures used to administer all study measures in a manner that minimized
14 potential sources of error/bias (including the study measure and its comparators)?
15 2
16 1
17 0
18
19 9. Were analyses conducted for each specific hypothesis or purpose?
20 2
21 1
22 0
23
24 10. Were appropriate statistical tests performed to obtain point estimates of the measurement
25 properties?
26 2
27 1
28 0
29
30 11. Were appropriate ancillary analyses done to quantify the confidence in the estimates of the clinical
31 measurement property (Precision/Confidence intervals; benchmark comparisons/ROC curves, alternate forms of
32 analysis like SEM/MID, etc.)?
33 2
34 1
35 0
36
37 12. Were clear, specific and accurate conclusions made about the clinical measurement properties; that
38 were associated with appropriate clinical measurement recommendations and supported by the study objectives,
39 analysis and results?
40 2
41
42
43
44
45
46
47

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

1
0
Subtotals (of column 1 and 2) Total Score (sum of subtotals/24*100)

APPENDIX D. Description of each performance battery from selected articles

Battery	Description of Tasks
<p>Relevant FCE Subtasks^{25,26,27,28,29,30}</p>	<p>Material Handling Tasks: All lifting tests were executed with a wooden crate (40 × 30 × 26 cm) of 2.5 kg, and four to five weight increments of 2.5 kg or 5 kg each were used until the maximum amount of weight was reached. Maximum performance was recorded in kg.</p> <p>Lifting floor to waist: Measured after five lifts of crate from floor to table and vice versa (time limit < 90 s): hands remained on the crate during the test. Increase weight in 4-5 steps until maximum is reached</p> <p>Overhead lift test: Five lifts from waist to crown height and vice versa within 90 s in standing position. Increase weight in 4–5 steps until maximum is reached</p> <p>Two-handed carrying: Carrying of a crate for a short distance measured after five carries of 1.5 m distance at waist height. Hands remain on the crate during the test.</p> <p>One-handed carrying: Carrying wooden crate for 15 m within 90 s beginning with the right hand and thereafter the left hand.</p> <p>Overhead working: Standing with hands at crown height for manipulation of nuts and bolts. The time that the position was held is recorded (sec).</p> <p>Repetitive reaching: fast horizontal movements of the upper extremity in a sitting position. Marbles are removed from bowls at arm length distance at table height from left to right and vice versa, with right and then left arm. The time taken to remove 30 marbles is recorded (sec).</p> <p>Overhead lift test: Five lifts from waist to crown height and vice versa within 90 s in standing position. Increase weight in 4–5 steps until maximum is reached</p>

36/bmjopen-2019-031242 on 24 November 2019. Downloaded from <http://bmjopen.bmj.com/> on April 15, 2024 by guest. Protected by copyright.

	<p>Repetitive bending and overhead reaching: 20 marbles in 2 bowls at table height and crown height. Standing in front of bowl of marbles and moving the marbles as fast as possible from table height to crown height.</p>
<p>A Physiotherapy Test Package^{33,34,35,36}</p>	<p>PILE Tests: “The lifting tests were performed standing in front of bookshelves with shelves at 0.76m and 1.37 m from the floor. Subjects were asked to lift weights in a plastic box from floor to waist level (0–0.76 m) for the lumbar PILE test, or from waist to shoulder height (0.76–1.37 m) for the cervical PILE test. The initial weight was 3.6 kg for women and 6.9 kg for men. A ‘lifting movement’ involved a single transfer from one level to the next and back again. After every four such lifting movements (= 20 s), the weight was increased by 2.5 kg for women and 4.5 kg for men. The weight managed during the last lifting movement was recorded and used as a test result, as well as this maximum weight divided by the ‘adjusted weight’”.</p> <p>2x20m WWB: “Subjects were asked to walk 20 m at a comfortable speed along a corridor, to turn around where 20 m was marked and then to walk 20 m back to the starting point. In the first walking test they carried no extra weight, but in the second they carried one carrier bag in each hand, containing 4 kg each for the women, 8 kg each for the men. The time taken was recorded to get the walking speed. The tests were discontinued after 50 s”.</p>
<p>BTEWS II³¹</p>	<p>“The protocol consisted of performing a series of shoulder functional tasks before and after a fatiguing activity. Functional tasks consisted of active shoulder range of motion (ROM) in both flexion and abduction and cumulative power output (PO) accumulated over 10s during a repetitive pushing/pulling task in a horizontal plane at shoulder level”.</p>
<p>FIT - HaNSA³²</p>	<p>“The FIT-HaNSA protocol consists of three timed tasks and each task is performed for a maximum of 300 seconds (s) with approximately 30 s pause between them (set-up time for next task). Task 1 (waist-up) requires the patient to alternately “grab, lift, move and place” three 1000 g containers located on waist level and 25 cm above waist level shelves, using their affected arm, at a metronome pace of 60 beats per minute for 300 s or until they felt unable to continue. The time to complete Task 1 is measured using a stopwatch. Task 2 (eye down) is identical to Task 1 except that the two shelves are placed at eye-level and 25 cm below. Task 3 (overhead work) requires a patient to repeatedly screw and unscrew bolts in a sagittal plane oriented plate</p>

36bmjopen-2019-034424 on 24 November 2019. Downloaded from <http://bmjopen.bmj.com/> on April 19, 2024 by guest. Protected by copyright.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

	positioned at eye-level using both arms". More complete description at https://srs-mcmaster.ca/wp-content/uploads/2015/04/FIT-HaNSAProtocol_April2007.pdf
--	--

For peer review only



PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	1
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	2
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	3
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	3
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	4
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	4
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	3-4
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	3-4
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	4
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	4
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	5
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	NA
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I ²) for each meta-analysis.	NA



PRISMA 2009 Checklist

Page 1 of 2

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	NA
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	NA
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	6-7
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICCO, follow-up period) and provide the citations.	6-7
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	6-10
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	6-10
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	6-10
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	6-10
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	NA
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	11-13
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	14-16
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	16
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	18

336bmjopen-2019-031242 on 24 November 2019. Downloaded from http://bmjopen.bmj.com/ on April 19, 2024 by guest. Protected by copyright.

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit: www.prisma-statement.org.

Page 2 of 2

For peer review only - <http://bmjopen.bmj.com/site/about/guidelines.xhtml>

BMJ Open

Systematic Review of the Measurement Properties of Performance-based Functional Tests in Patients with Neck Disorders

Journal:	<i>BMJ Open</i>
Manuscript ID	bmjopen-2019-031242.R2
Article Type:	Original research
Date Submitted by the Author:	08-Oct-2019
Complete List of Authors:	McGee, Steven; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Sipos, Taylor; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Allin, Thomas; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Chen, Celia; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Greco, Alexandra; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Bobos, Pavlos; Western University, Health and Rehabilitation Sciences; University of Toronto, Dalla Lana School of Public Health, Institute of Health Policy Management and Evaluation MacDermid, Joy ; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Group, CATWAD; Michele Sterling, Anne Söderlund, Michele Curatolo, Jim Elliott, David M Walton, Helge Kasch, Linda Carroll, Hans Westergren, Samuel McLean, Gwendolen Jull, Genevieve Grant Luke Connelly, Joy C MacDermid, Mandy Nielsen, Pierre Côté, Tonny Elmoose Andersen, Trudy Rebbeck Annick Maujean, Sarah Robins, Kenneth Chen, Julia Treleaven
Primary Subject Heading:	Rehabilitation medicine
Secondary Subject Heading:	Rehabilitation medicine
Keywords:	functional, psychometric properties, neck pain, cervical, outcome measures

SCHOLARONE™
Manuscripts

1
2
3 1 **Title:** Systematic Review of the Measurement Properties of Performance-based Functional
4 Tests in Patients with Neck Disorders

5
6
7 3 ¹Steven McGee, PT

8
9 4 ²Taylor Sipos, PT

10
11 5 ³Thomas Allin, PT

12
13 6 ⁴Celia Chen, PT

14
15 7 ⁵Alexandra Greco, PT

16
17 8 ⁶Pavlos Bobos, PT, PhD(c) (corresponding author)

18
19 9 ⁷Joy MacDermid, PT, PhD

20
21 10 ⁸CATWAD

22
23 11

12 **Authors' information**

24
25 13 ¹Steven McGee PT, School of Physical Therapy, Department of Health and Rehabilitation
26 Sciences, Western University, London, Ontario, Canada, (smcgee7@uwo.ca)

27
28
29 15

30
31 16 ²Taylor Sipos PT, School of Physical Therapy, Department of Health and Rehabilitation Sciences,
32 Western University, London, Ontario, Canada, (jsipos@uwo.ca)

33
34
35 18

36
37 19 ³Thomas Allin PT, School of Physical Therapy, Department of Health and Rehabilitation Sciences,
38 Western University, London, Ontario, Canada, (tallin@uwo.ca)

39
40
41 21

42
43 22 ⁴Celia Chen PT, School of Physical Therapy, Department of Health and Rehabilitation Sciences,
44 Western University, London, Ontario, Canada, (qchen224@uwo.ca)

45
46
47 24

48
49 25 ⁵Alexandra Greco PT, School of Physical Therapy, Department of Health and Rehabilitation
50 Sciences, Western University, London, Ontario, Canada, (agreco33@uwo.ca)

51
52
53 27

54
55 28 ⁶Pavlos Bobos PT, PhD(c), (corresponding author) Doctoral Candidate, Western's Bone and Joint
56 Institute, Department of Health and Rehabilitation Sciences, Western University, Elborn College,
57 1201 Western Road, N6G 1H1, London, Ontario, Dalla Lana School of Public Health, Institute of

31 Health Policy Management and Evaluation, Department of Clinical Epidemiology and Health Care
32 Research, University of Toronto, Canada, (pbobos@uwo.ca), tel: +1 519 661 2111 x88912

33
34 ⁷Joy C MacDermid BScPT, PhD, Professor, Physical Therapy and Surgery, Western University,
35 London, ON and Co-director Clinical Research Lab, Hand and Upper Limb Centre, St. Joseph's
36 Health Centre, London, Ontario; Professor Rehabilitation Science McMaster University,
37 Hamilton, ON, Canada (jmacderm@uwo.ca)

38
39 ⁵CATWAD Coauthors: Michele Sterling m.sterling@uq.edu.au, Anne Söderlund
40 anne.soderlund@mdh.se, Michele Curatolo curatolo@uw.edu, Jim Elliott [j-](mailto:j-elliott@northwestern.edu)
41 elliott@northwestern.edu, David M Walton dwalton5@uwo.ca, Helge Kasch helgkasc@rm.dk,
42 Linda Carroll linda.carroll@ualberta.ca, Hans Westergren Hans.Westergren@skane.se, Samuel A
43 McLean, Samuel_McLean@med.unc.edu, Gwendolen Jull g.jull@uq.edu.au, Genevieve Grant
44 genevieve.grant@monash.edu, Luke Connelly l.connelly@uq.edu.au, Joy C MacDermid,
45 jmacderm@uwo.ca, Mandy Nielsen mandy.nielsen@griffith.edu.au, Pierre Cote
46 pierre.cote@uoit.ca, Tonny Elmoose Andersen tandersen@health.sdu.dk, Trudy Rebbeck
47 trudy.rebbeck@sydney.edu.au, Annick Maujean a.maujean@uq.edu.au, Sarah Robins
48 s.robins1@uq.edu.au, Kenneth Chen k.chen8@uq.edu.au, Julia Treleaven
49 j.treleaven@uq.edu.au

50
51 **Key Words:** functional, psychometric properties, neck, cervical, outcome measures

52
53 **Word Count:** 4509

61 Abstract

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

65 **Design:** Systematic review

66 **Data Sources:** A literature search using PubMed, Scopus, CINAHL, EMBASE, COCHRANE,
67 Google Scholar, and a citation mapping strategy was conducted till July 2019

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

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

75 **Results:** The search obtained 12 articles which reported on 4 outcome measures (Functional
76 Capacity Evaluations (FCE), Baltimore Therapeutic Equipment Work Simulator II (BTEWS II),
77 Functional Impairment Test- Hand and Neck/Shoulder/Arm (FIT-HaNSA)) and a physiotherapy
78 test package, to assess the functional abilities in patients with mechanical neck pain. Of the selected
79 papers: 1 reports content validity, 5 construct validity, 4 reliability, 1 sensitivity to change, and 1
80 both reliability and construct validity. QACMRR scores ranged from 68% to 95%.

81 **Conclusions:** This review found very good quality evidence that the FIT-HaNSA has
82 excellent inter and intra-rater reliability and very weak to weak convergent validity. Excellent
83 quality evidence of fair test-retest reliability, weak convergent validity, and very weak known

84 groups validity for the BTEWS II test was found. Good to excellent quality evidence exists that an
85 FCE battery has poor to excellent reliability and very weak to strong validity. Good to excellent
86 quality of weak to strong validity and trivial to strong effect sizes were found for a physiotherapy
87 test package.

88 **Prospero registration:** CRD42018112358

91 **Strengths and limitations of this study**

- 92 • The psychometric properties of performance outcome measures for neck pain were
93 synthesized and critically appraised
- 94 • This study assessed the risk of bias and the quality of measurements properties
- 95 • The feasibility or usability of these tools was not assessed

97 **Introduction**

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

1
2
3 106 Outcome measures are a crucial component in monitoring patients with neck pain to
4
5 107 determine the effects of treatment[11,12], evaluation of interventions, guiding return to work, and
6
7
8 108 justifying treatment.[13,14] Several self-reported outcome measures currently exist to assess
9
10 109 disability and function in those with neck pain (e.g. the Neck Disability Index - NDI). [13]
11
12 110 Evidence-based clinical practice guidelines suggest that measures assessing physical performance
13
14 111 should also be used for people with neck pain.[15] Performance-based testing is where the
15
16 112 assessment is based on actual performance of a task or activity. Physical performance can be
17
18 113 assessed by testing a person's ability to execute a standardized activity in a standardized
19
20 114 environment (i.e. clinical setting).[16] Time to complete the activity, number of repetitions
21
22 115 performed, and weight lifted are frequently used to quantify the physical performance.[17]
23
24 116 Conversely, self-report measures examine patients' perception and experience of their ability to
25
26 117 perform functional tasks. [16] Previous research has demonstrated poor to fair relationships
27
28 118 between physical performance and self-report measures of ability in patients with various
29
30 119 musculoskeletal disorders suggesting that these measures assess different constructs of function.
31
32 120 [17,18] Consequently, physical performance tests and self-report measures complement each other
33
34 121 and may each contribute unique information about a patient's function. [19]
35
36
37
38
39

40 122 A fundamental component of monitoring outcomes is having reliable and valid tools with
41
42 123 known measurement properties.[13,20] While recent research has investigated the psychometric
43
44 124 properties of patient-reported outcomes in people with neck pain [13,21] there is a gap in
45
46 125 knowledge with respect to performance-based functional outcomes. The purpose of this systematic
47
48 126 review was to identify and synthesize clinical measurement studies that evaluate measurement
49
50 127 properties of performance-based functional tests in patients with neck disorders.
51
52
53

54 128
55
56
57
58
59
60

129 **METHODS**

130 **Patient and Public Involvement**

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

132

133 **Study Design and Protocol Registration**

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

137

138 **Search Strategy**

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

150

151 **Inclusion Criteria**

1
2
3 152 Articles were included in the final review if all of the following criteria were met:

- 4
5 153 • >50% of the study's patient population had neck pain or a musculoskeletal neck disorder
6
7
8 154 (e.g. whiplash associated disorder (WAD II))
9
10 155 • Patients in the study completed a functional-based test
11
12 156 • Clinometric properties of at least one performance-based test were reported.

13
14
15 157 A test was considered functional-based if it met the following criteria:

- 16
17 158 • assessment of a patient's ability to execute a standardized activity in a standardized
18
19 159 environment
20
21
22 160 • tests assessing muscular endurance (e.g. cervical flexion test) or proprioception were not
23
24 161 deemed functional-based as they are often not reflective of physical working conditions.

25
26 162 Both traumatic and non-traumatic origins of neck pain were considered. Definitions for the
27
28 163 properties can be found in **APPENDIX A**.

29
30
31 164

32 33 165 **Article Selection**

34
35
36 166 Titles and abstracts generated by the search strategy were screened by two authors (SM
37
38 167 and PB) independently. Articles that met the inclusion criteria and selected for a full text review
39
40 168 were also reviewed in pairs of authors. Disagreements were resolved by the most experienced
41
42 169 author (JCM)

43
44
45 170

46 47 171 **Data Extraction**

48
49 172 Data extraction and critical appraisal was performed in pairs of two raters among the authors, after
50
51 173 the completion of a calibration session in which the most experienced author (JCM) reviewed the
52
53 174 data extraction tools with the authors that performed the data extraction. When reviewers disagreed

1
2
3 175 during data extraction and/or critical appraisal, and consensus could not be met, a third author
4
5 176 arbitrated. A data extraction form [23] (**APPENDIX A and APPENDIX B**), developed by one of
6
7
8 177 the authors (JCM.), was used to ensure systematicity. Authors extracted sample size, patient
9
10 178 population characteristics, functional tests performed and reported psychometric properties. The
11
12 179 interpretation of ICC was as follows: $ICC < 0.50$ indicating poor, $0.50 \leq ICC < 0.75$ indicating
13
14 180 moderate, $0.75 \leq ICC < 0.9$ indicating good, and $ICC \geq 0.9$ indicating excellent reliability were used
15
16 181 as a common benchmark. [24] For validity estimates, correlation coefficient (Pearson's/Spearman)
17
18 182 and the 95% confidence intervals were extracted if were available. [23,25] Evan's guidelines to
19
20 183 interpret the strength of the correlation was used which included: 0.00–0.19 “very weak”, 0.20–
21
22 184 0.39 “weak”, 0.40–0.59 “moderate”, 0.60–0.79 “strong”, and 0.80–1.00 “very strong”. [26] To
23
24 185 assist clinical decision making, standard benchmark scores of trivial (< 0.20), small (≥ 0.20 to $<$
25
26 186 0.50), moderate (≥ 0.50 to < 0.80) or large (≥ 0.80), as proposed by Cohen, were used. [27] For
27
28 187 studies assessing construct validity specifically, results in accordance with pre-defined hypotheses
29
30
31 188 were evaluated to interpret the findings.
32
33
34
35
36
37
38
39

191 **Quality Appraisal for Clinical Measurement Research Reports Evaluation Form**

192 Pairs of authors critically appraised the quality of each study using a standardized 12-item
193 evaluation tool (QACMRR) designed to assess the quality of studies determining measurement
194 properties in outcome measures (**APPENDIX C**). If disagreement was present a third person (JM)
195 assist in resolving the discrepancy. [23] This tool has been found to have moderate to excellent
196 pre-consensus inter-rater reliability ($ICC: 0.69-0.91$, $\kappa = 0.62-1.00$) across a number of systematic
197 reviews.[23,25,28] The evaluation criteria of this tool included twelve items: 1) Thorough

1
2
3 198 literature review to define the research question; 2) Specific inclusion/exclusion criteria; 3)
4
5 199 Specific hypotheses; 4) Appropriate scope of psychometric properties; 5) Sample size; 6) Follow-
6
7
8 200 up; 7) The authors referenced specific procedures for administration, scoring, and interpretation of
9
10 201 procedures; 8) Measurement techniques were standardized; 9) Data were presented for each
11
12 202 hypothesis; 10) Appropriate statistics-point estimates; 11) Appropriate statistical error estimates;
13
14 203 and 12) Valid conclusions and recommendations. [23,25] Each item is scored from 0 to 2 with
15
16 204 (score=2) is the best; (score=1) is acceptable but suboptimal; (score=0) is not done/documentated,
17
18 205 substantially inadequate or inappropriate. An article's total score – quality - was calculated by the
19
20 206 sum of scores for each item, divided by the numbers of items and multiplied by 100%. [23,25]
21
22 207 Overall, the quality summary of appraised articles ranges from (0%-30%) Poor, (31%-50%) Fair,
23
24 208 (51%-70%) Good, (71%-90%) Very Good, and (>90%) Excellent
25
26
27
28
29
30
31
32

33 211 RESULTS

34
35 212 The search strategy resulted in 840 published articles. After duplications were removed, 31
36
37 213 articles were deemed relevant and were screened at full text. Overall, 12 articles met our inclusion
38
39 214 criteria (**FIGURE 1**). The excluded articles were removed due to inappropriate patient
40
41 215 populations, investigations into self-report measures or tests assessing proprioception/muscular
42
43 216 endurance rather than functional-based measures, or because the articles were found to be
44
45 217 systematic reviews. The characteristics of the included studies and the summary of psychometric
46
47 218 properties are presented in **TABLE 1**. The quality assessment is summarized and presented in
48
49 219 **TABLE 2**. Percent agreement was calculated for quality scores between the 2 raters and it was
50
51
52 220 90%.
53
54
55
56
57
58
59
60

221

222 **Participants**

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

230

231 **Functional-Based Tests**

232 The 12 articles that were included for review provided properties on the following
233 functional based tests: Functional Capacity Evaluations (FCE)[29–34], The Baltimore Therapeutic
234 Equipment Work Simulator II (BTEWS II) [35], Functional Impairment Test- Hand and
235 Neck/Shoulder/Arm (FIT-HaNSA) [36], as well as items off of a physiotherapy test package
236 including a cervical and lumbar Progressive Isoinertial Lifting Evaluation (PILE-C, PILE-L) test
237 [37–40] and 2 x 20 m with burden walking test (2x20M-WWB) [37–40]. Descriptions of all
238 functional-based tests and their relevant subtasks are provided in **APPENDIX D**.

239

240 **Functional Capacity Evaluations (FCE)**

241 Six articles reported measurement properties for an FCE battery. We identified multiple
242 versions of the FCE in the literature with one article reporting properties on the Workwell FCE
243 [30], two reporting on the Whiplash Associated Disorder (WAD) FCE [29,31] and three reporting

244 on the neck-FCE.[32–34] These test batteries include various combinations of muscular strength,
245 endurance and functional based tests. The measurement properties of the functional based tests
246 used by the FCE are outlined in **TABLE 3**.

247

248 *Individuals with Sub-acute to chronic WAD*

249 Trippolini et al. (2014)[30] evaluated the Workwell FCE test-retest reliability,
250 measurement error, convergent validity and predictive criterion validity of future work capacity in
251 workers diagnosed with WAD I or II. Interclass Correlation Coefficients (ICC) ranged from 0.66
252 to 0.96 (moderate to excellent). Limits of agreement relative to mean performance ranged from 21
253 to 57% for functional based sub-tests. Correlations between FCE sub scores and baseline work
254 capacity were very weak to weak ranging between $r=0.06$ and $r=0.39$. FCE sub scores did not
255 predict future work capacity at 1, 3, 6 and 12 months.

256 Trippolini et al. (2015)[29] assessed the WAD FCE (31) and evaluated convergent validity
257 and known-groups validity. FCE subscales showed very weak to strong correlations (0.15-0.68)
258 with each of: pain, self-reported functional ability, self-reported disability, anxiety and depression.
259 It was found that the FCE had known-group sex validity (males vs females) for 1 of 3 functional
260 subtests (lifting waist-overhead) and reported significant performance differences between culture
261 groups (German vs non-German language groups). To test construct validity, 29 a priori
262 formulated hypotheses were tested, 4 related to gender differences, 20 related associations with
263 other constructs, 5 related to cultural differences. In total 23 out of 29 hypotheses were confirmed
264 (79 %).

265

266 *Work-Related Neck Disorders*

1
2
3 267 Reesink et al. (2007)[34] developed an independent FCE for patients with musculoskeletal
4
5 268 neck disorders (neck FCE). They performed a review of epidemiological literature and identified
6
7
8 269 four physical risk factors for work-related neck disorders and used that information to develop an
9
10 270 FCE consisting of eight functional-based tests. Content validity was established by following
11
12 271 operational definitions of the risk factors when searching the literature and using current literature
13
14
15 272 to provide a rationale to guide their development of the tasks comprising the FCE.
16
17 273

19 274 *Chronic Neck Pain*

21 275 Reneman et al. (2017)[32] measured test-retest reliability of the subscales of the neck FCE
22
23 276 in patients with multifactorial neck pain. Test-retest ICC's ranged from poor to excellent (0.39-
24
25 277 0.96). Limits of agreement relative to mean performance range from 32.0% to 56.5% for functional
26
27
28 278 based sub tests. Convergent validity was performed against the Neck Disability Index (NDI) items
29
30 279 and total score.[33] The authors found weak to strong Pearson correlations (0.39-0.70) for the FCE
31
32 280 sub scores to both NDI individual items and the NDI total score.
33
34 281

37 282 **The Baltimore Therapeutic Equipment Work Simulator II (BTEWS II)**

39 283 *Chronic Neck Pain*

41 284 Lomond and Côté, (2011)[35] reported on the reliability, measurement error, minimum
42
43 285 detectable change (MDC) and validity of the power output (PO) task during the BTEWS II test in
44
45 286 patients with chronic neck and shoulder pain (TABLE 4). Test-retest reliability, measured with
46
47
48 287 Spearman Rank correlations and ICC's was moderate and measured at $\rho=0.37$ and $ICC_{2,1} = 0.54$,
49
50 288 respectively. The standard error of measurement (SEM) and the minimal detectable change at 90%
51
52
53 289 confidence (MDC₉₀) for the PO task were measured as 30.25 and 70.59, respectively. Weak
54
55
56
57
58
59
60

290 Spearman Rank correlations between the PO task and the NDI, Shoulder Pain and Disability Index
291 (SPADI) and Numeric Rating Scale (NRS) for pain tests were recorded. There were no significant
292 performance differences between control and pain groups for the PO task.

293

294 **Functional Impairment Test- Hand and Neck/Shoulder/Arm (Fit-HaNSA)**

295 *Sub-acute to chronic WAD*

296 Pierrynowski et al. (2016)[36] reported on the reliability, measurement error, MDC and
297 validity of the Fit-HaNSA test in a sample of people with WAD II following motor vehicle
298 collision (MVC) (**TABLE 5**). Intra-rater reliability ICC's for patient subtask and total scores were
299 moderate to good ranging between 0.70-0.78. [36] Inter-rater reliability ICC's for patient subtask
300 and total scores were moderate to good and ranged between 0.54-0.84. [36] The Bland and Altman
301 plot for the patient group showed a 26 seconds (s) bias in terms of improved performance on the
302 second test (possible learning effect). The standard deviation of difference was 124 s and 95%
303 Limits of Agreement (LoA₉₅) was 248 seconds. [36] The SEM for people with WAD II was
304 reported to be 76 s. The MDC₉₀ was measured as 176 s. [36]

305 Spearman rank correlations were also calculated between the Fit-HANSA, Numeric Pain
306 Rating Scale (NPRS), NDI, the disabilities of arm, hand and shoulder (DASH) and 6 cervical range
307 of motion measures. Most (59 of 78) of the correlations between performance and comparator
308 measures were very weak to weak ($r < 0.4$). [36] All correlations between total Fit-HaNSA scores
309 and subtask scores had good correlations ($r < 0.75$), except for Task 1-Task 3. [36] Significant
310 performance differences between WAD II and control groups (known group validity) were
311 recorded for the total Fit-HaNSA score and all 3 subtask scores. [36]

312

313 **Physiotherapy Test Package Subtests**

314 Ljungquist et al. published a series of articles[37–40] which evaluated the clinimetric
315 properties of a physiotherapy test package for patients with spinal pain (**TABLE 6**). This
316 package included muscular strength & endurance tests, submaximal endurance tests, and three
317 functional tests. These functional tests included the PILE-C, PILE-L, and 2x20M-WWB test.
318 Ljungquist’s series of articles reported on convergent validity, known-groups validity, reliability,
319 measurement error and sensitivity to change for these tests. [37–40]

320

321 *Undetermined duration of neck pain*

322 In a 1999 article [39], correlations between the tests of the package and pain (CR-10) and
323 perceived exertion (Borg RPE) were determined. All correlations were very weak to moderate
324 (0.10-0.48) except for moderate to strong correlations (0.55-0.65) between the PILE-C test and
325 pain intensity and between 2x20M-WWB test and pain intensity.

326 In a 2003 article[37], the PILE-C, PILE-L and 2x20M-WWB tests were tested to determine their
327 ability to discriminate between known-groups (neck pain vs back pain). Subjects with spinal pain
328 completed the CR-10, the University of Alabama Pain Behavior scale (UAB) and the Borg RPE
329 test. Specific cut points were used to distinguish patients with high vs. low pain intensity, high
330 vs. low pain behavior, and high vs. low perceived exertion in patients, respectively. Participants
331 then completed the test package and it was determined if each subtest could discriminate
332 between participants with high vs. low pain intensity. The PILE-C and the 2x20M-WWB tests
333 were hypothesized to be more difficult for persons with neck pain and the PILE-L was
334 hypothesized to be more difficult for persons with back pain. Subjects with neck pain performed
335 worse on the PILE-C test compared to those with back pain. Subjects with back pain did not

336 perform worse than those with neck pain on the PILE-L test and subjects with back pain
337 performed worse on the 2x20M-WWB test.

338 The functional tests were able to discriminate between all 3 subgroups with the exception of the
339 PILE-C being unable to discriminate between participants with high vs. low perceived exertion.

340 In a paper from 1999[39], the PILE-C, PILE-L and 2x20M-WWB tests were found to have
341 significant discriminative abilities in distinguishing healthy subjects from patients with spinal pain.
342 The sensitivity and specificity for this known group discrimination for the PILE-C test, were
343 reported to be 0.93 (very strong) and 0.69 (strong), respectively. The sensitivity and specificity for
344 the PILE-L test were reported to be 0.85 (very strong) and 0.65 (strong), respectively.

345 The inter and intra rater reliability were tested on participants with spinal pain.[38] Limits
346 of agreement were used to measure inter rater reliability and repeatability, defined as 2x the within-
347 subject standard deviation of each variable. Interrater agreement for 2 tests was deemed
348 “acceptable”, while all 3 functional tests had “clinically acceptable” intra-rater reliability.

349 Sensitivity-to-change was evaluated in the test package following 6 months of a
350 physiotherapy intervention. Using ROC curves, Wilcoxon sign ranked tests and spearman
351 correlation coefficients, only the 2x20m-WWB test and the PILE-C (women only) were deemed
352 to be sensitive to change. [40] Additionally, moderate to large effect sizes were found for all test
353 components.

354

355 **DISCUSSION**

356 This study synthesized 12 studies assessing clinometric properties of 4 different functional-
357 based assessments. Given the limited number of studies, the substantial variation in the types of
358 tests examined, the methods used to assess the clinical measurement properties, and the study

1
2
3 359 populations, the current state of knowledge does not allow firm conclusions regarding
4
5 360 recommendations for an optimal functional-based test at this time. Overall, the quality ranging
6
7
8 361 from good to excellent (67-92%) as determined by the QACMRR, for a range of properties of the
9
10 362 4 different assessments in patients with acute or chronic neck pain that is musculoskeletal in origin.
11
12 363 Studies obtaining higher percentages indicate research that has been consistent with best practice
13
14 364 where studies with lower percentages are more likely to be inadequate or inappropriate

17 365 **FCE**

19 366 The breadth of a functional-based test is variable and defined by the developers. An
20
21 367 advantage of the functional assessment designed by Reesink et al.[34] is that they mapped the
22
23 368 eight subtests to risk factors identified in the literature for work-related neck disorders. The eight
24
25 369 subtests consist of: material handling tasks, lifting floor to waist, overhead lift test, one-handed
26
27 370 and two-handed carrying, overhead working, repetitive reaching, overhead lifting, and repetitive
28
29 371 bending and overhead reaching. Given the systematic approach and rationale these authors used
30
31 372 in developing the FCE and this approach being used in previous research [41], we suggest that
32
33 373 this test has strong content validity.

37 374 Six articles address the clinical measurement properties of this FCE ranging from good to
38
39 375 excellent quality (67-92%). There was evidence that the FCE was stable over test-retest time of
40
41 376 7-14 days. [31,32] These measures demonstrate longer stability over time compared to self-report
42
43 377 measures such as the Neck Disability Index (NDI) which has demonstrated test-retest reliability
44
45 378 within only a short period of 0-3 days. [28] Whether this longer-term stability is a characteristic of
46
47 379 functional-based tests or reflects differences in study populations in context requires further
48
49 380 testing. These two studies had relatively lower quality scores on the QACMRR (67-75%)
50
51 381 compared to other studies in this review putting into question test-retest time. Although test-retest
52
53
54
55
56
57
58
59

1
2
3 382 reliability has been assessed, inter-rater and intra-rater reliability has yet to be researched. Unlike
4
5 383 self-report measures, we expect measurement error due to the evaluator and functional-based tests.
6
7
8 384 Thus, future research should explore these aspects of reliability.
9

10 385 Convergent validity is often examined in clinical measurement studies. We suggest that
11
12 386 this may be because these comparisons are easily performed by correlating different tests rather
13
14 387 than providing strong confidence in the validity of the measurement. Often convenient
15
16 388 comparisons are performed rather than those most relevant. Across many domains and measures
17
18 389 it has become clear that the relationship between self-reported function and performance-based
19
20 390 function or physical impairment is often very weak to moderate. Therefore, the value of assessment
21
22 391 of these relationships as a form of validation has limited value. Several studies of very good to
23
24 392 excellent quality have reported on the convergent validity of the FCE. [29,30,33] The highest
25
26 393 quality article determined by the QACMRR (92%) found the relationship between the FCE and
27
28 394 work capacity to be poorly associated with one another. [30] The same study found that the ability
29
30 395 of the FCE to predict future work capacity was poor. This may be considered a more important
31
32 396 comparison since ideally functional-based tests would relate to important outcomes like return to
33
34 397 work. No studies to our knowledge report the responsiveness or sensitivity to change of the FCE.
35
36 398 This is an important gap since the focus of rehabilitation is often to remediate limitations in goal
37
38 399 impairments or work capacity, and assessment of these changes is critical to clinical decision-
39
40 400 making and reporting outcomes. Thus, future research should evaluate the responsiveness of the
41
42 401 FCE to provide insight in the measure's ability to detect change after an intervention.
43
44
45
46
47
48

49 402 **FIT-HaNSA**

50
51 403 One study of very good quality (88%) assessed the FIT-HaNSA, a test consisting of two
52
53 404 reaching tasks (waist and eye-level) and sustained overhead task performance. [36] Overall, the
54
55
56
57
58
59
60

1
2
3 405 FIT-HaNSA demonstrated excellent inter-rater reliability (0.84) and intra-rater reliability (0.78).
4
5 406 The specific subtests included within the FIT-HaNSA similarly demonstrate fair to excellent (0.54-
6
7 407 0.80) and good (0.70-0.72) inter-rater and intra-rater reliability respectively. The FIT-HaNSA also
8
9 408 demonstrated a clear ability to distinguish between people with WAD 2 and healthy controls.
10
11 409 Correlations between the FIT-HaNSA and other patient self-report disability and functional
12
13 410 outcome measures (NPRS, NDI, DASH, CROM and FIT-HaNSA) were generally very weak to
14
15 411 weak ($\rho < 0.4$), consistent with other studies comparing performance and self-report. [17,18] The
16
17 412 largest limitation in critically synthesizing information for this test is that only a single study was
18
19 413 found that reported the measurement properties for people with neck disorders. It should be noted
20
21 414 however that it has been validated in other MSK disorders. [35,41] Although others have noted
22
23 415 the lag in development of functional-based measures in comparison to self-report measures, FIT-
24
25 416 HaNSA was recommended as a functional-based measure for people with shoulder disorders. [42]
26
27 417 Further research is necessary to investigate the responsiveness of the FIT-HaNSA.
28
29
30
31
32

33 418 **BTEWS II**

34
35 419 Another study of very good quality (88%) assessed the efficacy of the BTEWS II where
36
37 420 the participants performed a dynamic pushing and pulling task in which power output was recorded
38
39 421 over a 10 second sample.[35] While the convergent validity aspect of this paper was assessed as
40
41 422 consistent with best practice through the critical appraisal process, the relationship between the
42
43 423 power output on the BTEWS and measures of pain and disability (NDI, SPADI, NRS) were poorly
44
45 424 associated with each other. In addition, the power output component was not found to be
46
47 425 significantly different between people with neck pain and healthy controls which suggests it might
48
49 426 not be discriminative. Discrimination between patients and healthy controls is a low standard for
50
51 427 an outcome measure, and tests that cannot fulfil this benchmark should be viewed with caution.
52
53
54
55
56
57
58
59
60

1
2
3 428 Because of the weak measurement properties demonstrated by the power output component of the
4
5 429 BTEWS II, it does not appear to be a desirable functional-based measure to assess function in
6
7
8 430 people with neck pain. However, we acknowledge for all of the functional-based tests the evidence
9
10 431 pool is so shallow that there is high potential that future studies might lead to different conclusions.
11
12 432 Future research should also investigate the reliability and responsiveness of the BTEWS II.

14 433 **Physiotherapy Test Package Subtests**

16 434 Four studies ranging from good to very good quality (68-82%) assessed relevant items
17
18 435 from a physiotherapy test package, including a lift from floor-to-waist and a waist-to-shoulder task
19
20 436 and a two-handed carrying task. The properties of these assessment items include weak to
21
22 437 moderate correlations to pain, perceived exertion, and had “fair to good” reliability. The 2x20m-
23
24 438 WWB and PILE-C tests were found to be sensitive-to-change which is valuable information as no
25
26 439 other study has assessed this property in functional-based measures in patients with neck disorders.
27
28 440 Thus, this measure may be of value in clinical settings when assessing functional capacity before
29
30 441 and after a treatment intervention. All tests had discriminative ability for detecting participants
31
32 442 with spinal pain vs healthy controls. Most of the three tests demonstrated poor construct validity
33
34 443 in that they were poorly related to pain and perceived exertion and the results were not in
35
36 444 accordance with pre-defined hypotheses. Thus, further research is necessary to investigate these
37
38 445 constructs. Three of the four results from the studies assessing the physiotherapy test package had
39
40 446 a mixed sample of patients with various pain sites including back pain. While the majority of each
41
42 447 cohort in these studies had neck pain, careful consideration should be taken to apply these tests to
43
44 448 a neck pain specific population.

51 449 **Clinical Implications**

52
53
54
55
56
57
58
59
60

1
2
3 450 This study confirms that functional-based tests have had far less development and
4
5 451 evaluation than self-report measures. Limitations include the number of tests and insufficient body
6
7 452 of evidence to make confident recommendations with respect to functional-based testing. It is clear
8
9 453 that self-report and functional-based measures provide different perspectives. Theoretically,
10
11 454 functional-based tests are important to inform our understanding about the mechanisms of
12
13 455 intervention and how interventions increase capacity. Future research may benefit by also
14
15 456 comparing results from a functional-based measure to work capacity to when assessing construct
16
17 457 validity. Overall more work is required to further establish the psychometric properties of
18
19 458 functional-based tests in persons with neck disorders, including sensitivity-to-change,
20
21 459 responsiveness, and predictive validity.
22
23
24
25

26 460 The FCE evaluated patients with neck pain of varying origin including WAD, work-related
27
28 461 neck disorders, and chronic idiopathic neck pain. The BTEWs II evaluated functional capacity in
29
30 462 patients with chronic neck pain, the FIT-HaNSA evaluated patients with WAD, and the
31
32 463 physiotherapy test package did not specify the origin of musculoskeletal neck pain in their cohort.
33
34 464 Thus, specific functional-based measures may be more applicable depending on the origin of the
35
36 465 musculoskeletal neck pain being assessed.
37
38
39

40 466 The data presented suggest that the FIT-HaNSA has the strongest clinometric properties
41
42 467 though this is based on a single higher quality paper specific to neck disorder. [36] Importantly,
43
44 468 normative data have been published [43], it has been validated in multiple studies in patients with
45
46 469 shoulder conditions [44–46] and has been recommended when compared to other measures [42].
47
48 470 The FCE has a limited evidence base from which to draw, though it was developed with strong
49
50 471 content validity and further evaluation may demonstrate its usefulness.
51
52
53

54 472 **Limitations**

55
56
57
58
59
60

1
2
3 473 A challenge in synthesizing clinical measurement evidence is the wide range of properties
4
5 474 and indicators that need to be considered. Unlike effectiveness studies where one can focus on the
6
7
8 475 effect size of treatment there are many considerations that would affect the recommendations made
9
10 476 about outcome measures. This is further complicated when the pool of evidence is shallow.
11
12 477 Although the quality assessment tool (QACMRR) developed by one of the authors of this review
13
14 478 which assess the quality of design of individual studies were useful for interpreting the evidentiary
15
16
17 479 pool, there is no clear method to synthesize the extracted clinical measurement evidence. While
18
19 480 some systematic reviews on treatment might only report findings from high-quality studies, it is
20
21 481 important to see how outcome measures perform in different contexts. Further, the assessment of
22
23 482 quality is complicated given that clinical measurement studies have so many dimensions.
24
25
26 483 Therefore, exclusion of lower quality studies has questionable value. Thus, a more practical
27
28 484 approach is to consider quality when interpreting the findings, rather than excluding studies.

30
31 485 The QACMRR focuses on whether the authors made appropriate decisions in selecting the
32
33 486 scope and methods of their clinical measurement evaluations within a given study and provides
34
35 487 descriptors of poor fair or good design options. Quality focuses on issues that might affect risk of
36
37
38 488 bias or imprecision in estimates; whereas risk of bias assessments focusses on items that might
39
40 489 result in a biased estimate. For example, insufficient power is a precision (quality) issue, not a risk
41
42 490 of bias. Although it is difficult to interpret the meaning of the percentage of the QACMRR as there
43
44 491 are no established cut-offs for distinguishing good and poor-quality studies, it provides one way
45
46
47 492 of ranking the articles in order of quality. We did not use COSMIN checklist since it was developed
48
49 493 for PROMS and some of the components/steps that involved are not applicable to performance-
50
51 494 based tests.

1
2
3 495 Another limitation in this review was that the feasibility or usability of these tools was not
4
5 496 assessed. While feasibility was not the focus of this review, information on the practical
6
7 497 application of these functional-based measures provides valuable information to clinicians for
8
9 498 determining whether these tests are appropriate to use in their given setting. Thus, future research
10
11 499 should not only investigate further the psychometric properties of these tools, but also report the
12
13 500 feasibility of using these tests so that they may be used in clinical settings and to identify
14
15 501 limitations that restrict their application in practice.
16
17
18
19 502

21 503 **CONCLUSION**

24 504 This review found very good quality evidence that the FIT-HaNSA has excellent inter and
25
26 505 intra-rater reliability and very weak to weak convergent validity. Excellent quality evidence of fair
27
28 506 test-retest reliability, weak convergent validity, and very weak known groups validity for the
29
30 507 BTEWS II test was found. Good to excellent quality evidence exists that an FCE battery has poor
31
32 508 to excellent reliability and very weak to strong validity. Good to excellent quality of weak to strong
33
34 509 validity and trivial to strong effect sizes were found for a physiotherapy test package. Functional-
35
36 510 based evaluation in people with neck disorders is an area needing much research attention both to
37
38 511 establish the measurement properties of existing measures, potentially to develop innovative new
39
40 512 measures and to perform head-to-head comparisons of measures before an optimal functional-
41
42 513 based test can be identified.
43
44
45
46
47 514

49 515 **Authors' contributions**

51 516 SM contributed significantly to conception and design of the study, data extraction, critical
52
53 517 appraisal, interpretation of data and drafting of the manuscript. TS, TA, PB, and CC were involved
54
55 518 in literature search, critical appraisal and interpretation of data and drafting. AG was involved in
56
57
58
59
60

1
2
3 519 critical appraisal and drafting. JM was also involved in the conception and design of the study,
4
5 520 drafting, and revised the manuscript for important intellectual content. PB and CATWAD were
6
7 521 involved in the drafting and review of the manuscript. All authors have given their final approval
8
9 522 on the manuscript to be published

10 523

11 524 Declarations**12 525 Ethics approval and consent to participate**

13 526 Not applicable

14 527

15 528 Consent for publication

16 529 Not applicable

17 530

18 531 Availability of data and material19 532 Data sharing is not applicable to this article as no datasets were generated or analyzed during the
20
21 533 current study

22 534

23 535 Funding Statement24 536 This work was supported by the Canadian Institutes of Health Research (CIHR) with funding
25
26 537 reference number (FRN: SCA-145102).

27 538

28 539 Competing Interest Statement

29 540 None to report.

30 541

31 542 References32 543 1 Carroll LJ, Hogg-Johnson S, van der Velde G, *et al.* Course and Prognostic Factors for
33
34 544 Neck Pain in the General Population. Results of the Bone and Joint Decade 2000-2010
35
36 545 Task Force on Neck Pain and Its Associated Disorders. *J Manipulative Physiol Ther*
37
38 546 Published Online First: 2009. doi:10.1016/j.jmpt.2008.11.013

- 1
2
3 547 2 Croft PR, Lewis M, Papageorgiou AC, *et al.* Risk factors for neck pain: A longitudinal
4
5 548 study in the general population. *Pain* Published Online First: 2001. doi:10.1016/S0304-
6
7 549 3959(01)00334-7
8
9
10 550 3 Vos T, Allen C, Arora M, *et al.* Global, regional, and national incidence, prevalence, and
11
12 551 years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis
13
14 552 for the Global Burden of Disease Study 2015. *Lancet* Published Online First: 2016.
15
16 553 doi:10.1016/S0140-6736(16)31678-6
17
18
19 554 4 Blanpied PR, Gross AR, Elliott JM, *et al.* Neck Pain: Revision 2017. *J Orthop Sport Phys*
20
21 555 *Ther* Published Online First: 2017. doi:10.2519/jospt.2017.0302
22
23
24 556 5 Nygren A, Berglund A, von Koch M. Neck-and-shoulder pain, an increasing problem.
25
26 557 Strategies for using insurance material to follow trends. *Scand J Rehabil Med Suppl*
27
28 558 1995;**32**:107–12.<http://www.ncbi.nlm.nih.gov/pubmed/7784832> (accessed 19 Jul 2018).
29
30
31 559 6 Wright A, Mayer TG, Gatchel RJ. Outcomes of disabling cervical spine disorders in
32
33 560 compensation injuries: A prospective comparison to tertiary rehabilitation response for
34
35 561 chronic lumbar spinal disorders. *Spine (Phila Pa 1976)* Published Online First: 1999.
36
37 562 doi:10.1097/00007632-199901150-00020
38
39
40 563 7 Rempel DM, Harrison RJ, Barnhart S. Work-Related Cumulative Trauma Disorders of the
41
42 564 Upper Extremity. *JAMA J Am Med Assoc* Published Online First: 1992.
43
44 565 doi:10.1001/jama.1992.03480060084035
45
46
47 566 8 Borghouts JAJ, Koes BW, Vondeling H, *et al.* Cost-of-illness of neck pain in The
48
49 567 Netherlands in 1996. *Pain* Published Online First: 1999. doi:10.1016/S0304-
50
51 568 3959(98)00268-1
52
53
54 569 9 Hogg-Johnson S, van der Velde G, Carroll LJ, *et al.* The Burden and Determinants of

- 1
2
3 570 Neck Pain in the General Population. Results of the Bone and Joint Decade 2000-2010
4
5 571 Task Force on Neck Pain and Its Associated Disorders. *J Manipulative Physiol Ther*
6
7
8 572 Published Online First: 2009. doi:10.1016/j.jmpt.2008.11.010
9
10 573 10 Bobos P, Nazari G, Palimeris S, *et al.* The contribution of health and psychological factors
11
12 574 in patients with chronic neck pain and disability: A cross-sectional study. *J Clin*
13
14 575 *Diagnostic Res* Published Online First: 2018. doi:10.7860/JCDR/2018/31284.11203
15
16
17 576 11 Bobos P, Billis E, Papanikolaou D-T, *et al.* Does Deep Cervical Flexor Muscle Training
18
19 577 Affect Pain Pressure Thresholds of Myofascial Trigger Points in Patients with Chronic
20
21 578 Neck Pain? A Prospective Randomized Controlled Trial. *Rehabil Res Pract* Published
22
23 579 Online First: 2016. doi:10.1155/2016/6480826
24
25
26 580 12 Nazari G, Bobos P, Billis E, *et al.* Cervical flexor muscle training reduces pain, anxiety,
27
28 581 and depression levels in patients with chronic neck pain by a clinically important amount:
29
30 582 A prospective cohort study. *Physiother Res Int* 2018;**23**. doi:10.1002/pri.1712
31
32
33 583 13 Bobos P, MacDermid JC, Walton DM, *et al.* Patient-Reported Outcome Measures Used
34
35 584 for Neck Disorders: An Overview of Systematic Reviews. *J Orthop Sport Phys Ther*
36
37 585 2018;**48**:1–76. doi:10.2519/jospt.2018.8131
38
39
40 586 14 Macdermid JC, Walton DM, Bobos P, *et al.* The Open Orthopaedics Journal A Qualitative
41
42 587 Description of Chronic Neck Pain has Implications for Outcome Assessment and
43
44 588 Classification. *Open Orthop J* 2016;**10**:746–56. doi:10.2174/1874325001610010746
45
46
47 589 15 Childs JD, Cleland JA, Elliott JM, *et al.* Neck pain: Clinical practice guidelines linked to
48
49 590 the international classification of functioning, disability, and health from the orthopaedic
50
51 591 section of the american physical therapy association. *J. Orthop. Sports Phys. Ther.* 2008.
52
53 592 doi:10.2519/jospt.2008.0303
54
55
56
57
58
59
60

- 1
2
3 593 16 Kay TM, Huijbregts M. Physical Rehabilitation Outcome Measures: A Guide to Enhanced
4
5 594 Clinical Decision Making, Second Edition. *Physiother Canada* Published Online First:
6
7 595 2003. doi:10.2310/6640.2003.35271
8
9
10 596 17 Simmonds MJ, Olson SL, Jones S, *et al.* Psychometric characteristics and clinical
11
12 597 usefulness of physical performance tests in patients with low back pain. *Spine (Phila Pa*
13
14 598 *1976)* Published Online First: 1998. doi:10.1097/00007632-199811150-00011
15
16
17 599 18 Stratford PW, Kennedy D, Pagura SMC, *et al.* The relationship between self-report and
18
19 600 performance-related measures: questioning the content validity of timed tests. *Arthritis*
20
21 601 *Rheum* 2003;**49**:535–40. doi:10.1002/art.11196
22
23
24 602 19 Novy DM, Simmonds MJ, Lee CE. Physical performance tasks: what are the underlying
25
26 603 constructs? *Arch Phys Med Rehabil* 2002;**83**:44–
27
28 604 7.<http://www.ncbi.nlm.nih.gov/pubmed/11782832> (accessed 19 Jul 2018).
29
30
31 605 20 MacDermid JC, Stratford P. Applying evidence on outcome measures to hand therapy
32
33 606 practice. *J Hand Ther* Published Online First: 2004. doi:10.1197/j.jht.2004.02.005
34
35
36 607 21 Alreni ASE, Harrop D, Lowe A, *et al.* Measures of upper limb function for people with
37
38 608 neck pain. A systematic review of measurement and practical properties. *Musculoskelet*
39
40 609 *Sci Pract* 2017;**29**:155–63. doi:10.1016/j.msksp.2017.02.004
41
42
43 610 22 Moher D, Shamseer L, Clarke M, *et al.* Preferred reporting items for systematic review
44
45 611 and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev* Published Online
46
47 612 First: 2015. doi:10.1186/2046-4053-4-1
48
49 613 23 Law MC, MacDermid J. *Evidence-based rehabilitation : a guide to practice*. Thorofare,
50
51 614 NJ: : Slack Incorporated 2014.
52
53
54 615 24 Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation
55
56
57
58
59
60

- 1
2
3 616 Coefficients for Reliability Research. *J Chiropr Med* 2016;**15**:155–63.
4
5 617 doi:10.1016/j.jcm.2016.02.012
6
7
8 618 25 Roy JS, Desmeules F, MacDermid JC. Psychometric properties of presenteeism scales for
9
10 619 musculoskeletal disorders: A systematic review. *J Rehabil Med* Published Online First:
11
12 620 2011. doi:10.2340/16501977-0643
13
14
15 621 26 Divaris K, Vann WF, Baker AD, *et al.* Examining the accuracy of caregivers' assessments
16
17 622 of young children's oral health status. *J Am Dent Assoc* 2012;**143**:1237–47.
18
19 623 doi:10.14219/jada.archive.2012.0071
20
21
22 624 27 Cohen J. Statistical power analysis for the behavioral sciences. *Stat. Power Anal. Behav.*
23
24 625 *Sci.* 1988. doi:10.1234/12345678
25
26 626 28 MacDermid JC, Walton DM, Avery S, *et al.* Measurement properties of the neck
27
28 627 disability index: a systematic review. *J Orthop Sports Phys Ther* 2009;**39**:400–17.
29
30 628 doi:10.2519/jospt.2009.2930
31
32
33 629 29 Trippolini MA, Dijkstra PU, Geertzen JHB, *et al.* Construct Validity of Functional
34
35 630 Capacity Evaluation in Patients with Whiplash-Associated Disorders. *J Occup Rehabil*
36
37 631 2015;**25**:481–92. doi:10.1007/s10926-014-9555-0
38
39
40 632 30 Trippolini MA, Dijkstra PU, Côté P, *et al.* Can functional capacity tests predict future
41
42 633 work capacity in patients with whiplash-associated disorders? *Arch Phys Med Rehabil*
43
44 634 2014;**95**:2357–66. doi:10.1016/j.apmr.2014.07.406
45
46
47 635 31 Trippolini MA, Reneman MF, Jansen B, *et al.* Reliability and safety of functional capacity
48
49 636 evaluation in patients with whiplash associated disorders. *J Occup Rehabil* 2013;**23**:381–
50
51 637 90. doi:10.1007/s10926-012-9403-z
52
53
54 638 32 Reneman MF, Roelofs M, Schiphorst Preuper HR. Reliability and Agreement of Neck
55
56
57
58
59
60

- 1
2
3 639 Functional Capacity Evaluation Tests in Patients With Chronic Multifactorial Neck Pain.
4
5 640 *Arch Phys Med Rehabil* 2017;**98**:1476–9. doi:10.1016/j.apmr.2016.12.005
6
7
8 641 33 van der Meer S, Reneman MF, Verhoeven J, *et al.* Relationship between self-reported
9
10 642 disability and functional capacity in patients with whiplash associated disorder. *J Occup*
11
12 643 *Rehabil* 2014;**24**:419–24. doi:10.1007/s10926-013-9473-6
13
14
15 644 34 Reesink DD, Jorritsma W, Reneman MF. Basis for a functional capacity evaluation
16
17 645 methodology for patients with work-related neck disorders. *J Occup Rehabil*
18
19 646 2007;**17**:436–49. doi:10.1007/s10926-007-9086-z
20
21
22 647 35 Lomond K V, Côté JN. Shoulder functional assessments in persons with chronic
23
24 648 neck/shoulder pain and healthy subjects: Reliability and effects of movement repetition.
25
26 649 *Work* 2011;**38**:169–80. doi:10.3233/WOR-2011-1119
27
28
29 650 36 Pierrynowski M, McPhee C, P Mehta S, *et al.* Intra and Inter-Rater Reliability and
30
31 651 Convergent Validity of FIT-HaNSA in Individuals with Grade II Whiplash Associated
32
33 652 Disorder. *Open Orthop J* 2016;**10**:179–89. doi:10.2174/1874325001610010179
34
35
36 653 37 Ljungquist T, Jensen IB, Nygren A, *et al.* Physical performance tests for people with long-
37
38 654 term spinal pain: aspects of construct validity. *J Rehabil Med* 2003;**35**:69–
39
40 655 75. <http://www.ncbi.nlm.nih.gov/pubmed/12691336> (accessed 11 Jul 2018).
41
42
43 656 38 Ljungquist T, Harms-Ringdahl K, Nygren A, *et al.* Intra- and inter-rater reliability of an
44
45 657 11-test package for assessing dysfunction due to back or neck pain. *Physiother Res Int*
46
47 658 1999;**4**:214–32. <http://www.ncbi.nlm.nih.gov/pubmed/10581627> (accessed 11 Jul 2018).
48
49
50 659 39 Ljungquist T, Fransson B, Harms-Ringdahl K, *et al.* A physiotherapy test package for
51
52 660 assessing back and neck dysfunction--discriminative ability for patients versus healthy
53
54 661 control subjects. *Physiother Res Int* Published Online First: 1999. doi:10.1002/pri.158
55
56
57
58
59
60

- 1
2
3 662 40 Ljungquist T, Nygren Å, Jensen I, *et al.* Physical performance tests for people with spinal
4 pain - Sensitivity to change. *Disabil Rehabil* Published Online First: 2003.
5 663
6
7 664 doi:10.1080/0963828031000090579
8
9
10 665 41 Reneman MF, Dijkstra PU, Westmaas M, *et al.* Test-retest reliability of lifting and
11 carrying in a 2-day functional capacity evaluation. *J Occup Rehabil* 2002;**12**:269–
12 666
13 75.<http://www.ncbi.nlm.nih.gov/pubmed/12389478> (accessed 19 Jul 2018).
14 667
15
16 668 42 Hegedus EJ, Vidt ME, Tarara DT. The best combination of physical performance and self-
17 report measures to capture function in three patient groups. *Phys Ther Rev* 2014;**19**:196–
18 669
19 203. doi:10.1179/1743288X13Y.0000000121
20 670
21
22 671 43 Roy J-S, Macdermid JC, Boyd KU, *et al.* Rotational strength, range of motion, and
23 function in people with unaffected shoulders from various stages of life. *Sports Med*
24 672
25 *Arthrosc Rehabil Ther Technol* 2009;**1**:4. doi:10.1186/1758-2555-1-4
26 673
27
28 674 44 Kumta P, MacDermid JC, Mehta SP, *et al.* The FIT-HaNSA Demonstrates Reliability and
29 Convergent Validity of Functional Performance in Patients With Shoulder Disorders. *J*
30 675
31 *Orthop Sport Phys Ther* 2012;**42**:455–64. doi:10.2519/jospt.2012.3796
32 676
33
34 677 45 Macdermid JCJC, Ghobrial M, Badra Quirion K, *et al.* Validation of a new test that
35 assesses functional performance of the upper extremity and neck (FIT-HaNSA) in patients
36 678
37 with shoulder pathology. *BMC Musculoskelet Disord* 2007;**8**:42. doi:10.1186/1471-2474-
38 679
39 8-42
40 680
41
42 681 46 Hawkes DH, Alizadehkhayat O, Fisher AC, *et al.* Normal shoulder muscular activation
43 and co-ordination during a shoulder elevation task based on activities of daily living: An
44 682
45 electromyographic study. *J Orthop Res* 2012;**30**:53–60. doi:10.1002/jor.21482
46 683
47
48 684
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 685
4 686
5 687
6 688
7 689
8 690
9 691
10 692
11 693
12 694
13 695
14 696
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For peer review only

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. 1999	Neck pain (55%), back pain, multiple pain sites,	53	PILE-C, PILE-L	N/A	Good (68%)
Ljungquist et al. 1999	Neck pain (50%), lumbar pain, thoracic pain, shoulder pain, multiple pain sites,	68	PILE-C, PILE-L, 2 x 20m WWB	8 days	Very Good (79%)
Ljungquist et al. 2003	Neck pain, lumbar pain, thoracic pain, shoulder pain, lower extremity pain, multiple pain sites,	235	PILE-C, PILE-L, 2 x 20m WWB	N/A	Very Good (82%)
Ljungquist et al. 2003	cervical pain (25%), lumbar pain, cervical (25%) and lumbar pain, multiple pain sites,	186	PILE-C, PILE-L, 2 x 20m WWB	6 months	Very Good (79%)
Lomond and Cote. 2011	Chronic neck and shoulder pain (100%)	32	BTEWS II	9.5 days	Very Good (88%)
Pierrynowski et al. 2016	Sub-acute and chronic WAD II	66	FIT-HaNSA	2-7 days	Very Good (88%)
Reesink et al. 2007	N/A	N/A	Neck-FCE	N/A	N/A
Reneman et al. 2017	Chronic multifactorial neck pain	18	Neck-FCE	2 weeks	Good (67%)
Trippolini et al. 2013	Sub acute and chronic WAD I and II	32	WAD FCE	7 days	Very Good (75%)
Trippolini et al. 2014	Sub acute and chronic WAD I and II	267	Workwell FCE	N/A	Excellent (92%)

Trippolini et al. 2015	Sub acute and chronic WAD I and II	314	WAD FCE	N/A	Very Good (86%)
Van der Meer et al. 2013	Chronic WAD I and II	40	Neck FCE	N/A	Very Good (86%)

PILE-C, Progressive Isoinertial Lifting Evaluation-Cervical; PILE-L, Progressive Isoinertial Lifting Evaluation; CBT, Cognitive-Behavioural Therapy; PT, Physical Therapy; NRPS, Numeric Pain Rating Scale; BTEWS II, Baltimore Therapeutic Equipment Work Simulator II; WAD, Whiplash Associated Disorder; MVA, Motor Vehicle Accident; FIT-HaNSA, Functional Impairment Test-Hand and Neck/Shoulder/Arm; FCE, Functional Capacity Evaluation; EXP, Experimental; M, Male; F, Female; N/A, not applicable

For peer review only

36/bmjopen-2019-031242 on 24 November 2019. Downloaded from <http://bmjopen.bmj.com/> on April 19, 2024 by guest. Protected by copyright.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

TABLE 2. Quality of Studies on Psychometric Properties of Functional-based Tests Evaluated in Neck Disorder Patients

Study	Item Evaluation Criteria												Total (%)
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	
Trippolini et al, 2014	2	2	2	2	1	2	2	2	2		1	2	92%
Lomond and Cote, 2011	2	2	1	2	0	2	2	2	2		2	2	88%
Pierrynowski et al, 2016	2	2	1	2	0	2	2	2	2		2	2	88%
Trippolini et al, 2015	2	2	2	0	1	N/A	2	2	2		2	2	86%
Van der Meer et al, 2013	2	1	2	1	2	N/A	2	1	2		1	2	86%
Ljungquist et al 2003 KGV**	2	2	2	0	0	N/A	2	2	2		2	2	82%
Ljungquist et al 1999 Rel****	2	1	1	2	0	2	2	2	2		1	2	79%
Ljungquist et al 2003 STC***	1	1	1	2	1	1	2	2	2		2	2	79%
Trippolini et al, 2013	2	2	1	1	0	0	2	2	2		2	2	75%
Ljungquist et al 1999 KGV**	2	1	1	2	0	N/A	2	1	2		1	2	68%
Reneman et al, 2017	1	2	1	1	1	0	1	2	2		2	1	67%
Reesink, 2007*	-	-	-	-	-	-	-	-	-		-	-	N/A

1
2
3 12-item evaluation tool (QACMRR) designed to assess the quality of studies determining measurement properties in outcome
4 measures. Questions 1-12 in the tool evaluate aspects of study question, study design, measurements, analyses, and study
5 recommendations.

6 KGV, known-groups validity; rel, reliability; STC, sensitivity-to-change

7 *Paper is not applicable for completion of study quality tool
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

For peer review only

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	Weak
			SFS: 0.42-0.61	Moderate-strong
			NDI: 0.34-0.45	Weak-moderate
			HADS-A: 0.27-0.36	Weak
		HADS-D: 0.30-0.41	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	0.06-0.39	Very weak - Weak
		Linear Mixed Model Regression of All Predictors	$\beta=-0.04$, 95% CI: -0.15 – 0.06 p=0.428 (task 6)	Not Significant

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

*Pain measured via Numeric Rating Scale

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	76 s	
		LOA ₉₅	248 s	
		MDC ₉₀	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

Fit-HaNSA, Functional Impairment Test, Hand and Neck/Shoulder/Arm; ICC, Intraclass correlation coefficient; SEM, Standard Error of Measurement; LOA₉₅, 95% Limits of Agreement; MDC₉₀, 90% Minimal Detectable Change; WAD, Whiplash Associated Disorder; Mod, Moderate

*Correlations completed with Numeric Pain Rating Scale, Neck Disability Index, Disabilities of Arm, Shoulder, Hand and 6 cervical range of motion tests

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

ICC, Intraclass correlation coefficient; SEM, Standard Error of Measurement; MDC₉₀, 90% Minimal Detectable Change; ANOVA, Analysis of Variance

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

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 behavior	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 x 20m WWB	Inter-rater Reliability	Mean Difference LoA	0.05 -1.33 and 1.43	
2 x 20m WWB	Intra-rater Reliability	Repeatability % of Range	3.2 10.7%	
2 x 20m WWB	Convergent Validity	Spearman Correlation	CR-10: 0.55 - 0.65 RPE: 0.10 - 0.48	Moderate - Strong very weak – moderate
2 x 20m WWB	Discriminative: spinal pain vs control	Wilcoxon Sign Ranked Test	p=0.014	Significant
2 x 20m WWB	Discriminative: High vs. low pain intensity	Mann Whitney U	p<0.001	Significant
2 x 20m WWB	Discriminative: High vs. low pain behaviour	Mann Whitney U	p<0.001	Significant
2 x 20m WWB	Discriminative: High vs. low perceived exertion	Mann Whitney U	p<0.001	Significant
2 x 20m WWB	Sensitivity to change	Effect Size	Subjects improving: 0.38-0.78 Subjects deteriorating: 0.13-0.62	Small – Moderate Trivial – Moderate

1
2
3 PILE-C, Progressive Iso-inertial Lifting Evaluation – Cervical; PILE-L, Progressive Iso-inertial Lifting Evaluation – Lumbar; LoA,
4 Limits of Agreement; SD, Standard Deviation; M, Male; F, Female; RPE, Rating of perceived exertion; KGV, Known-groups
5 Validity; Neg., Negligible; Mod., Moderate, *CR-10: Measurement of pain construct
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

For peer review only

36bmjopen-2019-021242 on 24 November 2019. Downloaded from <http://bmjopen.bmj.com/> on April 19, 2024 by guest. Protected by copyright.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

For peer review only

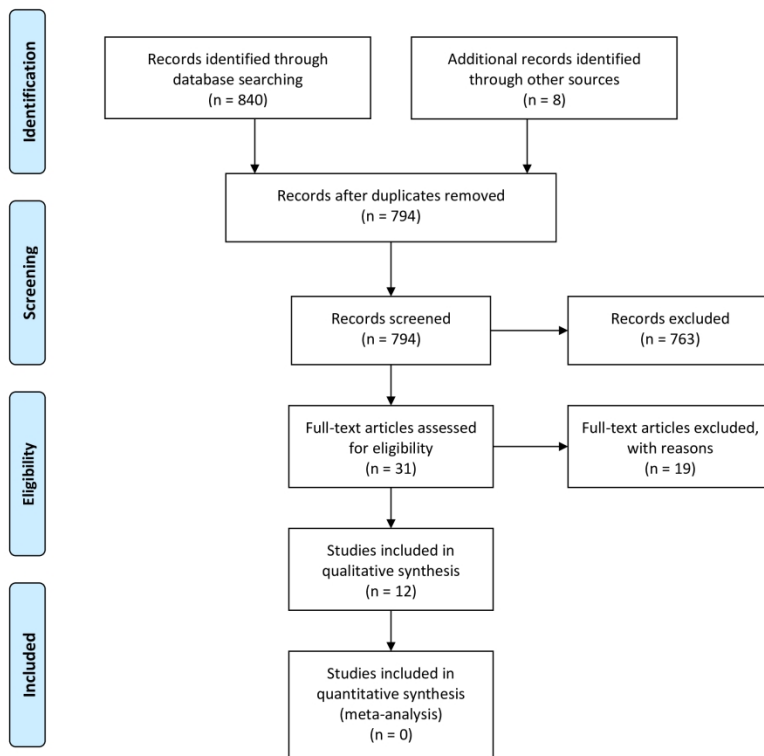


Figure 1

215x279mm (300 x 300 DPI)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Appendix 1: Search terms

EMBASE-OVID

1. exp "outcome and process assessment (health care)"/ or "outcome assessment (health care)"/ or treatment outcome/
2. outcome?.ti.
3. exp "Range of Motion, Articular"/
4. Pain Measurement/
5. exp disability evaluation/
6. "Recovery of Function"/
7. Questionnaires/
8. self-report.tw.
9. ((impairment or disability or function) adj2 (measure? or scale? or evaluation?)).tw.
10. range of motion.tw.
11. (strength adj2 (measure? or scale? or evaluation?)).tw.
12. (outcome? adj2 (measure* or scale? or indicator?)).tw.
13. or/1-12
14. "reproducibility of results"/
15. exp "Sensitivity and Specificity"/
16. reliability.mp.
17. validity.mp.
18. responsiveness.mp.
19. Psychometrics/
20. rasch.mp.
21. factor analysis, statistical/
22. factor analysis.tw.
23. differential functioning.mp.
24. (validity or validation).mp. [mp=title, original title, abstract, name of substance word, subject heading word, unique identifier]
25. (validity or validation).mp.
26. item difficulty.mp.
27. translation.tw.
28. or/14-27
29. 13 and 28
30. Neck Pain/
31. exp Brachial Plexus Neuropathies/
32. exp neck injuries/ or exp whiplash injuries/
33. cervical pain.mp.
34. neckache.mp.
35. whiplash.mp.
36. cervicodynia.mp.
37. cervicgia.mp.
38. brachialgia.mp.
39. brachial neuritis.mp.
40. brachial neuralgia.mp.
41. neck pain.mp.

- 1
- 2
- 3
- 4 42. neck injur*.mp.
- 5 43. brachial plexus neuropath*.mp.
- 6 44. brachial plexus neuritis.mp.
- 7 45. thoracic outlet syndrome/ or cervical rib syndrome/
- 8 46. Torticollis/
- 9 47. exp brachial plexus neuropathies/ or exp brachial plexus neuritis/
- 10 48. cervico brachial neuralgia.ti,ab.
- 11 49. cervicobrachial neuralgia.ti,ab.
- 12 50. (monoradicul* or monoradicl*).tw.
- 13 51. or/30-50
- 14 52. exp headache/ and cervic*.tw.
- 15 53. exp genital diseases, female/
- 16 54. genital disease*.mp.
- 17 55. or/53-54
- 18 56. 52 not 55
- 19 57. 51 or 56
- 20 58. neck/
- 21 59. neck muscles/
- 22 60. exp cervical plexus/
- 23 61. exp cervical vertebrae/
- 24 62. atlanto-axial joint/
- 25 63. atlanto-occipital joint/
- 26 64. Cervical Atlas/
- 27 65. spinal nerve roots/
- 28 66. exp brachial plexus/
- 29 67. (odontoid* or cervical or occip* or atlant*).tw.
- 30 68. axis/ or odontoid process/
- 31 69. Thoracic Vertebrae/
- 32 70. cervical vertebrae.mp.
- 33 71. cervical plexus.mp.
- 34 72. cervical spine.mp.
- 35 73. (neck adj3 muscles).mp.
- 36 74. (brachial adj3 plexus).mp.
- 37 75. (thoracic adj3 vertebrae).mp.
- 38 76. neck.mp.
- 39 77. (thoracic adj3 spine).mp.
- 40 78. (thoracic adj3 outlet).mp.
- 41 79. trapezius.mp.
- 42 80. cervical.mp.
- 43 81. cervico*.mp.
- 44 82. 80 or 81
- 45 83. exp genital diseases, female/
- 46 84. genital disease*.mp.
- 47 85. exp *Uterus/
- 48 86. 83 or 84 or 85
- 49 87. 82 not 86
- 50
- 51
- 52
- 53
- 54
- 55
- 56
- 57
- 58
- 59
- 60

- 1
- 2
- 3
- 4 88. 58 or 59 or 60 or 61 or 62 or 63 or 64 or 65 or 66 or 67 or 68 or 69 or 70 or 71 or 72 or 73 or
- 5 74 or 75 or 76 or 77 or 78 or 79 or 87
- 6 89. exp pain/
- 7 90. exp injuries/
- 8 91. pain.mp.
- 9 92. ache.mp.
- 10 93. sore.mp.
- 11 94. stiff.mp.
- 12 95. discomfort.mp.
- 13 96. injur*.mp.
- 14 97. neuropath*.mp.
- 15 98. or/89-97
- 16 99. 88 and 98
- 17 100. Radiculopathy/
- 18 101. exp temporomandibular joint disorders/ or exp temporomandibular joint dysfunction
- 19 syndrome/
- 20 102. myofascial pain syndromes/
- 21 103. exp "Sprains and Strains"/
- 22 104. exp Spinal Osteophytosis/
- 23 105. exp Neuritis/
- 24 106. Polyradiculopathy/
- 25 107. exp Arthritis/
- 26 108. Fibromyalgia/
- 27 109. spondylitis/ or discitis/
- 28 110. spondylosis/ or spondylolysis/ or spondylolisthesis/
- 29 111. radiculopathy.mp.
- 30 112. radiculitis.mp.
- 31 113. temporomandibular.mp.
- 32 114. myofascial pain syndrome*.mp.
- 33 115. thoracic outlet syndrome*.mp.
- 34 116. spinal osteophytosis.mp.
- 35 117. neuritis.mp.
- 36 118. spondylosis.mp.
- 37 119. spondylitis.mp.
- 38 120. spondylolisthesis.mp.
- 39 121. or/100-120
- 40 122. 88 and 121
- 41 123. exp neck/
- 42 124. exp cervical vertebrae/
- 43 125. Thoracic Vertebrae/
- 44 126. neck.mp.
- 45 127. (thoracic adj3 vertebrae).mp.
- 46 128. cervical.mp.
- 47 129. cervico*.mp.
- 48 130. 128 or 129
- 49 131. exp genital diseases, female/
- 50
- 51
- 52
- 53
- 54
- 55
- 56
- 57
- 58
- 59
- 60

132. genital disease*.mp.
133. exp *Uterus/
134. or/131-133
135. 130 not 134
136. (thoracic adj3 spine).mp.
137. cervical spine.mp.
138. 123 or 124 or 125 or 126 or 127 or 135 or 136 or 137
139. Intervertebral Disk/
140. (disc or discs).mp.
141. (disk or disks).mp.
142. 139 or 140 or 141
143. 138 and 142
144. herniat*.mp.
145. slipped.mp.
146. prolapse*.mp.
147. displace*.mp.
148. degenerat*.mp.
149. (bulge or bulged or bulging).mp.
150. 144 or 145 or 146 or 147 or 148 or 149
151. 143 and 150
152. intervertebral disk degeneration/ or intervertebral disk displacement/
153. intervertebral disk displacement.mp.
154. intervertebral disc displacement.mp.
155. intervertebral disk degeneration.mp.
156. intervertebral disc degeneration.mp.
157. 152 or 153 or 154 or 155 or 156
158. 138 and 157
159. 57 or 99 or 122 or 151 or 158
160. animals/ not (animals/ and humans/)
161. 159 not 160
162. exp *neoplasms/
163. exp *wounds, penetrating/
164. 162 or 163
165. 161 not 164
166. 29 and 165
167. guidelines as topic/
168. practice guidelines as topic/
169. guideline.pt.
170. practice guideline.pt.
171. (guideline? or guidance or recommendations).ti.
172. consensus.ti.
173. or/167-172
174. meta-analysis/
175. exp meta-analysis as topic/
176. (meta analy* or metaanaly* or met analy* or metanaly*).tw.
177. review literature as topic/

- 1
- 2
- 3 178. (collaborative research or collaborative review* or collaborative overview*).tw.
- 4 179. (integrative research or integrative review* or intergrative overview*).tw.
- 5 180. (quantitative adj3 (research or review* or overview*)).tw.
- 6 181. (research integration or research overview*).tw.
- 7 182. (systematic* adj3 (review* or overview*)).tw.
- 8 183. (methodologic* adj3 (review* or overview*)).tw.
- 9 184. exp technology assessment biomedical/
- 10 185. (hta or thas or technology assessment*).tw.
- 11 186. ((hand adj2 search*) or (manual* adj search*)).tw.
- 12 187. ((electronic adj database*) or (bibliographic* adj database*)).tw.
- 13 188. ((data adj2 abstract*) or (data adj2 extract*)).tw.
- 14 189. (analys* adj3 (pool or pooled or pooling)).tw.
- 15 190. mantel haenszel.tw.
- 16 191. (cochrane or pubmed or pub med or medline or embase or psycinfo or psychlit or psychinfo or
- 17 psychlit or cinahl or science citation indes).ab.
- 18 192. or/174-191
- 19 193. 173 or 192
- 20 194. 166 and 193
- 21
- 22
- 23
- 24
- 25
- 26
- 27
- 28
- 29
- 30
- 31
- 32
- 33
- 34
- 35
- 36
- 37
- 38
- 39
- 40
- 41
- 42
- 43
- 44
- 45
- 46
- 47
- 48
- 49
- 50
- 51
- 52
- 53
- 54
- 55
- 56
- 57
- 58
- 59
- 60

APPENDICES

APPENDIX A. Data extraction guide for studies evaluating the quality of studies evaluating the clinical measurement properties of outcome measures

Instructions

Clinical measurement studies may evaluate a wide spectrum of measurement properties; or evaluate aspects that relate to the implementability or interpretation of outcome measures. Individual clinical measurement studies cannot address every aspect of the measurement properties of an instrument. Ideally systematic reviews will synthesize the quality and content of research evidence addressing the clinical measurement properties of individual outcome measures. The summative knowledge about the measurement properties, cultural transferability, and utility across different contexts provides the scope of information needed to select an outcome measure for a specific patient (population), purpose and context.

This guide should facilitate extraction of data from individual clinical measurement studies. An explanation of the measurement property addressed in each item and how it might be measured within a given study is listed to facilitate finding and extracting that information. The accompanying extraction form can then be used to collect the specific information on these measurements or utility properties from specific studies.

The purpose of data extraction is to extract the specific information reported by authors within a study, not to evaluate the validity or value of that piece of information. Evaluation of the quality of the published version of the clinical measurement study (also called critical appraisal) is performed in a separate step. See the accompanying critical appraisal tool and guide. It is advisable to extract detailed specific information from the study; recognizing that this information may later be synthesized or subject to meta-analysis.

There is no standardized process for synthesizing clinical measurement information. Based on the findings of extraction you may elect to present the synthesized data in a descriptive way by creating a summary table of the data extracted in each category. If you find some studies with similar designs, you may be able to conduct a meta-analysis of some properties like clinically important difference (CID) or minimal detectable change (MDC); if appropriate given the sample and technique - this can be valuable as it may provide more stable estimates of these important properties.

<u>Population studied</u>		
Population	A description of the study population	Sample size, pathology/disorder, demographics, setting, acute vs. chronic, where subjects were chosen from. Report meaningful demographics and indicators of the population studied.
Intervention	Interventions (if applicable) applied during longitudinal studies	Description of the nature, frequency, intensity of the intervention and the follow-up interval.
<u>Reliability</u>		
Reliability Description	The extent to which scores for patients who have not changed are the same for repeated measurement under several conditions: for example, using different sets of items from the same health-related instrument (internal consistency), over time (test retest) by different persons on the same occasion (interrater) or by the same persons (i.e., raters or responders) on different occasions (intra-rater)	Test procedures or measures are typically reapplied on repeated occasions in individuals considered to have a stable condition during that time frame which repeated testing occurs. Repeated testing may be performed on different occasions (test-retest) for self-report measures, OR by the same rater (intra-rater) or different raters (inter-rater) if it is an observer-based scale. In some cases different test instruments (inter-instrument) are evaluated. The most common statistic used is the intraclass correlation coefficient for quantitative data (Shrout & Fleiss, 1979) and kappa (Landis & Koch, 1977) for nominal data. Standard error of measurement is used to present a quantitative estimate of the reliability—in the original units of measure. Report the type of reliability evaluated and coefficients obtained.
Measurement Error	The systematic and random error of a patient's score that is not attributed to true changes in the construct to be measured	This may be reported as 1. Standard error of measurement (in older articles you may see coefficient of variation) 2. Altman and Bland graphical technique (Bland & Altman, 1990; Bland & Altman, 1987; Bland & Altman, 1986) where the difference on repeated tests for each individual (limits of agreement) is plotted versus their

336bmjopen-2019-021722n14 November 2019 09:00:00. Copyright. For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

		mean score. The mean difference and the boundaries of 2SD are shown to define the limits of agreement.
Internal consistency	The extent to which items on a test or subscale are related (an indication of the consistency of the concept measured).	Cronbach's alpha is the inter-item correlation usually reported. Report alpha and whether it relates to the entire instrument or specific subscales.
<u>Validity</u>		
Content Validity	The degree to which the content of a health-related instrument is an adequate reflection of the construct to be measured	<p>A variety of techniques can be used to assess the extent to which items on a given measure reflected the necessary content to capture the concept of interest. Some of the techniques you will find are listed. Extract what was done to determine content validity and what was found.</p> <ol style="list-style-type: none"> 1) Patients and experts were involved during item selection/reduction - report how they were used and key decisions 2) Patients were consulted for reading and comprehension - report key findings 3) Cognitive interviews (Cibelli, 1994; Ojanen & Gogates, 2006) were done with patients to determine how items were interpreted by respondents; their perceptions of the items - report key findings 4) Expert panels or Delphi procedures were used to select items or evaluate the validity of the instrument - report key findings and decisions 5) During translation specific study, the meaning of the questions to another cultural or language group was studied - report key findings and decisions 6) ICF linking (Cieza et al., 2002) or other coding of content was performed - report the results which may include the distribution of content across ICF domains, or the distribution of specific codes
Construct Validity	The degree to which the scores of a health-related instrument are consistent with hypotheses (for instance with regard to internal	When extracting data about correlational validity, the pre-constructed hypothesis and whether it is supported should be documented. For correlational construct

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

36/bmjopen-2019-002222-n1411-g001-01-20240711 11:11:43 AM
For peer review only - http://bmjopen.bmj.com/ on April 19, 2024 by guest. Protected by copyright.

	relationships, relationships to scores of other instruments, or differences between relevant groups) based on the assumption that the health-related instrument validly measures the construct to be measured	validity, this will be the nature and strength of the prespecified relationship and the correlations that support that. Relation to other indices/constructs that are similar (convergent) or different (divergent) can be reported. Ideally, hypotheses are formulated/reported and supported by correlations that are in accordance with the hypotheses. Note that there is no consistent agreement on what subjective term should be applied to validity correlations. Note that there is no consistent agreement on what subjective term should be applied to validity correlations. Some authors use subjective terminology defined for reliability such as: strong (>0.70) and moderate (0.40-0.70) correlations; others use the correlations like effect size benchmarks that 0.4 indicates a moderate effect and 0.6 a large effect. For validity assessment is more important than correlations prespecified constructed hypotheses, although not all papers are written clearly with respect to this.
Structural Validity/Hypothesis Testing	The degree to which the scores of a health-related instrument are an adequate reflection of the dimensionality of the construct to be measured	Extract test names, prespecified expected relationship and correlations observed.
Structural validity - discriminative	discriminative analysis supports the validity of a measure by demonstrating that the measurement is able to differentiate between groups that are prespecified and <u>known</u> to be different on the construct being assessed.	Data extraction should include the nature of the subgroups and the size of the difference observed between them (and its statistical significance). Typically, statistical tests of difference are performed. Since known groups analysis can provide data that is useful in clinical practice as benchmarks for comparing these known groups, it is a more practical form of construct validity than correlational. Data extraction/presentation should reflect this by presenting the group central tendency, their margins and statistical significance in an accessible manner.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

36/bmjopen-2019-061220 on November 2019. Downloaded from <http://bmjopen.bmj.com/> on April 19, 2024 by guest. Protected by copyright.

<p>Criterion validity</p>	<p>Criterion validation is determined by comparing a given outcome measure to an accepted standard of measure. For subjective constructs like pain and disability, it can be argued that there is no criterion since there is no external gold standard. Therefore, for self-report measures, validation focuses on construct validity.</p> <p>For performance measures, it is common to have a criterion measure that is considered to be highly precise and rigorous as the criterion comparator.</p>	<p>Authors will state that their measure is being compared against a specific instrument and report the correlation or agreement between the measures. Extract the test names and results: correlations or other as reported.</p>
<p>Responsiveness/Clinical Change</p>		
<p>Responsiveness</p>	<p>The ability of a health-related instrument to detect change over time in the construct to be measured</p>	<p>Extract indicators of responsiveness include: effect size, standard response mean and the method for assessing whether patients were improved, stable or worse. (Beaton, 2000)</p>
<p>Interpretability</p>		
<p>Interpretability</p>	<p>The degree to which one can assign qualitative meaning that is, clinical or commonly understood connotations to an instrument's quantitative scores or change in scores.</p>	

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

APPENDIX B. Data extraction form for studies evaluating the clinical measurement properties of outcome measures

Authors: _____ Year: _____ Rater: _____

Instructions

When using the data extraction form, it is important to realize that the purpose of data extraction is to remove or extract the specific information reported by authors within a study, not to evaluate the validity or value of that piece of information. To make data extraction as useful as possible, and to avoid the need for repeated data extractions, it is advisable to read the accompanying guide and then be as specific as possible when extracting information.

	DATA EXTRACTED
	Population studied
Population	
Intervention	
	Reliability
Reliability (relative)	
Reliability (absolute)	
Minimum Detectable Change	
	Content/structural validity
Internal consistency	
Content Validity	

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

36/bmjopen-2019-031242 on 24 November 2019. Downloaded from <http://bmjopen.bmj.com/> on April 19, 2024 by guest. Protected by copyright.

Floor-Ceiling Effects	
Factorial validity	
Item response /Rasch Analyses	
Construct/Criterion Validity	
Known groups	
Convergent	
Divergent	
Longitudinal Validity	
Concurrent criterion	
Predictive criterion	
Responsiveness/Clinical Change	
Responsiveness	

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

Minimally Clinical Important Difference	

For peer review only

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

APPENDIX C. Quality Appraisal for Clinical Measurement Research Reports Evaluation Form

Rater (Group) _____

Author(s) (Study Author(s)) _____

Year (Year of publication) _____

1. Was the relevant background work cited to define what is currently known about the measurement properties of measures under study, and the potential contributions of the current research question to informing that knowledge base?

2

1

0

2. Were appropriate inclusion/exclusion criteria defined? *

2

1

0

3. Were specific clinical measurement questions/hypotheses identified?

2

1

0

4. Was an appropriate scope of measurement properties considered?

2

1

0

5. Was an appropriate sample size used?

2

1

0

6. Was appropriate retention/follow-up obtained? (for studies involving retesting; otherwise n/a)

2

1

0

- 1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
7. Were specific descriptions provided of the measure under study and the method(s) used to administer it?
- 2
1
0
8. Were standardized procedures used to administer all study measures in a manner that minimized potential sources of error/bias (including the study measure and its comparators)?
- 2
1
0
9. Were analyses conducted for each specific hypothesis or purpose?
- 2
1
0
10. Were appropriate statistical tests performed to obtain point estimates of the measurement properties?
- 2
1
0
11. Were appropriate ancillary analyses done to quantify the confidence in the estimates of the clinical measurement property (Precision/Confidence intervals; benchmark comparisons/ROC curves, alternate forms of analysis like SEM/MID, etc.)?
- 2
1
0
12. Were clear, specific and accurate conclusions made about the clinical measurement properties; that were associated with appropriate clinical measurement recommendations and supported by the study objectives, analysis and results?
- 2
1
0
- Subtotals (of column 1 and 2) Total Score (sum of subtotals/24*100)

APPENDIX D. Description of each performance battery from selected articles

Battery	Description of Tasks
Relevant FCE Subtasks ^{25,26,27,28,29,30}	<p>Material Handling Tasks: All lifting tests were executed with a wooden crate (40 × 30 × 26 cm) of 2.5 kg, and four to five weight increments of 2.5 kg or 5 kg each were used until the maximum amount of weight was reached. Maximum performance was recorded in kg.</p> <p>Lifting floor to waist: Measured after five lifts of crate from floor to table and vice versa (time limit < 90 s): hands remained on the crate during the test. Increase weight in 4-5 steps until maximum is reached</p> <p>Overhead lift test: Five lifts from waist to crown height and vice versa within 90 s in standing position. Increase weight in 4–5 steps until maximum is reached</p> <p>Two-handed carrying: Carrying of a crate for a short distance measured after five carries of 1.5 m distance at waist height. Hands remain on the crate during the test.</p> <p>One-handed carrying: Carrying wooden crate for 15 m within 90 s beginning with the right hand and thereafter the left hand.</p> <p>Overhead working: Standing with hands at crown height for manipulation of nuts and bolts. The time that the position was held is recorded (sec).</p> <p>Repetitive reaching: fast horizontal movements of the upper extremity in a sitting position. Marbles are removed from bowls at arm length distance at table height from left to right and vice versa, with right and then left arm. The time taken to remove 30 marbles is recorded (sec).</p> <p>Overhead lift test: Five lifts from waist to crown height and vice versa within 90 s in standing position. Increase weight in 4–5 steps until maximum is reached</p> <p>Repetitive bending and overhead reaching: 20 marbles in 2 bowls at table height and crown height. Standing in front of bowl of marbles and moving the marbles as fast as possible from table height to crown height.</p>

336/bmjopen-2019-033436 on 24 November 2019. Downloaded from <http://bmjopen.bmj.com/>. Protected by copyright.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

<p>A Physiotherapy Test Package^{33,34,35,36}</p>	<p>PILE Tests: “The lifting tests were performed standing in front of bookshelves with shelves at 0.76m and 1.37 m from the floor. Subjects were asked to lift weights in a plastic box from floor to waist level (0–0.76 m) for the lumbar PILE test, or from waist to shoulder height (0.76–1.37 m) for the cervical PILE test. The initial weight was 3.6 kg for women and 4.9 kg for men. A ‘lifting movement’ involved a single transfer from one level to the next and back again. After every four such lifting movements (= 20 s), the weight was increased by 2.25 kg for women and 4.5 kg for men. The weight managed during the last lifting movement was recorded and used as a test result, as well as this maximum weight divided by the ‘adjusted weight’”.</p> <p>2x20m WWB: “Subjects were asked to walk 20 m at a comfortable speed along a corridor, to turn around where 20 m was marked and then to walk 20 m back to the starting point. In the first walking test they carried no extra weight, but in the second they carried one carrier bag in each hand, containing 4 kg each for the women, 8 kg each for the men. The time taken was recorded to get the walking speed. The tests were discontinued after 50 s”.</p>
<p>BTEWS II³¹</p>	<p>“The protocol consisted of performing a series of shoulder functional tasks before and after a fatiguing activity. Functional tasks consisted of active shoulder range of motion (ROM) in both flexion and abduction and cumulative power output (PO) accumulated over 10s during a repetitive pushing/pulling task in a horizontal plane at shoulder level”.</p>
<p>FIT - HaNSA³²</p>	<p>“The FIT-HaNSA protocol consists of three timed tasks and each task is performed for a maximum of 300 seconds (s) with approximately 30 s pause between them (set-up time for next task). Task 1 (waist-up) requires the patient to alternately “grab, lift, move and place” three 1000 g containers located on waist level and 25 cm above waist level shelves, using their affected arm, at a metronome pace of 60 beats per minute for 300 s or until they felt unable to continue. The time to complete Task 1 is measured using a stopwatch. Task 2 (eye-down) is identical to Task 1 except that the two shelves are placed at eye-level and 25 cm below. Task 3 (overhead work) requires a patient to repeatedly screw and unscrew bolts in a sagittal plane oriented plate positioned at eye-level using both arms”. More complete description at https://srs-mcmaster.ca/wp-content/uploads/2015/04/FIT-HaNSAProtocol_April2007.pdf</p>



PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	1
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	2
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	3
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	3
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	4
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	4
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	3-4
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	3-4
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	4
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	4
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	5
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	NA
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	NA



PRISMA 2009 Checklist

Page 1 of 2

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	NA
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	NA
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	6-7
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICCO, follow-up period) and provide the citations.	6-7
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	6-10
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	6-10
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	6-10
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	6-10
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	NA
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	11-13
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	14-16
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	16
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	18

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit: www.prisma-statement.org.

Page 2 of 2

For peer review only - <http://bmjopen.bmj.com/site/about/guidelines.xhtml>