

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**

# Psychometric Properties of the Global Rating of Change Scales in Patients with Neck Disorders: A Systematic Review with Meta-Analysis and Meta-Regression

Journal:	BMJ Open
Manuscript ID	bmjopen-2019-033909
Article Type:	Original research
Date Submitted by the Author:	27-Aug-2019
Complete List of Authors:	Bobos, Pavlos; Western University, Health and Rehabilitation Sciences; University of Toronto, Institute of Health Policy Management and Evaluation MacDermid, Joy; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Nazari, Goris; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Furtado, Rochelle; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Group, CATWAD; Michele Sterling Anne Söderlund, Michele Curatolo, James M Elliott, David M Walton, Helge Kasch, Linda Carroll, Hans Westergren, Gwendolen Jull, Eva-Maj Malmström, Luke B Connelly, Joy C MacDermid, Mandy Nielsen, Pierre Côté, Tonny Elmose Andersen, Trudy Rebbeck, Annick Maujean, Sarah Robins, Kenneth Chen, Julia Treleaven
Keywords:	neck pain, global assessment, psychometric properties, systematic review

SCHOLARONE™ Manuscripts

- 1 Psychometric Properties of the Global Rating of Change Scales in Patients with Neck
- 2 Disorders: A Systematic Review with Meta-Analysis and Meta-Regression
- 3 Pavlos Bobos<sup>1</sup>, Joy C MacDermid<sup>2</sup>, Goris Nazari<sup>3</sup>, Rochelle Furtado<sup>4</sup> and CATWAD co-authors<sup>5</sup>
- <sup>1</sup>Pavlos Bobos PT, PhD(c), (corresponding author) Doctoral Candidate, Western's Bone and Joint
- 6 Institute, Department of Health and Rehabilitation Sciences, Western University, Elborn College,
- 7 1201 Western Road, N6G 1H1, London, Ontario, Dalla Lana School of Public Health, Institute of
- 8 Health Policy Management and Evaluation, Department of Clinical Epidemiology and Health Care
- 9 Research, University of Toronto, Canada, (pbobos@uwo.ca), tel: +1 519 661 2111 x88912
- <sup>2</sup>Joy C MacDermid BScPT, PhD, Professor, Physical Therapy and Surgery, Western University,
- London, ON and Co-director Clinical Research Lab, Hand and Upper Limb Centre, St. Joseph's
- 12 Health Centre, London, Ontario; Professor Rehabilitation Science McMaster University,
- Hamilton, ON, Canada (<u>imacderm@uwo.ca</u>)
- <sup>3</sup>Goris Nazari PT, PhD(c) Doctoral Candidate, Western's Bone and Joint Institute, School of
- 15 Physical Therapy, Department of Health and Rehabilitation Sciences, Western University,
- 16 London, Ontario, Canada, (gnazari@uwo.ca)
- <sup>4</sup>Rochelle Furtado MSc Western's Bone and Joint Institute, School of Physical Therapy,
- 18 Department of Health and Rehabilitation Sciences, Western University, London, Ontario, Canada,
- 19 (<u>rfurtad5@uwo.ca</u>)
- <sup>5</sup>CATWAD: Michele Sterling <u>m.sterling@uq.edu.au</u>, Anne Söderlund <u>anne.soderlund@mdh.se</u>,
- 21 Michele Curatolo, <u>curatolo@uw.edu</u>, James M Elliott <u>j-elliott@northwestern.edu</u>, David Walton
- dwalton5@uwo.ca, Helge Kasch helgkasc@rm.dk, Linda Carroll linda.carroll@ualberta.ca,
- Hans Westergren <u>Hans.Westergren@skane.se</u>, Gwendolen Jull <u>g.jull@uq.edu.au</u>, Eva-Maj
- 24 Malmström eva-maj.malmstrom@med.lu.se, Luke B Connelly l.connelly@uq.edu.au, Joy C
- 25 MacDermid <u>imacderm@uwo.ca</u>, Mandy Nielsen <u>mandy.nielsen@griffith.edu.au</u>, Pierre Côté
- 26 pierre.cote@uoit.ca, Tonny Elmose Andersen tandersen@health.sdu.dk, Trudy Rebbeck
- 27 trudy.rebbeck@sydney.edu.au, Annick Maujean a.maujean@uq.edu.au, Sarah Robins
- s.robins1@uq.edu.au, Kenneth Chen k.chen8@uq.edu.au, Julia Treleaven j.treleaven@uq.edu.au
- **Kewords:** neck pain, global assessment, psychometric properties, systematic review
- **30 Word count: 3908**

- Objective: The purpose of this systematic review was to critically appraise and synthesize the psychometric properties of Global Rating of Change (GROC) scales for assessment of patients
- with neck pain.
- **Design:** Systematic review
- 36 Data sources: A search was performed in 4 databases (MEDLINE, EMBASE, CINAHL,
- 37 SCOPUS) until February 2019.
- **Data extraction and synthesis:** Eligible articles were appraised using Consensus-based Standards
- 39 for the selection of health Measurement Instruments (COSMIN) checklist and the Quality
- 40 Appraisal for Clinical Measurement Research Reports Evaluation Form.
- **Results:** The search obtained 16 eligible studies and included in total 1533 patients with neck pain.
- 42 Test-retest reliability of Global Perceived Effect (GPE) was very high (Intra-class correlation
- coefficient (ICC) = 0.80 to 0.92) for patients with whiplash. Pooled data of Pearson's r indicated
- that GROC scores were moderately correlated with neck disability change scores (0.53, 95% CI:
- 45 0.47 to 0.59). Pooled data of Spearman's correlations indicated that GROC scores were moderately
- 46 correlated with neck disability change scores (0.56, 95% CI: 0.41 to 0.68).
- **Conclusions:** This study found excellent quality evidence of very good to excellent test-retest
- 48 reliability of GPE for patients with Whiplash Associated Disorders. Evidence from very good-to-
- 49 excellent quality studies found that GROC scores are moderately correlated to an external criterion
- patient-reported outcome (PROM) measure evaluated pre-post treatment in patients with neck
- pain. No studies were found that addressed the optimal form of GROC scales for patients with
- neck disorders or compared the GROC to other options for single-item global assessment.
- **Prospero registration number:** CRD 42018117874

- We rated the quality of individual studies and the overall risk of bias using two standardized approaches
- Our focus on neck pain increased the specificity of results but are not necessarily applicable to other musculoskeletal conditions
- Conceptual concerns about global ratings of change being affected by recall bias are not adequately addressed by psychometric evidence
- No studies addressing the optimal form of global rating were found.

#### Introduction

Neck pain is the 4<sup>th</sup> leading cause of disability and approximately half of adult the population with neck pain will experience a clinically important episode once in their lifetime. [1–3] The annual prevalence of neck pain it is estimated between 15% and 50%, with females having a higher prevalence rate than males. [2,3] Neck pain has been associated with many other comorbidities such as headaches, dizziness, anxiety, depression, back pain and arthralgias.[3–6] Several different methods for classifying neck pain have been described, using indicators such as duration (acute, sub-acute or chronic), degree of interference (low, moderate, severe) or most likely structure at fault (e.g. neuropathy vs. mechanical). [7]

As part of a patient-centric approach to care, clinicians will commonly evaluate response to intervention by asking the patient directly whether they feel better, worse, or the same since the prior encounter. While direct questioning can provide a qualitative indicator of change in status, many best practice guidelines endorse use of some form of quantified patient-reported outcome (PRO) as an adjunct to oral self-report. PROs are available to quantify several different constructs in people with neck pain, including pain severity, disability and neck function. [8] Any PRO

intended to provide an estimate of change over time should be responsive to subtle shifts in the patient's condition. To facilitate interpretation of change scores, a common property of many such tools is the minimum clinically important difference (MCID), which is a change threshold that corresponds to the minimum shift in scale values that most patients would indicate corresponds to an important change in their overall condition. A well-recognized approach to establishing an MCID for a PRO is to compare the magnitude of change against an anchor, most commonly a Global Rating of Change (GROC) scale. These scales allow patients or study participants to indicate whether their condition has gotten worse, better, or stayed the same and to quantify the magnitude of that change. As they have been adopted as a sort of 'standard' against which change in other tools is compared, the GROC can also be used on its own as an omnibus generic indicator of change. [8]

Despite being accepted as a standard measure, there is considerable variation in how the GROC has been constructed and implemented in research in neck pain. Some are 15 points, some 11 points, and others are 7 points. The common structure across these is the use of a middle '0' score corresponding to 'no change', with negative values indicating magnitudes of worsening while positive values indicate improvement.[9] Variations of the GROC (in name or structure) include the "Global Perceived Effect", "Patient Global Impression of Change", "Transition Ratings", and "Global Scale". [9]

A critical component of monitoring changes in health outcomes is having valid, reliable and responsive tools with strong psychometric properties. While recent research [8] has examined the psychometric properties of the most commonly reported PROs for neck disorders, to date there has been no systematic review to summarize the measurement properties of GROC scales themselves in patients with neck disorders. Therefore, this systematic review aims to critically

60

1 2		
3	103	appraise and synthesize the psychometric properties of the GROC scales in patients with neck
5 6	104	disorders.
7 8	105	
9 10 11	106	METHODS
12 13	107	Patient and Public Involvement
14 15	108	There was no patient or public involvement in the design or planning of this study.
16 17	109	
18 19 20	110	Study Design and Protocol Registration
21 22	111	We conducted a systematic review to evaluate the psychometric properties of GROC scales in
23 24	112	patients with neck disorders. The protocol was registered in PROSPERO register database with
25 26	113	registration number: CRD 42018117874
27 28 29	114	
30 31	115	Eligibility Criteria
32 33	116	We included studies in this systematic review if the following criteria were met [10–12]:
34 35 36	117	<ul> <li>Design: psychometric testing, randomized/ cohort studies</li> </ul>
37 38	118	• Participants: > 50% of the study's patient population with neck conditions/disorders,
39 40	119	• Intervention/Comparison: studies that reported on the psychometric properties (reliability,
41 42	120	validity, responsiveness) of GROC, Global Perceived Effect (GPE) and Patient Global
43 44 45	121	Impression of Change (PGIC),
46 47	122	<ul> <li>Outcomes: GROC, GPE and PGIC.</li> </ul>
48 49	123	Studies with no data on the GROC scales' psychometric properties, and conference
50 51		abstract/posters were excluded from this systematic review.
52 53	124	abstract/posters were excluded from this systematic review.
54 55 56	125	
57 58		5

Information Sources

To identify studies on the psychometric properties (reliability, validity, responsiveness) of the GROC, GPE and PGIC we searched the Medline, EMBASE, Scopus and CINAHL databases from inception till February 2019, using a combination of keywords. Furthermore, we identified additional studies by examining the reference list of each of the selected studies. The full list with keyword strategy is presented in **APPENDIX 1**.

## Study Selection

Two investigators (PB and GN) performed the systematic electronic searches independently in each database. The same investigators then proceeded to identify and remove the duplicate studies. In the next stage, we performed the independent screening of the titles and abstracts and any full-text article marked as include or uncertain were obtained. In the final stage, the same two independent authors performed the full text reviews independently to assess final article eligibility. In case of disagreement, a third reviewer; the most experienced member (JM), facilitated a consensus through discussion.

### Data Extraction

The fourth author (RF) performed the data extractions. The extracted data were then cross-checked by another author (PB). Data extraction included the author, year, study population/condition, setting, sample size, age, properties evaluated, retest-interval, and the intervention protocol (if used to assess responsiveness parameters). [13,14] For reliability estimates, Standard Error of Measurement (SEM), Intra-class Correlation Coefficient (ICC), Minimal Detectable Change (MDC) and 95% confidence intervals were extracted. [13,14] The ICC interpretation of ICC < 0.40

indicating poor,  $0.40 \le ICC < 0.75$  indicating fair-to-good and  $ICC \ge 0.75$  indicating excellent reliability were used as a common benchmark. For validity estimates, correlation coefficient (Pearson's/Spearman) and the 95% confidence intervals were extracted. [13,14] Evan's guidelines to interpret the strength of the correlation was used which included: 0.00–0.19 "very weak", 0.20– 0.39 "weak", 0.40–0.59 "moderate", 0.60–0.79 "strong", and 0.80–1.00 "very strong". [15] For responsiveness estimates, the Effect Size (ES), Standardized Response Mean (SRM), Clinically Important Difference (CID), and/or Minimal Clinically Important Difference (MCID) including the method of MCID estimation – Anchor-/Distribution-based methods, and 95% confidence intervals were extracted. [13,14] To assist clinical decision making, standard benchmark scores of trivial (< 0.20), small ( $\ge 0.20$  to < 0.50), moderate ( $\ge 0.50$  to < 0.80) or large ( $\ge 0.80$ ), as proposed by Cohen, were used. [16] When insufficient data were presented, PB contacted the authors by email and requested further data.

Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) assesses the risk of bias for the psychometric properties reported on a property-by-property basis. A score for the risk of bias in estimates of psychometric properties was assessed by two authors (PB) and (RF) using the new (COSMIN) checklist.[17] If disagreement was present a third person (JM) assist in resolving the discrepancy. Each study was scored on the 4-point scale as "very good", "adequate", "doubtful" or "inadequate" for each of the checklist criteria for relevant measurement properties (e.g. reliability, responsiveness, etc.). To determine the overall score for each measurement property, the worst score counts method was used wherein the lowest score for the checklist criteria of the relevant property was taken as the overall score. [18] We then assessed

the result of individual studies on a measurement property against the updated criteria for good measurement properties. This involved the evaluation of results of included studies as either sufficient (+), insufficient (-), or indeterminate (?). [17]

Quality Appraisal for Clinical Measurement Research Reports Evaluation Form

A summary score for the overall quality of individual studies was appraised independently by the authors (PB) and (RF) using a structured clinical measurement specific appraisal tool. [13,14] In case of disagreement a third person was consulted (JM) to resolve the conflict. The evaluation criteria of this tool included twelve items: 1) Thorough literature review to define the research question; 2) Specific inclusion/exclusion criteria; 3) Specific hypotheses; 4) Appropriate scope of psychometric properties; 5) Sample size; 6) Follow-up; 7) The authors referenced specific procedures for administration, scoring, and interpretation of procedures; 8) Measurement techniques were standardized; 9) Data were presented for each hypothesis; 10) Appropriate statistics-point estimates; 11) Appropriate statistical error estimates; and 12) Valid conclusions and recommendations. [13,14] An article's total score – quality - was calculated by the sum of scores for each item, divided by the numbers of items and multiplied by 100%. [13,14] Overall, the quality summary of appraised articles range from (0%-30%) Poor, (31%-50%) Fair, (51%-70%) Good, (71%-90%) Very Good, and (>90%) Excellent. [13,14]

#### Synthesis of Results

A qualitative synthesis was conducted to report findings on test-retest reliability statistics. A metaanalysis of Pearson's and Spearman's correlation was performed in Comprehensive Meta-Analysis 3.3 software (Englewood, NJ). The meta-analyses were conducted using a random effect

model and the correlation coefficients were converted to z values. Heterogeneity was deemed substantial if I<sup>2</sup> values were more than 50%. [19] A Meta-regression was planned to explore the sources of unexplained heterogeneity by considering the following factors: a. neck pain with or without radicular symptoms, b. acute or chronic, c. age and d. sex. Forest plots were created using means and 95% confidence intervals for correlation coefficients. We summarize the main results of the included articles based on the neck disorders, reported psychometric estimate and the study quality ratings. 

#### **RESULTS**

Study Selection

Our search yielded 123 articles. After removal of duplicates, 106 studies remained and were screened using their title and abstract; leaving 28 articles selected for full-text review. Of these, 17 studies were considered eligible. [20,21,30–35,22–29] The flow of the study selection process is presented in Figure 1.

#### Study Characteristics

The 16 eligible studies were conducted between 2006 and 2017 and included 1533 participants with neck pain/disorders (mean of 96 participants per study). [20,21,30,32–35,22–29] Study size ranged from 29 to 200 participants. A summary description of all the studies included is displayed in **Table 1.** Concurrent validity was evaluated in 14 studies by comparing the difference of pain intensity, disability and function scores with the score of GROC scales. Two studies [24,29] examined the test-retest reliability of a 7-point and an 11-point GPE scale for patients with whiplash-associated disorders (WAD). One study [22] examined whether occurrences of within-

data mining, Al training, and similar technologies

Protected by copyright, including for uses related to text

and between-session changes were significantly associated with functional outcomes, pain, and self-report of recovery in patients at discharge who were treated with manual therapy for mechanical neck pain.

- COSMIN Risk of Bias rating and Quality appraisal of the Included Studies
- Regarding the risk of bias, all studies were rated as very good (**Table 2**). The quality of the studies ranged from 88% to 96% (**Table 3**). The most common flaws were 1) lack of/inadequate sample size calculations, 2) missing data (i.e. inadequate follow up), and 3) inconsistencies between the data presented and hypothesis stated.

- Reported GROC scales
- The most commonly reported GROC scale (n=6 studies) was a 15-point scale with the most frequent anchors being "-7 (a very great deal worse) to zero (about the same) to +7 (a very great deal better)". A 7-point scale was reported in 5 studies, 11- and 5-point scales were reported in 2 studies and a 9-point scale in one study. The anchors in those scales varied greatly and are presented in Table 1. Only 6 studies [24,29–31,33,34] reported full detail regarding the specific questions asked of the patients with neck disorder when a GROC scale was administered. Those questions that were reported are presented in **Box 1**.

- 237 Reliability Measures
- Two studies were included that examined test-retest reliability of GPE for patients with WAD.
- Kamper et al. (2010) [24] examined the [time interval] test-retest reliability of an 11-point GPE
- scale in 134 patients with chronic WAD and reported an Intra-class Correlation Coefficient (ICC)

of 0.99 (95% CI 0.99 to 0.99) at baseline, 0.96 (0.95 to 0.97) at 6 weeks, and 0.92 (0.89 to 0.94) at 12 months. (**Table 4**). Ngo et al. (2010) assessed the test-retest reliability of a 7-point scale of GPE in patients with acute WAD at 3 to 5 days. [29] The ICC and 95% confidence intervals (CI) were used to determine the test-retest reliability of the two versions of the perceived recovery questions using their original seven-item responses. Ngo et al. also computed weighted kappa coefficients and 95% CI using quadratic weights to determine whether the distribution of responses influenced the reliability as measured by the ICC. An ICC for general recovery of 0.70 (0.60 to 0.80) () and an ICC for neck pain questions of 0.80 (0.72 to 0.87) were found. A weighted Kappa was also calculated (Kappa = 0.70 (0.42 to 0.98)) at six weeks for general recovery and at six weeks Kappa = 0.80 (0.51 to 1.0) for neck pain questions (**Table 4**).

# Validity Measures

We found 14 studies that examined concurrent validity measures between GROC and another PRO (**Table 5**). Bjorklund et al. compared the validity of GROC with ProFitMAP-neck change scores (moderate correlations: rho = 0.47, (p<0.05) and the Neck Disability Index (NDI) (moderate correlations: rho = 0.59, (p<0.05) in patients with non-specific neck-shoulder pain.[30] Cleland et al. compared the validity of GROC with NDI change scores (very weak correlations: r = 0.19) and with Patient Specific Functional Scale change scores (PSFS) (very strong correlations: r = 0.82) in 38 patients with cervical radiculopathy.[20] Cleland et al. compared the GROC with NDI change scores (moderate correlations: r = 0.58) and with Numeric Pain Rating Scale (NPRS) scores (moderate correlations: r = 0.57) in 137 patients with neck pain.[21] Farooq et al. compared the GROC with the Urdu version of NDI change scores, and indicated moderate correlations r = 0.50 in 106 patients with neck pain.[36] Guzy et al. compared the GROC with NDI change scores

and reported moderate to strong correlations r = -0.73 at two weeks and -0.56 at four weeks, in 95

Meta-Analysis and Meta-Regression of Correlations between Disability change scores and GROC scores Five studies [21,23,32,35,36] of very good-to-excellent quality reported the Pearson correlation coefficients between neck disability change scores and the GROC scores and were pooled together. We found that GROC was positively correlated with disability change scores (r = 0.53, 95% CI: 0.47 to 0.59,  $I^2 = 0\%$ ). Six studies [25–28,30,34] of very good-to-excellent quality reported the Spearman correlation coefficients between neck disability changes scores and the GROC scores and were pooled together. We found that GROC was moderately correlated with disability change scores (rho = 0.56, 95% CI: 0.41 to 0.68,  $I^2 = 85\%$ ). The forest plots with correlation coefficients with 95% CIs are presented in Figure 2-3. Our meta-regression showed that age was found as a significant factor in influencing Fisher's Z scores ( $\beta = -0.034$ , 95% CI -0.05 to -0.01, p = 0.001). The model explained 68% of the variance ( $R^2 = 0.68$ ) (Figure 4). *Area under the curve (AUC) – Sensitivity and Specificity* Cook et al. [22] found that between-session NPRS- pain changes were associated with greater than 3-point change on the GROC at 96-hours (AUC=0.76). The pain change associated with GROC 

**DISCUSSION** 

who did not achieve a 36.7% change in pain (Table 4).

was more specific (Specificity=79.2%, range: 62.2 - 91.1) than sensitive (Sensitivity=65.6%,

range: 57.9 to 74.6). Those with a 36.7% between-sessions change in pain were also 7.3 times

more likely to report an improvement of greater than 3 points change on the GROC than those

This review has synthesized the current research from 17 studies that aimed to evaluate the psychometric properties of GROC scales for patients with neck disorders, with the goal to provide evidence for clinicians and researchers concerning its use within clinical practice and research. From the 17 included studies, only 2 studies [24,29] reported test-retest reliability statistics of the 7- and 11-points item GPE scales for patients with WAD only. We were able to pool data from 12 studies regarding concurrent validity of GROC scales and neck disability change scores at one time point after the interventions.<sup>3</sup> Themes influencing interpretation of the GROC were explored in a study [31] that evaluated the factors that contribute to how patients respond to a question on global perceived effect. This study found that treatment process, biomechanical performance, self-efficacy and the nature of the condition may influence the responses on global perceived effect, which is consistent with what we would expect for patients with neck pain. This suggests that change is a complex multifactorial global concept. A strength of GROC is that it is intended as a global assessment, and it can be assumed that it reflects the aspects of change important to the individual patient.

Reliability can be defined as the degree to which a measure produces consecutive results with the least amount of random error when the status of the population remains unchanged. The reliability of GPE displayed an excellent test-retest reliability of ICC>0.90 over an interval of 6 weeks and 12 months for patients with WAD. Conducting an assessment with a long test-retest interval (e.g. 12 months), can provide challenges as there is higher risk of individuals with WAD being symptomatically unstable.[9] Determining if patients are symptomatically-stable can be achieved by administering another PRO such as the Single Assessment Numeric Evaluation (SANE)[37], however, the 7- and 11- points GPE scales still demonstrated good stability properties

at long test intervals (i.e., of 6 weeks and 12 months). Therefore, the measurements of the reliability parameters of the GPE may be very useful during longer test intervals in clinical trials.

The psychometric property of validity is defined as the degree to which a PRO measures what it is intended to measure. Pooled data from 11 studies overall suggest that post-treatment changes of on validated disability outcome measures were moderately (Pearson's r = 0.51, 95% CI: 0.43 to 0.58; Spearman's rho = 0.56, 95% CI: 0.41 to 0.68) correlated to change in perceived effect) (Figure 2-3). This finding suggests that GROC scores taken at one point in time were related to scores in pain and disability in patients with neck disorders, as measured by standardized measures taken at 2 points in time. We identified one study [22] that found a 36.7% change in pain for within- and between- session changes was associated with a 50% reduction in the NDI and an improvement of >3 points on a 15-points GROC scale for patients with neck pain. This quantified predictive change value may have clinical utility for use in clinical practice.

Previous studies [9,38] have indicated serious concerns about the conceptual validity of the global rating of change. The review by Kamper et al.[9] clearly showed that GROC was related to final status more than change and was least related to baseline health status. This result undermines the premise of what the global rating of change actually measures. For this reason, we conclude that the 0.50 pooled correlation across 12 studies between the GROC and other PROM change scores (e.g. NDI scores) may reflect a relationship between follow-up status and change rather than supporting the contention that GROC actually measures change. This would also explain why only 25% of the variation in GROC change scores was explained by changes scores from a PROM change score measured at 2 points in time. In all studies, participants completed the GROC scale at one time point after the intervention, and hence recall bias is a cause for concern. However, another potential factor for moderate correlations is that the PROM, used as a

A unique aspect of this study was that it focused on global rating of change scales in a neck pain patient population. Our study appraisal suggests that future studies concerning GROC should include adequate sample sizes, maintain a rigorous follow up and report appropriate statistical error estimates, since these were often inadequate. Various critical appraisal tools exist, and the perspectives and ratings may differ across instruments. We used 2 different critical appraisal tools to evaluate quality from 2 perspectives. The COSMIN risk of bias assessments reflects the level of confidence in the conclusions and pooled estimates. The quality appraisal tool focuses on design issues in the studies and reflects gaps in research designs that should be considered in interpretation of current research and improved in future studies. Substantial heterogeneity was detected (I<sup>2</sup>>50%) in pooled Spearman's correlation coefficients which is a concern when pooling data. Our univariate meta-regression analysis indicated that age across the studies explained 68% of the variance (Figure 4). Other factors such as type of neck pain (with or without radicular symptoms), acute or chronic and sex did not explain the remaining heterogeneity (not statically significant). Furthermore, the scope of our literature search was focused on identifying full-text papers written in English only.

While this study included 16 studies, only 2 of these reported reliability statistics for GROC scales for patients with chronic WAD. Therefore, the applicability of our study is mostly limited to patients with chronic WAD. For validity measurements, GROC scales were mostly investigated by correlation analyses to evaluate the external responsiveness of another PRO measure over a

specific time point. From our meta-analysis, we can be confident that the GROC scores were 

moderately correlated with neck disability change scores. However, more robust psychometric design studies to test the measurement properties of GROC scales as the primary outcome of investigation are highly needed. Future studies should aim to test to what extent the different range of items (e.g. 7-point scale vs 11-point scale), the anchors (e.g. much worse vs much better) may affect the measurement properties of GROC scales for patients with neck disorders.

# **CONCLUSIONS**

This study found excellent quality evidence of very good to excellent test-retest reliability of GPE for patients with WAD. Evidence of very good to excellent quality studies found that GROC scores are moderately correlated to an external criterion PROM measure measured pre-post treatment in patients with neck disorders. Studies addressing the optimal form of GROC scales for patients with neck disorders or comparing the GROC to other options for single-item global assessment of change were not found.

#### **Authors' contributions**

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

399		
400	Decl	arations
401	Ethic	cs approval and consent to participate
402	Not a	applicable
403	Cons	sent for publication
404	Not a	applicable
405	Avai	lability of data and material
406	Data	sharing is not applicable to this article as no datasets were generated or analyzed during the
407	curre	ent study
408	Func	ling Statement
409	This	work was supported by the Canadian Institutes of Health Research (CIHR) with funding
410	refer	ence number (FRN: SCA-145102).
411	Com	peting Interest Statement
412	None	e to report
413		
414		
415	Refe	rences
416	1	Murray CJL, Abraham J, Ali MK, et al. The State of US health, 1990-2010: Burden of diseases,
417		injuries, and risk factors. JAMA - J Am Med Assoc Published Online First: 2013.
418		doi:10.1001/jama.2013.13805
419	2	Fejer R, Kyvik KO, Hartvigsen J. The prevalence of neck pain in the world population: A
420		systematic critical review of the literature. Eur. Spine J. 2006. doi:10.1007/s00586-004-0864-4
421	3	Hogg-Johnson S, van der Velde G, Carroll LJ, et al. The Burden and Determinants of Neck Pain in
422		the General Population. Results of the Bone and Joint Decade 2000-2010 Task Force on Neck

1 2			
3 4	423		Pain and Its Associated Disorders. <i>J Manipulative Physiol Ther</i> Published Online First: 2009.
5 6	424		doi:10.1016/j.jmpt.2008.11.010
7 8	425	4	Bobos P, Nazari G, Palimeris S, et al. The contribution of health and psychological factors in
9	426		patients with chronic neck pain and disability: A cross-sectional study. J Clin Diagnostic Res
11 12	427		Published Online First: 2018. doi:10.7860/JCDR/2018/31284.11203
13 14	428	5	Macdermid JC, Walton DM, Bobos P, et al. The Open Orthopaedics Journal A Qualitative
15 16 17	429		Description of Chronic Neck Pain has Implications for Outcome Assessment and Classification.
17 18 19	430		Open Orthop J 2016;10:746–56. doi:10.2174/1874325001610010746
20 21	431	6	Treleaven J. Sensorimotor disturbances in neck disorders affecting postural stability, head and eye
22 23	432		movement control-Part 2: Case studies. Man Ther 2008;13:266–75.
24 25	433		doi:10.1016/j.math.2007.11.002
26 27	434	7	Cohen SP. Epidemiology, diagnosis, and treatment of neck pain. <i>Mayo Clin Proc</i> 2015; <b>90</b> :284–99.
28 29	435		doi:10.1016/j.mayocp.2014.09.008
30 31	436	8	Bobos P, MacDermid JC, Walton DM, et al. Patient-Reported Outcome Measures Used for Neck
32 33	437		Disorders: An Overview of Systematic Reviews. J Orthop Sport Phys Ther 2018;48:775–88.
34 35	438		doi:10.2519/jospt.2018.8131
36 37 38	439	9	Kamper SJ, Maher CG, Mackay G. Global Rating of Change Scales: A Review of Strengths and
39 40	440		Weaknesses and Considerations for Design. J Man Manip Ther 2009;17:163–70.
41 42	441		doi:10.1002/mus.21062
43 44	442	10	Nazari G, Bobos P, MacDermid JC, et al. The Effectiveness of Instrument-Assisted Soft Tissue
45 46	443		Mobilization in Athletes, Participants Without Extremity or Spinal Conditions, and Individuals
47 48	444		with Upper Extremity, Lower Extremity, and Spinal Conditions: A Systematic Review. Arch Phys
49 50	445		Med Rehabil Published Online First: February 2019. doi:10.1016/j.apmr.2019.01.017
51 52	446	11	Bobos P, Nazari G, Szekeres M, et al. The effectiveness of joint-protection programs on pain,
53 54	447		hand function, and grip strength levels in patients with hand arthritis: A systematic review and
55 56 57	448		meta-analysis. <i>J Hand Ther</i> 2018; <b>32</b> :194–211. doi:10.1016/j.jht.2018.09.012
58 59			19

449	12	Nazari G, Bobos P, MacDermid JC, et al. Psychometric properties of the Zephyr bioharness
450		device: A systematic review. BMC Sports Sci Med Rehabil 2018;10. doi:10.1186/s13102-018-
451		0094-4
452	13	Law MC, MacDermid J. Evidence-based rehabilitation: a guide to practice. Thorofare, NJ::
453		Slack Incorporated 2014.
454	14	Roy JS, Desmeules F, MacDermid JC. Psychometric properties of presenteeism scales for
455		musculoskeletal disorders: A systematic review. J Rehabil Med Published Online First: 2011.
456		doi:10.2340/16501977-0643
457	15	Wuensch KL, Evans JD. Straightforward Statistics for the Behavioral Sciences. J Am Stat Assoc
458		Published Online First: 2006. doi:10.2307/2291607
459	16	Cohen J. Statistical power analysis for the behavioral sciences. Stat. Power Anal. Behav. Sci.
460		1988. doi:10.1234/12345678
461	17	Mokkink LB, de Vet HCW, Prinsen CAC, et al. COSMIN Risk of Bias checklist for systematic
462		reviews of Patient-Reported Outcome Measures. Qual Life Res Published Online First: 2018.
463		doi:10.1007/s11136-017-1765-4
464	18	Terwee CB, Mokkink LB, Knol DL, et al. Rating the methodological quality in systematic reviews
465		of studies on measurement properties: a scoring system for the COSMIN checklist. 2012;:651–7.
466		doi:10.1007/s11136-011-9960-1
467	19	Higgins JPT, Thompson SG, Deeks JJ, et al. Measuring inconsistency in meta-analyses. BMJ
468		Published Online First: 2003. doi:10.1136/bmj.327.7414.557
469	20	Cleland J, Fritz J, Whitman J, et al. The reliability and construct validity of the Neck Disability
470		Index and Patient Specific Functional Scale. Spine (Phila Pa 1976) 2006;31:598-602.
471	21	Cleland JA, Childs JD, Whitman JM. Psychometric Properties of the Neck Disability Index and
472		Numeric Pain Rating Scale in Patients With Mechanical Neck Pain. Arch Phys Med Rehabil
473		2008; <b>89</b> :69–74. doi:10.1016/j.apmr.2007.08.126
474	22	Cook C, Lawrence J, Michalak K, et al. Is there preliminary value to a within- and/or between-
		20

475		session change for determining short-term outcomes of manual therapy on mechanical neck pain?
476		J Man Manip Ther 2014; <b>22</b> :173–80. doi:10.1179/2042618614y.0000000071
477	23	Guzy G, Vernon H, Polczyk R, et al. Psychometric validation of the authorized Polish version of
478		the Neck Disability Index. <i>Disabil Rehabil</i> 2013; <b>35</b> :2132–7. doi:10.3109/09638288.2013.771706
479	24	Kamper SJ, Ostelo RWJG, Knol DL, et al. Global Perceived Effect scales provided reliable
480		assessments of health transition in people with musculoskeletal disorders, but ratings are strongly
481		influenced by current status. J Clin Epidemiol 2010;63:760-766.e1.
482		doi:10.1016/j.jclinepi.2009.099
483	25	Shaheen AAM, Omar MTA, Vernon H. Cross-cultural adaptation, reliability, and validity of the
484		arabic version of neck disability index in patients with neck pain. Spine (Phila Pa 1976)
485		2013; <b>38</b> :609–15. doi:10.1097/BRS.0b013e31828b2d09
486	26	Takeshita K, Hosono N, Kawaguchi Y, et al. Validity, reliability and responsiveness of the
487		Japanese version of the Neck Disability Index. J Orthop Sci 2013;18:14-21. doi:10.1007/s00776-
488		012-0304-y
489	27	Trouli MN, Vernon HT, Kakavelakis KN, et al. Translation of the Neck Disability Index and
403	_ ,	
490	-,	validation of the Greek version in a sample of neck pain patients. BMC Musculoskelet Disord
	2,	
490	28	validation of the Greek version in a sample of neck pain patients. BMC Musculoskelet Disord
490 491		validation of the Greek version in a sample of neck pain patients. <i>BMC Musculoskelet Disord</i> 2008; <b>9</b> :1–8. doi:10.1186/1471-2474-9-106
490 491 492		validation of the Greek version in a sample of neck pain patients. <i>BMC Musculoskelet Disord</i> 2008; <b>9</b> :1–8. doi:10.1186/1471-2474-9-106  Tuttle N, Laakso L, Barrett R. Change in impairments in the first two treatments predicts outcome
490 491 492 493		validation of the Greek version in a sample of neck pain patients. <i>BMC Musculoskelet Disord</i> 2008; <b>9</b> :1–8. doi:10.1186/1471-2474-9-106  Tuttle N, Laakso L, Barrett R. Change in impairments in the first two treatments predicts outcome in impairments, but not in activity limitations, in subacute neck pain: An observational study. <i>Aust</i>
490 491 492 493 494	28	validation of the Greek version in a sample of neck pain patients. <i>BMC Musculoskelet Disord</i> 2008; <b>9</b> :1–8. doi:10.1186/1471-2474-9-106  Tuttle N, Laakso L, Barrett R. Change in impairments in the first two treatments predicts outcome in impairments, but not in activity limitations, in subacute neck pain: An observational study. <i>Aust J Physiother</i> 2006; <b>52</b> :281–5. doi:10.1016/S0004-9514(06)70008-3
490 491 492 493 494 495	28	validation of the Greek version in a sample of neck pain patients. <i>BMC Musculoskelet Disord</i> 2008; <b>9</b> :1–8. doi:10.1186/1471-2474-9-106  Tuttle N, Laakso L, Barrett R. Change in impairments in the first two treatments predicts outcome in impairments, but not in activity limitations, in subacute neck pain: An observational study. <i>Aust J Physiother</i> 2006; <b>52</b> :281–5. doi:10.1016/S0004-9514(06)70008-3  Ngo Trung, Stupar Maja, Co^te' Pierre, Boyle Eleanor, Shearer Heather. A study of the test –
490 491 492 493 494 495 496	28	validation of the Greek version in a sample of neck pain patients. <i>BMC Musculoskelet Disord</i> 2008; <b>9</b> :1–8. doi:10.1186/1471-2474-9-106  Tuttle N, Laakso L, Barrett R. Change in impairments in the first two treatments predicts outcome in impairments, but not in activity limitations, in subacute neck pain: An observational study. <i>Aust J Physiother</i> 2006; <b>52</b> :281–5. doi:10.1016/S0004-9514(06)70008-3  Ngo Trung, Stupar Maja, Co^te′ Pierre, Boyle Eleanor, Shearer Heather. A study of the test – retest reliability of the self-perceived general recovery and self-perceived change in neck pain
490 491 492 493 494 495 496 497	28	validation of the Greek version in a sample of neck pain patients. <i>BMC Musculoskelet Disord</i> 2008; <b>9</b> :1–8. doi:10.1186/1471-2474-9-106  Tuttle N, Laakso L, Barrett R. Change in impairments in the first two treatments predicts outcome in impairments, but not in activity limitations, in subacute neck pain: An observational study. <i>Aust J Physiother</i> 2006; <b>52</b> :281–5. doi:10.1016/S0004-9514(06)70008-3  Ngo Trung, Stupar Maja, Co^te' Pierre, Boyle Eleanor, Shearer Heather. A study of the test – retest reliability of the self-perceived general recovery and self-perceived change in neck pain questions in patients with recent whiplash-associated disorders. 2010;:957–62.
490 491 492 493 494 495 496 497 498	28 29	validation of the Greek version in a sample of neck pain patients. <i>BMC Musculoskelet Disord</i> 2008; <b>9</b> :1–8. doi:10.1186/1471-2474-9-106  Tuttle N, Laakso L, Barrett R. Change in impairments in the first two treatments predicts outcome in impairments, but not in activity limitations, in subacute neck pain: An observational study. <i>Aust J Physiother</i> 2006; <b>52</b> :281–5. doi:10.1016/S0004-9514(06)70008-3  Ngo Trung, Stupar Maja, Co^te' Pierre, Boyle Eleanor, Shearer Heather. A study of the test – retest reliability of the self-perceived general recovery and self-perceived change in neck pain questions in patients with recent whiplash-associated disorders. 2010;:957–62. doi:10.1007/s00586-010-1289-x

501		Qual Life Res 2017; <b>26</b> :161–70. doi:10.1007/s11136-016-1373-8
502	31	Evans R, Bronfort G, Maiers M, et al. "I know it" s changed": a mixed-methods study of the
503		meaning of Global Perceived Effect in chronic neck pain patients. 2014;:888-97.
504		doi:10.1007/s00586-013-3149-y
505	32	Jorritsma W, Dijkstra PU, De Vries GE, et al. Detecting relevant changes and responsiveness of
506		Neck Pain and Disability Scale and Neck Disability Index. Eur Spine J 2012;21:2550–7.
507		doi:10.1007/s00586-012-2407-8
508	33	Monticone M, Frigau L, Vernon H, et al. Responsiveness and minimal important change of the
509		NeckPix © in subjects with chronic neck pain undergoing rehabilitation. Eur Spine $J$
510		2018; <b>27</b> :1324–31. doi:10.1007/s00586-017-5343-9
511	34	Monticone M, Ambrosini E, Vernon H, et al. Responsiveness and minimal important changes for
512		the Neck Disability Index and the Neck Pain Disability Scale in Italian subjects with chronic neck
513		pain. Eur Spine J 2015; <b>24</b> :2821–7. doi:10.1007/s00586-015-3785-5
514	35	Young BA, Walker MJ, Strunce JB, et al. Responsiveness of the Neck Disability Index in patients
515		with mechanical neck disorders. Spine J 2009;9:802–8. doi:10.1016/j.spinee.2009.06.002
516	36	Farooq MN, Mohseni-Bandpei MA, Gilani SA, et al. Urdu version of the neck disability index: A
517		reliability and validity study. BMC Musculoskelet Disord 2017;18:1-11. doi:10.1186/s12891-017-
518		1469-5
519	37	Williams GN, Gangel TJ, Arciero RA, et al. Comparison of the single assessment numeric
520		evaluation method and two shoulder rating scales. Outcomes measures after shoulder surgery. Am
521		J Sports Med 1999;27:214-21. doi:10.1177/03635465990270021701
522	38	Schmitt J, Abbott JH. Global Ratings of Change Do Not Accurately Reflect Functional Change
523		Over Time in Clinical Practice. J Orthop Sport Phys Ther 2015;45:106–11.
524		doi:10.2519/jospt.2015.5247
525		
526		

**Population** 

Patients with cervical

Patients with neck

Patients with any

Patients with neck

Patients with neck

Patients with chronic

non-specific neck

Patients with any

whiplash-associated

Patients with chronic

Patients with chronic

Women with non-

specific neck-

shoulder pain

radiculopathy

pain only

neck pain

pain

pain

pain

disorder.

neck pain

neck pain

**Properties Evaluated** 

Validity (correlation)

Between NDI and GRoC

Validity (correlation)

Between NDI and GRoC

Between PSFS and GRoC

Validity (correlation)

Between NDI and GRoC

Between NPRS and GRoC

ROC curves and AUC to

measure sensitivity and

specificity. Binomial logistic

regression analysis was also

calculated to determine

overall effect.

Validity (correlation)

Between NDI-U and GRoC

Validity (correlation)

Between NDI-P and GRoC

Validity (correlation)

Between NDI and GRoC

Between NPAD and GRoC

Test-retest reliability

Validity (correlation)

Between NeckPix and GPE

Validity (correlation)

Between NDI and GPE

Between NPDS and GPE

Sample

Size

104

38

137

56

106

95

76

134

153

200

Setting

Not specified

Hospital

5 Outpatient

physical

therapy

clinics

Academic

locations in

Northeast

Ohio

Physical

therapy

clinics

Outpatient

rehabilitation

clinic

Tertiary

university

center for

rehabilitation

Physical

therapy

clinics

Outpatient

Rehabilitatio

n Unit

Outpatient

Rehabilitatio

n Unit

2

 Table 1. Study Characteristics

5

Study

6 Bjorklund

7 et al (2017)

9

10

12 leland et 1al (2006)

14 15 Cleland et

**16**l. (2008) 17

18 15 ook et al

262014)

21 22

23 24Faroog et

291. (2017) 26

27 28 Guzy et al.

292013) 30

31 32 33. Jorritsma et

34. (2012) 35

36 3 Kamper et

38l. (2010)

39 40

41 Monticone

42<sub>t al. 2017</sub>

43

44

45

4) Monticone

et al. 2015

51

52 53 54

59

60

49 50

527

528

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

23

BMJ Open: first published as

Interval

GRoC scale

week)

administered only

one time point (1

after intervention at

GRoC was completed

at follow up. Within a

week over the perio

GRoC was comp

Baseline and at

follow up 48- an 49

hours post baseline

GRoC was completed at three weeks after 2019.

GRoC scale was

After completion

the program vary

from 3 to 5 mon

patients filled the

Baseline, 6 week

and 12 months

At the end of

follow-up

treatment (8 weeks)

and one year befare

At the end of treatment 8 weels

bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de

**GPE** 

and at 4 weeks

completed at 2 work

at follow up. Wikin

of 7 weeks.

week

**GROC** evaluated

**GRoC 7-points** 

1. Very much worse; 2. Much

worse; 3. Minimally worse; 4.

No change; 5. Minimally

improved; 6. Much improved; 7.

Very much improved.

**GRoC 15-points** 

-7 (a very great deal worse) to

zero (about the same) to +7 (a

very great deal better)

**GRoC 15-points** 

-7 (a very great deal worse) to

zero (about the same) to +7 (a

very great deal better)

**GRoC 15-points** 

-7 (a very great deal worse) to

zero (about the same) to +7 (a

very great deal better)

GRoC 15-points

-7 (a very great deal worse) to

zero (about the same) to +7 (a

very great deal better)

**GRoC 7-points** 

'complete recovery' over "no

change" to "my complaints are

worse than ever"

GPE 7-points

3 (completely recovered) to zero

(no change) to -3 (worse than

ever)

GPE 11-points

-5 (vastly worse) to zero

(unchanged) to +5 (completely

recovered)

GPE 5-points

(helped a lot = 1, helped = 2),

one no change level (helped only

a little = 3), and two worsening

levels (did not help = 4, made

things worse = 5)

GPE 5-points

(helped a lot = 1, helped = 2),

one no change level (helped only

a little = 3), and two worsening

levels (did not help = 4, made things worse = 5)

The proper of the pain for more than 2 physiotherap weeks pain for more than 2 physiotherap y clinics between PSFS and GPE Completely recovered Seventh PSFS and GPE Between PSFS	4 (2010) M 5 (6 6 W 7 8 9 10 11 12 1\$\frac{1}{3}\text{haheen et Pa}{14}\text{. (2015) pa}{15} 15 1\$\frac{1}{4}\text{tal. pa}{182014) ra} 19	ratients with WAD. Most participants 69.6%) had grade II VAD.  ratients with neck ain lasting more nan 3 months ratients with neck ain, cervical adiculopathy and/or ervical myelopathy ratients with neck	Interviewed by person or by telephone in Ontario  3 primary health centers  Variety of clinics and hospital settings Primary	70	Validity (correlation) Between NDI-Ar and GRoC  Validity (correlation) Between NDI-J and GRoC  Validity (correlation)	GPE 7-points  1. General recovery question Completely better Much improved Slightly improved No change Slightly worse Much worse Worse than ever  2. Change in neck pain question: very much better, better, slightly better, no change, slightly worse, worse, or very much worse GROC 15-points  -7 (a very great deal worse) to zero (about the same) to +7 (a very great deal better) PGIC 7-points much better, better, slightly better, unchanged, slightly worse, worse and much worse GROC 15-points	Protected by copyright, in Within 2 months late 1 week for test-rates are the second at the second a
Worse at all) and from '(a very great deal better) to 1 (almost the same, hardly any better at all)  Tutle et al. Patients with neck physiotherap weeks   Private physiotherap y clinics   Between NDI and GPE   Setween NDI	13. (2015) pa 14. (2015) pa th 15 16 akeshita pa 18 2014) ra 19 ce 20 rouli et al. Pa 2 (2008) pa	ain lasting more nan 3 months ratients with neck ain, cervical adiculopathy and/or ervical myelopathy ratients with neck	Variety of clinics and hospital settings  Primary healthcare	130	Validity (correlation) Between NDI-J and GRoC  Validity (correlation)	-7 (a very great deal worse) to zero (about the same) to +7 (a very great deal better)  PGIC 7-points much better, better, slightly better, unchanged, slightly worse, worse and much worse  GRoC 15-points -7 (a very great deal worse) to -1	Over 8 weeks  Over 8 weeks  Within 2 months to 1 week for test-reger
ABES). (ABES).	23 24 25 Tuttle et al. Pa 26 2006) pa 27 w 28 29 30 Young et Pa	ain for more than 2 yeeks	Private physiotherap y clinics		Between NDI and GPE Between PSFS and GPE Between VAS and GPE Between ROM and GPE	worse at all) and from 7 (a very great deal better) to 1 (almost the same, hardly any better at all)  GPE 11-points  -5 is vastly worse and +5 is completely recovered  GRoC 15-points	6 weeks related to the state of
r	31. (2009) w 31 pa		therapy			0 ("about the same") to +7 ("a very great deal better")	ta mining,

1 Study	Item Evaluation Criteria*											Op		
Study 3	1	2	3	4	5	6	7	8	9	10	11	12	Total (%)	Quality Summary =
4Bjorklund et al (2017)	2	2	2	2	2	1	2	2	2	2	2	2	96	Excellent g
Cleland et al. (2008)	2	2	2	2	1	2	2	2	2	2	2	2	96	Excellent 5
7Trouli et al. (2008)	2	2	2	2	1	2	2	2	2	2	2	2	96	Excellent
<sup>8</sup> Tuttle et al. (2006)	2	2	2	2	1	2	2	2	2	2	2	2	96	Excellent &
1Kgamper et al. (2010)	2	2	2	2	1	2	2	2	2	2	2	2	96	Excellent P 2
12 12 12 12	2	2	2	2	1	2	2	2	1	2	2	2	92	Excellent 6 36/
13 531 <b>TABLE 2.</b>	Summa	ry of	Psyc	home	etric P	roperti	es Re	porte	ed in S	Studie	s and	COSM	IN Risk of Bias	s (RoB)

**TABLE 2.** Summary of Psychometric Properties Reported in Studies and COSMIN Risk of Bias (RoB) and Quality studies

Study	<b>Psychometric</b>	COSMIN	COSMIN	<b>Quality of</b>
	<b>Properties Reported</b>	RoB	Rating*§	Studies**
			(Criteria)	(QACMRR)
Bjorklund et al (2017)	Validity (correlation)	Very Good	?	Excellent
Cleland et al (2006)	Validity (correlation)	Very Good	+	Excellent
Cleland et al. (2008)	Validity (correlation)	Very Good	-	Excellent
Cook et al (2014)	Sensitivity Specificity	Very Good Very Good	+	Excellent
Farooq et al. (2017)	Validity (correlation)	Very Good	+	Excellent
Guzy et al. (2013)	Validity (correlation)	Very Good	?	Very good
Jorritsma et al. (2012)	Validity (correlation)	Very Good	?	Excellent
Kamper et al. (2010)	Test-retest reliability	Very Good	+	Excellent
Monticone et al. (2017)	Validity (correlation)	Very Good	?	Excellent
Monticone et al. (2015)	Validity (correlation	Very Good	?	Excellent
Ngo et al. (2010)	Test-retest reliability	Very Good	4	Excellent
Shaheen et al. (2015)	Validity (correlation)	Very Good	?	Excellent
Takeshita et al. (2014)	Validity (correlation)	Very Good	?	Very good
Trouli et al. (2008)	Validity (correlation)	Very Good	+	Excellent
Tuttle et al. (2006)	Validity (correlation)	Very Good	?	Excellent
Young et al. (2009)	Validity (correlation)	Very Good	?	Excellent

COSMIN, Consensus-based Standards for the Selection of health Measurement Instruments, Criteria for good measurement properties: '+' sufficient; '-' insufficient; '?' indeterminate. §§ The grading for the quality of the evidence based on the modified GRADE approach is not applicable. \*\*Quality Appraisal for Clinical Measurement Research Reports Evaluation Form (QACMRR).

2														
<sup>3</sup> Jorritsma et al. (2012)	2	2	2	2	1	1	2	2	2	2	2	2	92	Excellent
5Cleland et al (2006)	2	2	2	2	1	1	2	2	2	2	2	2	92	Excellent
Monticone et al. (2017)	2	2	2	2	1	1	2	2	2	2	2	2	92	Excellent
8Monticone et al. (2015)	2	2	2	2	1	1	2	2	2	2	2	2	92	Excellent
<sup>9</sup> Ngo et al. (2010)	2	2	2	2	2	2	2	2	1	2	1	2	92	Excellent
15 haheen et al. (2013)	2	2	2	2	2	2	2	2	2	2	1	1	92	Excellent of
12arooq et al. (2017)	2	2	1	2	2	2	2	2	1	2	2	2	92	Excellent 🛱
14 oung et al. (2009)	2	2	2	2	1	1	2	2	2	2	2	2	92	Excellent §
16uzy et al. (2013)	2	2	1	2	1	2	2	2	1	2	2	2	88	Very good
16 17akeshita et al. (2014) 17 537	2	2	1	1	1	2	2	2	2	2	2	2	88	Very good 2.
18		, and the second				•		, and the second	, and the second			•		τ,

\*Item Evaluation Criteria: 1. Thorough literature review to define the research question; 2. Specific inclusion/exclusion criteria; 3. Specific hypotheses; 4. Appropriate scope of psychometric properties; 5. Sample size; 6. Follow-up; 7. The authors referenced specific procedures for administration, scoring, and interpretation of procedures; 8. Measurement techniques were standardized; 9. Data were presented for each hypothesis; 10. Appropriate statistics-point estimates; 11. Appropriate statistical error estimates; 12. Valid conclusions and clinical recommendations.

Total score = (sum of subtotals  $\div$  24 × 100). If for a specific paper an item is deemed NA (Not Applicable), then, Total score = (sum of subtotals  $\div$  (2 × number of Applicable items) × 100).

NA – Not Applicable. The subsections no. 6, asks for percentage of retention/follow up. This subsection only applies to reliability test-retest studies

Quality Summary: Poor (0%-30%), Fair (31%-50%), Good (51%-70%), Very good (71%-90%), Excellent (>90%):

# **TABLE 5. SUMMARY OF VALIDITY PROPERTIES OF GRoC SCALES**

Study	Type of Validity	Validity Estimates	COSMIN	Quality
<u> </u>	Spearman's correlation		Very Good	Excellent
joakBredetal \$39MM	between the change scores  IARY OF REPIABILITY GROC and ProFitMap-neck	Y PROPERTIES THE GROGO LES  k  rho = 0.59, (p<0.05)		
Tyne	GRoC and NDI of Reliability ns (Pearson r)	Reliability Estimates	Very Good	COSMIN Excellent
<u>1 ypc</u>	between change scores	Renaulity Estimates	very Good	
leland et al. (2006)	PSFS and GRoC	Intra-class correlating coefficients (ICC) 0.99 (0.99 0.09) – baseline		Very Good
leland et al. (2008)	Correlations (Pearson r) between change scores	0.96 (0.95 - 0.97) – at six weeks $0.92 (0.89 - 0.94)$ at twelve months.	Very Good	Excellent
	NDI and GRoC NRS and GRoC	Intra-class cortel at 58 coefficients (ICC) 0.70 (0.60–0.80) – at six weeks (General recovery)		Very Good
ook et al. (2014)	Receiver operator characteristics (ROC) Within-session change Between-session change Between session change of Pain and GROC Sensitivity	0.80 (0.72–0.87) – at six weeks (neck pain questions)  Weighted kappa  0.70 (0.42=0.98) – at six weeks (neck pain questions)  0.80 (0.51–1.0) – at six weeks (neck pain questions)	Very Good	Excellent  Very Good  Excellent  Excellent
arooq et al. (2017)		0.81 (0.64-0.99) when defined as "completely better" or "n r 迪加姆ved		Excellent
uzy et al. (2013)		Dichotomized response epicintered effange in heck pain question Four-week internali(s = -0.56)	ons Good	Very good
orritsma et al. (2012)	Test-retest Correlation between change scores of NPAD and GPE	0.46 (0.20-0.74) when "recovered" was defined as "very m	uehery Good	Very good  Excellent
Ionticone et al. (2017)	the NeckPix©	the kappa coefficient was FloppaHiepants who remembered previous answers to the general recovery question; 0.88 (0.64–	1) for	Excellent
Ionticone et al. (2015)	Correlation (Spearman <sup>th</sup> between change scores NDI-I and GPE	ose who did not remember and 0.50 (0.02–0.98) for participal  were not specifically beginning the second of the se		Excellent
haheen et al. (2013)	Correlations (Spearman)	evious answers to the change in neck pain question; 0.74 (0.41 those who did not remember and \$2.40 (0.22-1) for participants	-t) for Good	Excellent
akeshita et al. (2014)	Correlations NDI and PGIC	were not asked the question.  rho = 0.47, p<0.001	Very Good	Very good
1: . 1 (2000)	NDI-J and PGIC Correlation (Spearman's)	rho = 0.59, p<0.001	Very Good	Excellent
rouli et al. (2008)  549	GROC vs Gr-NDI	rho = 0.30, p=0.02		
				27

Page 28 of 40

	Correlations (Spearman's) NDI vs GPE (post 1, minus pre-1)		Very Good	Excellent
	NDI vs GPE (post 2, minus pre-1)			
	NDI vs GPE (post 2, minus pre-2)	1 0.17		
	•	$     \text{rho} = 0.17 \\     \text{rho} = 0.01 $		
	PSFS vs GPE (post 1,	rho = 0.01		
	minus pre-1) PSFS vs GPE (post 2,			
	minus pre-1)	rho = 0.06		
	PSFS vs GPE (post 2,	rho = 0.03 rho = 0.03		
uttle et al. (2006)	minus pre-2)	1110 – 0.03		
(= * * * )	Pain Intensity (post 1,	rho = 0.00		
	minus pre-1)	rho = 0.05		
	Pain Intensity (post 2,	rho = 0.01		
	minus pre-1)	rho = 0.03		
	Pain Intensity (post 2,	rho = 0.01		
	minus pre-2)	rho = 0.00		
	Total ROM (post 1, minus			
	pre-1)			
	Total ROM (post 2, minus pre-1)			
	Total ROM (post 2, minus			
	pre-2)			
oung et al. (2009)	Correlations (Pearson's) between change scores NDI and GRoC	r=0.52 (p<0.01)	Very Good	Excellent
550				
551				
552				
552				
553				
554				
555				
<b></b>				
556				
557				
558				
559				
560				
561				
562				
				28

Box 1. Questions of Global Rating of Change (GROC) scales

Author	GROC item- scale	Patients with neck disorders were asked:
Bjorklund et		"Compared to before the treatment of the study started, my overall
al. (2017)	GROC 7-points	status is now"
		"Compared to before the treatment of the study started, my status
		regarding my neck-shoulder problem is now''
Evans et al		"Overall, how much has your neck pain changed since you started
(2014)	GPE 9-points	treatment in the study?''
Kamper et al.		"With respect to your whiplash injury how would you describe yourself
(2010)	GPE 11-points	now compared to immediately after your accident"
Monticone et		"Overall, how much did the treatment you received help your fear of
al. (2017)	GPE 5-points	movement due to current neck pain?
		"Overall, how much did the treatment you delivered help your
		subject's fear of movement due to her/his current neck pain?"
Monticone et		"Overall, how much did the treatment you received help your neck
al. (2015)	GPE 5-points	problem?''
Ngo et al.		"How well do you feel you are recovering from your injuries?"
(2010)	GPE 7-points	"How do you feel your neck pain has changed since the injury?"

Figure 1. Flow diagram of included studies

**Figure 2**. Meta-analysis of Pearson's correlation coefficients between neck disability change scores and GROC scores in patients with neck disorders based on 5 very good to excellent quality studies.

**Figure 3**. Meta-analysis of Spearman's correlation coefficients between neck disability change scores and GROC scores in patients with neck disorders based on 6 very good to excellent quality studies.

**Figure 4**. Random effects univariate meta-regression between age and the Fisher's Z estimates. Each circle represents a study and the size of the circle indicates the influence of that study on the model. The regression prediction is illustrated by the straight line and the curved lines represent the 95% confidence intervals. Age explained 68% of the variance in the model ( $R^2$ =0.68).

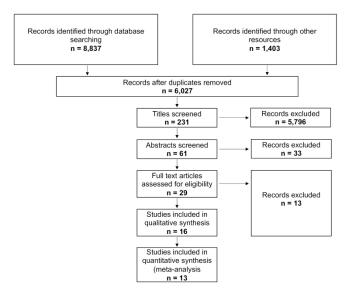


Figure 1. Flow diagram of included studies 338x190mm (300 x 300 DPI)

Study name	Statistics for each study					Correlation and 95% CI				
	Correlation	Lower limit	Upper limit	Z-Value	p-Value					
Cleland et al 2008	0.580	0.457	0.681	7.669	0.000				+=-	
Farooq et al 2017	0.500	0.342	0.631	5.575	0.000				-	
Guzy et al 2013	0.560	0.402	0.686	6.004	0.000					
Jorritsma et al 2012	0.490	0.297	0.644	4.580	0.000				-	
Young et al 2009	0.520	0.352	0.656	5.407	0.000				-	
	0.536	0.470	0.596	13.225	0.000				<b>+</b>	
						-1.00	-0.50	0.00	0.50	1.00

Figure 2. Meta-analysis of Pearson's correlation coefficients between neck disability change scores and GROC scores in patients with neck disorders based on 5 very good to excellent quality studies.

215x279mm (300 x 300 DPI)



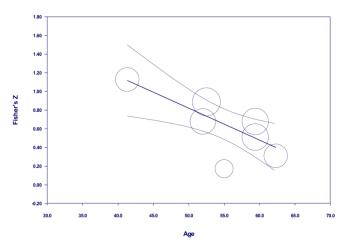


Figure 4. Random effects univariate meta-regression between age and the Fisher's Z estimates. Each circle represents a study and the size of the circle indicates the influence of that study on the model. The regression prediction is illustrated by the straight line and the curved lines represent the 95% confidence intervals. Age explained 68% of the variance in the model (R2=0.68).

215x279mm (300 x 300 DPI)

Protected by copyright, including for uses related to text

data mining, Al training, and similar technologies

60

## **Appendix 1: Search terms**

#### **MEDLINE-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. cervicalgia.mp.
- 38. brachialgia.mp.
- 39. brachial neuritis.mp.
- 40. brachial neuralgia.mp.
- 41. neck pain.mp.

ining, Al training, and similar technologies

Protected by copyright, including for uses related

```
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
```

85. exp \*Uterus/

86. 83 or 84 or 85

87. 82 not 86

42. neck injur\*.mp. 43. brachial plexus neuropath\*.mp. 44. brachial plexus neuritis.mp. 45. thoracic outlet syndrome/ or cervical rib syndrome/ 46. Torticollis/ 47. exp brachial plexus neuropathies/ or exp brachial plexus neuritis/ 48. cervico brachial neuralgia.ti,ab. 49. cervicobrachial neuralgia.ti,ab. 50. (monoradicul\* or monoradicl\*).tw. 51. or/30-50 52. exp headache/ and cervic\*.tw. 53. exp genital diseases, female/ mital disea.
r/53-54
i2 not 55
51 or 56
neck/
neck muscles/
exp cervical plexus/
exp cervical vertebrae/
2. atlanto-axial joint/
i3. atlanto-occipital joint/
i4. Cervical Atlas/
65. spinal nerve roots/
66. exp brachial plexus/
67. (odontoid\* or cervical or occip\* or atlant\*).tw.
68. axis/ or odontoid process/
9 Thoracic Vertebrae/
ival vertebrae.mp. 54. genital disease\*.mp. 77. (thoracic adj3 spine).mp. 78. (thoracic adj3 outlet).mp. 79. trapezius.mp. 80. cervical.mp. 81. cervico\*.mp. 82. 80 or 81 83. exp genital diseases, female/ 84. genital disease\*.mp.

BMJ Open: first published as 10.1136/bmjopen-2019-033909 on 25 November 2019. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de Enseignement Superieur (ABES) .

ining, Al training, and similar technologies

Protected by copyright, including for uses related

```
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 74 or 75 or 76 or 77 or 78 or 79 or 87
```

- 89. exp pain/
- 90. exp injuries/
- 91. pain.mp.
- 92. ache.mp.
- 93. sore.mp.
- 94. stiff.mp.
- 95. discomfort.mp.
- 96. injur\*.mp.
- 97. neuropath\*.mp.
- 98. or/89-97
- 99. 88 and 98
- 100. Radiculopathy/
- 101. exp temporomandibular joint disorders/ or exp temporomandibular joint dysfunction syndrome/
- 102. myofascial pain syndromes/
- 103. exp "Sprains and Strains"/
- 104. exp Spinal Osteophytosis/
- 105. exp Neuritis/
- 106. Polyradiculopathy/
- 107. exp Arthritis/
- 108. Fibromyalgia/
- 109. spondylitis/ or discitis/
- 110. spondylosis/ or spondylolysis/ or spondylolisthesis/
- 111. radiculopathy.mp.
- 112. radiculitis.mp.
- 113. temporomandibular.mp.
- 114. myofascial pain syndrome\*.mp.
- 115. thoracic outlet syndrome\*.mp.
- 116. spinal osteophytosis.mp.
- 117. neuritis.mp.
- 118. spondylosis.mp.
- 119. spondylitis.mp.
- 120. spondylolisthesis.mp.
- 121. or/100-120
- 122. 88 and 121
- 123. exp neck/
- 124. exp cervical vertebrae/
- 125. Thoracic Vertebrae/
- 126. neck.mp.
- 127. (thoracic adj3 vertebrae).mp.
- 128. cervical.mp.
- 129. cervico\*.mp.
- 130. 128 or 129
- 131. exp genital diseases, female/

```
2
3
             132. genital disease*.mp.
4
             133. exp *Uterus/
5
             134. or/131-133
6
             135. 130 not 134
7
             136. (thoracic adj3 spine).mp.
8
9
             137. cervical spine.mp.
10
             138. 123 or 124 or 125 or 126 or 127 or 135 or 136 or 137
11
             139. Intervertebral Disk/
12
             140. (disc or discs).mp.
13
             141. (disk or disks).mp.
14
             142. 139 or 140 or 141
15
             143. 138 and 142
16
17
             144. herniat*.mp.
18
             145. slipped.mp.
19
             146. prolapse*.mp.
20
             147. displace*.mp.
21
             148. degenerat*.mp.
22
             149. (bulge or bulged or bulging).mp.
23
             150. 144 or 145 or 146 or 147 or 148 or 149
24
25
             151. 143 and 150
26
             152. intervertebral disk degeneration/ or intervertebral disk displacement/
27
             153. intervertebral disk displacement.mp.
28
             154. intervertebral disc displacement.mp.
29
             155. intervertebral disk degeneration.mp.
30
             156. intervertebral disc degeneration.mp.
31
32
             157. 152 or 153 or 154 or 155 or 156
33
             158, 138 and 157
34
             159. 57 or 99 or 122 or 151 or 158
35
             160. animals/ not (animals/ and humans/)
36
             161. 159 not 160
37
             162. exp *neoplasms/
38
             163. exp *wounds, penetrating/
39
40
             164. 162 or 163
41
             165. 161 not 164
42
             166. 29 and 165
43
             167. guidelines as topic/
44
             168. practice guidelines as topic/
45
             169. guideline.pt.
46
47
             170. practice guideline.pt.
48
             171. (guideline? or guidance or recommendations).ti.
49
             172. consensus.ti.
50
             173. or/167-172
51
             174. meta-analysis/
52
             175. exp meta-analysis as topic/
53
             176. (meta analy* or metaanaly* or met analy* or metanaly*).tw.
54
55
             177. review literature as topic/
```

- 179. (integrative research or integrative review\* or intergrative overview\*).tw.
- 180. (quantitative adj3 (research or review\* or overview\*)).tw.
- 181. (research integration or research overview\*).tw.
- 182. (systematic\* adj3 (review\* or overview\*)).tw.
- 183. (methodologic\* adj3 (review\* or overview\*)).tw.
- 184. exp technology assessment biomedical/
- 185. (hta or thas or technology assessment\*).tw.
- 186. ((hand adj2 search\*) or (manual\* adj search\*)).tw.
- 187. ((electronic adj database\*) or (bibliographic\* adj database\*)).tw.
- 188. ((data adj2 abstract\*) or (data adj2 extract\*)).tw.
- 189. (analys\* adj3 (pool or pooled or pooling)).tw.
- 190. mantel haenszel.tw.
- 191. (cohrane or pubmed or pub med or medline or embase or psycinfo or psyclit or psychinfo or psychlit or cinahl or science citation indes).ab.
- 192. or/174-191
- 193. 173 or 192
- 194. 166 and 193



# PRISMA 2009 Checklist

Page 39 of 40		BMJ Open ct by c c	
PRISMA 2	009	by соругіднь, ореп-2019-03	
Section/topic	#	Checklist item 3909 on	Reported on page #
7 TITLE		g 75 fo z	
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT		es re	
11 12 Structured summary 13 14	2	Provide a structured summary including, as applicable: background; objectives; data sources study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; in the conclusions and implications of key findings; systematic review registration number.	2
INTRODUCTION		x upe	
Rationale	3	Describe the rationale for the review in the context of what is already known.	3-5
8 Objectives	4	Provide an explicit statement of questions being addressed with reference to participants hererventions, comparisons, outcomes, and study design (PICOS).	4-5
METHODS		ing,	
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.	5
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.	5
27 Information sources 28	7	Describe all information sources (e.g., databases with dates of coverage, contact with story authors to identify additional studies) in the search and date last searched.	6
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	Appendix1
32 Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic we and, if applicable, included in the meta-analysis).	6
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in diplicate) and any processes for obtaining and confirming data from investigators.	6-7
36 37 Data items 38	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	6-7
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.	6-7
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	8-9
43 Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including necessarily assures of consistency (e.g., I²) for each meta-analysis. For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml	8-9



## **PRISMA 2009 Checklist**

		BMJ Open Cted by C	Page 40 of 40
PRISMA	<b>A</b> 2009	· · · · · · · · · · · · · · · · · · ·	
2 3		Checklist  Page 1 of 2  Page 1 of 2	
Section/topic	#	Checklist item  Checklist item	Reported on page #
Risk of bias across stud	dies 15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	8-9
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-realization), if done, indicating which were pre-specified.	8=9
RESULTS		men t to	
14 Study selection 15	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with a flow diagram.	9
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, Pichos, follow-up period) and provide the citations.	9-10
Risk of bias within studi	ies 19	Present data on risk of bias of each study and, if available, any outcome level assessme	10
Results of individual stu	ıdies 20	For all outcomes considered (benefits or harms), present, for each study: (a) simple suntained data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot	10-12
23 Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measure.	13
Risk of bias across stud	dies 22	Present results of any assessment of risk of bias across studies (see Item 15).	10
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-	13
DISCUSSION	<u> </u>	imi imi	
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).	14-15
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., ingomplete retrieval of identified research, reporting bias).	16
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	14-15
<sup>36</sup> FUNDING	<u>'</u>	A A G	
37 38 Funding 39	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
40 41 <i>From:</i> Moher D, Liberati A, 42 doi:10.1371/journal.pmed100	Tetzlaff J, Altma	an DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med For more information, visit: www.prisma-statement.org.	6(7): e1000097.
43 44 45 46		Page 2 of 2 For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml	
47			

# **BMJ** Open

## Psychometric Properties of the Global Rating of Change Scales in Patients with Neck Disorders: A Systematic Review with Meta-Analysis and Meta-Regression

Journal:	BMJ Open
Manuscript ID	bmjopen-2019-033909.R1
Article Type:	Original research
Date Submitted by the Author:	11-Oct-2019
Complete List of Authors:	Bobos, Pavlos; Western University, Health and Rehabilitation Sciences; University of Toronto, Institute of Health Policy Management and Evaluation MacDermid, Joy; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Nazari, Goris; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Furtado, Rochelle; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Group, CATWAD; Michele Sterling, Anne Söderlund, Michele Curatolo, James M Elliott, David M Walton, Helge Kasch, Linda Carroll, Hans Westergren, Gwendolen Jull, Eva-Maj Malmström, Luke B Connelly, Joy C MacDermid, Mandy Nielsen, Pierre Côté, Tonny Elmose Andersen, Trudy Rebbeck, Annick Maujean, Sarah Robins, Kenneth Chen, Julia Treleaven
<b>Primary Subject Heading</b> :	Rehabilitation medicine
Secondary Subject Heading:	Rehabilitation medicine
Keywords:	neck pain, global assessment, psychometric properties, systematic review

SCHOLARONE™ Manuscripts

- 1 Psychometric Properties of the Global Rating of Change Scales in Patients with Neck
- 2 Disorders: A Systematic Review with Meta-Analysis and Meta-Regression
- 3 Pavlos Bobos<sup>1</sup>, Joy C MacDermid<sup>2</sup>, Goris Nazari<sup>3</sup>, Rochelle Furtado<sup>4</sup> and CATWAD co-authors<sup>5</sup>
- <sup>1</sup>Pavlos Bobos PT, PhD(c), (corresponding author) Doctoral Candidate, Western's Bone and Joint
- 6 Institute, Department of Health and Rehabilitation Sciences, Western University, Elborn College,
- 7 1201 Western Road, N6G 1H1, London, Ontario, Dalla Lana School of Public Health, Institute of
- 8 Health Policy Management and Evaluation, Department of Clinical Epidemiology and Health Care
- 9 Research, University of Toronto, Canada, (pbobos@uwo.ca), tel: +1 519 661 2111 x88912
- <sup>2</sup>Joy C MacDermid BScPT, PhD, Professor, Physical Therapy and Surgery, Western University,
- London, ON and Co-director Clinical Research Lab, Hand and Upper Limb Centre, St. Joseph's
- 12 Health Centre, London, Ontario; Professor Rehabilitation Science McMaster University,
- Hamilton, ON, Canada (<u>imacderm@uwo.ca</u>)
- <sup>3</sup>Goris Nazari PT, PhD(c) Doctoral Candidate, Western's Bone and Joint Institute, School of
- 15 Physical Therapy, Department of Health and Rehabilitation Sciences, Western University,
- 16 London, Ontario, Canada, (gnazari@uwo.ca)
- <sup>4</sup>Rochelle Furtado MSc Western's Bone and Joint Institute, School of Physical Therapy,
- Department of Health and Rehabilitation Sciences, Western University, London, Ontario, Canada,
- 19 (<u>rfurtad5@uwo.ca</u>)
- <sup>5</sup>CATWAD: Michele Sterling <u>m.sterling@uq.edu.au</u>, Anne Söderlund <u>anne.soderlund@mdh.se</u>,
- 21 Michele Curatolo, curatolo@uw.edu, James M Elliott j-elliott@northwestern.edu, David Walton
- dwalton5@uwo.ca, Helge Kasch helgkasc@rm.dk, Linda Carroll linda.carroll@ualberta.ca,
- Hans Westergren <u>Hans.Westergren@skane.se</u>, Gwendolen Jull <u>g.jull@uq.edu.au</u>, Eva-Maj
- 24 Malmström eva-maj.malmstrom@med.lu.se, Luke B Connelly l.connelly@uq.edu.au, Joy C
- 25 MacDermid <u>imacderm@uwo.ca</u>, Mandy Nielsen <u>mandy.nielsen@griffith.edu.au</u>, Pierre Côté
- 26 <u>pierre.cote@uoit.ca</u>, Tonny Elmose Andersen <u>tandersen@health.sdu.dk</u>, Trudy Rebbeck
- 27 <u>trudy.rebbeck@sydney.edu.au</u>, Annick Maujean <u>a.maujean@uq.edu.au</u>, Sarah Robins
- 28 s.robins1@uq.edu.au, Kenneth Chen k.chen8@uq.edu.au, Julia Treleaven j.treleaven@uq.edu.au
- **Kewords:** neck pain, global assessment, psychometric properties, systematic review
- **30 Word count: 3908**

- Objective: The purpose of this systematic review was to critically appraise and synthesize the psychometric properties of Global Rating of Change (GROC) scales for assessment of patients
- with neck pain.
- **Design:** Systematic review
- 36 Data sources: A search was performed in 4 databases (MEDLINE, EMBASE, CINAHL,
- 37 SCOPUS) until February 2019.
- **Data extraction and synthesis:** Eligible articles were appraised using Consensus-based Standards
- 39 for the selection of health Measurement Instruments (COSMIN) checklist and the Quality
- 40 Appraisal for Clinical Measurement Research Reports Evaluation Form.
- **Results:** The search obtained 16 eligible studies and included in total 1533 patients with neck pain.
- 42 Test-retest reliability of Global Perceived Effect (GPE) was very high (Intra-class correlation
- coefficient (ICC) = 0.80 to 0.92) for patients with whiplash. Pooled data of Pearson's r indicated
- that GROC scores were moderately correlated with neck disability change scores (0.53, 95% CI:
- 45 0.47 to 0.59). Pooled data of Spearman's correlations indicated that GROC scores were moderately
- 46 correlated with neck disability change scores (0.56, 95% CI: 0.41 to 0.68).
- **Conclusions:** This study found excellent quality evidence of very good to excellent test-retest
- 48 reliability of GPE for patients with Whiplash Associated Disorders. Evidence from very good-to-
- 49 excellent quality studies found that GROC scores are moderately correlated to an external criterion
- patient-reported outcome (PROM) measure evaluated pre-post treatment in patients with neck
- pain. No studies were found that addressed the optimal form of GROC scales for patients with
- neck disorders or compared the GROC to other options for single-item global assessment.
- Prospero registration number: CRD 42018117874

## Strengths and limitations of this study

- We rated the quality of individual studies and the overall risk of bias using two standardized approaches
- Our focus on neck pain increased the specificity of results but are not necessarily
   applicable to other musculoskeletal conditions
- Conceptual concerns about global ratings of change being affected by recall bias are not adequately addressed by psychometric evidence
- No studies addressing the optimal form of global rating were found.

#### Introduction

Neck pain is the 4<sup>th</sup> leading cause of disability and approximately half of adult the population with neck pain will experience a clinically important episode once in their lifetime. [1–3] The annual prevalence of neck pain it is estimated between 15% and 50%, with females having a higher prevalence rate than males. [2,3] Neck pain has been associated with many other comorbidities such as headaches, dizziness, anxiety, depression, back pain and arthralgias.[3–6] Several different methods for classifying neck pain have been described, using indicators such as duration (acute, sub-acute or chronic), degree of interference (low, moderate, severe) or most likely structure at fault (e.g. neuropathy vs. mechanical). [7]

As part of a patient-centric approach to care, clinicians will commonly evaluate response to intervention by asking the patient directly whether they feel better, worse, or the same since the prior encounter. While direct questioning can provide a qualitative indicator of change in status, many best practice guidelines endorse use of some form of quantified patient-reported outcome (PRO) as an adjunct to oral self-report. PROs are available to quantify several different constructs

in people with neck pain, including pain severity, disability and neck function. [8] Any PRO intended to provide an estimate of change over time should be responsive to subtle shifts in the patient's condition. To facilitate interpretation of change scores, a common property of many such tools is the minimum clinically important difference (MCID), which is a change threshold that corresponds to the minimum shift in scale values that most patients would indicate corresponds to an important change in their overall condition. A well-recognized approach to establishing an MCID for a PRO is to compare the magnitude of change against an anchor, most commonly a Global Rating of Change (GROC) scale. These scales allow patients or study participants to indicate whether their condition has gotten worse, better, or stayed the same and to quantify the magnitude of that change. As they have been adopted as a sort of 'standard' against which change in other tools is compared, the GROC can also be used on its own as an omnibus generic indicator of change. [8]

Despite being accepted as a standard measure, there is considerable variation in how the GROC has been constructed and implemented in research in neck pain. Some are 15 points, some 11 points, and others are 7 points. The common structure across these is the use of a middle '0' score corresponding to 'no change', with negative values indicating magnitudes of worsening while positive values indicate improvement.[9] Variations of the GROC (in name or structure) include the "Global Perceived Effect", "Patient Global Impression of Change", "Transition Ratings", and "Global Scale". [9]

A well-established component of health outcomes is having a tool with strong psychometric properties of validity, reliability and responsiveness to be able to monitor change. While recent research [8] has examined the psychometric properties of the most commonly reported PROs for neck disorders, to date there has been no systematic review to summarize the

1 2		
3	102	measurement properties of GROC scales themselves in patients with neck disorders. Therefore,
5 6	103	this systematic review aims to critically appraise and synthesize the psychometric properties of the
7 8 9	104	GROC scales in patients with neck disorders.
9 10 11	105	
12 13	106	METHODS
14 15	107	Patient and Public Involvement
16 17 18	108	There was no patient or public involvement in the design or planning of this study.
19 20	109	
21 22	110	Study Design and Protocol Registration
23 24	111	We conducted a systematic review to evaluate the psychometric properties of GROC scales in
25 26 27	112	patients with neck disorders. The protocol was registered in PROSPERO register database with
28 29	113	registration number: CRD 42018117874
30 31	114	
32 33 34	115	Eligibility Criteria
35 36	116	We included studies in this systematic review if the following criteria were met [10–12]:
37 38	117	Design: psychometric testing, randomized/ cohort studies
39 40 41	118	• Participants: > 50% of the study's patient population with neck conditions/disorders,
42 43	119	• Intervention/Comparison: studies that reported on the psychometric properties (reliability,
44 45	120	validity, responsiveness) of GROC, Global Perceived Effect (GPE) and Patient Global
46 47	121	Impression of Change (PGIC),
48 49 50	122	Outcomes: GROC, GPE and PGIC
51 52	123	Articles were written in English language only
53 54		
55 56		
57 58		5

Studies with no data on the GROC scales' psychometric properties, and conference abstract/posters were excluded from this systematic review.

## Information Sources

To identify studies on the psychometric properties (reliability, validity, responsiveness) of the GROC, GPE and PGIC we searched the Medline, EMBASE, Scopus and CINAHL databases from inception till February 2019, using a combination of keywords. Furthermore, we identified additional studies by examining the reference list of each of the selected studies. The full list with keyword strategy is presented in **APPENDIX 1**.

## Study Selection

Two investigators (PB and GN) performed the systematic electronic searches independently in each database. The same investigators then proceeded to identify and remove the duplicate studies. In the next stage, we performed the independent screening of the titles and abstracts and any full-text article marked as include or uncertain were obtained. In the final stage, the same two independent authors performed the full text reviews independently to assess final article eligibility. In case of disagreement, a third reviewer; the most experienced member (JM), facilitated a consensus through discussion.

#### Data Extraction

The fourth author (RF) performed the data extractions. The extracted data were then cross-checked by another author (PB). Data extraction included the author, year, study population/condition, setting, sample size, age, properties evaluated, retest-interval, and the intervention protocol (if used

to assess responsiveness parameters). [13,14] For reliability estimates, Standard Error of Measurement (SEM), Intra-class Correlation Coefficient (ICC), Minimal Detectable Change (MDC) and 95% confidence intervals were extracted. [13,14] The ICC interpretation of ICC < 0.40 indicating poor, 0.40 ≤ ICC < 0.75 indicating fair-to-good and ICC ≥ 0.75 indicating excellent reliability were used as a common benchmark.[15] For validity estimates, correlation coefficient (Pearson's/Spearman) and the 95% confidence intervals were extracted. [13,14] Evan's guidelines to interpret the strength of the correlation was used which included: 0.00–0.19 "very weak", 0.20– 0.39 "weak", 0.40–0.59 "moderate", 0.60–0.79 "strong", and 0.80–1.00 "very strong". [16] For responsiveness estimates, the Effect Size (ES), Standardized Response Mean (SRM), Clinically Important Difference (CID), and/or Minimal Clinically Important Difference (MCID) including the method of MCID estimation – Anchor-/Distribution-based methods, and 95% confidence intervals were extracted. [13,14] To assist clinical decision making, standard benchmark scores of trivial (< 0.20), small ( $\geq$  0.20 to < 0.50), moderate ( $\geq$  0.50 to < 0.80) or large ( $\geq$  0.80), as proposed by Cohen, were used. [17] When insufficient data were presented, PB contacted the authors by email and requested further data.

Consensus-based Standards for the selection of health Measurement Instruments (COSMIN)

Consensus-based Standards for the selection of health Measurement Instruments (COSMIN)

assesses the risk of bias for the psychometric properties reported on a property-by-property basis.

A score for the risk of bias in estimates of psychometric properties was assessed by two authors (PB) and (RF) using the new (COSMIN) checklist.[18] If disagreement was present a third person (JM) assist in resolving the discrepancy. Each study was assessed by COSMIN on the 4-point scale

as "very good", "adequate", "doubtful" or "inadequate" for each of the checklist criteria for

relevant measurement properties (e.g. reliability, responsiveness, etc.). According to COSMIN, when determining the overall score for each measurement property, the worst score counts method was used wherein the lowest score for the checklist criteria of the relevant property was taken as the overall score. [19] We then assessed the result of individual studies on a measurement property against the updated criteria for good measurement properties. This involved the evaluation of results of included studies as either sufficient (+), insufficient (-), or indeterminate (?). [18]

Quality Appraisal for Clinical Measurement Research Reports Evaluation Form

A summary score for the overall quality of individual studies was appraised independently by the authors (PB) and (RF) using a structured clinical measurement specific appraisal tool. [13,14] In case of disagreement a third person was consulted (JM) to resolve the conflict. The evaluation criteria of this tool included twelve items: 1) Thorough literature review to define the research question; 2) Specific inclusion/exclusion criteria; 3) Specific hypotheses; 4) Appropriate scope of psychometric properties; 5) Sample size; 6) Follow-up; 7) The authors referenced specific procedures for administration, scoring, and interpretation of procedures; 8) Measurement techniques were standardized; 9) Data were presented for each hypothesis; 10) Appropriate statistics-point estimates; 11) Appropriate statistical error estimates; and 12) Valid conclusions and recommendations. [13,14] An article's total score – quality - was calculated by the sum of scores for each item, divided by the numbers of items and multiplied by 100%. [13,14] Overall, the quality summary of appraised articles range from (0%-30%) Poor, (31%-50%) Fair, (51%-70%) Good, (71%-90%) Very Good, and (>90%) Excellent. [13,14]

Synthesis of Results

A qualitative synthesis was conducted to report findings on test-retest reliability statistics. A meta-analysis of Pearson's and Spearman's correlation was performed in R (version 3.6.1) with metaphor package. [20] The meta-analyses were conducted using a random effect model and the correlation coefficients were converted to z values. Heterogeneity was deemed substantial if I<sup>2</sup> values were more than 50%. [21] A Meta-regression was planned to explore the sources of unexplained heterogeneity by considering the following factors: a. neck pain with or without radicular symptoms, b. acute or chronic, c. age and d. sex. Forest plots were created using means and 95% confidence intervals for correlation coefficients. We summarize the main results of the included articles based on the neck disorders, reported psychometric estimate and the study quality ratings. 

#### **RESULTS**

Study Selection

Our search yielded 8,837 articles. After removal of duplicates, 6,027 studies remained and were screened using their title and abstract; leaving 29 articles selected for full-text review. Of these, 16 studies were considered eligible. [22,23,24–31,32–37] The flow of the study selection process is presented in Figure 1.

O. C.

## Study Characteristics

The 16 eligible studies were conducted between 2006 and 2017 and included 1533 participants with neck pain/disorders (mean of 96 participants per study). [22,23,24–31,32,34–37,] Study size ranged from 29 to 200 participants. A summary description of all the studies included is displayed in **Table 1.** Concurrent validity was evaluated in 14 studies by comparing the difference of pain

intensity, disability and function scores with the score of GROC scales. Two studies [26,31] examined the test-retest reliability of a 7-point and an 11-point GPE scale for patients with whiplash-associated disorders (WAD). One study [24] examined whether occurrences of within-and between-session changes were significantly associated with functional outcomes, pain, and self-report of recovery in patients at discharge who were treated with manual therapy for mechanical neck pain.

COSMIN Risk of Bias rating and Quality appraisal of the Included Studies

Regarding the risk of bias, all studies were rated as very good (**Table 2**). The quality of the studies ranged from 88% to 96% (**Table 3**). The most common flaws were 1) lack of/inadequate sample size calculations, 2) missing data (i.e. inadequate follow up), and 3) inconsistencies between the data presented and hypothesis stated.

## Reported GROC scales

The most commonly reported GROC scale (n=6 studies) was a 15-point scale with the most frequent anchors being "-7 (a very great deal worse) to zero (about the same) to +7 (a very great deal better)". A 7-point scale was reported in 5 studies, 11- and 5-point scales were reported in 2 studies and a 9-point scale in one study. The anchors in those scales varied greatly and are presented in Table 1. Only 6 studies [26,31–33,35,36] reported full detail regarding the specific questions asked of the patients with neck disorder when a GROC scale was administered. Those questions that were reported are presented in **Box 1**.

Two studies were included that examined test-retest reliability of GPE for patients with WAD. Kamper et al. (2010) [26] examined the [time interval] test-retest reliability of an 11-point GPE scale in 134 patients with chronic WAD and reported an Intra-class Correlation Coefficient (ICC) of 0.99 (95% CI 0.99 to 0.99) at baseline, 0.96 (0.95 to 0.97) at 6 weeks, and 0.92 (0.89 to 0.94) at 12 months (**Table 4**). Ngo et al. (2010) assessed the test-retest reliability of a 7-point scale of GPE in patients with acute WAD at 3 to 5 days. [31] The ICC and 95% confidence intervals (CI) were used to determine the test-retest reliability of the two versions of the perceived recovery questions using their original seven-item responses. Ngo et al. also computed weighted kappa coefficients and 95% CI using quadratic weights to determine whether the distribution of responses influenced the reliability as measured by the ICC. An ICC for general recovery of 0.70 (0.60 to 0.80) and an ICC for neck pain questions of 0.80 (0.72 to 0.87) were found. A weighted Kappa was also calculated (Kappa = 0.70 (0.42 to 0.98)) at six weeks for general recovery and at six weeks Kappa = 0.80 (0.51 to 1.0) for neck pain questions (**Table 4**).

## Validity Measures

We found 14 studies that examined concurrent validity measures between GROC and another PRO.[22,23,25,27–30,32,34,35,36–38] Correlations of Pearson's and Spearman's coefficients between GROC and another PRO were ranging from very weak to very strong correlations. The validity measures are presented and summarized in Table 5.

Meta-Analysis and Meta-Regression of Correlations between Disability change scores and GROC scores Five studies [23,25,34,37,38] of very good-to-excellent quality reported the Pearson correlation coefficients between neck disability change scores and the GROC scores and were pooled together. We found that GROC was positively correlated with disability change scores (r = 0.53, 95% CI: 0.47 to 0.59,  $I^2 = 0\%$ ). Six studies [27–30,32,36] of very good-to-excellent quality reported the Spearman correlation coefficients between neck disability changes scores and the GROC scores and were pooled together. We found that GROC was moderately correlated with disability change scores (rho = 0.56, 95% CI: 0.41 to 0.68,  $I^2 = 85\%$ ). The forest plots with correlation coefficients with 95% CIs are presented in Figure 2-3. Our meta-regression showed that age was found as a significant factor in influencing Fisher's Z scores ( $\beta = -0.034$ , 95% CI -0.05 to -0.01, p = 0.001). The model explained 68% of the variance ( $R^2 = 0.68$ ) (Figure 4). *Area under the curve (AUC) – Sensitivity and Specificity* Cook et al. [24] found that between-session NPRS- pain changes were associated with greater than 3-point change on the GROC at 96-hours (AUC=0.76). The pain change associated with GROC was more specific (Specificity=79.2%, range: 62.2 - 91.1) than sensitive (Sensitivity=65.6%,

DISCUSSION

range: 57.9 to 74.6). Those with a 36.7% between-sessions change in pain were also 7.3 times

more likely to report an improvement of greater than 3 points change on the GROC than those

who did not achieve a 36.7% change in pain (Table 4).

This review has synthesized the current research from 16 studies that aimed to evaluate the psychometric properties of GROC scales for patients with neck disorders, with the goal to provide evidence for clinicians and researchers concerning its use within clinical practice and research. From the 16 included studies, only 2 studies [26,31] reported test-retest reliability statistics of the 7- and 11-points item GPE scales for patients with WAD only. We were able to pool data from 12 studies regarding concurrent validity of GROC scales and neck disability change scores at one time point after the interventions. Themes influencing interpretation of the GROC were explored in a study [33] that evaluated the factors that contribute to how patients respond to a question on global perceived effect. This study found that treatment process, biomechanical performance, self-efficacy and the nature of the condition may influence the responses on global perceived effect, which is consistent with what we would expect for patients with neck pain. This suggests that change is a complex multifactorial global concept. A strength of GROC is that it is intended as a global assessment, and it can be assumed that it reflects the aspects of change important to the individual patient.

Reliability can be defined as the degree to which a measure produces consecutive results with the least amount of random error when the status of the population remains unchanged. The reliability of GPE displayed an excellent test-retest reliability of ICC>0.90 over an interval of 6 weeks and 12 months for patients with WAD. Conducting an assessment with a long test-retest interval (e.g. 12 months), can provide challenges as there is higher risk of individuals with WAD being symptomatically unstable.[9] Determining if patients are symptomatically-stable can be achieved by administering another PRO such as the Single Assessment Numeric Evaluation (SANE)[39], however, the 7- and 11- points GPE scales still demonstrated good stability properties

at long test intervals (i.e., of 6 weeks and 12 months).[26] Therefore, the measurements of the reliability parameters of the GPE may be very useful during longer test intervals in clinical trials.

The psychometric property of validity is defined as the degree to which a PRO measures what it is intended to measure. Pooled data from 11 studies overall suggest that post-treatment changes of on validated disability outcome measures were moderately (Pearson's r = 0.51, 95% CI: 0.43 to 0.58; Spearman's rho = 0.56, 95% CI: 0.41 to 0.68) correlated to change in perceived effect) (Figure 2-3). This finding suggests that GROC scores taken at one point in time were related to scores in pain and disability in patients with neck disorders, as measured by standardized measures taken at 2 points in time. We identified one study [24] that found a 36.7% change in pain for within- and between- session changes was associated with a 50% reduction in the NDI and an improvement of >3 points on a 15-points GROC scale for patients with neck pain. This quantified predictive change value may have clinical utility for use in clinical practice.

Previous studies [9,40] have indicated serious concerns about the conceptual validity of the global rating of change. The review by Kamper et al.[9] clearly showed that GROC was related to final status more than change and was least related to baseline health status. This result undermines the premise of what the global rating of change actually measures. For this reason, we conclude that the 0.50 pooled correlation across 12 studies between the GROC and other PROM change scores (e.g. NDI scores) may reflect a relationship between follow-up status and change rather than supporting the contention that GROC actually measures change. This would also explain why only 25% of the variation in GROC change scores was explained by changes scores from a PROM change score measured at 2 points in time. In all studies, participants completed the GROC scale at one time point after the intervention, and hence recall bias is a cause for concern. However, another potential factor for moderate correlations is that the PROMs that have been used

as the comparator with GROC scores may not reflect priorities that are important to patients. That is, the field has largely been driven by assumptions that the GROC is a 'gold standard' for evaluating true change in a respondent's condition or status, and that all items on the comparator PROM are of equal importance to all people with that condition. The work presented herein challenges the valorization of the GROC as a gold standard for change, and prior work has challenged the notions that all PROM items are equally important.[9,41,42] It is therefore possible that the very constructs being evaluated require greater critical discourse before authors can say, with confidence, that one scale functions well or poorly based on its associations with another scale. Since no studies compared a retrospective global assessment of the GROC to pre-post single item global PROM e.g. the SANE, we do not know the extent to which these two factors contributed to moderate correlation.

A unique aspect of this study was that it focused on global rating of change scales in a neck pain patient population. Our study appraisal suggests that future studies concerning GROC should include adequate sample sizes, maintain a rigorous follow up and report appropriate statistical error estimates, since these were often inadequate. Various critical appraisal tools exist, and the perspectives and ratings may differ across instruments. We used 2 different critical appraisal tools to evaluate quality from 2 perspectives. The COSMIN risk of bias assessments reflects the level of confidence in the conclusions and pooled estimates. The quality appraisal tool focuses on design issues in the studies and reflects gaps in research designs that should be considered in interpretation of current research and improved in future studies. Substantial heterogeneity was detected (12>50%) in pooled Spearman's correlation coefficients which is a concern when pooling data. Our univariate meta-regression analysis indicated that age across the studies explained 68% of the variance (Figure 4). Other factors such as type of neck pain (with or without radicular symptoms),

While this study included 16 studies, only 2 of these reported reliability statistics for GROC scales for patients with chronic WAD. Therefore, the applicability of our study is mostly limited to patients with chronic WAD. For validity measurements, GROC scales were mostly investigated by correlation analyses to evaluate the external responsiveness of another PRO measure over a specific time point. From our meta-analysis, we can be confident that the GROC scores were moderately correlated with neck disability change scores. However, more robust psychometric design studies to test the measurement properties of GROC scales as the primary outcome of investigation are highly needed. Future studies should aim to test to what extent the different range of items (e.g. 7-point scale vs 11-point scale), the anchors (e.g. much worse vs much better) may affect the measurement properties of GROC scales for patients with neck disorders. Also, it is important to indicate that most outcome measures are ordinal and assume that additive scores of ordinal items can be treated as interval level. This potentially could lead to scaling problems even in the face of strong psychometric properties. The main protection we have is to create new scales or retrofit existing scales based on Rasch analysis.

#### **CONCLUSIONS**

This study found excellent quality evidence of very good to excellent test-retest reliability of GPE for patients with WAD. Evidence of very good to excellent quality studies found that GROC scores are moderately correlated to an external criterion PROM measure measured pre-post treatment in patients with neck disorders. Studies addressing the optimal form of GROC scales for patients with

neck disorders or comparing the GROC to other options for single-item global assessment of change were not found.

## **Authors' contributions**

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

#### **Declarations**

## **Ethics approval and consent to participate**

Not applicable

## **Consent for publication**

390 Not applicable

### Availability of data and material

Data sharing is not applicable to this article as no datasets were generated or analyzed during the

393 current study

## **Funding Statement**

This work was supported by the Canadian Institutes of Health Research (CIHR) with funding

reference number (FRN: SCA-145102).

397	Con	npeting Interest Statement
398		e to report
399	1 (011	
400		
400	Dofo	rences
401		
	1	Murray CJL, Abraham J, Ali MK, et al. The State of US health, 1990-2010: Burden of diseases,
403		injuries, and risk factors. <i>JAMA - J Am Med Assoc</i> Published Online First: 2013.
404		doi:10.1001/jama.2013.13805
405	2	Fejer R, Kyvik KO, Hartvigsen J. The prevalence of neck pain in the world population: A
406		systematic critical review of the literature. Eur. Spine J. 2006. doi:10.1007/s00586-004-0864-4
407	3	Hogg-Johnson S, van der Velde G, Carroll LJ, et al. The Burden and Determinants of Neck Pain in
408		the General Population. Results of the Bone and Joint Decade 2000-2010 Task Force on Neck
409		Pain and Its Associated Disorders. J Manipulative Physiol Ther Published Online First: 2009.
410		doi:10.1016/j.jmpt.2008.11.010
411	4	Bobos P, Nazari G, Palimeris S, et al. The contribution of health and psychological factors in
412		patients with chronic neck pain and disability: A cross-sectional study. J Clin Diagnostic Res
413		2018; <b>12</b> :YC04–7. doi:10.7860/JCDR/2018/31284.11203
414	5	MacDermid JC, Walton DM, Bobos P, et al. A Qualitative Description of Chronic Neck Pain has
415		Implications for Outcome Assessment and Classification. Open Orthop J 2017;10:746–56.
416		doi:10.2174/1874325001610010746
417	6	Treleaven J. Sensorimotor disturbances in neck disorders affecting postural stability, head and eye
418		movement control-Part 2: Case studies. <i>Man Ther</i> 2008; <b>13</b> :266–75.
419		doi:10.1016/j.math.2007.11.002
420	7	Cohen SP. Epidemiology, diagnosis, and treatment of neck pain. Mayo Clin Proc 2015;90:284–99.
421		doi:10.1016/j.mayocp.2014.09.008

1 2			
3	422	8	Bobos P, Macdermid JC, Walton DM, et al. Patient-reported outcome measures used for neck
5 6	423		disorders: An overview of systematic reviews. J. Orthop. Sports Phys. Ther. 2018;48:775–88.
7 8	424		doi:10.2519/jospt.2018.8131
9 10	425	9	Kamper SJ, Maher CG, Mackay G. Global Rating of Change Scales: A Review of Strengths and
11 12	426		Weaknesses and Considerations for Design. <i>J Man Manip Ther</i> 2009; <b>17</b> :163–70.
13 14	427		doi:10.1002/mus.21062
15 16	428	10	Nazari G, Bobos P, MacDermid JC, et al. The Effectiveness of Instrument-Assisted Soft Tissue
17 18	429		Mobilization in Athletes, Participants Without Extremity or Spinal Conditions, and Individuals
19 20 21	430		with Upper Extremity, Lower Extremity, and Spinal Conditions: A Systematic Review. Arch Phys
22 23	431		Med Rehabil Published Online First: February 2019. doi:10.1016/j.apmr.2019.01.017
24 25	432	11	Bobos P, Nazari G, Szekeres M, et al. The effectiveness of joint-protection programs on pain,
26 27	433		hand function, and grip strength levels in patients with hand arthritis: A systematic review and
28 29	434		meta-analysis. J Hand Ther 2018; <b>32</b> :194–211. doi:10.1016/j.jht.2018.09.012
30 31	435	12	Nazari G, Bobos P, MacDermid JC, et al. Psychometric properties of the Zephyr bioharness
32 33	436		device: A systematic review. BMC Sports Sci Med Rehabil 2018;10. doi:10.1186/s13102-018-
34 35	437		0094-4
36 37	438	13	Law MC, MacDermid J. Evidence-based rehabilitation: a guide to practice. Thorofare, NJ::
38 39 40	439		Slack Incorporated 2014.
40 41 42	440	14	Roy JS, Desmeules F, MacDermid JC. Psychometric properties of presenteeism scales for
43 44	441		musculoskeletal disorders: A systematic review. J Rehabil Med Published Online First: 2011.
45 46	442		doi:10.2340/16501977-0643
47 48	443	15	Sańchez J. Rosner, B.: Fundamentals of Biostatistics, third edition. PWS-Kent, Boston 1990, xv,
49 50	444		655 pp., ISBN 0-534-91973-1. <i>Biometrical J</i> 1993; <b>35</b> :150. doi:10.1002/bimj.4710350205
51 52	445	16	Wuensch KL, Evans JD. Straightforward Statistics for the Behavioral Sciences. J Am Stat Assoc
53 54	446		Published Online First: 2006. doi:10.2307/2291607
55 56	447	17	Cohen J. Statistical power analysis for the behavioral sciences. Stat. Power Anal. Behav. Sci.
57 58 59			19
59 60			For peer review only - http://bmiopen.bmi.com/site/about/guidelines.xhtml

448		1988. doi:10.1234/12345678
449	18	Mokkink LB, de Vet HCW, Prinsen CAC, et al. COSMIN Risk of Bias checklist for systematic
450		reviews of Patient-Reported Outcome Measures. Qual Life Res Published Online First: 2018.
451		doi:10.1007/s11136-017-1765-4
452	19	Terwee CB, Mokkink LB, Knol DL, et al. Rating the methodological quality in systematic reviews
453		of studies on measurement properties: a scoring system for the COSMIN checklist. 2012;:651–7.
454		doi:10.1007/s11136-011-9960-1
455	20	Viechtbauer W. Conducting meta-analisys in R with metafor package. <i>J Stat Softw</i> 2010; <b>36</b> :1–48.
456	21	Higgins JPT, Thompson SG, Deeks JJ, et al. Measuring inconsistency in meta-analyses. BMJ
457		Published Online First: 2003. doi:10.1136/bmj.327.7414.557
458	22	Cleland J, Fritz J, Whitman J, et al. The reliability and construct validity of the Neck Disability
459		Index and Patient Specific Functional Scale. Spine (Phila Pa 1976) 2006;31:598-602.
460	23	Cleland JA, Childs JD, Whitman JM. Psychometric Properties of the Neck Disability Index and
461		Numeric Pain Rating Scale in Patients With Mechanical Neck Pain. Arch Phys Med Rehabil
462		2008; <b>89</b> :69–74. doi:10.1016/j.apmr.2007.08.126
463	24	Cook C, Lawrence J, Michalak K, et al. Is there preliminary value to a within- and/or between-
464		session change for determining short-term outcomes of manual therapy on mechanical neck pain?
465		J Man Manip Ther 2014; <b>22</b> :173–80. doi:10.1179/2042618614y.0000000071
466	25	Guzy G, Vernon H, Polczyk R, et al. Psychometric validation of the authorized Polish version of
467		the Neck Disability Index. <i>Disabil Rehabil</i> 2013; <b>35</b> :2132–7. doi:10.3109/09638288.2013.771706
468	26	Kamper SJ, Ostelo RWJG, Knol DL, et al. Global Perceived Effect scales provided reliable
469		assessments of health transition in people with musculoskeletal disorders, but ratings are strongly
470		influenced by current status. J Clin Epidemiol 2010;63:760-766.e1.
471		doi:10.1016/j.jclinepi.2009.09.009
472	27	Shaheen AAM, Omar MTA, Vernon H. Cross-cultural adaptation, reliability, and validity of the
473		arabic version of neck disability index in patients with neck pain. Spine (Phila Pa 1976)

1 2				
3 4	474		2013; <b>38</b> :609–15. doi:10.1097/BRS.0b013e31828b2d09	
5 6	475	28	Takeshita K, Hosono N, Kawaguchi Y, et al. Validity, reliability and responsiveness of the	
7 8	476		Japanese version of the Neck Disability Index. J Orthop Sci 2013;18:14–21. doi:10.1007/s00776-	
9 10	477		012-0304-y	
11 12	478	29	Trouli MN, Vernon HT, Kakavelakis KN, et al. Translation of the Neck Disability Index and	
13 14	479		validation of the Greek version in a sample of neck pain patients. BMC Musculoskelet Disord	
15 16	480		2008; <b>9</b> :1–8. doi:10.1186/1471-2474-9-106	
17 18	481	30	Tuttle N, Laakso L, Barrett R. Change in impairments in the first two treatments predicts outcome	9
19 20 21	482		in impairments, but not in activity limitations, in subacute neck pain: An observational study. Aust	t
22 23	483		J Physiother 2006; <b>52</b> :281–5. doi:10.1016/S0004-9514(06)70008-3	
24 25	484	31	Ngo Trung, Stupar Maja, Co^te' Pierre, Boyle Eleanor, Shearer Heather. A study of the test –	
26 27	485		retest reliability of the self-perceived general recovery and self-perceived change in neck pain	
28 29	486		questions in patients with recent whiplash-associated disorders. 2010;:957-62.	
30 31	487		doi:10.1007/s00586-010-1289-x	
32 33	488	32	Björklund M, Wiitavaara B, Heiden M. Responsiveness and minimal important change for the	
34 35	489		ProFitMap-neck questionnaire and the Neck Disability Index in women with neck-shoulder pain.	
36 37	490		Qual Life Res 2017; <b>26</b> :161–70. doi:10.1007/s11136-016-1373-8	
38 39 40	491	33	Evans R, Bronfort G, Maiers M, et al. "I know it" s changed ": a mixed-methods study of the	
40 41 42	492		meaning of Global Perceived Effect in chronic neck pain patients. 2014;:888–97.	
43 44	493		doi:10.1007/s00586-013-3149-y	
45 46	494	34	Jorritsma W, Dijkstra PU, De Vries GE, et al. Detecting relevant changes and responsiveness of	
47 48	495		Neck Pain and Disability Scale and Neck Disability Index. Eur Spine J 2012;21:2550–7.	
49 50	496		doi:10.1007/s00586-012-2407-8	
51 52	497	35	Monticone M, Frigau L, Vernon H, et al. Responsiveness and minimal important change of the	
53 54	498		NeckPix @ in subjects with chronic neck pain undergoing rehabilitation. Eur Spine J	
55 56	499		2018; <b>27</b> :1324–31. doi:10.1007/s00586-017-5343-9	
57 58			2	21

500	36	Monticone M, Ambrosini E, Vernon H, et al. Responsiveness and minimal important changes for
501		the Neck Disability Index and the Neck Pain Disability Scale in Italian subjects with chronic neck
502		pain. Eur Spine J 2015; <b>24</b> :2821–7. doi:10.1007/s00586-015-3785-5
503	37	Young BA, Walker MJ, Strunce JB, et al. Responsiveness of the Neck Disability Index in patients
504		with mechanical neck disorders. <i>Spine J</i> 2009; <b>9</b> :802–8. doi:10.1016/j.spinee.2009.06.002
505	38	Farooq MN, Mohseni-Bandpei MA, Gilani SA, et al. Urdu version of the neck disability index: A
506		reliability and validity study. <i>BMC Musculoskelet Disord</i> 2017; <b>18</b> :1–11. doi:10.1186/s12891-017-
507		1469-5
508	39	Williams GN, Gangel TJ, Arciero RA, et al. Comparison of the single assessment numeric
509	3)	evaluation method and two shoulder rating scales. Outcomes measures after shoulder surgery. Am
510		J Sports Med 1999; <b>27</b> :214–21. doi:10.1177/03635465990270021701
	40	
511	40	Schmitt J, Abbott JH. Global Ratings of Change Do Not Accurately Reflect Functional Change
512		Over Time in Clinical Practice. J Orthop Sport Phys Ther 2015;45:106–11.
513		doi:10.2519/jospt.2015.5247
514	41	Chiarotto A, Ostelo RW, Boers M, et al. A systematic review highlights the need to investigate the
515		content validity of patient-reported outcome measures for physical functioning in patients with
516		low back pain. <i>J Clin Epidemiol</i> 2018; <b>95</b> :73–93. doi:10.1016/j.jclinepi.2017.11.005
517	42	Ailliet L, Knol DL, Rubinstein SM, et al. Definition of the construct to be measured is a
518		prerequisite for the assessment of validity. the Neck Disability Index as an example. J Clin
519		Epidemiol 2013; <b>66</b> :775-782.e2. doi:10.1016/j.jclinepi.2013.02.005
520		
521		
522		
523		
524		
525		
323		

, Al training, and similar technologies

Protected by copyright, including for uses related

Figure 1. Flow diagram of included studie	ļ	Figure	1. Flov	v diagram	of included	studies
---	---	--------	---------	-----------	-------------	---------

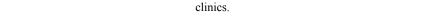
- **Figure 2**. Meta-analysis of Pearson's correlation coefficients between neck disability change scores and GROC scores in patients with neck disorders based on 5 very good to excellent quality studies.
- **Figure 3**. Meta-analysis of Spearman's correlation coefficients between neck disability change scores and GROC scores in patients with neck disorders based on 6 very good to excellent quality studies.

Figure 4. Random effects univariate meta-regression between age and the Fisher's Z estimates. Each circle represents a study and the size of the circle indicates the influence of that study on the model. The illustra.
68% of the v. regression prediction is illustrated by the straight line and the curved lines represent the 95% confidence intervals. Age explained 68% of the variance in the model (R<sup>2</sup>=0.68)

Study	Population	Setting	Sample Size	<b>Properties Evaluated</b>	GROC evaluated	Interval
jorklund	Women with non-	Not specified	104	Validity (correlation)	GRoC 7-points	GRoC scale
t al (2017)	specific neck- shoulder pain			Between NDI and GRoC	1. Very much worse; 2. Much worse; 3. Minimally worse; 4. No change; 5. Minimally improved; 6. Much improved; 7. Very much improved.	administered only after intervention a one time point (1 week)
Cleland et	Patients with cervical	Hospital	38	Validity (correlation)	GRoC 15-points	GRoC was completed following Wife
(2006)	radiculopathy			Between NDI and GRoC	-7 (a very great deal worse) to	week over the peool
				Between PSFS and GRoC	zero (about the same) to +7 (a very great deal better)	of 7 weeks.
leland et	Patients with neck	5 Outpatient	137	Validity (correlation)	GRoC 15-points	GRoC was complete
. (2008)	pain only	physical therapy		Between NDI and GRoC	-7 (a very great deal worse) to	at follow up. Within week
		clinics		Between NPRS and GRoC	zero (about the same) to +7 (a very great deal better)	GRoC was completed the follow up. With week over the pedd of 7 weeks.  GRoC was completed the follow up. With week week week including the follow up. Baseline and at ing
ook et al	Patients with any	Academic	56	ROC curves and AUC to	GRoC 15-points	Baseline and at
2014)	neck pain	locations in Northeast		measure sensitivity and specificity. Binomial logistic	-7 (a very great deal worse) to	follow up 48- and 9 hours post baseline
		Ohio		regression analysis was also	zero (about the same) to +7 (a very great deal better)	use
				calculated to determine overall effect.	very great dear better)	hours post baseline
arooq et	Patients with neck	Physical	106	Validity (correlation)	GRoC 15-points	GRoC was comp
. (2017)	pain	therapy clinics		Between NDI-U and GRoC	-7 (a very great deal worse) to	at three weeks are
					zero (about the same) to +7 (a very great deal better)	intervention in intervention i
uzy et al.	Patients with neck	Outpatient	95	Validity (correlation)	GRoC 7-points	GRoC scale was
2013)	pain	rehabilitation clinic		Between NDI-P and GRoC	'complete recovery' over 'no	completed at 2 weeks and at 4 weeks
		CHITIC			change'' to ''my complaints are worse than ever''	ning
orritsma et	Patients with chronic	Tertiary	76	Validity (correlation)	GPE 7-points	After completions
. (2012)	non-specific neck	university center for		Between NDI and GRoC	3 (completely recovered) to zero	the program vary from 3 to 5 mon 5
	r	rehabilitation		Between NPAD and GRoC	(no change) to -3 (worse than ever)	patients filled the GPE
amper et	Patients with any	Physical	134	Test-retest reliability	GPE 11-points	Baseline, 6 weeks,
1. (2010)	whiplash-associated disorder.	therapy clinics			-5 (vastly worse) to zero	and 12 months
					(unchanged) to +5 (completely recovered)	echno
Ionticone	Patients with chronic	Outpatient	153	Validity (correlation)	GPE 5-points	At the end of
t al. 2017	neck pain	Rehabilitatio n Unit		Between NeckPix and GPE	(helped a lot = $1$ , helped = $2$ ),	treatment (8 weeks
					one no change level (helped only a little = 3), and two worsening	follow-up
						After completion to the program varying from 3 to 5 months patients filled the GPE  Baseline, 6 weeks, and 12 months  At the end of treatment (8 weeks and one year before follow-up)
						24
	_			://bmjopen.bmj.com/site/abo		

				BMJ Open			BMJ OF
					levels (did not help = 4, made things worse = 5)		eeks
nticone	Patients with chronic	Outpatient	200	Validity (correlation)	GPE 5-points	At the end of	lblish Asildi
2015	neck pain	Rehabilitatio n Unit		Between NDI and GPE	(helped a lot = 1, helped = $2$ ),	treatment 8 we	eks ō a
				Between NPDS and GPE	one no change level (helped only a little = 3), and two worsening levels (did not help = 4, made things worse = 5)		7
t al.	Patients with WAD.	Interviewed	46	Test-retest reliability	GPE 7-points	3-5 days	ted t
	Most participants (69.6%) had grade II	by person or by telephone			1. General recovery question		by co
	WAD.	in Ontario			Completely better Much improved Slightly improved No change		Protected by copyright, including
					Slightly worse Much worse		nclu
					Worse than ever		on 25 ding
					2. Change in neck pain question:		for u
					very much better, better, slightly better, no change, slightly worse, worse, or very much worse		ember 201 Enseigner ses relate
een et	Patients with neck	3 primary	70	Validity (correlation)	GRoC 15-points	1 week	nent d to t
)15)	pain lasting more than 3 months	health centers		Between NDI-Ar and GRoC	-7 (a very great deal worse) to zero (about the same) to +7 (a very great deal better)		0.1136/bmJopen-2019-033909 on 25 November 2019. Downloaded Enseignement Superieur Protected by copyright, including for uses related to text and da
shita	Patients with neck	Variety of	130	Validity (correlation)	PGIC 7-points	Over 8 weeks	a;∑⇒ ≓
4)	pain, cervical radiculopathy and/or cervical myelopathy	clinics and hospital settings		Between NDI-J and GRoC	much better, better, slightly better, unchanged, slightly worse, worse and much worse		m http://br ES) . nining, Al t
li et al. 8)	Patients with neck	Primary	68	Validity (correlation)	GRoC 15-points	Within 2 mont	hsabut e
	pain	healthcare clinic		Between NDI-Gr and GRoC	-7 (a very great deal worse) to -1 (almost the same, hardly any worse at all) and from 7 (a very great deal better) to 1 (almost the same, hardly any better at all)	Within 2 mont 1 week for test 6 weeks 25	en.bmJ.com/ on est ang, and similar
le et al.	Patients with neck pain for more than 2	Private physiotherap	29	Validity (correlation)	GPE 11-points	6 weeks	Jun
))	weeks	y clinics		Between NDI and GPE	-5 is vastly worse and +5 is completely recovered		e 11, nolo
				Between PSFS and GPE	completely recovered		gies.
				Between VAS and GPE			· at
				Between ROM and GPE			∖gen
ng et 2009)	Patients presenting with mechanical neck pain	Outpatient physical therapy	91	Validity (correlation)	GRoC 15-points  -7 ("a very great deal worse") to 0 ("about the same") to +7 ("a	3 weeks	ce Bibliogi
						25	raphic

very great deal better")



**TABLE 2.** Summary of Psychometric Properties Reported in Studies and COSMIN Risk of Bias (RoB) and Quality studies

Study	Psychometric Properties Reported	COSMIN RoB	COSMIN Rating*§	Quality of Studies**
			(Criteria)	(QACMRR)
Bjorklund et al (2017)	Validity (correlation)	Very Good	?	Excellent
Cleland et al (2006)	Validity (correlation)	Very Good	+	Excellent
Cleland et al. (2008)	Validity (correlation)	Very Good	-	Excellent
Cook et al (2014)	Sensitivity Specificity	Very Good Very Good	+	Excellent
Farooq et al. (2017)	Validity (correlation)	Very Good	+	Excellent
Guzy et al. (2013)	Validity (correlation)	Very Good	?	Very good
Jorritsma et al. (2012)	Validity (correlation)	Very Good	?	Excellent
Kamper et al. (2010)	Test-retest reliability	Very Good	+	Excellent
Monticone et al. (2017)	Validity (correlation)	Very Good	?	Excellent
Monticone et al. (2015)	Validity (correlation	Very Good	?	Excellent
Ngo et al. (2010)	Test-retest reliability	Very Good	+	Excellent
Shaheen et al. (2015)	Validity (correlation)	Very Good	?	Excellent
Takeshita et al. (2014)	Validity (correlation)	Very Good	?	Very good
Trouli et al. (2008)	Validity (correlation)	Very Good	+	Excellent
Tuttle et al. (2006)	Validity (correlation)	Very Good	?	Excellent
Young et al. (2009)	Validity (correlation)	Very Good	?	Excellent

COSMIN, Consensus-based Standards for the Selection of health Measurement Instruments, Criteria for good measurement properties: '+' sufficient; '-'insufficient; '?' indeterminate. §§ The grading for the quality of the evidence based on the modified GRADE approach is not applicable. \*\*Quality Appraisal for Clinical Measurement Research Reports Evaluation Form (QACMRR).

**TABLE 3**. Quality Appraisal for Clinical Measurement Research Reports Evaluation Form

				It	em E	valu	ation	Crite	eria*					
Study	1	2	3	4	5	6	7	8	9	10	11	12	Total (%)	Quality Summary
Bjorklund et al (2017)	2	2	2	2	2	1	2	2	2	2	2	2	96	Excellent
Cleland et al. (2008)	2	2	2	2	1	2	2	2	2	2	2	2	96	Excellent
Trouli et al. (2008)	2	2	2	2	1	2	2	2	2	2	2	2	96	Excellent
Tuttle et al. (2006)	2	2	2	2	1	2	2	2	2	2	2	2	96	Excellent
Kamper et al. (2010)	2	2	2	2	1	2	2	2	2	2	2	2	96	Excellent
Cook et al (2014)	2	2	2	2	1	2	2	2	1	2	2	2	92	Excellent
Jorritsma et al. (2012)	2	2	2	2	1	1	2	2	2	2	2	2	92	Excellent
Cleland et al (2006)	2	2	2	2	1	1	2	2	2	2	2	2	92	Excellent
Monticone et al. (2017)	2	2	2	2	1	1	2	2	2	2	2	2	92	Excellent
Monticone et al. (2015)	2	2	2	2	1	1	2	2	2	2	2	2	92	Excellent
Ngo et al. (2010)	2	2	2	2	2	2	2	2	1	2	1	2	92	Excellent
Shaheen et al. (2013)	2	2	2	2	2	2	2	2	2	2	1	1	92	Excellent
Farooq et al. (2017)	2	2	1	2	2	2	2	2	1	2	2	2	92	Excellent
Young et al. (2009)	2	2	2	2	1	1	2	2	2	2	2	2	92	Excellent
Guzy et al. (2013)	2	2	1	2	1	2	2	2	1	2	2	2	88	Very good
Takeshita et al. (2014)  Item Evaluation Co	2	2	1	1	1	2	2	2	2	2	2	2	88	Very good

\*Item Evaluation Criteria: 1. Thorough literature review to define the research question; 2. Specific inclusion/exclusion criteria; 3. Specific hypotheses; 4. Appropriate scope of psychometric properties; 5. Sample size; 6. Follow-up; 7. The authors referenced specific procedures for administration, scoring, and interpretation of procedures; 8. Measurement techniques were standardized; 9. Data were presented for each hypothesis; 10. Appropriate statistics-point estimates; 11. Appropriate statistical error estimates; 12. Valid conclusions and clinical recommendations.

, Al training, and similar technologies

Protected by copyright, including for uses related to text

<u> </u>	
3	579
<del>]</del> 5	580
5	581
7	582
3	583
)  0	584
1	585
2	303
3	586
4  5	F07
6	587
7	588
8	
19 20	589
21	590
22	330
23	591
24 25	592
26	592
27	593
28 29	
30	594
31	595
32	
33 34	596
35	597
36	337
37	598
38 39	F00
10	599
11	600
12	
13 14	601
<del>14</del> 15	602
16	002
17	603
18 19	CO4
50	604
51	605
52	
53 54	606
55	607
6	

57 58

59

60

$Total\ score = (sum\ of\ subtotals \div 24 \times 100).\ If\ for\ a\ specific\ paper\ an\ item\ is\ deemed\ NA\ (Not\ Applicable),\ then,\ Total\ score\ score\ score\ subtotals\ score\ score$	ore
= (sum of subtotals $\div$ (2 × number of Applicable items) × 100).	

NA – Not Applicable. The subsections no. 6, asks for percentage of retention/follow up. This subsection only applies to reliability test-retest studies

Quality Summary: Poor (0%-30%), Fair (31%-50%), Good (51%-70%), Very good (71%-90%), Excellent (>90%): TO BEEL ENEMONA

		BMJ Open		Page 30
<b>ABLE 4</b> . St	ummary of reli	ability properties of GRoC scales		
Study	Type of Reliability	Reliability Estimates	COSMIN	Quality of Studies
Kamper et al. (2010)	Test-retest	Intra-class correlation coefficients (ICC) $0.99 (0.99 - 0.99)$ – baseline $0.96 (0.95 - 0.97)$ – at six weeks $0.92 (0.89 - 0.94)$ at twelve months.	Very Good	Excellent
Ngo et al. (2010)	Test-retest	Intra-class correlation coefficients (ICC)  0.70 (0.60–0.80) – at six weeks (General recovery)  0.80 (0.72–0.87) – at six weeks (neck pain questions)  Weighted Kappa  0.70 (0.42–0.98) – at six weeks (General recovery)  0.80 (0.51–1.0) – at six weeks (neck pain questions)  Dichotomized response options for recovery (K statistics)  0.85 (0.64–1) when "recovered" was defined  "completely better"  0.81 (0.64–0.99) when defined as "completely better" or  "much improved  Dichotomized response options for change in neck pain questions (K statistics)  0.46 (0.20–0.74) when "recovered" was defined as "very much better"  Recall questions (K statistics)  the kappa coefficient was 1 for participants who remembered their previous answers to the general recovery question; 0.88 (0.64–1) for those who did not remember and 0.50 (0.02–0.98) for participants who were not asked the question.  The kappa coefficient was 1 for participants who remembered their previous answers to the change in neck pain question; 0.74 (0.41–1) for those who did not remember and 0.66 (0.22–1) for participants who were not asked the question.	Very Good	Excellent
		w only - http://bmjopen.bmj.com/site/about/guidelines		30

**TABLE 5**. Summary of validity properties of GRoC scales 

Cleland et al. (2006)   Correlations (Pearson r)   between change scores   r = 0.19   Very Good   Excellent	Study	Type of Reliability	Validity Estimates	<u>COSMIN</u>	Quality of Studies
Cleland et al. (2006)   NDI and GRoC   PSFS and GRoC   r = 0.82   NDI and GRoC   r = 0.82   NDI and GRoC   r = 0.82   NDI and GRoC   r = 0.58   NDI and GRoC   Receiver operator characteristics (ROC)   Within-session change of Pain and GROC   AUC = 0.61   AUC = 0.66, -336.7%   Change in pain   Very Good   Excellent (2014)   Retween session change of Pain and GROC   Odds ratio = 7.3 (2.1		between the change scores of GRoC and ProFitMap-neck		Very Good	Excellent
Cleland et al. (2008)  Cleland et al. (2008)  Receiver operator characteristics (ROC)  NRS and GRoC  NRS and GRoC  NRS and GRoC  NRS and GRoC  Receiver operator characteristics (ROC)  Within-session change Between-session change		between change scores NDI and GRoC		Very Good	Excellent
Within-session change   Between-session change   Between-session change   Between-session change   Cook et al. (2014)   Between session change of Pain and GROC   Odds ratio = 7.3 (2.1,		between change scores NDI and GRoC	r = 0.58	Very Good	Excellent
Farooq et al. (2017) NDI-U $r=0.50$ Very Good Excellent (2017) NDI-U $r=0.50$ Very Good Excellent (2017) NDI-U $r=0.50$ Very Good Very good (2013) NDI vs GROC Four-week interval ( $r=-0.56$ ) Very Good Very good (2013) NDI vs GROC Four-week interval ( $r=-0.56$ ) Very Good Excellent (2012) NPAD and GPE $r=0.49$ (95 % CI 0.30— between change scores of NPAD and GPE $r=0.49$ (95 % CI 0.30— between change scores of the NeckPix© and GPE $r=0.49$ (95 % CI 0.30— between change scores of the NeckPix© and GPE $r=0.49$ (95 % CI 0.30— Very Good Excellent (2017) $r=0.49$ (95 % CI 0.30— Very Good Excellent (2017) $r=0.49$ (95 % CI 0.30— Very Good Excellent (2017) $r=0.49$ (95 % CI 0.30— Very Good Excellent (2017) $r=0.49$ (95 % CI 0.30— Very Good Excellent (2017) $r=0.49$ (95 % CI 0.30— Very Good Excellent (2018) $r=0.49$ (95		Receiver operator characteristics (ROC) Within-session change Between-session change Between session change of Pain and GROC Sensitivity	AUC = 0.61 AUC = 0.76, >36.7% change in pain Odds ratio = 7.3 (2.1, 24.7) 65.6% (57.9, 74.6)	Very Good	Quality of Studies  Excellent  Excellent  Excellent  Very good
Guzy et al. (2013)				Very Good	Excellent
between change scores of al. (2012)  Monticone et al. (2017)  Monticone et al. (2017)  Monticone et al. (2015)  Monticone et al. (2016)  NDI-I and GPE  NDPS and GPE  Shaheen et al. (2013)  Takeshita et al. (2014)  Torrelations  Monticone et between change scores rho = 0.71, p<0.01  Torelations (Spearman's)  NDI-Ar and GROC  Takeshita et al. (2014)  Monticone et between change scores rho = 0.71, p<0.01  Torelations (Spearman's)  NDI-Ar and GROC  Trouli et al. (2014)  Trouli et al. (2014)  Monticone et between change scores of the NeckPix©  Trouli et al. (2016)  Trouli et al. (2018)  GROC vs Gr-NDI  Tuttle et al. (2018)  NDI vs GPE (post 1, minus pre-1)  NDI vs GPE (post 2, minus pre-2)  PSFS vs GPE (post 1, minus pre-2)  Trho = 0.06	Guzy et al.	· /	0.73) Four-week interval (r = -	Very Good	Very good
$\begin{array}{c ccccccccccccccccccccccccccccccccccc$		between change scores of	`	Very Good	Excellent
Correlation (Spearman)		Correlations (Spearman) between change scores of the NeckPix©	rho = 0.69 - 0.82	Very Good	Excellent
Shaheen et al. (2013) Correlations (Spearman's) rho = 0.81, p<0.001 Very Good Excellent (2013) NDI-Ar and GROC rho = 0.81, p<0.001 Very Good Takeshita et al. (2014) NDI and PGIC rho = 0.47, p<0.001 Very Good Very good NDI-J and PGIC rho = 0.59, p<0.001 Very Good Excellent (2008) GROC vs Gr-NDI rho = 0.30, p=0.02 Correlations (Spearman's) Very Good Excellent NDI vs GPE (post 1, minus pre-1) rho = 0.17 NDI vs GPE (post 2, minus pre-1) rho = 0.01 Tuttle et al. NDI vs GPE (post 2, minus pre-2) rho = 0.03 (2006) PSFS vs GPE (post 1, minus pre-1) rho = 0.06		between change scores NDI-I and GPE		Very Good	Excellent  Excellent  Very good  Excellent
Takeshita et al. (2014)  NDI and PGIC  NDI-J and PGIC  NDI-J and PGIC  NDI-J and PGIC  Tho = 0.47, p<0.001  Tho = 0.59, p<0.001  Very Good  Excellent (2008)  Correlation (Spearman's)  Correlations (Spearman's)  NDI vs GPE (post 1, minus pre-1)  NDI vs GPE (post 2, minus pre-1)  Tuttle et al.  NDI vs GPE (post 2, minus pre-2)  PSFS vs GPE (post 1, minus pre-1)  Tho = 0.06		Correlations (Spearman's)	rho = 0.81, p<0.001	Very Good	Excellent
$ \begin{array}{c ccccccccccccccccccccccccccccccccccc$		NDI and PGIC	rho = 0.47, p < 0.001	Very Good	Very good
$ \begin{array}{c ccccccccccccccccccccccccccccccccccc$		Correlation (Spearman's)	•	Very Good	Excellent
	Tuttle et al.	Correlations (Spearman's) NDI vs GPE (post 1, minus pre-1) NDI vs GPE (post 2, minus pre-1) NDI vs GPE (post 2, minus pre-2)	rho = 0.17 rho = 0.01 rho = 0.03	Very Good	Excellent
PSFS vs GPE (post 2, minus pre-1)		PSFS vs GPE (post 2, minus pre-1)	rho = 0.03		

	Pain Intensity (post 1, minus pre-1) Pain Intensity (post 2, minus pre-1) Pain Intensity (post 2, minus pre-2) Total ROM (post 1, minus pre-1)	rho = 0.00 $rho = 0.05$ $rho = 0.01$ $rho = 0.03$		
	Total ROM (post 2, minus pre-1)	rho = 0.01		
Young et al. (2009)	Total ROM (post 2, minus pre-2)  Correlations (Pearson's) between change scores  NDI and GRoC	rho = 0.00 r =0.52 (p<0.01)	Very Good	Excellent
Monticone et al. (2015)	Correlation (Spearman) between change scores NDI-I and GPE	rho = 0.71, p<0.01 rho = 0.59, p<0.01	Very Good	Excellent
	NDPS and GPE			

Box 1. Questions of Global Rating of Change (GROC) scales

Author	GROC item- scale	Patients with neck disorders were asked:
Bjorklund et al. (2017)	GROC 7-points	"Compared to before the treatment of the study started, my overall status is now"
		"Compared to before the treatment of the study started, my status regarding my neck-shoulder problem is now"
Evans et al (2014)	GPE 9-points	''Overall, how much has your neck pain changed since you started treatment in the study?''
Kamper et al. (2010)	GPE 11-points	"With respect to your whiplash injury how would you describe yourself now compared to immediately after your accident"
Monticone et al. (2017)	GPE 5-points	"Overall, how much did the treatment you received help your fear of movement due to current neck pain?  "Overall, how much did the treatment you delivered help your subject's fear of movement due to her/ his current neck pain?"
Monticone et al. (2015)	GPE 5-points	"Overall, how much did the treatment you received help your neck problem?"
Ngo et al. (2010)	GPE 7-points	"How well do you feel you are recovering from your injuries?"  "How do you feel your neck pain has changed since the injury?"

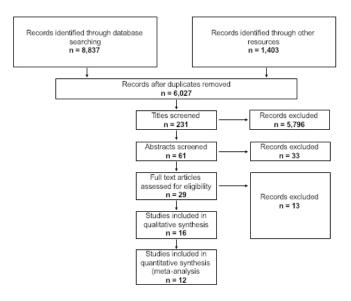


Figure 1. Flow diagram of included studies  $60x34mm (300 \times 300 DPI)$ 

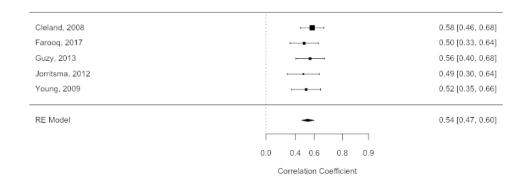


Figure 2. Meta-analysis of Pearson's correlation coefficients between neck disability change scores and GROC scores in patients with neck disorders based on 5 very good to excellent quality studies.

67x34mm (300 x 300 DPI)

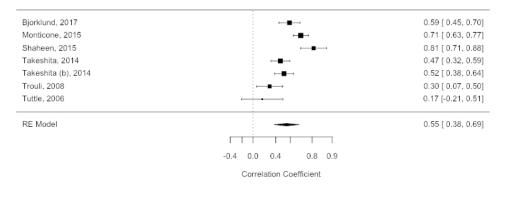


Figure 3. Meta-analysis of Spearman's correlation coefficients between neck disability change scores and GROC scores in patients with neck disorders based on 6 very good to excellent quality studies.

67x34mm (300 x 300 DPI)

#### Regression of Fisher's Z on Age

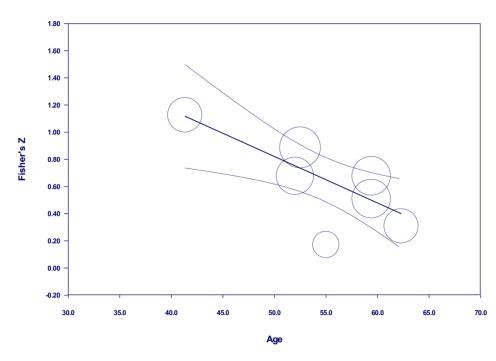


Figure 4. Random effects univariate meta-regression between age and the Fisher's Z estimates. Each circle represents a study and the size of the circle indicates the influence of that study on the model. The regression prediction is illustrated by the straight line and the curved lines represent the 95% confidence intervals. Age explained 68% of the variance in the model (R2=0.68)

160x118mm (300 x 300 DPI)

# Appendix 1

#### Search terms

#### **MEDLINE-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. cervicalgia.mp.
- 38. brachialgia.mp.
- 39. brachial neuritis.mp.

data mining, Al training, and similar technologies

Protected by copyright, including for uses related to text

```
2
3
             40. brachial neuralgia.mp.
4
             41. neck pain.mp.
5
             42. neck injur*.mp.
6
             43. brachial plexus neuropath*.mp.
7
             44. brachial plexus neuritis.mp.
8
```

- 45. thoracic outlet syndrome/ or cervical rib syndrome/
- 46. Torticollis/
- 47. exp brachial plexus neuropathies/ or exp brachial plexus neuritis/
- 48. cervico brachial neuralgia.ti,ab.
- 49. cervicobrachial neuralgia.ti,ab.
- 50. (monoradicul\* or monoradicl\*).tw.
- 51. or/30-50
- 52. exp headache/ and cervic\*.tw.
- 53. exp genital diseases, female/

- genital disease\*.mp.
  or/53-54
  52 not 55
  .51 or 56
  . neck/
  ). neck muscles/
  ). exp cervical plexus/
  1. exp cervical vertebrae/
  32. atlanto-axial joint/
  53. atlanto-occipital joint/
  64. Cervical Atlas/
  65. spinal nerve roots/
  66. exp brachial plexus/
  67. (odontoid\* or cervical or occip\* or atlant\*).tw.
  68. axis/ or odontoid process/
  Thoracic Vertebrae/

- 74. (brachial adj3 plexus).mp.
- 75. (thoracic adj3 vertebrae).mp.
- 76. neck.mp.
- 77. (thoracic adj3 spine).mp.
- 78. (thoracic adj3 outlet).mp.
- 79. trapezius.mp.
- 80. cervical.mp.
- 81. cervico\*.mp.
- 82.80 or 81
- 83. exp genital diseases, female/
- 84. genital disease\*.mp.
- 85. exp \*Uterus/

BMJ Open: first published as 10.1136/bmjopen-2019-033909 on 25 November 2019. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de Enseignement Superieur (ABES) .

ining, Al training, and similar technologies

Protected by copyright, including for uses related

```
86. 83 or 84 or 85
87. 82 not 86
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
74 or 75 or 76 or 77 or 78 or 79 or 87
89. exp pain/
90. exp injuries/
91. pain.mp.
92. ache.mp.
93. sore.mp.
94. stiff.mp.
95. discomfort.mp.
96. injur*.mp.
97. neuropath*.mp.
98. or/89-97
99. 88 and 98
100. Radiculopathy/
101. exp temporomandibular joint disorders/ or exp temporomandibular joint dysfunction
syndrome/
102. myofascial pain syndromes/
103. exp "Sprains and Strains"/
104. exp Spinal Osteophytosis/
105. exp Neuritis/
106. Polyradiculopathy/
107. exp Arthritis/
108. Fibromyalgia/
109. spondylitis/ or discitis/
110. spondylosis/ or spondylolysis/ or spondylolisthesis/
111. radiculopathy.mp.
112. radiculitis.mp.
113. temporomandibular.mp.
114. myofascial pain syndrome*.mp.
115. thoracic outlet syndrome*.mp.
116. spinal osteophytosis.mp.
117. neuritis.mp.
118. spondylosis.mp.
119. spondylitis.mp.
120. spondylolisthesis.mp.
121. or/100-120
122. 88 and 121
```

- 123. exp neck/
- 124. exp cervical vertebrae/
- 125. Thoracic Vertebrae/
- 126. neck.mp.
- 127. (thoracic adj3 vertebrae).mp.
- 128. cervical.mp.
- 129. cervico\*.mp.

60

BMJ Open: first published as 10.1136/bmjopen-2019-033909 on 25 November 2019. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de Enseignement Superieur (ABES) . Protected by copyright, including for uses related ining, Al training, and similar technologies

```
2
3
             130. 128 or 129
4
             131. exp genital diseases, female/
5
             132. genital disease*.mp.
6
             133. exp *Uterus/
7
             134. or/131-133
8
9
             135. 130 not 134
10
             136. (thoracic adj3 spine).mp.
11
             137. cervical spine.mp.
12
             138, 123 or 124 or 125 or 126 or 127 or 135 or 136 or 137
13
             139. Intervertebral Disk/
14
             140. (disc or discs).mp.
15
             141. (disk or disks).mp.
16
17
             142. 139 or 140 or 141
18
             143. 138 and 142
19
             144. herniat*.mp.
20
             145. slipped.mp.
21
             146. prolapse*.mp.
22
             147. displace*.mp.
23
24
             148. degenerat*.mp.
25
             149. (bulge or bulged or bulging).mp.
26
             150. 144 or 145 or 146 or 147 or 148 or 149
27
             151. 143 and 150
28
             152. intervertebral disk degeneration/ or intervertebral disk displacement/
29
             153. intervertebral disk displacement.mp.
30
             154. intervertebral disc displacement.mp.
31
32
             155. intervertebral disk degeneration.mp.
33
             156. intervertebral disc degeneration.mp.
34
             157. 152 or 153 or 154 or 155 or 156
35
             158. 138 and 157
36
             159. 57 or 99 or 122 or 151 or 158
37
             160. animals/ not (animals/ and humans/)
38
             161. 159 not 160
39
40
             162. exp *neoplasms/
41
             163. exp *wounds, penetrating/
42
             164. 162 or 163
43
             165. 161 not 164
44
             166. 29 and 165
45
             167. guidelines as topic/
46
             168. practice guidelines as topic/
47
48
             169. guideline.pt.
49
             170. practice guideline.pt.
50
             171. (guideline? or guidance or recommendations).ti.
51
             172. consensus.ti.
52
             173. or/167-172
53
             174. meta-analysis/
54
```

175. exp meta-analysis as topic/

Protected by copyright, including for uses related to text and

data mining, Al training, and similar technologies

- 176. (meta analy\* or metaanaly\* or met analy\* or metanaly\*).tw.
- 177. review literature as topic/
- 178. (collaborative research or collaborative review\* or collaborative overview\*).tw.
- 179. (integrative research or integrative review\* or integrative overview\*).tw.
- 180. (quantitative adj3 (research or review\* or overview\*)).tw.
- 181. (research integration or research overview\*).tw.
- 182. (systematic\* adj3 (review\* or overview\*)).tw.
- 183. (methodologic\* adj3 (review\* or overview\*)).tw.
- 184. exp technology assessment biomedical/
- 185. (hta or thas or technology assessment\*).tw.
- 186. ((hand adj2 search\*) or (manual\* adj search\*)).tw.
- 187. ((electronic adj database\*) or (bibliographic\* adj database\*)).tw.
- 188. ((data adj2 abstract\*) or (data adj2 extract\*)).tw.
- 189. (analys\* adj3 (pool or pooled or pooling)).tw.
- 190. mantel haenszel.tw.
- 191. (cohrane or pubmed or pub med or medline or embase or psycinfo or psychinfo or psychlit or cinahl or science citation indes).ab.

- 192. or/174-191
- 193. 173 or 192
- 194. 166 and 193

#### **Quality Appraisal for Clinical Measurement Research Reports**

#### **Evaluation Form**

Authors:	Year:	Rater:

	·				
<u>Qualit</u>	y Appraisal for Clinical Measurer	ment Research Reports			
	<u>Evaluation Form</u>	1			
Authors:	Year:	Rater:			
each item on your quality ch study you are evaluating. It	nality of a clinical measurement st hecklist, pick the descriptor that so tems rank descriptors are provide thesis are available from develope	ounds <u>most</u> like what was r d in the guide. (Forms and <u>c</u>	eported in the guides to extro		
	Evaluation criteria			Scor	
Study question			2		0
	und work cited to define what is c	urrently known about the			_
measurement properties of n research question to informir	measures under study, and the pong that knowledge base?	otential contributions of the	current		
Study Design	7				
2. Were appropriate inclusion	n/exclusion criteria defined?	4			
3. Were specific clinical meas	surement questions/hypotheses i	dentified?			
	of measurement properties cons				
4. Was an appropriate scope	or measurement properties come	idered?			
		idered?			
5. Was an appropriate sample		4	vise n/a)		
5. Was an appropriate sample	e size used?	4	vise n/a)		
5. Was an appropriate sample 6. Was appropriate retention  Measurements 7. Were specific descriptions	e size used?	s involving retesting; otherw	. ,		
5. Was an appropriate sample 6. Was appropriate retention  Measurements 7. Were specific descriptions administer it? 8. Were standardized proces	e size used? n/follow-up obtained? (for studies	s involving retesting; otherwise tudy and the method(s) use	ed to		

BMJ Open	Page 44 of
. Were analyses conducted for each specific hypothesis or purpose?	
.0. Were appropriate statistical tests performed to obtain point estimates of the measurement properties?	
1. Were appropriate ancillary analyses done to quantify the confidence in the estimates of the linical measurement property (Precision/Confidence intervals; benchmark comparisons/ROC urves, alternate forms of analysis like SEM/MID, etc.)?	Protecte
Recommendations	d by c
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?	Protected by copyright, including for uses related to text and da
Subtotals (of columns 1 and 2)	uding
Total score (sum of subtotals/24*100);	for us
f for a specific paper or topic an item is deemed inappropriate then you can sum of tems/2*number of items *100	Enseignem ses related
	sur (ABES) . I data mining, Al training, and similar technologies.
For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml	

#### **Quality Appraisal of a Clinical Measurement Study**

#### **Interpretation Guide**

3		BMJ Open	
		Quality Appraisal of a Clinical Measurement Study	
		<u>Interpretation Guide</u>	
			7
		ide which score to provide for each item on your quality checklist, read the following descriptors.	Flotected by copyright, illicituding for uses related
		e descriptor that sounds <u>most</u> like the study you were evaluating with respect to a given item. If s no documentation about any specific aspect of an item; then you must evaluate assuming that it	u by
		t done. Given the diversity in clinical measurement properties and design options, the evaluator	0
		make judgments using the criteria below and extend the principles to specific aspects that may	yilg
ı	not be	covered in these brief exemplars. In many cases, the study will not look exactly like the	ί,
		otor so there will be some interpretation as to which level of optimal methods for clinical	Ciuc
		rement studies have been achieved. In such cases, the evaluator can use the general approach	
		this study research design and conduct is consistent with best practice (score=2); is acceptable	2
	out sut (score=	poptimal (score=1); is not done/documented, substantially inadequate or inappropriate =0).	uses uses
'	(30010		161
			נפט נס נפ
			lo lext a
		Descriptors	
Stud	ly ques		lo lext a
		stion	to text and data milling,
Scor		stion	to text and data milling,
Scor	e	stion	to text and data milling,
Scor	e	stion	to text and data milling,
Scor	e	stion	to text and data milling,
Scor	e	stion	to text and data milling,
Scor	e	stion	to text and data milling,
Scor	e	stion	to text and data milling,
Scor Scor	e		to text and data milling,
Scor	re 2	stion	to text and data milling,
Scor	2 2	The authors:  - performed a thorough literature review indicating what is currently known, and not known, about the clinical measurement properties of the instruments or tests under study - presented a critical, and unbiased view of what is known about the current measurement properties - indicated how the current research question fills a gap in the current knowledge base - established a research question based on the above.  All of the above criteria were not fulfilled, but a sound rationale was provided for the research question	to text and data milling,
Scor	2 2	The authors:  - performed a thorough literature review indicating what is currently known, and not known, about the clinical measurement properties of the instruments or tests under study - presented a critical, and unbiased view of what is known about the current measurement properties - indicated how the current research question fills a gap in the current knowledge base - established a research question based on the above.  All of the above criteria were not fulfilled, but a sound rationale was provided for the research question previous literature.	to text and data milling,
Scor	1 0	The authors:  - performed a thorough literature review indicating what is currently known, and not known, about the clinical measurement properties of the instruments or tests under study - presented a critical, and unbiased view of what is known about the current measurement properties - indicated how the current research question fills a gap in the current knowledge base - established a research question based on the above.  All of the above criteria were not fulfilled, but a sound rationale was provided for the research question previous literature.	to text and data milling,
Scor	1 0	The authors:  - performed a thorough literature review indicating what is currently known, and not known, about the clinical measurement properties of the instruments or tests under study - presented a critical, and unbiased view of what is known about the current measurement properties - indicated how the current research question fills a gap in the current knowledge base - established a research question based on the above.  All of the above criteria were not fulfilled, but a sound rationale was provided for the research question previous literature.	to text and data milling,
Scor	1 0	The authors:  - performed a thorough literature review indicating what is currently known, and not known, about the clinical measurement properties of the instruments or tests under study - presented a critical, and unbiased view of what is known about the current measurement properties - indicated how the current research question fills a gap in the current knowledge base - established a research question based on the above.  All of the above criteria were not fulfilled, but a sound rationale was provided for the research question previous literature.	to text and data milling,

	2	Specific inclusion/exclusion criteria for the study were defined, that described the patients enrolled. The subjects were described in terms of health condition/demographics, key relevant outcome mediators and the recruitment context (setting).
-	1	Some information on participants and place is provided (not all of above). For example, age/sex/diagnosis and the name or type of the practice is listed; but no additional information.
-	0	No information on type of clinical settings or study participants is provided (other than number/mean age).
	2	Specific hypotheses or research questions are provided. The stated study purpose provides specific research questions or hypotheses that indicate which specific measurement properties will be evaluated. This should include the specific type of reliability (intra/inter-rater or test-retest) being tested or the type of validity (construct/criterion/content; longitudinal/concurrent; convergent/divergent) being tested. A prior hypothesis should describe the level of reliability expected; and for validity, expected relationships (strength of associations) or constructs.
-	1	The types of reliability and validity being tested were apparent in the methods/title, but clear and specific research questions or hypotheses were not specified
-	0	Specific types of reliability or validity under evaluation were not clearly defined nor were specific hypotheses on reliability and validity stated. ("The purpose of this study was to investigate the reliability of the purpose of this study was to investigate the reliability of the purpose of this study was to investigate the reliability of the purpose of this study was to investigate the reliability of the purpose of this study was to investigate the reliability of the purpose of this study was to investigate the reliability of the purpose of this study was to investigate the reliability of the purpose of this study was to investigate the reliability of the purpose of this study was to investigate the reliability of the purpose of this study was to investigate the reliability of the purpose of this study was to investigate the reliability of the purpose of this study was to investigate the reliability of the purpose of this study was to investigate the reliability of the purpose of this study was to investigate the reliability of the purpose of the purpo
		hypotheses on reliability and validity stated. ("The purpose of this study was to investigate the reliability of and validity of" can be rated as zero if no further detail on the types of reliability and validity or the nature of specific hypotheses is stated).
	2	An appropriate scope of clinical measurement properties would be indicated by  1. A detailed focus on reliability that included multiple forms of reliability (at least two of – intrarater, inter-rater, test retest); as well as both relative and absolute reliability (e.g., ICCs and SEM/MID or limits of agreement)  2. A detailed focus on validity that included multiple forms of validity (content (judgmental); structured (e.g., expert review/survey, qualitative interviews, ICF linking) or structural (e.g., factor analyses or Rasch), construct (known group differences; convergent/divergent associations), criterion (concurrent/predictive), responsiveness; predictive, evaluative or discriminative properties were established  3. Three or more indicators of reliability and validity were examined concurrently and provide a rice.
	2	An appropriate scope of clinical measurement properties would be indicated by  1. A detailed focus on reliability that included multiple forms of reliability (at least two of – intrarater, inter-rater, test retest); as well as both relative and absolute reliability (e.g., ICCs and SEM/MID or limits of agreement)  2. A detailed focus on validity that included multiple forms of validity (content (judgmental); structured (e.g., expert review/survey, qualitative interviews, ICF linking) or structural (e.g., factor analyses or Rasch), construct (known group differences; convergent/divergent associations), criterion (concurrent/predictive), responsiveness; predictive, evaluative or discriminative properties were established  3. Three or more indicators of reliability and validity were examined concurrently and provide a rice.

53		BMJ Open
5	2	Authors performed a sample size calculation and obtained their recruitment targets. Post-doc power analyses and/or confidence intervals confirm that the sample size was sufficient to define relatively precise estimates of reliability or validity.
	1	The authors provide an acceptable rationale for the number of subjects included in the study, but did no
		present specific sample size calculations or post-doc power analyses (or had a sample >100 but no justification).
	0	Size of the sample was not rationalized or is clearly underpowered.
<u> </u>	2	
	1	90% or more of the patients enrolled for study were re-evaluated.  70% or more of the enrolled patients were re-evaluated.
	0	
Иe	asurem	Less than 70% of the patients enrolled in the study were re-evaluated  nents
7	2	
		Documentation is provided for how the studied test is performed. This includes adequate description of the measure/test and how it is administered or scored. The authors may provide or reference a
		published manual/article that outlines specific procedures for administration, scoring (including scoring
		algorithms, handling of missing data) and interpretation that included any necessary information about
		positioning/active participation of the client, any special equipment required, calibration of equipment in
		necessary, training required, cost, examiner procedures/actions. If no manual is available, then the text
		describes key details of procedures in sufficient detail so they could be replicated.
		describes key details of procedures in sufficient detail so they could be replicated.
	1	The test(s) and its administration procedures are referenced; but there is inadequate description of the test procedures.
		test procedures.
	0	Minimal description of test procedures without appropriate references.
		Minimal description of test procedures without appropriate references.
		For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml
		For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

	BMJ Open	Page 48 of
2	This item addresses the overall study procedures for administering all study measures (study and its comparators) in an unbiased way. Test procedures should not introduce systematic er estimation of the clinical measurement properties. This includes standardized procedures for completed or administered the measures. For self-report, this includes order of presentation, completed at what time interval; handling of missing items. If relevant, then the paper should how cultural literacy issues were handled (e.g., exclusion, assisted or surrogate completion). I impairment measures, procedures would include calibration of any equipment; use of consist measurement tools and scoring, a priori exclusion of any participants likely to give invalid rest to complete testing (not exclusion of after enrollment); use of standardized instructions and the procedures. This can include order of administration of test and quality checking of scores. For testing, the appropriate retest interval will depend on the nature of the condition; but for accomplication is the majority require retesting within 48 hours; whereas chronic/stable conditions are conditions it may require retesting within 48 hours; whereas chronic/stable conditions are conditions it may require retesting within 48 hours; whereas chronic/stable conditions are conditions it may require retesting within 48 hours; whereas chronic/stable conditions are conditions it may require retesting within 48 hours; whereas chronic/stable conditions are conditions it may require retesting within 48 hours; whereas chronic/stable conditions are conditions.	rors in the who who who dinclude For Protected by control of test with the control of the contro
1	No obvious sources of bias in the study test protocol or how tests were performed/administeral apparent; but there were suboptimal procedures or an inadequate description of the measure protocol to be insured control of bias or that procedures were standardized.	red is
0	No description of the overall procedures for administering study tests; OR an obvious source data collection methods.	of bias in to
nalyses		nd da
2	Authors clearly defined which specific analyses were conducted for each of the stated specific hypotheses/questions of the study. This may be accomplished through organization of the respecific subheadings or by demarcating which analyses addressed specific clinical measurement properties. Data was presented for each hypothesis/research question posed.  Data was presented that addressed each of the measurement questions posed, but authors of the stated specific analyses were conducted for each of the stated specific hypotheses/question of the stated specific hypotheses/question of the respective specific clinical measurement properties.	sults under 📆 .
1	Data was presented that addressed each of the measurement questions posed, but authors of specific analyses to specific research questions or hypotheses.	did not linkg, and si
0	Data was not presented for every hypothesis or clinical measurement property outlined in the or methods.	e purposes milar tech
		nologies.
	For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml	

3		BMJ Open
10 2	2	<u>Tests selected</u> - Appropriate statistical tests were conducted to calculate a point estimate for clinical measurement properties. Examples are provided below; but are not exhaustive.
		1. Reliability (Relative=ICCs (Shrout & Fleiss, 1979) for quantitative, Kappa (Landis & Koch, 1977) for nominal data); absolute (SEM or plot of score differences vs. average score showing mean and 2SD limit – as per Altman and Bland) (Bland & Altman, 1986; Bland & Altman, 1987)
		2. Clinical relevance - minimal detectable change, clinically important difference (Jaeschke, Singer, & Guyatt, 1989; Beaton et al., 2001; Wells et al., 2001)  3. Validity  a. Validity associations - Pearson correlations for normally distributed data, Spearman rank correlations for ordinal data; or other correlations, if appropriate
		3. Validity
		a. Validity associations - Pearson correlations for normally distributed data, Spearman rank correlations for ordinal data; or other correlations, if appropriate
		b. Validity tests of significant difference - an appropriate global test like analysis of variance was used where indicated, with post-hoc tests that adjusted for multiple testing
		c. Validity of items scaling/responses - Rasch analysis or item response (Baylor et al., 2011; Pallant & Tennant, 2007; Kyngdon, 2006; Cipriani, Fox, Khuder, & Boudreau, 2005; Smith, Jr., Conrad, Chang, & Piazza, 2002)
		4. Responsiveness (Beaton, Bombardier, Katz, & Wright, 2001)- standardized response means or effect sizes or other recognized responsiveness indices were used.
-	1	Appropriate statistical tests were used in some instances; but suboptimal choices were made in other analyses.
(	0	Inappropriate use of statistical tests - incorrect tests for type of data; or a lack of analysis
11 2	2	The study goes beyond a single statistical point estimate of a clinical measurement property and providing supporting statistical analyses that increases confidence in the findings in terms of precision of the (key) indicator; or provide an alternate form of analysis of the clinical measurement property. The evaluator decides if these analyses are appropriate and informative. For example, with reliability, at least 2 of the following would constitute appropriate and informative analysis beyond a point estimate a reliability coefficient: 1. confidence intervals around the point estimate; 2. Comparison to appropriate, referenced benchmarks or standards; or 3. SEM or MDC. For correlations, tests of significance or confidence intervals were presented and indicators of the criterion benchmarks were provided. For studies involving cross-cultural validation, the analyses should compare multiple clinical measurement properties previously established for the measure and explain the extent to which the translated version is in accordance with these previously reported properties on the source measure.
		For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

	BMJ Open	Page 50 of !
		-
1	Either precision definition (confidence intervals) or appropriate benchmark comparison were used both. OR Some analyses were associated with indicators of precision or alternate form of analysis not all key indicators.	
0	Inappropriate use of benchmarks or confidence intervals; or indicators of precision or alternate fo absent	rm are
Recomme	Authors made specific conclusions and clinical measurement recommendations that were clearly in	otect
.2 2	Authors made specific conclusions and clinical measurement recommendations that were clearly to each hypotheses/question posed in the study and that were supported by the data presented. recommendations would state the estimated status of the clinical measurement property, the confidence in the estimate and the context for which those apply. To achieve a 2, the conclusion be specific; and conclusions cannot overstate the clinical measurement properties observed the state of the context for which those apply.	Ideal copyright,
1	Authors made conclusions and clinical measurement recommendations that were basically true (supported by study data); but vague. That is, they do not specify the extent, confidence or contex the findings. (The measure is "reliable and valid") OR authors made specific clinical measurement recommendations; but for only some of the study hypotheses.	Enseignem
0	Authors did not make conclusions about clinical measurement; OR made recommendations that we contradiction to the actual data presented	rere in <b>o</b> te sup
		(ABES) . ta mining, Al training, and similar technologies.
	For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml	

1. Abbott et al 2014	Ineligible population
2. Beattie et al 2011	Ineligible population (less than 50%)
3. Hoeskstra et al 2014	No properties for GRoC scales
4. Chansirinukor 2019	No properties for GRoC scales
5. <u>Chien et al 2015</u>	No properties for GRoC scales
6. <u>Cruz et al. 2015</u>	No properties for GRoC scales
7. Foroutani et al 2018	No English (Persian language)
8. Gagnon et al 2018	Ineligible population
9. Hefford et al 2012	Ineligible population
10. Hung et al 2019	Ineligible population
11. <u>Sharma et al 2017</u>	Ineligible population
12. Stevens et al 2019	Ineligible population
13. <u>Meyer et al 2014</u>	Ineligible population

Page 52 of 53



47

# PRISMA 2009 Checklist

- 3		Jht,	
Section/topic	#	Checklist item	Reported on page #
TITLE		g of N	
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT		is re	
Structured summary 13 14	2	Provide a structured summary including, as applicable: background; objectives; data south structured summary including, as applicable: background; objectives; data south structured summary including, as applicable: background; objectives; data south structured summary including, as applicable: background; objectives; data south structured summary including, as applicable: background; objectives; data south structured summary including, as applicable: background; objectives; data south structured summary including, as applicable: background; objectives; data south structured summary including, as applicable: background; objectives; data south structured summary including, as applicable: background; objectives; data south structured summary including, as applicable: background; objectives; data south structured summary including, as applicable: background; objectives; data south structured summary including, as applicable: background; objectives; data south structured summary including, as applicable: background; objectives; data south structured summary including, as applicable: background; objectives; data south structured summary including, as applicable: background; objectives; data south structured summary including structured summary inclu	2
15 INTRODUCTION		xt a a a a a a a a a a a a a a a a a a a	
16 17 Rationale	3	Describe the rationale for the review in the context of what is already known.	3-5
18 Objectives 19	4	Provide an explicit statement of questions being addressed with reference to participants hererventions, comparisons, outcomes, and study design (PICOS).	4-5
METHODS		ng,	
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.	5
24 25 Eligibility criteria 26	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.	5
27 Information sources 28	7	Describe all information sources (e.g., databases with dates of coverage, contact with story authors to identify additional studies) in the search and date last searched.	6
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	Appendix1
32 Study selection 33	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic very, and, if applicable, included in the meta-analysis).	6
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in diplicate) and any processes for obtaining and confirming data from investigators.	6-7
Data items 38	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and simplifications made.	6-7
Risk of bias in individual	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.	6-7
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	8-9
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I²) for each meta-analysis.  For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml	8-9

BMJ Open



46 47

# **PRISMA 2009 Checklist**

Page 53 of 53		BMJ Open	
PRISMA 20	09	Checklist  Page 1 of 2  Cted by copyright, in	
3 4		Page 1 of 2 ; ; 33 ; ; 9	
Section/topic	#	Checklist item 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).	8-9
10 Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-region), if done, indicating which were pre-specified.	8=9
13 RESULTS		men to	
14 Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, wৰ্ছাদুৰ্ভ asons for exclusions at each stage, ideally with a flow diagram.	9
17 Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, Pich (e.g., follow-up period) and provide the citations.	9-10
19 Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessme	10
Results of individual studies 22	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple suntin data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot	10-12
23 Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measure of consistency.	13
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	10
26 Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-	13
28 DISCUSSION	<u> </u>	simi	
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).	14-15
32 Limitations 33	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).	16
Conclusions		Provide a general interpretation of the results in the context of other evidence, and implications for future research.	14-15
36 FUNDING			
38 Funding 39 40	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

40
41 From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The GRISMA Statement. PLoS Med 6(7): e1000097.
42 doi:10.1371/journal.pmed1000097
43 For more information, visit: www.prisma-statement.org.
44 Page 2 of 2
45 For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml
46

# **BMJ Open**

# Psychometric Properties of the Global Rating of Change Scales in Patients with Neck Disorders: A Systematic Review with Meta-Analysis and Meta-Regression

Journal:	BMJ Open
Manuscript ID	bmjopen-2019-033909.R2
Article Type:	Original research
Date Submitted by the Author:	28-Oct-2019
Complete List of Authors:	Bobos, Pavlos; Western University, Health and Rehabilitation Sciences; University of Toronto, Institute of Health Policy Management and Evaluation MacDermid, Joy; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Nazari, Goris; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Furtado, Rochelle; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Group, CATWAD; Michele Sterling, Anne Söderlund, Michele Curatolo, James M Elliott, David M Walton, Helge Kasch, Linda Carroll, Hans Westergren, Gwendolen Jull, Eva-Maj Malmström, Luke B Connelly, Joy C MacDermid, Mandy Nielsen, Pierre Côté, Tonny Elmose Andersen, Trudy Rebbeck, Annick Maujean, Sarah Robins, Kenneth Chen, Julia Treleaven
<b>Primary Subject Heading</b> :	Rehabilitation medicine
Secondary Subject Heading:	Rehabilitation medicine
Keywords:	neck pain, global assessment, psychometric properties, systematic review

SCHOLARONE™ Manuscripts

- 1 Psychometric Properties of the Global Rating of Change Scales in Patients with Neck
- 2 Disorders: A Systematic Review with Meta-Analysis and Meta-Regression
- 3 Pavlos Bobos<sup>1</sup>, Joy C MacDermid<sup>2</sup>, Goris Nazari<sup>3</sup>, Rochelle Furtado<sup>4</sup> and CATWAD co-authors<sup>5</sup>
- <sup>1</sup>Pavlos Bobos PT, PhD(c), (corresponding author) Doctoral Candidate, Western's Bone and Joint
- 6 Institute, Department of Health and Rehabilitation Sciences, Western University, Elborn College,
- 7 1201 Western Road, N6G 1H1, London, Ontario, Dalla Lana School of Public Health, Institute of
- 8 Health Policy Management and Evaluation, Department of Clinical Epidemiology and Health Care
- 9 Research, University of Toronto, Canada, (pbobos@uwo.ca), tel: +1 519 661 2111 x88912
- <sup>2</sup>Joy C MacDermid BScPT, PhD, Professor, Physical Therapy and Surgery, Western University,
- London, ON and Co-director Clinical Research Lab, Hand and Upper Limb Centre, St. Joseph's
- 12 Health Centre, London, Ontario; Professor Rehabilitation Science McMaster University,
- Hamilton, ON, Canada (<u>imacderm@uwo.ca</u>)
- <sup>3</sup>Goris Nazari PT, PhD(c) Doctoral Candidate, Western's Bone and Joint Institute, School of
- 15 Physical Therapy, Department of Health and Rehabilitation Sciences, Western University,
- 16 London, Ontario, Canada, (gnazari@uwo.ca)
- <sup>4</sup>Rochelle Furtado MSc Western's Bone and Joint Institute, School of Physical Therapy,
- Department of Health and Rehabilitation Sciences, Western University, London, Ontario, Canada,
- 19 (<u>rfurtad5@uwo.ca</u>)
- <sup>5</sup>CATWAD: Michele Sterling <u>m.sterling@uq.edu.au</u>, Anne Söderlund <u>anne.soderlund@mdh.se</u>,
- 21 Michele Curatolo, curatolo@uw.edu, James M Elliott j-elliott@northwestern.edu, David Walton
- dwalton5@uwo.ca, Helge Kasch helgkasc@rm.dk, Linda Carroll linda.carroll@ualberta.ca,
- Hans Westergren <u>Hans.Westergren@skane.se</u>, Gwendolen Jull <u>g.jull@uq.edu.au</u>, Eva-Maj
- 24 Malmström eva-maj.malmstrom@med.lu.se, Luke B Connelly l.connelly@uq.edu.au, Joy C
- 25 MacDermid <u>imacderm@uwo.ca</u>, Mandy Nielsen <u>mandy.nielsen@griffith.edu.au</u>, Pierre Côté
- 26 <u>pierre.cote@uoit.ca</u>, Tonny Elmose Andersen <u>tandersen@health.sdu.dk</u>, Trudy Rebbeck
- 27 <u>trudy.rebbeck@sydney.edu.au</u>, Annick Maujean <u>a.maujean@uq.edu.au</u>, Sarah Robins
- 28 s.robins1@uq.edu.au, Kenneth Chen k.chen8@uq.edu.au, Julia Treleaven j.treleaven@uq.edu.au
- **Kewords:** neck pain, global assessment, psychometric properties, systematic review
- **30 Word count: 3908**

- Objective: The purpose of this systematic review was to critically appraise and synthesize the psychometric properties of Global Rating of Change (GROC) scales for assessment of patients
- with neck pain.
- **Design:** Systematic review
- 36 Data sources: A search was performed in 4 databases (MEDLINE, EMBASE, CINAHL,
- 37 SCOPUS) until February 2019.
- **Data extraction and synthesis:** Eligible articles were appraised using Consensus-based Standards
- 39 for the selection of health Measurement Instruments (COSMIN) checklist and the Quality
- 40 Appraisal for Clinical Measurement Research Reports Evaluation Form.
- **Results:** The search obtained 16 eligible studies and included in total 1533 patients with neck pain.
- 42 Test-retest reliability of Global Perceived Effect (GPE) was very high (Intra-class correlation
- coefficient (ICC) = 0.80 to 0.92) for patients with whiplash. Pooled data of Pearson's r indicated
- that GROC scores were moderately correlated with neck disability change scores (0.53, 95% CI:
- 45 0.47 to 0.59). Pooled data of Spearman's correlations indicated that GROC scores were moderately
- 46 correlated with neck disability change scores (0.56, 95% CI: 0.41 to 0.68).
- **Conclusions:** This study found excellent quality evidence of very good to excellent test-retest
- 48 reliability of GPE for patients with Whiplash Associated Disorders. Evidence from very good-to-
- 49 excellent quality studies found that GROC scores are moderately correlated to an external criterion
- patient-reported outcome (PROM) measure evaluated pre-post treatment in patients with neck
- pain. No studies were found that addressed the optimal form of GROC scales for patients with
- neck disorders or compared the GROC to other options for single-item global assessment.
- Prospero registration number: CRD 42018117874

# Strengths and limitations of this study

- We rated the quality of individual studies and the overall risk of bias using two standardized approaches
- Our focus on neck pain increased the specificity of results but are not necessarily
   applicable to other musculoskeletal conditions
- Conceptual concerns about global ratings of change being affected by recall bias are not adequately addressed by psychometric evidence
- No studies addressing the optimal form of global rating were found.

#### Introduction

Neck pain is the 4<sup>th</sup> leading cause of disability and approximately half of adult the population with neck pain will experience a clinically important episode once in their lifetime. [1–3] The annual prevalence of neck pain it is estimated between 15% and 50%, with females having a higher prevalence rate than males. [2,3] Neck pain has been associated with many other comorbidities such as headaches, dizziness, anxiety, depression, back pain and arthralgias.[3–6] Several different methods for classifying neck pain have been described, using indicators such as duration (acute, sub-acute or chronic), degree of interference (low, moderate, severe) or most likely structure at fault (e.g. neuropathy vs. mechanical). [7]

As part of a patient-centric approach to care, clinicians will commonly evaluate response to intervention by asking the patient directly whether they feel better, worse, or the same since the prior encounter. While direct questioning can provide a qualitative indicator of change in status, many best practice guidelines endorse use of some form of quantified patient-reported outcome (PRO) as an adjunct to oral self-report. PROs are available to quantify several different constructs

in people with neck pain, including pain severity, disability and neck function. [8] Any PRO intended to provide an estimate of change over time should be responsive to subtle shifts in the patient's condition. To facilitate interpretation of change scores, a common property of many such tools is the minimum clinically important difference (MCID), which is a change threshold that corresponds to the minimum shift in scale values that most patients would indicate corresponds to an important change in their overall condition. A well-recognized approach to establishing an MCID for a PRO is to compare the magnitude of change against an anchor, most commonly a Global Rating of Change (GROC) scale. These scales allow patients or study participants to indicate whether their condition has gotten worse, better, or stayed the same and to quantify the magnitude of that change. As they have been adopted as a sort of 'standard' against which change in other tools is compared, the GROC can also be used on its own as an omnibus generic indicator of change. [8]

Despite being accepted as a standard measure, there is considerable variation in how the GROC has been constructed and implemented in research in neck pain. GROC scales consist of ordered categories which may have different ranked levels (some have 15 levels, some 11 levels, and others have 7 levels). The common structure across these is the use of a middle '0' score corresponding to 'no change', with negative values indicating magnitudes of worsening while positive values indicate improvement.[9] Variations of the GROC (in name or structure) include the "Global Perceived Effect", "Patient Global Impression of Change", "Transition Ratings", and "Global Scale". [9]

A well-established component of health outcomes is having a tool with strong psychometric properties of validity, reliability and responsiveness to be able to monitor change.

While recent research [8] has examined the psychometric properties of the most commonly

59

60

1 2						
3 4	102	reported PROs for neck disorders, to date there has been no systematic review to summarize the				
5 6	103	measurement properties of GROC scales themselves in patients with neck disorders. Therefore,				
7 8	104	this systematic review aims to critically appraise and synthesize the psychometric properties of the				
9 10 11	105	GROC scales in patients with neck disorders.				
12 13	106					
14 15	107	METHODS				
16 17	108	Patient and Public Involvement				
18 19 20	109	There was no patient or public involvement in the design or planning of this study.				
21 22	110					
23 24 25	111	Study Design and Protocol Registration				
26 27	112	We conducted a systematic review to evaluate the psychometric properties of GROC scales in				
28 29	113	patients with neck disorders. The protocol was registered in PROSPERO register database with				
30 31	114	registration number: CRD 42018117874				
32 33 34	115					
35 36	116	Eligibility Criteria				
37 38	117	We included studies in this systematic review if the following criteria were met [10–12]:				
39 40 41	118	Design: psychometric testing, randomized/ cohort studies				
42 43	119	• Participants: > 50% of the study's patient population with neck conditions/disorders,				
44 45	120	• Intervention/Comparison: studies that reported on the psychometric properties (reliability,				
46 47	121	validity, responsiveness) of GROC, Global Perceived Effect (GPE) and Patient Global				
48 49 50	122	Impression of Change (PGIC),				
51 52	123	Outcomes: GROC, GPE and PGIC				
53 54 55 56	124	Articles were written in English language only				
<b>E</b> 7						

Studies with no data on the GROC scales' psychometric properties, and conference abstract/posters were excluded from this systematic review.

## Information Sources

To identify studies on the psychometric properties (reliability, validity, responsiveness) of the GROC, GPE and PGIC we searched the Medline, EMBASE, Scopus and CINAHL databases from inception till February 2019, using a combination of keywords. Furthermore, we identified additional studies by examining the reference list of each of the selected studies. The full list with keyword strategy is presented in **APPENDIX 1**.

# Study Selection

Two investigators (PB and GN) performed the systematic electronic searches independently in each database. The same investigators then proceeded to identify and remove the duplicate studies. In the next stage, we performed the independent screening of the titles and abstracts and any full-text article marked as include or uncertain were obtained. In the final stage, the same two independent authors performed the full text reviews independently to assess final article eligibility. In case of disagreement, a third reviewer; the most experienced member (JM), facilitated a consensus through discussion.

#### Data Extraction

The fourth author (RF) performed the data extractions. The extracted data were then cross-checked by another author (PB). Data extraction included the author, year, study population/condition, setting, sample size, age, properties evaluated, retest-interval, and the intervention protocol (if used

to assess responsiveness parameters). [13,14] For reliability estimates, Standard Error of Measurement (SEM), Intra-class Correlation Coefficient (ICC), Minimal Detectable Change (MDC) and 95% confidence intervals were extracted. [13,14] The ICC interpretation of ICC < 0.40 indicating poor, 0.40 ≤ ICC < 0.75 indicating fair-to-good and ICC ≥ 0.75 indicating excellent reliability were used as a common benchmark.[15] For validity estimates, correlation coefficient (Pearson's/Spearman) and the 95% confidence intervals were extracted. [13,14] Evan's guidelines to interpret the strength of the correlation was used which included: 0.00–0.19 "very weak", 0.20– 0.39 "weak", 0.40–0.59 "moderate", 0.60–0.79 "strong", and 0.80–1.00 "very strong". [16] For responsiveness estimates, the Effect Size (ES), Standardized Response Mean (SRM), Clinically Important Difference (CID), and/or Minimal Clinically Important Difference (MCID) including the method of MCID estimation – Anchor-/Distribution-based methods, and 95% confidence intervals were extracted. [13,14] To assist clinical decision making, standard benchmark scores of trivial (< 0.20), small ( $\geq$  0.20 to < 0.50), moderate ( $\geq$  0.50 to < 0.80) or large ( $\geq$  0.80), as proposed by Cohen, were used. [17] When insufficient data were presented, PB contacted the authors by email and requested further data.

164 Consensus-based Standards for the selection of health Measurement Instruments (COSMIN)

Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) assesses the risk of bias for the psychometric properties reported on a property-by-property basis. A score for the risk of bias in estimates of psychometric properties was assessed by two authors (PB) and (RF) using the new (COSMIN) checklist.[18] If disagreement was present a third person (JM) assist in resolving the discrepancy. Each study was assessed by COSMIN on the 4-point scale

as "very good", "adequate", "doubtful" or "inadequate" for each of the checklist criteria for

relevant measurement properties (e.g. reliability, responsiveness, etc.). According to COSMIN, when determining the overall score for each measurement property, the worst score counts method was used wherein the lowest score for the checklist criteria of the relevant property was taken as the overall score. [19] We then assessed the result of individual studies on a measurement property against the updated criteria for good measurement properties. This involved the evaluation of results of included studies as either sufficient (+), insufficient (-), or indeterminate (?). [18] Quality Appraisal for Clinical Measurement Research Reports Evaluation Form

A summary score for the overall quality of individual studies was appraised independently by the authors (PB) and (RF) using a structured clinical measurement specific appraisal tool. [13,14] In case of disagreement a third person was consulted (JM) to resolve the conflict. The evaluation criteria of this tool included twelve items: 1) Thorough literature review to define the research question; 2) Specific inclusion/exclusion criteria; 3) Specific hypotheses; 4) Appropriate scope of psychometric properties; 5) Sample size; 6) Follow-up; 7) The authors referenced specific procedures for administration, scoring, and interpretation of procedures; 8) Measurement techniques were standardized; 9) Data were presented for each hypothesis; 10) Appropriate statistics-point estimates; 11) Appropriate statistical error estimates; and 12) Valid conclusions and recommendations. [13,14] An article's total score – quality - was calculated by the sum of scores for each item, divided by the numbers of items and multiplied by 100%. [13,14] Overall, the quality summary of appraised articles range from (0%-30%) Poor, (31%-50%) Fair, (51%-70%) Good, (71%-90%) Very Good, and (>90%) Excellent. [13,14]

Synthesis of Results

A qualitative synthesis was conducted to report findings on test-retest reliability statistics. A meta-analysis of Pearson's and Spearman's correlation was performed in R (version 3.6.1) with metaphor package. [20] The meta-analyses were conducted using a random effect model and the correlation coefficients were converted to z values. Heterogeneity was deemed substantial if I<sup>2</sup> values were more than 50%. [21] A Meta-regression was planned to explore the sources of unexplained heterogeneity by considering the following factors: a. neck pain with or without radicular symptoms, b. acute or chronic, c. age and d. sex. Forest plots were created using means and 95% confidence intervals for correlation coefficients. We summarize the main results of the included articles based on the neck disorders, reported psychometric estimate and the study quality ratings. 

#### **RESULTS**

Study Selection

Our search yielded 8,837 articles. After removal of duplicates, 6,027 studies remained and were screened using their title and abstract; leaving 29 articles selected for full-text review. Of these, 16 studies were considered eligible. [22,23,24–31,32–37] The flow of the study selection process is presented in Figure 1.

O. C.

# Study Characteristics

The 16 eligible studies were conducted between 2006 and 2017 and included 1533 participants with neck pain/disorders (mean of 96 participants per study). [22,23,24–31,32,34–37,] Study size ranged from 29 to 200 participants. A summary description of all the studies included is displayed in **Table 1.** Concurrent validity was evaluated in 14 studies by comparing the difference of pain

intensity, disability and function scores with the score of GROC scales. Two studies [26,31] examined the test-retest reliability of a 7-point and an 11-point GPE scale for patients with whiplash-associated disorders (WAD). One study [24] examined whether occurrences of within-and between-session changes were significantly associated with functional outcomes, pain, and self-report of recovery in patients at discharge who were treated with manual therapy for mechanical neck pain.

COSMIN Risk of Bias rating and Quality appraisal of the Included Studies

Regarding the risk of bias, all studies were rated as very good (**Table 2**). The quality of the studies ranged from 88% to 96% (**Table 3**). The most common flaws were 1) lack of/inadequate sample size calculations, 2) missing data (i.e. inadequate follow up), and 3) inconsistencies between the data presented and hypothesis stated.

### Reported GROC scales

The most commonly reported GROC scale (n=6 studies) was a 15-point scale with the most frequent anchors being "-7 (a very great deal worse) to zero (about the same) to +7 (a very great deal better)". A 7-point scale was reported in 5 studies, 11- and 5-point scales were reported in 2 studies and a 9-point scale in one study. The anchors in those scales varied greatly and are presented in Table 1. Only 6 studies [26,31–33,35,36] reported full detail regarding the specific questions asked of the patients with neck disorder when a GROC scale was administered. Those questions that were reported are presented in **Box 1**.

Two studies were included that examined test-retest reliability of GPE for patients with WAD. Kamper et al. (2010) [26] examined the [time interval] test-retest reliability of an 11-point GPE scale in 134 patients with chronic WAD and reported an Intra-class Correlation Coefficient (ICC) of 0.99 (95% CI 0.99 to 0.99) at baseline, 0.96 (0.95 to 0.97) at 6 weeks, and 0.92 (0.89 to 0.94) at 12 months (**Table 4**). Ngo et al. (2010) assessed the test-retest reliability of a 7-point scale of GPE in patients with acute WAD at 3 to 5 days. [31] The ICC and 95% confidence intervals (CI) were used to determine the test-retest reliability of the two versions of the perceived recovery questions using their original seven-item responses. Ngo et al. also computed weighted kappa coefficients and 95% CI using quadratic weights to determine whether the distribution of responses influenced the reliability as measured by the ICC. An ICC for general recovery of 0.70 (0.60 to 0.80) and an ICC for neck pain questions of 0.80 (0.72 to 0.87) were found. A weighted Kappa was also calculated (Kappa = 0.70 (0.42 to 0.98)) at six weeks for general recovery and at six weeks Kappa = 0.80 (0.51 to 1.0) for neck pain questions (**Table 4**).

# Validity Measures

We found 14 studies that examined concurrent validity measures between GROC and another PRO.[22,23,25,27–30,32,34,35,36–38] Correlations of Pearson's and Spearman's coefficients between GROC and another PRO were ranging from very weak to very strong correlations. The validity measures are presented and summarized in Table 5.

Meta-Analysis and Meta-Regression of Correlations between Disability change scores and GROC

Five studies [23,25,34,37,38] of very good-to-excellent quality reported the Pearson correlation coefficients between neck disability change scores and the GROC scores and were pooled together. We found that GROC was positively correlated with disability change scores (r=0.53, 95% CI: 0.47 to 0.59,  $I^2=0\%$ ). Six studies [27–30,32,36] of very good-to-excellent quality reported the Spearman correlation coefficients between neck disability changes scores and the GROC scores and were pooled together. We found that GROC was moderately correlated with disability change scores (rho = 0.56, 95% CI: 0.41 to 0.68,  $I^2=85\%$ ). The forest plots with correlation coefficients with 95% CIs are presented in Figure 2-3. Our meta-regression showed that age was found as a significant factor in influencing Fisher's Z scores ( $\beta=-0.034, 95\%$  CI -0.05 to -0.01, p=0.001). The model explained 68% of the variance ( $I^2=0.68$ ) (Figure 4).

Cook et al. [24] found that between-session NPRS- pain changes were associated with greater than 3-point change on the GROC at 96-hours (AUC=0.76). The pain change associated with GROC was more specific (Specificity=79.2%, range: 62.2 - 91.1) than sensitive (Sensitivity=65.6%, range: 57.9 to 74.6). Those with a 36.7% between-sessions change in pain were also 7.3 times more likely to report an improvement of greater than 3 points change on the GROC than those who did not achieve a 36.7% change in pain (**Table 4**).

**DISCUSSION** 

This review has synthesized the current research from 16 studies that aimed to evaluate the psychometric properties of GROC scales for patients with neck disorders, with the goal to provide evidence for clinicians and researchers concerning its use within clinical practice and research. From the 16 included studies, only 2 studies [26,31] reported test-retest reliability statistics of the 7- and 11-ranked categories of GPE scales for patients with WAD only. We were able to pool data from 12 studies regarding concurrent validity of GROC scales and neck disability change scores at one time point after the interventions. Themes influencing interpretation of the GROC were explored in a study [33] that evaluated the factors that contribute to how patients respond to a question on global perceived effect. This study found that treatment process, biomechanical performance, self-efficacy and the nature of the condition may influence the responses on global perceived effect, which is consistent with what we would expect for patients with neck pain. This suggests that change is a complex multifactorial global concept. A strength of GROC is that it is intended as a global assessment, and it can be assumed that it reflects the aspects of change important to the individual patient.

Reliability can be defined as the degree to which a measure produces consecutive results with the least amount of random error when the status of the population remains unchanged. The reliability of GPE displayed an excellent test-retest reliability of ICC>0.90 over an interval of 6 weeks and 12 months for patients with WAD. Conducting an assessment with a long test-retest interval (e.g. 12 months), can provide challenges as there is higher risk of individuals with WAD being symptomatically unstable.[9] Determining if patients are symptomatically-stable can be achieved by administering another PRO such as the Single Assessment Numeric Evaluation (SANE)[39], however, the 7- and 11- ranked categories of GPE scales still demonstrated good stability properties at long test intervals (i.e., of 6 weeks and 12 months).[26] Therefore, the

measurements of the reliability parameters of the GPE may be very useful during longer test intervals in clinical trials.

The psychometric property of validity is defined as the degree to which a PRO measures what it is intended to measure. Pooled data from 11 studies overall suggest that post-treatment changes of on validated disability outcome measures were moderately (Pearson's r = 0.51, 95% CI: 0.43 to 0.58; Spearman's rho = 0.56, 95% CI: 0.41 to 0.68) correlated to change in perceived effect) (Figure 2-3). This finding suggests that GROC scores taken at one point in time were related to scores in pain and disability in patients with neck disorders, as measured by standardized measures taken at 2 points in time. We identified one study [24] that found a 36.7% change in pain for within- and between- session changes was associated with a 50% reduction in the NDI and an improvement of >3 levels on a 15-ordinal level GROC scale for patients with neck pain. This quantified predictive change value may have clinical utility for use in clinical practice.

Previous studies [9,40] have indicated serious concerns about the conceptual validity of the global rating of change. The review by Kamper et al.[9] clearly showed that GROC was related to final status more than change and was least related to baseline health status. This result undermines the premise of what the global rating of change actually measures. For this reason, we conclude that the 0.50 pooled correlation across 12 studies between the GROC and other PROM change scores (e.g. Neck Disability Index (NDI) scores) may reflect a relationship between follow-up status and change rather than supporting the contention that GROC actually measures change. This would also explain why only 25% of the variation in GROC change scores was explained by changes scores from a PROM change score measured at 2 points in time. In all studies, participants completed the GROC scale at one time point after the intervention, and hence recall bias is a cause for concern. However, another potential factor for moderate correlations is that the PROMs that

have been used as the comparator with GROC scores may not reflect priorities that are important to patients. That is, the field has largely been driven by assumptions that the GROC is a 'gold standard' for evaluating true change in a respondent's condition or status, and that all items on the comparator PROM are of equal importance to all people with that condition. The work presented herein challenges the valorization of the GROC as a gold standard for change, and prior work has challenged the notions that all PROM items are equally important.[9,41,42] It is therefore possible that the very constructs being evaluated require greater critical discourse before authors can say, with confidence, that one scale functions well or poorly based on its associations with another scale. Since no studies compared a retrospective global assessment of the GROC to pre-post single item global PROM e.g. the SANE, we do not know the extent to which these two factors contributed to moderate correlation.

A unique aspect of this study was that it focused on global rating of change scales in a neck

A unique aspect of this study was that it focused on global rating of change scales in a neck pain patient population. Our study appraisal suggests that future studies concerning GROC should include adequate sample sizes, maintain a rigorous follow up and report appropriate statistical error estimates, since these were often inadequate. Various critical appraisal tools exist, and the perspectives and ratings may differ across instruments. COSMIN is just one methodology that can be used to synthesize or evaluate outcome measures and other methods might be equally valid or provide different perspectives. We used 2 different critical appraisal tools to evaluate quality from 2 perspectives. The COSMIN risk of bias assessments reflects the level of confidence in the conclusions and pooled estimates. The quality appraisal tool focuses on design issues in the studies and reflects gaps in research designs that should be considered in interpretation of current research and improved in future studies. Substantial heterogeneity was detected (12>50%) in pooled Spearman's correlation coefficients which is a concern when pooling data. Sources of the observed

heterogeneity were identified in our meta-regression results. Our univariate meta-regression analysis indicated that age across the studies explained 68% of the variance (**Figure 4**). Other factors such as type of neck pain (with or without radicular symptoms), acute or chronic and sex did not explain the remaining heterogeneity (not statically significant). In our meta-regression, we used a patient level characteristic to identify the observed heterogeneity and therefore, our model may be vulnerable to aggregation bias. Furthermore, the scope of our literature search was focused on identifying full-text papers written in English only.

While this study included 16 studies, only 2 of these reported reliability statistics for GROC scales for patients with chronic WAD. Therefore, the applicability of our study is mostly limited to patients with chronic WAD. For validity measurements, GROC scales were mostly investigated by correlation analyses to evaluate the external responsiveness of another PRO measure over a specific time point. From our meta-analysis, we can be confident that the GROC scores were moderately correlated with neck disability change scores. However, more robust psychometric design studies to test the measurement properties of GROC scales as the primary outcome of investigation are highly needed. Future studies should aim to test to what extent the different range of items (e.g. 7-level scale vs 11-level scale), the anchors (e.g. much worse vs much better) may affect the measurement properties of GROC scales for patients with neck disorders. Also, it is important to indicate that most outcome measures are ordinal and assume that additive scores of ordinal items can be treated as interval level. This potentially could lead to scaling problems even in the face of strong psychometric properties. The main protection we have is to create new scales or retrofit existing scales based on Rasch analysis. Also, we acknowledge that the majority of work done on the GROC scales has been performed using statistical approaches that are most appropriate to linear rather than ordinal data

#### **CONCLUSIONS**

This study found excellent quality evidence of very good to excellent test-retest reliability of GPE for patients with WAD. Evidence of very good to excellent quality studies found that GROC scores are moderately correlated to an external criterion PROM measure measured pre-post treatment in patients with neck disorders. Studies addressing the optimal form of GROC scales for patients with neck disorders or comparing the GROC to other options for single-item global assessment of change were not found.

## **Authors' contributions**

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

### **Declarations**

## Ethics approval and consent to participate

Not applicable

#### **Consent for publication**

Not applicable

#### Availability of data and material

Data sharing is not applicable to this article as no datasets were generated or analyzed during the

400	curre	ent study
401	Func	ding Statement
402	This	work was supported by the Canadian Institutes of Health Research (CIHR) with funding
403	refer	ence number (FRN: SCA-145102).
404	Com	peting Interest Statement
405	None	e to report
406		
407		
408	Refe	rences
409	1	Murray CJL, Abraham J, Ali MK, et al. The State of US health, 1990-2010: Burden of diseases,
410		injuries, and risk factors. JAMA - J Am Med Assoc Published Online First: 2013.
411		doi:10.1001/jama.2013.13805
412	2	Fejer R, Kyvik KO, Hartvigsen J. The prevalence of neck pain in the world population: A
413		systematic critical review of the literature. Eur. Spine J. 2006. doi:10.1007/s00586-004-0864-4
414	3	Hogg-Johnson S, van der Velde G, Carroll LJ, et al. The Burden and Determinants of Neck Pain in
415		the General Population. Results of the Bone and Joint Decade 2000-2010 Task Force on Neck
416		Pain and Its Associated Disorders. J Manipulative Physiol Ther Published Online First: 2009.
417		doi:10.1016/j.jmpt.2008.11.010
418	4	Bobos P, Nazari G, Palimeris S, et al. The contribution of health and psychological factors in
419		patients with chronic neck pain and disability: A cross-sectional study. J Clin Diagnostic Res
420		2018; <b>12</b> :YC04–7. doi:10.7860/JCDR/2018/31284.11203
421	5	MacDermid JC, Walton DM, Bobos P, et al. A Qualitative Description of Chronic Neck Pain has
422		Implications for Outcome Assessment and Classification. <i>Open Orthop J</i> 2017; <b>10</b> :746–56.
423		doi:10.2174/1874325001610010746
424	6	Treleaven J. Sensorimotor disturbances in neck disorders affecting postural stability, head and eye

1 2				
3	425		movement control-Part 2: Case studies. Man Ther 2008;13:266–75.	
5 6	426		doi:10.1016/j.math.2007.11.002	
7 8	427	7	Cohen SP. Epidemiology, diagnosis, and treatment of neck pain. Mayo Clin Proc 2015;90:284–9	<del>)</del> 9.
9 10	428		doi:10.1016/j.mayocp.2014.09.008	
11 12	429	8	Bobos P, Macdermid JC, Walton DM, et al. Patient-reported outcome measures used for neck	
13 14	430		disorders: An overview of systematic reviews. J. Orthop. Sports Phys. Ther. 2018;48:775–88.	
15 16	431		doi:10.2519/jospt.2018.8131	
17 18	432	9	Kamper SJ, Maher CG, Mackay G. Global Rating of Change Scales: A Review of Strengths and	
19 20	433		Weaknesses and Considerations for Design. J Man Manip Ther 2009;17:163–70.	
21 22	434		doi:10.1002/mus.21062	
23 24 25	435	10	Nazari G, Bobos P, MacDermid JC, et al. The Effectiveness of Instrument-Assisted Soft Tissue	
26 27	436		Mobilization in Athletes, Participants Without Extremity or Spinal Conditions, and Individuals	
28 29	437		with Upper Extremity, Lower Extremity, and Spinal Conditions: A Systematic Review. Arch Phy	ys
30 31	438		Med Rehabil Published Online First: February 2019. doi:10.1016/j.apmr.2019.01.017	
32 33	439	11	Bobos P, Nazari G, Szekeres M, et al. The effectiveness of joint-protection programs on pain,	
34 35	440		hand function, and grip strength levels in patients with hand arthritis: A systematic review and	
36 37	441		meta-analysis. <i>J Hand Ther</i> 2018; <b>32</b> :194–211. doi:10.1016/j.jht.2018.09.012	
38 39	442	12	Nazari G, Bobos P, MacDermid JC, et al. Psychometric properties of the Zephyr bioharness	
40 41	443		device: A systematic review. BMC Sports Sci Med Rehabil 2018;10. doi:10.1186/s13102-018-	
42 43 44	444		0094-4	
45 46	445	13	Law MC, MacDermid J. Evidence-based rehabilitation: a guide to practice. Thorofare, NJ::	
47 48	446		Slack Incorporated 2014.	
49 50	447	14	Roy JS, Desmeules F, MacDermid JC. Psychometric properties of presenteeism scales for	
51 52	448		musculoskeletal disorders: A systematic review. J Rehabil Med Published Online First: 2011.	
53 54	449		doi:10.2340/16501977-0643	
55 56	450	15	Sańchez J. Rosner, B.: Fundamentals of Biostatistics, third edition. PWS-Kent, Boston 1990, xv,	,
57 58				19
59 60			For peer review only - http://bmiopen.bmi.com/site/about/quidelines.xhtml	

451		655 pp., ISBN 0-534-91973-1. <i>Biometrical J</i> 1993; <b>35</b> :150. doi:10.1002/bimj.4710350205
452	16	Wuensch KL, Evans JD. Straightforward Statistics for the Behavioral Sciences. J Am Stat Assoc
453		Published Online First: 2006. doi:10.2307/2291607
454	17	Cohen J. Statistical power analysis for the behavioral sciences. Stat. Power Anal. Behav. Sci.
455		1988. doi:10.1234/12345678
456	18	Mokkink LB, de Vet HCW, Prinsen CAC, et al. COSMIN Risk of Bias checklist for systematic
457		reviews of Patient-Reported Outcome Measures. Qual Life Res Published Online First: 2018.
458		doi:10.1007/s11136-017-1765-4
459	19	Terwee CB, Mokkink LB, Knol DL, et al. Rating the methodological quality in systematic reviews
460		of studies on measurement properties: a scoring system for the COSMIN checklist. 2012;:651–7.
461		doi:10.1007/s11136-011-9960-1
462	20	Viechtbauer W. Conducting meta-analisys in R with metafor package. J Stat Softw 2010;36:1–48.
463	21	Higgins JPT, Thompson SG, Deeks JJ, et al. Measuring inconsistency in meta-analyses. BMJ
464		Published Online First: 2003. doi:10.1136/bmj.327.7414.557
465	22	Cleland J, Fritz J, Whitman J, et al. The reliability and construct validity of the Neck Disability
466		Index and Patient Specific Functional Scale. Spine (Phila Pa 1976) 2006;31:598-602.
467	23	Cleland JA, Childs JD, Whitman JM. Psychometric Properties of the Neck Disability Index and
468		Numeric Pain Rating Scale in Patients With Mechanical Neck Pain. Arch Phys Med Rehabil
469		2008; <b>89</b> :69–74. doi:10.1016/j.apmr.2007.08.126
470	24	Cook C, Lawrence J, Michalak K, et al. Is there preliminary value to a within- and/or between-
471		session change for determining short-term outcomes of manual therapy on mechanical neck pain?
472		J Man Manip Ther 2014; <b>22</b> :173–80. doi:10.1179/2042618614y.0000000071
473	25	Guzy G, Vernon H, Polczyk R, et al. Psychometric validation of the authorized Polish version of
474		the Neck Disability Index. Disabil Rehabil 2013;35:2132–7. doi:10.3109/09638288.2013.771706
475	26	Kamper SJ, Ostelo RWJG, Knol DL, et al. Global Perceived Effect scales provided reliable
476		assessments of health transition in people with musculoskeletal disorders, but ratings are strongly

59

1 2			
3	477		influenced by current status. J Clin Epidemiol 2010;63:760-766.e1.
5 6	478		doi:10.1016/j.jclinepi.2009.09.009
7 8	479	27	Shaheen AAM, Omar MTA, Vernon H. Cross-cultural adaptation, reliability, and validity of the
9 10	480		arabic version of neck disability index in patients with neck pain. Spine (Phila Pa 1976)
11 12	481		2013; <b>38</b> :609–15. doi:10.1097/BRS.0b013e31828b2d09
13 14	482	28	Takeshita K, Hosono N, Kawaguchi Y, et al. Validity, reliability and responsiveness of the
15 16	483		Japanese version of the Neck Disability Index. J Orthop Sci 2013;18:14–21. doi:10.1007/s00776-
17 18	484		012-0304-у
19 20 21	485	29	Trouli MN, Vernon HT, Kakavelakis KN, et al. Translation of the Neck Disability Index and
22 23	486		validation of the Greek version in a sample of neck pain patients. BMC Musculoskelet Disord
24 25	487		2008; <b>9</b> :1–8. doi:10.1186/1471-2474-9-106
26 27	488	30	Tuttle N, Laakso L, Barrett R. Change in impairments in the first two treatments predicts outcome
28 29	489		in impairments, but not in activity limitations, in subacute neck pain: An observational study. Aust
30 31	490		J Physiother 2006; <b>52</b> :281–5. doi:10.1016/S0004-9514(06)70008-3
32 33	491	31	Ngo Trung, Stupar Maja, Co^te' Pierre, Boyle Eleanor, Shearer Heather. A study of the test –
34 35	492		retest reliability of the self-perceived general recovery and self-perceived change in neck pain
36 37	493		questions in patients with recent whiplash-associated disorders. 2010;:957-62.
38 39 40	494		doi:10.1007/s00586-010-1289-x
41 42	495	32	Björklund M, Wiitavaara B, Heiden M. Responsiveness and minimal important change for the
43 44	496		ProFitMap-neck questionnaire and the Neck Disability Index in women with neck-shoulder pain.
45 46	497		Qual Life Res 2017; <b>26</b> :161–70. doi:10.1007/s11136-016-1373-8
47 48	498	33	Evans R, Bronfort G, Maiers M, et al. "I know it" s changed ": a mixed-methods study of the
49 50	499		meaning of Global Perceived Effect in chronic neck pain patients. 2014;:888-97.
51 52	500		doi:10.1007/s00586-013-3149-y
53 54	501	34	Jorritsma W, Dijkstra PU, De Vries GE, et al. Detecting relevant changes and responsiveness of
55 56 57	502		Neck Pain and Disability Scale and Neck Disability Index. <i>Eur Spine J</i> 2012; <b>21</b> :2550–7.

503		doi:10.1007/s00586-012-2407-8
504	35	Monticone M, Frigau L, Vernon H, et al. Responsiveness and minimal important change of the
505		NeckPix © in subjects with chronic neck pain undergoing rehabilitation. Eur Spine $J$
506		2018; <b>27</b> :1324–31. doi:10.1007/s00586-017-5343-9
507	36	Monticone M, Ambrosini E, Vernon H, et al. Responsiveness and minimal important changes for
508		the Neck Disability Index and the Neck Pain Disability Scale in Italian subjects with chronic neck
509		pain. Eur Spine J 2015; <b>24</b> :2821–7. doi:10.1007/s00586-015-3785-5
510	37	Young BA, Walker MJ, Strunce JB, et al. Responsiveness of the Neck Disability Index in patients
511		with mechanical neck disorders. Spine J 2009;9:802–8. doi:10.1016/j.spinee.2009.06.002
512	38	Farooq MN, Mohseni-Bandpei MA, Gilani SA, et al. Urdu version of the neck disability index: A
513		reliability and validity study. BMC Musculoskelet Disord 2017;18:1–11. doi:10.1186/s12891-017-
514		1469-5
515	39	Williams GN, Gangel TJ, Arciero RA, et al. Comparison of the single assessment numeric
516		evaluation method and two shoulder rating scales. Outcomes measures after shoulder surgery. Am
517		J Sports Med 1999; <b>27</b> :214–21. doi:10.1177/03635465990270021701
518	40	Schmitt J, Abbott JH. Global Ratings of Change Do Not Accurately Reflect Functional Change
519		Over Time in Clinical Practice. J Orthop Sport Phys Ther 2015;45:106–11.
520		doi:10.2519/jospt.2015.5247
521	41	Chiarotto A, Ostelo RW, Boers M, et al. A systematic review highlights the need to investigate the
522		content validity of patient-reported outcome measures for physical functioning in patients with
523		low back pain. <i>J Clin Epidemiol</i> 2018; <b>95</b> :73–93. doi:10.1016/j.jclinepi.2017.11.005
524	42	Ailliet L, Knol DL, Rubinstein SM, et al. Definition of the construct to be measured is a
525		prerequisite for the assessment of validity. the Neck Disability Index as an example. J Clin
526		Epidemiol 2013; <b>66</b> :775-782.e2. doi:10.1016/j.jclinepi.2013.02.005
527		
528		

1		
2	529	
4	323	
5 6	530	
7	531	
8 9	F22	
10 11	532	
12	533	ı
13 14	534	ı
15	535	(
16 17	536	ı
18	537	·
19 20	538	ı
21	539	r
22	540	r
23 24	541	i
25 26	542	
27	543	
28 29		
30	544	
31 32	545	
33	546	
34 35	547	
36		
37 38	548	
39	549	
40 41	550	
42 43	551	
44	552	
45 46		
47	553	
48 49	554	
50	555	
51 52	556	
53		
54	557	
55 56	558	

59

60

Figure 1. Flow diagram of included studies

**Figure 2**. Meta-analysis of Pearson's correlation coefficients between neck disability change scores and GROC scores in patients with neck disorders based on 5 very good to excellent quality studies.

**Figure 3**. Meta-analysis of Spearman's correlation coefficients between neck disability change scores and GROC scores in patients with neck disorders based on 6 very good to excellent quality studies.

**Figure 4**. Random effects univariate meta-regression between age and the Fisher's Z estimates. Each circle represents a study and the size of the circle indicates the influence of that study on the model. The regression prediction is illustrated by the straight line and the curved lines represent the 95% confidence intervals. Age explained 68% of the variance in the model ( $R^2$ =0.68)

				BMJ Open		Page 24 of
1 2 3 559 4 5 560 6 7 <b>Table 1</b> . St	tudy Characteristics					-
8 9 Study	Donulation	Catting	Sample	Duomoution Evaluated	GROC evaluated	Intowal
10 11	Population	Setting	Size	<b>Properties Evaluated</b>	(ranked categories)	interval P
<b>½</b> jorklund <b>3</b> t al (2017) 4 5 6 7	Women with non- specific neck- shoulder pain	Not specified	104	Validity (correlation) Between NDI and GRoC	GRoC (7)  1. Very much worse; 2. Much worse; 3. Minimally worse; 4. No change; 5. Minimally improved; 6. Much improved; 7.	GRoC scale administered only after intervention (account one time point one time point (account one time point one time point one time point (account one time point one time point one time point one time point (account one time point (account one time point one time point one time point one time point (account one time point one time point (account one time point one time point one time point (account one time point one ti
8 Seleland et 201 (2006)	Patients with cervical radiculopathy	Hospital	38	Validity (correlation)	Very much improved.  GRoC (15)	GRoC was completed at follow up. Within a
1 2 3				Between NDI and GRoC Between PSFS and GRoC	-7 (a very great deal worse) to zero (about the same) to +7 (a very great deal better)	week over the period of 7 weeks.
©leland et 51. (2008) 6 7	Patients with neck pain only	5 Outpatient physical therapy clinics	137	Validity (correlation)  Between NDI and GRoC  Between NPRS and GRoC	GRoC (15)  -7 (a very great deal worse) to zero (about the same) to +7 (a very great deal better)	GRoC was completed at follow up. Wishing week elated to
€ook et al	Patients with any neck pain	Academic locations in Northeast Ohio	56	ROC curves and AUC to measure sensitivity and specificity. Binomial logistic regression analysis was also calculated to determine overall effect.	GRoC (15)  -7 (a very great deal worse) to zero (about the same) to +7 (a very great deal better)	Baseline and at to the follow up 48- and the follow up 48- and fol
Farooq et 6al. (2017) 7 8	Patients with neck pain	Physical therapy clinics	106	Validity (correlation)  Between NDI-U and GRoC	GRoC (15)  -7 (a very great deal worse) to zero (about the same) to +7 (a very great deal better)	GRoC was completed at three weeks after intervention  GRoC scale was completed at 2 weeks and at 4 weeks
0 Guzy et al. (2013) 2 3	Patients with neck pain	Outpatient rehabilitation clinic	95	Validity (correlation)  Between NDI-P and GRoC	GRoC (7)  'complete recovery' over 'no change' to 'my complaints are worse than ever'	GRoC scale was a completed at 2 weeks and at 4 weeks
5 Jorritsma et al. (2012) 8 9	Patients with chronic non-specific neck pain	Tertiary university center for rehabilitation	76	Validity (correlation)  Between NDI and GRoC  Between NPAD and GRoC	GPE (7)  3 (completely recovered) to zero (no change) to -3 (worse than ever)	After completion of
1 561 2 3 562 4 5						from 3 to 5 months patients filled the GPE
57 58 59 60	F	For peer review	only - http:	//bmjopen.bmj.com/site/abo	out/guidelines.xhtml	24 .

				BMJ Open			BMJ Open: f
	Patients with any whiplash-associated disorder.	Physical therapy clinics	134	Test-retest reliability	GPE (11)  -5 (vastly worse) to zero (unchanged) to +5 (completely recovered)	Baseline, 6 weeks, and 12 months	BMJ Open: first published
ne 17	Patients with chronic	Outpatient	153	Validity (correlation)	GPE (5)	At the end of	as
	neck pain	Rehabilitatio n Unit		Between NeckPix and GPE	(helped a lot = 1, helped = 2), one no change level (helped only a little = 3), and two worsening levels (did not help = 4, made things worse = 5)	At the end of treatment (8 weeks) and one year before follow-up of treatment 8 weeks and one year before follow-up of treatment 8 weeks and one year before follow-up of treatment 8 weeks and data means and data means are supported to text and data means are supported to	10.1136/bmjope
cone	Patients with chronic	Outpatient Rehabilitatio	200	Validity (correlation)	GPE (5)	At the end of	h-20
)13	neck pain	n Unit		Between NDI and GPE Between NPDS and GPE	(helped a lot = 1, helped = 2), one no change level (helped only a little = 3), and two worsening levels (did not help = 4, made	realment o weep ight, include	19-033909 o
et al.	Detients id WAD	Total in 1	16	To all makes at mall all little	things worse = 5)	ing	n 25
et ai. 0)	Patients with WAD.  Most participants	Interviewed by person or	46	Test-retest reliability	GPE (7)	3-5 days of	Nov
	(69.6%) had grade II WAD.	by telephone in Ontario			General recovery question     Completely better Much improved Slightly improved No change	ses related t	ember 2019.
					Slightly worse Much worse	o tex	Dow
					Worse than ever	ct and	/nloa
					Change in neck pain question:     very much better, better, slightly better, no change, slightly worse, worse, or very much worse	ini	n h
en et	Patients with neck	3 primary	70	Validity (correlation)	GRoC (15)	1 week ≥	d//:d∷
015)	pain lasting more than 3 months	health centers		Between NDI-Ar and GRoC	-7 (a very great deal worse) to zero (about the same) to +7 (a very great deal better)	Over 8 weeks  Over 8 weeks  Within 2 monthsol  week for test-regies.	om http://bmjopen.bmj.com/ on June
shita	Patients with neck	Variety of	130	Validity (correlation)	PGIC (7)	Over 8 weeks	mj.cc
·)	pain, cervical radiculopathy and/or cervical myelopathy	clinics and hospital settings		Between NDI-J and GRoC	much better, better, slightly better, unchanged, slightly worse, worse and much worse	imilar tecl	m/ on Jur
ıli et al. 8)	Patients with neck	Primary	68	Validity (correlation)	GRoC (15)	Within 2 months bu	ıt <u>1</u>
	pain	healthcare clinic		Between NDI-Gr and GRoC	-7 (a very great deal worse) to -1 (almost the same, hardly any worse at all) and from 7 (a very great deal better) to 1 (almost the same, hardly any better at all)	r week for test-reges	, 2025 at Agence
		For peer review o	nly - http:	//bmjopen.bmj.com/site/abo	out/guidelines.xhtml	25	2025 at Agence Bibliographique de l

**TABLE 2.** Summary of Psychometric Properties Reported in Studies and COSMIN Risk of Bias (RoB) and Quality studies

Study	Psychometric Properties Reported	COSMIN RoB	COSMIN Rating*§	Quality of Studies**
			(Criteria)	(QACMRR)
Bjorklund et al (2017)	Validity (correlation)	Very Good	?	Excellent
Cleland et al (2006)	Validity (correlation)	Very Good	+	Excellent
Cleland et al. (2008)	Validity (correlation)	Very Good	-	Excellent
Cook et al (2014)	Sensitivity Specificity	Very Good Very Good	+	Excellent
Farooq et al. (2017)	Validity (correlation)	Very Good	+	Excellent
Guzy et al. (2013)	Validity (correlation)	Very Good	?	Very good
Jorritsma et al. (2012)	Validity (correlation)	Very Good	?	Excellent
Kamper et al. (2010)	Test-retest reliability	Very Good	+	Excellent
Monticone et al. (2017)	Validity (correlation)	Very Good	?	Excellent
Monticone et al. (2015)	Validity (correlation	Very Good	?	Excellent
Ngo et al. (2010)	Test-retest reliability	Very Good	+	Excellent
Shaheen et al. (2015)	Validity (correlation)	Very Good	?	Excellent
Takeshita et al. (2014)	Validity (correlation)	Very Good	?	Very good
Trouli et al. (2008)	Validity (correlation)	Very Good	+	Excellent
Tuttle et al. (2006)	Validity (correlation)	Very Good	?	Excellent
Young et al. (2009)	Validity (correlation)	Very Good	?	Excellent

COSMIN, Consensus-based Standards for the Selection of health Measurement Instruments, Criteria for good measurement properties: '+' sufficient; '-'insufficient; '?' indeterminate. §§ The grading for the quality of the evidence based on the modified GRADE approach is not applicable. \*\*Quality Appraisal for Clinical Measurement Research Reports Evaluation Form (QACMRR).

**TABLE 3**. Quality Appraisal for Clinical Measurement Research Reports Evaluation Form

	Item Evaluation Criteria*													
Study	1	2	3	4	5	6	7	8	9	10	11	12	Total (%)	Quality Summary
Bjorklund et al (2017)	2	2	2	2	2	1	2	2	2	2	2	2	96	Excellent
Cleland et al. (2008)	2	2	2	2	1	2	2	2	2	2	2	2	96	Excellent
Trouli et al. (2008)	2	2	2	2	1	2	2	2	2	2	2	2	96	Excellent
Tuttle et al. (2006)	2	2	2	2	1	2	2	2	2	2	2	2	96	Excellent
Kamper et al. (2010)	2	2	2	2	1	2	2	2	2	2	2	2	96	Excellent
Cook et al (2014)	2	2	2	2	1	2	2	2	1	2	2	2	92	Excellent
Jorritsma et al. (2012)	2	2	2	2	1	1	2	2	2	2	2	2	92	Excellent
Cleland et al (2006)	2	2	2	2	1	1	2	2	2	2	2	2	92	Excellent
Monticone et al. (2017)	2	2	2	2	1	1	2	2	2	2	2	2	92	Excellent
Monticone et al. (2015)	2	2	2	2	1	1	2	2	2	2	2	2	92	Excellent
Ngo et al. (2010)	2	2	2	2	2	2	2	2	1	2	1	2	92	Excellent
Shaheen et al. (2013)	2	2	2	2	2	2	2	2	2	2	1	1	92	Excellent
Farooq et al. (2017)	2	2	1	2	2	2	2	2	1	2	2	2	92	Excellent
Young et al. (2009)	2	2	2	2	1	1	2	2	2	2	2	2	92	Excellent
Guzy et al. (2013)	2	2	1	2	1	2	2	2	1	2	2	2	88	Very good
Takeshita et al. (2014)	2	2	1	1	1	2	2	2	2	2	2	2	88	Very good

\*Item Evaluation Criteria: 1. Thorough literature review to define the research question; 2. Specific inclusion/exclusion criteria; 3. Specific hypotheses; 4. Appropriate scope of psychometric properties; 5. Sample size; 6. Follow-up; 7. The authors referenced specific procedures for administration, scoring, and interpretation of procedures; 8. Measurement techniques were standardized; 9. Data were presented for each hypothesis; 10. Appropriate statistics-point estimates; 11. Appropriate statistical error estimates; 12. Valid conclusions and clinical recommendations.

data mining, Al training, and similar technologies

Protected by copyright, including for uses related to text and

1		
2		
3 4	588	To
<del>4</del> 5	589	=
6	590	N
7	591	re
8	592	Qı
9	593	Qί
10 11	E0.4	
12	594	
13	595	
14		
15 16	596	
17	597	
18	337	
19	598	
20 21	500	
22	599	
23	600	
24		
25 26	601	
27	602	
28	002	
29	603	
30 31	604	
32	604	
33	605	
34		
35 36	606	
30 37	607	
38	007	
39	608	
40	600	
41 42	609	
43	610	
44		
45	611	
46 47	612	
48	012	
49	613	
50		
51 52	614	
52 53	615	
54	3-3	
55	616	
56 57		
J/		

58

59

60

Total score = (sum of subtotals  $\div$  24 × 100). If for a specific paper an item is deemed NA (Not Applicable), then, Total score = (sum of subtotals  $\div$  (2 × number of Applicable items) × 100).

NA – Not Applicable. The subsections no. 6, asks for percentage of retention/follow up. This subsection only applies to reliability test-retest studies

Quality Summary: Poor (0%-30%), Fair (31%-50%), Good (51%-70%), Very good (71%-90%), Excellent (>90%):

TO BEEL CHON ONL

		BMJ Open		Page 30
<b>ABLE 4</b> . St	ummary of reli	ability properties of GRoC scales		
Study	Type of Reliability	Reliability Estimates	COSMIN	Quality of Studies
Kamper et al. (2010)	Test-retest	Intra-class correlation coefficients (ICC) $0.99 (0.99 - 0.99)$ – baseline $0.96 (0.95 - 0.97)$ – at six weeks $0.92 (0.89 - 0.94)$ at twelve months.	Very Good	Excellent
Ngo et al. (2010)	Test-retest	Intra-class correlation coefficients (ICC)  0.70 (0.60–0.80) – at six weeks (General recovery)  0.80 (0.72–0.87) – at six weeks (neck pain questions)  Weighted Kappa  0.70 (0.42–0.98) – at six weeks (General recovery)  0.80 (0.51–1.0) – at six weeks (neck pain questions)  Dichotomized response options for recovery (K statistics)  0.85 (0.64–1) when "recovered" was defined  "completely better"  0.81 (0.64–0.99) when defined as "completely better" or  "much improved  Dichotomized response options for change in neck pain questions (K statistics)  0.46 (0.20–0.74) when "recovered" was defined as "very much better"  Recall questions (K statistics)  the kappa coefficient was 1 for participants who remembered their previous answers to the general recovery question; 0.88 (0.64–1) for those who did not remember and 0.50 (0.02–0.98) for participants who were not asked the question.  The kappa coefficient was 1 for participants who remembered their previous answers to the change in neck pain question; 0.74 (0.41–1) for those who did not remember and 0.66 (0.22–1) for participants who were not asked the question.	Very Good	Excellent
		w only - http://bmjopen.bmj.com/site/about/guidelines		30

Cleland et al. (2006)   Correlations (Pearson r)   between change scores   r = 0.19   Very Good   Excellent	Study	Type of Reliability	Validity Estimates	<u>COSMIN</u>	Quality of Studies
Cleland et al. (2006)   NDI and GRoC   PSFS and GRoC   r = 0.82   NDI and GRoC   r = 0.82   NDI and GRoC   r = 0.82   NDI and GRoC   r = 0.58   NDI and GRoC   Receiver operator characteristics (ROC)   Within-session change of Pain and GROC   AUC = 0.61   AUC = 0.66, -336.7%   Change in pain   Very Good   Excellent (2014)   Retween session change of Pain and GROC   Odds ratio = 7.3 (2.1		between the change scores of GRoC and ProFitMap-neck		Very Good	Excellent
Cleland et al. (2008)  Cleland et al. (2008)  Receiver operator characteristics (ROC)  NRS and GRoC  NRS and GRoC  NRS and GRoC  NRS and GRoC  Cook et al.  (2014)  Receiver operator characteristics (ROC)  Within-session change Between-session change Between-session change Between-session change Between-session change Between session session roble session session roble session session roble session rob		between change scores NDI and GRoC		Very Good	Excellent
Within-session change   Between-session change   Between-session change   Between-session change   Cook et al. (2014)   Between session change of Pain and GROC   Odds ratio = 7.3 (2.1,		between change scores NDI and GRoC	r = 0.58	Very Good	Excellent
Farooq et al. (2017) NDI-U $r=0.50$ Very Good Excellent (2017) NDI-U $r=0.50$ Very Good Excellent (2017) NDI-U $r=0.50$ Very Good Very good (2013) NDI vs GROC Four-week interval ( $r=-0.56$ ) Very Good Very good (2013) NDI vs GROC Four-week interval ( $r=-0.56$ ) Very Good Excellent (2012) NPAD and GPE $r=0.49$ (95 % CI 0.30— between change scores of NPAD and GPE $r=0.49$ (95 % CI 0.30— between change scores of the NeckPix© and GPE $r=0.49$ (95 % CI 0.30— between change scores of the NeckPix© and GPE $r=0.49$ (95 % CI 0.30— Very Good Excellent (2017) $r=0.49$ (95 % CI 0.30— Very Good Excellent (2017) $r=0.49$ (95 % CI 0.30— Very Good Excellent (2017) $r=0.49$ (95 % CI 0.30— Very Good Excellent (2017) $r=0.49$ (95 % CI 0.30— Very Good Excellent (2017) $r=0.49$ (95 % CI 0.30— Very Good Excellent (2018) $r=0.49$ (95		Receiver operator characteristics (ROC) Within-session change Between-session change Between session change of Pain and GROC Sensitivity	AUC = 0.61 AUC = 0.76, >36.7% change in pain Odds ratio = 7.3 (2.1, 24.7) 65.6% (57.9, 74.6)	Very Good	Quality of Studies  Excellent  Excellent  Excellent  Very good
Guzy et al. (2013)				Very Good	Excellent
between change scores of al. (2012)  Monticone et al. (2017)  Monticone et al. (2017)  Monticone et al. (2015)  NDI-I and GPE  NDPS and GPE  Shaheen et al. (2013)  Takeshita et al. (2014)  Torrelations  Monticone et al. (2014)  Torrelations (Spearman's)  NDI-Ar and GROC  Torrelations  Monticone et al. (2014)  NDI-J and PGIC  Trouli et al. (2014)  Trouli et al. (2014)  Monticone et between change scores of the NeckPix© rho = 0.71, p<0.01  Trouli et al. (2014)  Trouli et al. (2014)  Monticone et between change scores of the NeckPix© rho = 0.71, p<0.01  Trouli et al. (2014)  Monticone et between change scores of the NeckPix© rho = 0.81, p<0.01  Trouli et al. (2014)  Monticone et al. (2014)  NDI and GPE  NDI and GPE  NDI and PGIC  Trouli et al. (2014)  Monticone et between change scores of the NeckPix© rho = 0.81, p<0.01  Torelations (Spearman's)  NDI and PGIC  Trouli et al. (2014)  Monticone et al. (2014)  NDI and GPE  NDI and GPE  NDI and PGIC  Trouli et al. (2014)  Monticone rho = 0.47, p<0.001  Trouli et al. (2014)  NDI and PGIC  Trouli et al. (201	Guzy et al.	· /	0.73) Four-week interval (r = -	Very Good	Very good
$\begin{array}{c ccccccccccccccccccccccccccccccccccc$		between change scores of	`	Very Good	Excellent
Correlation (Spearman)		Correlations (Spearman) between change scores of the NeckPix©	rho = 0.69 - 0.82	Very Good	Excellent
Shaheen et al. (2013) Correlations (Spearman's) rho = 0.81, p<0.001 Very Good Excellent (2013) NDI-Ar and GROC rho = 0.81, p<0.001 Very Good Takeshita et al. (2014) NDI and PGIC rho = 0.47, p<0.001 Very Good Very good NDI-J and PGIC rho = 0.59, p<0.001 Very Good Excellent (2008) GROC vs Gr-NDI rho = 0.30, p=0.02 Correlations (Spearman's) Very Good Excellent NDI vs GPE (post 1, minus pre-1) rho = 0.17 NDI vs GPE (post 2, minus pre-1) rho = 0.01 Tuttle et al. NDI vs GPE (post 2, minus pre-2) rho = 0.03 (2006) PSFS vs GPE (post 1, minus pre-1) rho = 0.06		between change scores NDI-I and GPE		Very Good	Excellent  Excellent  Very good  Excellent
Takeshita et al. (2014)  NDI and PGIC  NDI-J and PGIC  NDI-J and PGIC  NDI-J and PGIC  Tho = 0.47, p<0.001  Tho = 0.59, p<0.001  Very Good  Excellent (2008)  Correlation (Spearman's)  Correlations (Spearman's)  NDI vs GPE (post 1, minus pre-1)  NDI vs GPE (post 2, minus pre-1)  Tuttle et al.  NDI vs GPE (post 2, minus pre-2)  PSFS vs GPE (post 1, minus pre-1)  Tho = 0.06		Correlations (Spearman's)	rho = 0.81, p<0.001	Very Good	Excellent
$ \begin{array}{c ccccccccccccccccccccccccccccccccccc$		NDI and PGIC	rho = 0.47, p < 0.001	Very Good	Very good
$ \begin{array}{c ccccccccccccccccccccccccccccccccccc$		Correlation (Spearman's)	•	Very Good	Excellent
	Tuttle et al.	Correlations (Spearman's) NDI vs GPE (post 1, minus pre-1) NDI vs GPE (post 2, minus pre-1) NDI vs GPE (post 2, minus pre-2)	rho = 0.17 rho = 0.01 rho = 0.03	Very Good	Excellent
PSFS vs GPE (post 2, minus pre-1)		PSFS vs GPE (post 2, minus pre-1)	rho = 0.03		

Excellent
Excellent

Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies

## Box 1. Questions of Global Rating of Change (GROC) scales

GROC (7)	"Compared to before the treatment of the study started, my overall status is now"  "Compared to before the treatment of the study started, my status
GROC (7)	
	"Compared to before the treatment of the study started, my status
	regarding my neck-shoulder problem is now'
	"Overall, how much has your neck pain changed since you started
GPE (9)	treatment in the study?''
	"With respect to your whiplash injury how would you describe
GPE (11)	yourself now compared to immediately after your accident"
	"Overall, how much did the treatment you received help your fear of
GPE (5)	movement due to current neck pain?
	"Overall, how much did the treatment you delivered help your
	subject's fear of movement due to her/his current neck pain?"
	"Overall, how much did the treatment you received help your neck
GPE (5)	problem?''
	''How well do you feel you are recovering from your injuries?''
GPE (7)	"How do you feel your neck pain has changed since the injury?"
	GPE (11)  GPE (5)

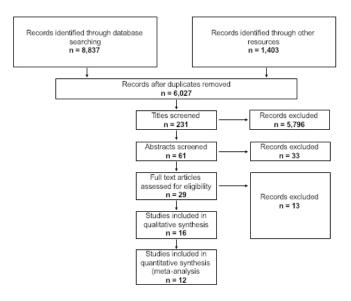


Figure 1. Flow diagram of included studies  $60x34mm (300 \times 300 DPI)$ 

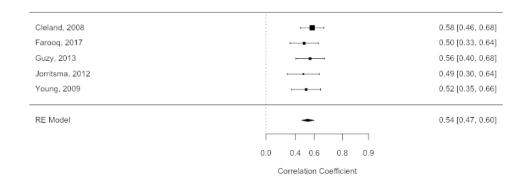


Figure 2. Meta-analysis of Pearson's correlation coefficients between neck disability change scores and GROC scores in patients with neck disorders based on 5 very good to excellent quality studies.

67x34mm (300 x 300 DPI)

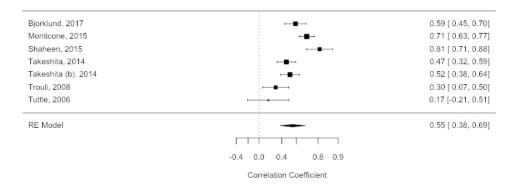


Figure 3. Meta-analysis of Spearman's correlation coefficients between neck disability change scores and GROC scores in patients with neck disorders based on 6 very good to excellent quality studies.

67x34mm (300 x 300 DPI)

#### Regression of Fisher's Z on Age

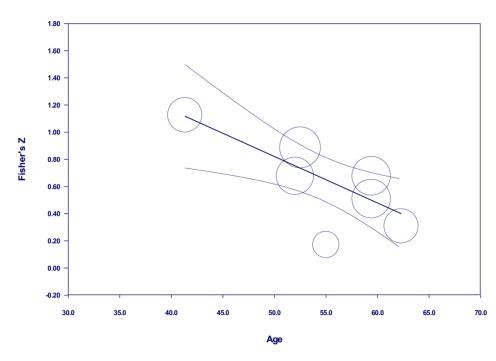


Figure 4. Random effects univariate meta-regression between age and the Fisher's Z estimates. Each circle represents a study and the size of the circle indicates the influence of that study on the model. The regression prediction is illustrated by the straight line and the curved lines represent the 95% confidence intervals. Age explained 68% of the variance in the model (R2=0.68)

160x118mm (300 x 300 DPI)

## Appendix 1

#### Search terms

#### **MEDLINE-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. cervicalgia.mp.
- 38. brachialgia.mp.
- 39. brachial neuritis.mp.

data mining, Al training, and similar technologies

Protected by copyright, including for uses related to text

```
2
3
             40. brachial neuralgia.mp.
4
             41. neck pain.mp.
5
             42. neck injur*.mp.
6
             43. brachial plexus neuropath*.mp.
7
             44. brachial plexus neuritis.mp.
8
```

- 45. thoracic outlet syndrome/ or cervical rib syndrome/
- 46. Torticollis/
- 47. exp brachial plexus neuropathies/ or exp brachial plexus neuritis/
- 48. cervico brachial neuralgia.ti,ab.
- 49. cervicobrachial neuralgia.ti,ab.
- 50. (monoradicul\* or monoradicl\*).tw.
- 51. or/30-50
- 52. exp headache/ and cervic\*.tw.
- 53. exp genital diseases, female/

- genital disease\*.mp.
  or/53-54
  52 not 55
  .51 or 56
  . neck/
  ). neck muscles/
  ). exp cervical plexus/
  1. exp cervical vertebrae/
  32. atlanto-axial joint/
  53. atlanto-occipital joint/
  64. Cervical Atlas/
  65. spinal nerve roots/
  66. exp brachial plexus/
  67. (odontoid\* or cervical or occip\* or atlant\*).tw.
  68. axis/ or odontoid process/
  Thoracic Vertebrae/

- 74. (brachial adj3 plexus).mp.
- 75. (thoracic adj3 vertebrae).mp.
- 76. neck.mp.
- 77. (thoracic adj3 spine).mp.
- 78. (thoracic adj3 outlet).mp.
- 79. trapezius.mp.
- 80. cervical.mp.
- 81. cervico\*.mp.
- 82.80 or 81
- 83. exp genital diseases, female/
- 84. genital disease\*.mp.
- 85. exp \*Uterus/

Protected by copyright, including for uses related

ining, Al training, and similar technologies

```
86. 83 or 84 or 85
87. 82 not 86
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
74 or 75 or 76 or 77 or 78 or 79 or 87
89. exp pain/
90. exp injuries/
91. pain.mp.
92. ache.mp.
93. sore.mp.
94. stiff.mp.
95. discomfort.mp.
96. injur*.mp.
97. neuropath*.mp.
98. or/89-97
99.88 and 98
100. Radiculopathy/
101. exp temporomandibular joint disorders/ or exp temporomandibular joint dysfunction
syndrome/
102. myofascial pain syndromes/
103. exp "Sprains and Strains"/
104. exp Spinal Osteophytosis/
105. exp Neuritis/
106. Polyradiculopathy/
107. exp Arthritis/
108. Fibromyalgia/
109. spondylitis/ or discitis/
110. spondylosis/ or spondylolysis/ or spondylolisthesis/
111. radiculopathy.mp.
112. radiculitis.mp.
113. temporomandibular.mp.
114. myofascial pain syndrome*.mp.
115. thoracic outlet syndrome*.mp.
116. spinal osteophytosis.mp.
117. neuritis.mp.
118. spondylosis.mp.
119. spondylitis.mp.
120. spondylolisthesis.mp.
121. or/100-120
122. 88 and 121
123. exp neck/
124. exp cervical vertebrae/
125. Thoracic Vertebrae/
```

- 127. (thoracic adj3 vertebrae).mp.
- 128. cervical.mp.
- 129. cervico\*.mp.

60

BMJ Open: first published as 10.1136/bmjopen-2019-033909 on 25 November 2019. Downloaded from http://bmjopen.bmj.com/ on June 11, 2025 at Agence Bibliographique de Enseignement Superieur (ABES) . Protected by copyright, including for uses related ining, Al training, and similar technologies

```
2
3
             130. 128 or 129
4
             131. exp genital diseases, female/
5
             132. genital disease*.mp.
6
             133. exp *Uterus/
7
             134. or/131-133
8
9
             135. 130 not 134
10
             136. (thoracic adj3 spine).mp.
11
             137. cervical spine.mp.
12
             138, 123 or 124 or 125 or 126 or 127 or 135 or 136 or 137
13
             139. Intervertebral Disk/
14
             140. (disc or discs).mp.
15
             141. (disk or disks).mp.
16
17
             142. 139 or 140 or 141
18
             143. 138 and 142
19
             144. herniat*.mp.
20
             145. slipped.mp.
21
             146. prolapse*.mp.
22
             147. displace*.mp.
23
24
             148. degenerat*.mp.
25
             149. (bulge or bulged or bulging).mp.
26
             150. 144 or 145 or 146 or 147 or 148 or 149
27
             151. 143 and 150
28
             152. intervertebral disk degeneration/ or intervertebral disk displacement/
29
             153. intervertebral disk displacement.mp.
30
             154. intervertebral disc displacement.mp.
31
32
             155. intervertebral disk degeneration.mp.
33
             156. intervertebral disc degeneration.mp.
34
             157. 152 or 153 or 154 or 155 or 156
35
             158. 138 and 157
36
             159. 57 or 99 or 122 or 151 or 158
37
             160. animals/ not (animals/ and humans/)
38
             161. 159 not 160
39
40
             162. exp *neoplasms/
41
             163. exp *wounds, penetrating/
42
             164. 162 or 163
43
             165. 161 not 164
44
             166. 29 and 165
45
             167. guidelines as topic/
46
             168. practice guidelines as topic/
47
48
             169. guideline.pt.
49
             170. practice guideline.pt.
50
             171. (guideline? or guidance or recommendations).ti.
51
             172. consensus.ti.
52
             173. or/167-172
53
             174. meta-analysis/
54
```

175. exp meta-analysis as topic/

Protected by copyright, including for uses related to text and

data mining, Al training, and similar technologies

- 176. (meta analy\* or metaanaly\* or met analy\* or metanaly\*).tw.
- 177. review literature as topic/
- 178. (collaborative research or collaborative review\* or collaborative overview\*).tw.
- 179. (integrative research or integrative review\* or integrative overview\*).tw.
- 180. (quantitative adj3 (research or review\* or overview\*)).tw.
- 181. (research integration or research overview\*).tw.
- 182. (systematic\* adj3 (review\* or overview\*)).tw.
- 183. (methodologic\* adj3 (review\* or overview\*)).tw.
- 184. exp technology assessment biomedical/
- 185. (hta or thas or technology assessment\*).tw.
- 186. ((hand adj2 search\*) or (manual\* adj search\*)).tw.
- 187. ((electronic adj database\*) or (bibliographic\* adj database\*)).tw.
- 188. ((data adj2 abstract\*) or (data adj2 extract\*)).tw.
- 189. (analys\* adj3 (pool or pooled or pooling)).tw.
- 190. mantel haenszel.tw.
- 191. (cohrane or pubmed or pub med or medline or embase or psycinfo or psychinfo or psychlit or cinahl or science citation indes).ab.

- 192. or/174-191
- 193. 173 or 192
- 194. 166 and 193

## **Quality Appraisal for Clinical Measurement Research Reports**

#### **Evaluation Form**

Authors:	Year:	Rater:

	·				
<u>Qualit</u>	y Appraisal for Clinical Measurer	ment Research Reports			
	<u>Evaluation Form</u>	1			
Authors:	Year:	Rater:			
each item on your quality ch study you are evaluating. It	nality of a clinical measurement st hecklist, pick the descriptor that so tems rank descriptors are provide thesis are available from develope	ounds <u>most</u> like what was r d in the guide. (Forms and <u>c</u>	eported in the guides to extro		
	Evaluation criteria			Scor	
Study question			2		0
	und work cited to define what is c	urrently known about the			_
measurement properties of n research question to informir	measures under study, and the pong that knowledge base?	otential contributions of the	current		
Study Design	7				
2. Were appropriate inclusion	n/exclusion criteria defined?	4			
3. Were specific clinical meas	surement questions/hypotheses i	dentified?			
	of measurement properties cons				
4. Was an appropriate scope	or measurement properties come	idered?			
		idered?			
5. Was an appropriate sample		1	vise n/a)		
5. Was an appropriate sample	e size used?	4	vise n/a)		
5. Was an appropriate sample 6. Was appropriate retention  Measurements 7. Were specific descriptions	e size used?	s involving retesting; otherw	. ,		
5. Was an appropriate sample 6. Was appropriate retention  Measurements 7. Were specific descriptions administer it? 8. Were standardized proced	e size used? n/follow-up obtained? (for studies	s involving retesting; otherwise tudy and the method(s) use	ed to		

BMJ Open	Page 44 of
. Were analyses conducted for each specific hypothesis or purpose?	
.0. Were appropriate statistical tests performed to obtain point estimates of the measurement properties?	
1. Were appropriate ancillary analyses done to quantify the confidence in the estimates of the linical measurement property (Precision/Confidence intervals; benchmark comparisons/ROC urves, alternate forms of analysis like SEM/MID, etc.)?	Protecte
Recommendations	d by c
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?	Protected by copyright, including for uses related to text and da
Subtotals (of columns 1 and 2)	uding
Total score (sum of subtotals/24*100);	for us
f for a specific paper or topic an item is deemed inappropriate then you can sum of tems/2*number of items *100	Enseignem ses related
	sur (ABES) . I data mining, Al training, and similar technologies.
For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml	

#### **Quality Appraisal of a Clinical Measurement Study**

#### **Interpretation Guide**

3		BMJ Open	
		Quality Appraisal of a Clinical Measurement Study	
		<u>Interpretation Guide</u>	
			7
		ide which score to provide for each item on your quality checklist, read the following descriptors.	Flotected by copyright, illicituding for uses related
		e descriptor that sounds <u>most</u> like the study you were evaluating with respect to a given item. If s no documentation about any specific aspect of an item; then you must evaluate assuming that it	u by
		t done. Given the diversity in clinical measurement properties and design options, the evaluator	0
		make judgments using the criteria below and extend the principles to specific aspects that may	yilg
ı	not be	covered in these brief exemplars. In many cases, the study will not look exactly like the	ί,
		otor so there will be some interpretation as to which level of optimal methods for clinical	Ciuc
		rement studies have been achieved. In such cases, the evaluator can use the general approach	
		this study research design and conduct is consistent with best practice (score=2); is acceptable	2
	out sut (score=	poptimal (score=1); is not done/documented, substantially inadequate or inappropriate =0).	uses uses
'	(30010		161
			נפט נס נפ
			lo lext a
		Descriptors	
Stud	ly ques		lo lext a
		stion	to text and data milling,
Scor		stion	to text and data milling,
Scor	e	stion	to text and data milling,
Scor	e	stion	to text and data milling,
Scor	e	stion	to text and data milling,
Scor	e	stion	to text and data milling,
Scor	e	stion	to text and data milling,
Scor	e	stion	to text and data milling,
Scor Scor	e		to text and data milling,
Scor	re 2	stion	to text and data milling,
Scor	2 2	The authors:  - performed a thorough literature review indicating what is currently known, and not known, about the clinical measurement properties of the instruments or tests under study - presented a critical, and unbiased view of what is known about the current measurement properties - indicated how the current research question fills a gap in the current knowledge base - established a research question based on the above.  All of the above criteria were not fulfilled, but a sound rationale was provided for the research question	to text and data milling,
Scor	2 2	The authors:  - performed a thorough literature review indicating what is currently known, and not known, about the clinical measurement properties of the instruments or tests under study - presented a critical, and unbiased view of what is known about the current measurement properties - indicated how the current research question fills a gap in the current knowledge base - established a research question based on the above.  All of the above criteria were not fulfilled, but a sound rationale was provided for the research question previous literature.	to text and data milling,
Scor	1 0	The authors:  - performed a thorough literature review indicating what is currently known, and not known, about the clinical measurement properties of the instruments or tests under study - presented a critical, and unbiased view of what is known about the current measurement properties - indicated how the current research question fills a gap in the current knowledge base - established a research question based on the above.  All of the above criteria were not fulfilled, but a sound rationale was provided for the research question previous literature.	to text and data milling,
Scor	1 0	The authors:  - performed a thorough literature review indicating what is currently known, and not known, about the clinical measurement properties of the instruments or tests under study - presented a critical, and unbiased view of what is known about the current measurement properties - indicated how the current research question fills a gap in the current knowledge base - established a research question based on the above.  All of the above criteria were not fulfilled, but a sound rationale was provided for the research question previous literature.	to text and data milling,
Scor	1 0	The authors:  - performed a thorough literature review indicating what is currently known, and not known, about the clinical measurement properties of the instruments or tests under study - presented a critical, and unbiased view of what is known about the current measurement properties - indicated how the current research question fills a gap in the current knowledge base - established a research question based on the above.  All of the above criteria were not fulfilled, but a sound rationale was provided for the research question previous literature.	to text and data milling,

	2	Specific inclusion/exclusion criteria for the study were defined, that described the patients enrolled. The subjects were described in terms of health condition/demographics, key relevant outcome mediators and the recruitment context (setting).
-	1	Some information on participants and place is provided (not all of above). For example, age/sex/diagnosis and the name or type of the practice is listed; but no additional information.
-	0	No information on type of clinical settings or study participants is provided (other than number/mean age).
	2	Specific hypotheses or research questions are provided. The stated study purpose provides specific research questions or hypotheses that indicate which specific measurement properties will be evaluated. This should include the specific type of reliability (intra/inter-rater or test-retest) being tested or the type of validity (construct/criterion/content; longitudinal/concurrent; convergent/divergent) being tested. A prior hypothesis should describe the level of reliability expected; and for validity, expected relationships (strength of associations) or constructs.
-	1	The types of reliability and validity being tested were apparent in the methods/title, but clear and specific research questions or hypotheses were not specified
-	0	Specific types of reliability or validity under evaluation were not clearly defined nor were specific hypotheses on reliability and validity stated. ("The purpose of this study was to investigate the reliability of the purpose of this study was to investigate the reliability of the purpose of this study was to investigate the reliability of the purpose of this study was to investigate the reliability of the purpose of this study was to investigate the reliability of the purpose of this study was to investigate the reliability of the purpose of this study was to investigate the reliability of the purpose of this study was to investigate the reliability of the purpose of this study was to investigate the reliability of the purpose of this study was to investigate the reliability of the purpose of this study was to investigate the reliability of the purpose of this study was to investigate the reliability of the purpose of this study was to investigate the reliability of the purpose of this study was to investigate the reliability of the purpose of the purpo
		hypotheses on reliability and validity stated. ("The purpose of this study was to investigate the reliability of and validity of" can be rated as zero if no further detail on the types of reliability and validity or the nature of specific hypotheses is stated).
	2	An appropriate scope of clinical measurement properties would be indicated by  1. A detailed focus on reliability that included multiple forms of reliability (at least two of – intrarater, inter-rater, test retest); as well as both relative and absolute reliability (e.g., ICCs and SEM/MID or limits of agreement)  2. A detailed focus on validity that included multiple forms of validity (content (judgmental); structured (e.g., expert review/survey, qualitative interviews, ICF linking) or structural (e.g., factor analyses or Rasch), construct (known group differences; convergent/divergent associations), criterion (concurrent/predictive), responsiveness; predictive, evaluative or discriminative properties were established  3. Three or more indicators of reliability and validity were examined concurrently and provide a rice.
	2	An appropriate scope of clinical measurement properties would be indicated by  1. A detailed focus on reliability that included multiple forms of reliability (at least two of – intrarater, inter-rater, test retest); as well as both relative and absolute reliability (e.g., ICCs and SEM/MID or limits of agreement)  2. A detailed focus on validity that included multiple forms of validity (content (judgmental); structured (e.g., expert review/survey, qualitative interviews, ICF linking) or structural (e.g., factor analyses or Rasch), construct (known group differences; convergent/divergent associations), criterion (concurrent/predictive), responsiveness; predictive, evaluative or discriminative properties were established  3. Three or more indicators of reliability and validity were examined concurrently and provide a rice.

53		BMJ Open
5	2	Authors performed a sample size calculation and obtained their recruitment targets. Post-doc power analyses and/or confidence intervals confirm that the sample size was sufficient to define relatively precise estimates of reliability or validity.
	1	The authors provide an acceptable rationale for the number of subjects included in the study, but did no
		present specific sample size calculations or post-doc power analyses (or had a sample >100 but no justification).
	0	Size of the sample was not rationalized or is clearly underpowered.
<u> </u>	2	
	1	90% or more of the patients enrolled for study were re-evaluated.  70% or more of the enrolled patients were re-evaluated.
	0	
Иe	asurem	Less than 70% of the patients enrolled in the study were re-evaluated  nents
7	2	
		Documentation is provided for how the studied test is performed. This includes adequate description of the measure/test and how it is administered or scored. The authors may provide or reference a
		published manual/article that outlines specific procedures for administration, scoring (including scoring
		algorithms, handling of missing data) and interpretation that included any necessary information about
		positioning/active participation of the client, any special equipment required, calibration of equipment in
		necessary, training required, cost, examiner procedures/actions. If no manual is available, then the text
		describes key details of procedures in sufficient detail so they could be replicated.
		describes key details of procedures in sufficient detail so they could be replicated.
	1	The test(s) and its administration procedures are referenced; but there is inadequate description of the test procedures.
		test procedures.
	0	Minimal description of test procedures without appropriate references.
		Minimal description of test procedures without appropriate references.
		For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml
		For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

	BMJ Open	Page 48 of
2	This item addresses the overall study procedures for administering all study measures (study and its comparators) in an unbiased way. Test procedures should not introduce systematic er estimation of the clinical measurement properties. This includes standardized procedures for completed or administered the measures. For self-report, this includes order of presentation, completed at what time interval; handling of missing items. If relevant, then the paper should how cultural literacy issues were handled (e.g., exclusion, assisted or surrogate completion). I impairment measures, procedures would include calibration of any equipment; use of consist measurement tools and scoring, a priori exclusion of any participants likely to give invalid rest to complete testing (not exclusion of after enrollment); use of standardized instructions and the procedures. This can include order of administration of test and quality checking of scores. For testing, the appropriate retest interval will depend on the nature of the condition; but for accomplication is the majority require retesting within 48 hours; whereas chronic/stable conditions are conditions it may require retesting within 48 hours; whereas chronic/stable conditions are conditions it may require retesting within 48 hours; whereas chronic/stable conditions are conditions it may require retesting within 48 hours; whereas chronic/stable conditions are conditions it may require retesting within 48 hours; whereas chronic/stable conditions are conditions it may require retesting within 48 hours; whereas chronic/stable conditions are conditions.	rors in the who who who dinclude For Protected by control of test with the control of the contro
1	No obvious sources of bias in the study test protocol or how tests were performed/administeral apparent; but there were suboptimal procedures or an inadequate description of the measure protocol to be insured control of bias or that procedures were standardized.	red is
0	No description of the overall procedures for administering study tests; OR an obvious source data collection methods.	of bias in to
nalyses		nd da
2	Authors clearly defined which specific analyses were conducted for each of the stated specific hypotheses/questions of the study. This may be accomplished through organization of the respecific subheadings or by demarcating which analyses addressed specific clinical measurement properties. Data was presented for each hypothesis/research question posed.  Data was presented that addressed each of the measurement questions posed, but authors of the stated specific analyses were conducted for each of the stated specific hypotheses/question of the stated specific hypotheses/question of the respective specific clinical measurement properties.	sults under 📆 .
1	Data was presented that addressed each of the measurement questions posed, but authors of specific analyses to specific research questions or hypotheses.	did not linkg, and si
0	Data was not presented for every hypothesis or clinical measurement property outlined in the or methods.	e purposes milar tech
		nologies.
	For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml	

3		BMJ Open
10 2	2	<u>Tests selected</u> - Appropriate statistical tests were conducted to calculate a point estimate for clinical measurement properties. Examples are provided below; but are not exhaustive.
		1. Reliability (Relative=ICCs (Shrout & Fleiss, 1979) for quantitative, Kappa (Landis & Koch, 1977) for nominal data); absolute (SEM or plot of score differences vs. average score showing mean and 2SD limit – as per Altman and Bland) (Bland & Altman, 1986; Bland & Altman, 1987)
		2. Clinical relevance - minimal detectable change, clinically important difference (Jaeschke, Singer, & Guyatt, 1989; Beaton et al., 2001; Wells et al., 2001)  3. Validity  a. Validity associations - Pearson correlations for normally distributed data, Spearman rank correlations for ordinal data; or other correlations, if appropriate
		3. Validity
		a. Validity associations - Pearson correlations for normally distributed data, Spearman rank correlations for ordinal data; or other correlations, if appropriate
		b. Validity tests of significant difference - an appropriate global test like analysis of variance was used where indicated, with post-hoc tests that adjusted for multiple testing
		c. Validity of items scaling/responses - Rasch analysis or item response (Baylor et al., 2011; Pallant & Tennant, 2007; Kyngdon, 2006; Cipriani, Fox, Khuder, & Boudreau, 2005; Smith, Jr., Conrad, Chang, & Piazza, 2002)
		4. Responsiveness (Beaton, Bombardier, Katz, & Wright, 2001)- standardized response means or effect sizes or other recognized responsiveness indices were used.
-	1	Appropriate statistical tests were used in some instances; but suboptimal choices were made in other analyses.
(	0	Inappropriate use of statistical tests - incorrect tests for type of data; or a lack of analysis
11 2	2	The study goes beyond a single statistical point estimate of a clinical measurement property and providing supporting statistical analyses that increases confidence in the findings in terms of precision of the (key) indicator; or provide an alternate form of analysis of the clinical measurement property. The evaluator decides if these analyses are appropriate and informative. For example, with reliability, at least 2 of the following would constitute appropriate and informative analysis beyond a point estimate a reliability coefficient: 1. confidence intervals around the point estimate; 2. Comparison to appropriate, referenced benchmarks or standards; or 3. SEM or MDC. For correlations, tests of significance or confidence intervals were presented and indicators of the criterion benchmarks were provided. For studies involving cross-cultural validation, the analyses should compare multiple clinical measurement properties previously established for the measure and explain the extent to which the translated version is in accordance with these previously reported properties on the source measure.
		For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

	BMJ Open Pa	ige 50 of
1	Either precision definition (confidence intervals) or appropriate benchmark comparison were used -both. OR Some analyses were associated with indicators of precision or alternate form of analysis -b not all key indicators.	
0	Inappropriate use of benchmarks or confidence intervals; or indicators of precision or alternate formabsent	n are
Recomm	endations  Authors made specific conclusions and clinical measurement recommendations that were clearly re	otect
12 2	Authors made specific conclusions and clinical measurement recommendations that were clearly re to each hypotheses/question posed in the study and that were supported by the data presented. It recommendations would state the estimated status of the clinical measurement property, the confidence in the estimate and the context for which those apply. To achieve a 2, the conclusion m be specific; and conclusions cannot overstate the clinical measurement properties observed the stunor ignore suboptimal measurement properties found.	leal copyright,
1	Authors made conclusions and clinical measurement recommendations that were basically true (supported by study data); but vague. That is, they do not specify the extent, confidence or context the findings. (The measure is "reliable and valid") OR authors made specific clinical measurement recommendations; but for only some of the study hypotheses.	ng for uses related
0	Authors did not make conclusions about clinical measurement; OR made recommendations that we contradiction to the actual data presented	re int Sup
		(ABES) . ta mining, Al training, and similar technologies.
	For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml	

1. Abbott et al 2014	Ineligible population
2. Beattie et al 2011	Ineligible population (less than 50%)
3. Hoeskstra et al 2014	No properties for GRoC scales
4. Chansirinukor 2019	No properties for GRoC scales
5. <u>Chien et al 2015</u>	No properties for GRoC scales
6. <u>Cruz et al. 2015</u>	No properties for GRoC scales
7. Foroutani et al 2018	No English (Persian language)
8. Gagnon et al 2018	Ineligible population
9. Hefford et al 2012	Ineligible population
10. Hung et al 2019	Ineligible population
11. <u>Sharma et al 2017</u>	Ineligible population
12. Stevens et al 2019	Ineligible population
13. <u>Meyer et al 2014</u>	Ineligible population

Page 52 of 53



47

# PRISMA 2009 Checklist

- 3		Jht,	
Section/topic	#	Checklist item	Reported on page #
TITLE		g 75 fo z	
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT		is re	
Structured summary 13 14	2	Provide a structured summary including, as applicable: background; objectives; data south structured summary including, as applicable: background; objectives; data south structured summary including, as applicable: background; objectives; data south structured summary including, as applicable: background; objectives; data south structured summary including, as applicable: background; objectives; data south structured summary including, as applicable: background; objectives; data south structured summary including, as applicable: background; objectives; data south structured summary including, as applicable: background; objectives; data south structured summary including, as applicable: background; objectives; data south structured summary including, as applicable: background; objectives; data south structured summary including, as applicable: background; objectives; data south structured summary including, as applicable: background; objectives; data south structured summary including, as applicable: background; objectives; data south structured summary including, as applicable: background; objectives; data south structured summary including, as applicable: background; objectives; data south structured summary including structured summary inclu	2
15 INTRODUCTION		xt a a a a a a a a a a a a a a a a a a a	
16 17 Rationale	3	Describe the rationale for the review in the context of what is already known.	3-5
18 Objectives 19	4	Provide an explicit statement of questions being addressed with reference to participants heterventions, comparisons, outcomes, and study design (PICOS).	4-5
METHODS		ng,	
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.	5
24 25 Eligibility criteria 26	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.	5
27 Information sources 28	7	Describe all information sources (e.g., databases with dates of coverage, contact with story authors to identify additional studies) in the search and date last searched.	6
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	Appendix1
32 Study selection 33	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic very, and, if applicable, included in the meta-analysis).	6
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in diplicate) and any processes for obtaining and confirming data from investigators.	6-7
Data items 38	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and simplifications made.	6-7
Risk of bias in individual	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.	6-7
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	8-9
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I²) for each meta-analysis.  For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml	8-9

BMJ Open



46 47

## **PRISMA 2009 Checklist**

Page 53	of 53		BMJ Open dd t	
Page 53 of 53  PRISMA 2009 Checklist  Page 1 of 2  BMJ Open  Page 1 of 2				
3 4			Page 1 of 2 ; 33 ; 9	
5 6 7	tion/topic	#	Checklist item 25	Reported on page #
8 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).	8-9
10 Addi 11	itional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-registrical pre-specified.	8=9
13 RES	BULTS		d to	
14 Stud 15	ly selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with a screened, assessed for eligibility, and included in the review, with a screened, assessed for eligibility, and included in the review, with a screened, assessed for eligibility, and included in the review, with a screened, assessed for eligibility, and included in the review, with a screened, assessed for eligibility, and included in the review, with a screened, assessed for eligibility, and included in the review, with a screened, assessed for eligibility, and included in the review, with a screened, assessed for eligibility, and included in the review, with a screened, assessed for eligibility and included in the review, with a screened in the review of the screened in the scr	9
16 17 Stud 18	ly characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, Provide the citations.	9-10
19 Risk	of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessme	10
21 Resu	ults of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple suntained data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot	10-12
23 Synt	thesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	13
25 Risk	of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	10
26 Addi	itional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-	13
28 <b>DIS</b>	CUSSION		'	
29 30 31	nmary 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).	14-15
32 Limit	tations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., ingomplete retrieval of identified research, reporting bias).	16
34 Cond	clusions	26	Provide a general interpretation of the results in the context of other evidence, and implifiations for future research.	14-15
36 FLINDING				
37 38 Fund 39 40	ding	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

40
41 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.
42 doi:10.1371/journal.pmed1000097
43
For more information, visit: www.prisma-statement.org.
44
Page 2 of 2
For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml
46