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

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

Pavlos Bobos[1], Joy C MacDermid[2], Goris Nazari[3], Rochelle Furtado[4] and CATWAD co-authors[5]

[1]Pavlos Bobos PT, PhD(c), (corresponding author) Doctoral Candidate, Western's Bone and Joint Institute, Department of Health and Rehabilitation Sciences, Western University, Elborn College, 1201 Western Road, N6G 1H1, London, Ontario, Dalla Lana School of Public Health, Institute of Health Policy Management and Evaluation, Department of Clinical Epidemiology and Health Care Research, University of Toronto, Canada, (pbobos@uwo.ca), tel: +1 519 661 2111 x88912

[2]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 Health Centre, London, Ontario; Professor Rehabilitation Science McMaster University, Hamilton, ON, Canada (jmacderm@uwo.ca)

[3]Goris Nazari PT, PhD(c) Doctoral Candidate, Western's Bone and Joint Institute, School of Physical Therapy, Department of Health and Rehabilitation Sciences, Western University, London, Ontario, Canada, (gnazari@uwo.ca)

[4]Rochelle Furtado MSc Western's Bone and Joint Institute, School of Physical Therapy, Department of Health and Rehabilitation Sciences, Western University, London, Ontario, Canada, (rfurtad5@uwo.ca)

[5]CATWAD: Michele Sterling m.sterling@uq.edu.au, Anne Söderlund anne.soderlund@mdh.se, Michele Curatolo, curatolo@uw.edu, James M Elliott j-elliott@northwestern.edu, David Walton dwalton5@uwo.ca, Helge Kasch helgkasc@rm.dk, Linda Carroll linda.carroll@ualberta.ca, Hans Westergren Hans.Westergren@skane.se, Gwendolen Jull g.jull@uq.edu.au, Eva-Maj Malmström eva-maj.malmstrom@med.lu.se, Luke B Connelly l.connelly@uq.edu.au, Joy C MacDermid jmacderm@uwo.ca, Mandy Nielsen mandy.nielsen@griffith.edu.au, Pierre Côté pierre.cote@uoit.ca, Tonny Elmose Andersen tandersen@health.sdu.dk, Trudy Rebbeck trudy.rebbeck@sydney.edu.au, Annick Maujean a.maujean@uq.edu.au, Sarah Robins s.robins1@uq.edu.au, Kenneth Chen k.chen8@uq.edu.au, Julia Treleaven j.treleaven@uq.edu.au

Kewords: neck pain, global assessment, psychometric properties, systematic review

Word count: 3908

ABSTRACT

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

Data sources: A search was performed in 4 databases (MEDLINE, EMBASE, CINAHL, SCOPUS) until February 2019.

Data extraction and synthesis: Eligible articles were appraised using Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) checklist and the Quality Appraisal for Clinical Measurement Research Reports Evaluation Form.

Results: The search obtained 16 eligible studies and included in total 1533 patients with neck pain. 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: 0.47 to 0.59). Pooled data of Spearman's correlations indicated that GROC scores were moderately 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 reliability of GPE for patients with Whiplash Associated Disorders. Evidence from very good-to-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 4th 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 appraise and synthesize the psychometric properties of the GROC scales in patients with neck disorders.

METHODS

Patient and Public Involvement

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

Study Design and Protocol Registration

We conducted a systematic review to evaluate the psychometric properties of GROC scales in patients with neck disorders. The protocol was registered in PROSPERO register database with registration number: CRD 42018117874

Eligibility Criteria

We included studies in this systematic review if the following criteria were met [10–12]:

- Design: psychometric testing, randomized/ cohort studies
- Participants: > 50% of the study's patient population with neck conditions/disorders,
- Intervention/Comparison: studies that reported on the psychometric properties (reliability, validity, responsiveness) of GROC, Global Perceived Effect (GPE) and Patient Global Impression of Change (PGIC),
- Outcomes: GROC, GPE and PGIC.

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, $0.40 \leq \text{ICC} < 0.75$ indicating fair-to-good and $\text{ICC} \geq 0.75$ 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 (≥ 0.20 to < 0.50), moderate (≥ 0.50 to < 0.80) or large (≥ 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 meta-analysis 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^2 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-

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.**

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 patients with neck pain.[23] Jorritsma et al. compared the validity of GPE with Neck Pain and Disability Scale change scores (NPAD) (moderate correlations: r = 0.49 (95% CI 0.30 to 0.64) in patients with chronic non-specific neck pain. [32] Monticone et al. compared the GPE with NeckPix change scores (strong correlations: rho = 0.69 to 0.82) in patients with chronic neck pain.[33] Monticone et al. compared the GPE with the Italian version NDI change scores (moderate correlations: Spearman's coefficient = 0.59) in patients with chronic neck pain. [34] Shaheen et al. compared the validity of GROC with the Arabic version of NDI change scores and indicated very strong correlations: *r* coefficient = 0.81, in 70 patients with neck pain lasting more than three months.[25] Takeshita et al. compared the validity of PGIC with the original NDI and the Japanese version of NDI-J change scores and reported moderate correlations: *r* coefficient = 0.47, and r = 0.59 in 130 patients with neck pain, cervical radiculopathy and/or cervical myelopathy respectively.[26] Trouli et al. compared the validity of the GROC with the Greek version of NDI change scores and reported weak correlations: *r* coefficient = 0.30, in 68 patients with neck pain.[27] Tuttle et al. compared the validity of GPE with NDI (*r* coefficient range: 0.01 to 0.17; very weak correlations), with PSFS (*r* coefficient range: 0.03 to 0.06; very weak correlations), with pain intensity (*r* coefficient range: 0.00 to 0.05; very weak correlations), and with ROM (*r* coefficient range: 0.00 to 0.03; very weak correlations), in 29 patients with neck pain for more than two weeks.[28] Young et al. compared the validity of GROC with NDI change scores and reported moderate correlations (*r* coefficient = 0.52) in patients with mechanical neck pain.

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 (β = -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 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 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.[3] 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 comparator, may not reflect the issues or priorities that are important to patients. 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 (I^2>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

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 current study

Funding Statement

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

Competing Interest Statement

None to report

References

1 Murray CJL, Abraham J, Ali MK, *et al.* The State of US health, 1990-2010: Burden of diseases, injuries, and risk factors. *JAMA - J Am Med Assoc* Published Online First: 2013. doi:10.1001/jama.2013.13805

2 Fejer R, Kyvik KO, Hartvigsen J. The prevalence of neck pain in the world population: A systematic critical review of the literature. Eur. Spine J. 2006. doi:10.1007/s00586-004-0864-4

3 Hogg-Johnson S, van der Velde G, Carroll LJ, *et al.* The Burden and Determinants of Neck Pain in the General Population. Results of the Bone and Joint Decade 2000-2010 Task Force on Neck

Pain and Its Associated Disorders. *J Manipulative Physiol Ther* Published Online First: 2009. doi:10.1016/j.jmpt.2008.11.010

4 Bobos P, Nazari G, Palimeris S, *et al.* The contribution of health and psychological factors in patients with chronic neck pain and disability: A cross-sectional study. *J Clin Diagnostic Res* Published Online First: 2018. doi:10.7860/JCDR/2018/31284.11203

5 Macdermid JC, Walton DM, Bobos P, *et al.* The Open Orthopaedics Journal A Qualitative Description of Chronic Neck Pain has Implications for Outcome Assessment and Classification. *Open Orthop J* 2016;**10**:746–56. doi:10.2174/1874325001610010746

6 Treleaven J. Sensorimotor disturbances in neck disorders affecting postural stability, head and eye movement control-Part 2: Case studies. *Man Ther* 2008;**13**:266–75. doi:10.1016/j.math.2007.11.002

7 Cohen SP. Epidemiology, diagnosis, and treatment of neck pain. *Mayo Clin Proc* 2015;**90**:284–99. doi:10.1016/j.mayocp.2014.09.008

8 Bobos P, MacDermid JC, Walton DM, *et al.* Patient-Reported Outcome Measures Used for Neck Disorders: An Overview of Systematic Reviews. *J Orthop Sport Phys Ther* 2018;**48**:775–88. doi:10.2519/jospt.2018.8131

9 Kamper SJ, Maher CG, Mackay G. Global Rating of Change Scales: A Review of Strengths and Weaknesses and Considerations for Design. *J Man Manip Ther* 2009;**17**:163–70. doi:10.1002/mus.21062

10 Nazari G, Bobos P, MacDermid JC, *et al.* The Effectiveness of Instrument-Assisted Soft Tissue Mobilization in Athletes, Participants Without Extremity or Spinal Conditions, and Individuals with Upper Extremity, Lower Extremity, and Spinal Conditions: A Systematic Review. *Arch Phys Med Rehabil* Published Online First: February 2019. doi:10.1016/j.apmr.2019.01.017

11 Bobos P, Nazari G, Szekeres M, *et al.* The effectiveness of joint-protection programs on pain, hand function, and grip strength levels in patients with hand arthritis: A systematic review and meta-analysis. *J Hand Ther* 2018;**32**:194–211. doi:10.1016/j.jht.2018.09.012

12 Nazari G, Bobos P, MacDermid JC, *et al.* Psychometric properties of the Zephyr bioharness device: A systematic review. *BMC Sports Sci Med Rehabil* 2018;**10**. doi:10.1186/s13102-018-0094-4

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

14 Roy JS, Desmeules F, MacDermid JC. Psychometric properties of presenteeism scales for musculoskeletal disorders: A systematic review. *J Rehabil Med* Published Online First: 2011. doi:10.2340/16501977-0643

15 Wuensch KL, Evans JD. Straightforward Statistics for the Behavioral Sciences. *J Am Stat Assoc* Published Online First: 2006. doi:10.2307/2291607

16 Cohen J. Statistical power analysis for the behavioral sciences. Stat. Power Anal. Behav. Sci. 1988. doi:10.1234/12345678

17 Mokkink LB, de Vet HCW, Prinsen CAC, *et al.* COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. *Qual Life Res* Published Online First: 2018. doi:10.1007/s11136-017-1765-4

18 Terwee CB, Mokkink LB, Knol DL, *et al.* Rating the methodological quality in systematic reviews of studies on measurement properties : a scoring system for the COSMIN checklist. 2012;:651–7. doi:10.1007/s11136-011-9960-1

19 Higgins JPT, Thompson SG, Deeks JJ, *et al.* Measuring inconsistency in meta-analyses. *BMJ* Published Online First: 2003. doi:10.1136/bmj.327.7414.557

20 Cleland J, Fritz J, Whitman J, *et al.* The reliability and construct validity of the Neck Disability Index and Patient Specific Functional Scale. *Spine (Phila Pa 1976)* 2006;**31**:598–602.

21 Cleland JA, Childs JD, Whitman JM. Psychometric Properties of the Neck Disability Index and Numeric Pain Rating Scale in Patients With Mechanical Neck Pain. *Arch Phys Med Rehabil* 2008;**89**:69–74. doi:10.1016/j.apmr.2007.08.126

22 Cook C, Lawrence J, Michalak K, *et al.* Is there preliminary value to a within- and/or between-

session change for determining short-term outcomes of manual therapy on mechanical neck pain? *J Man Manip Ther* 2014;**22**:173–80. doi:10.1179/2042618614y.0000000071

23 Guzy G, Vernon H, Polczyk R, *et al.* Psychometric validation of the authorized Polish version of the Neck Disability Index. *Disabil Rehabil* 2013;**35**:2132–7. doi:10.3109/09638288.2013.771706

24 Kamper SJ, Ostelo RWJG, Knol DL, *et al.* Global Perceived Effect scales provided reliable assessments of health transition in people with musculoskeletal disorders, but ratings are strongly influenced by current status. *J Clin Epidemiol* 2010;**63**:760-766.e1. doi:10.1016/j.jclinepi.2009.09.009

25 Shaheen AAM, Omar MTA, Vernon H. Cross-cultural adaptation, reliability, and validity of the arabic version of neck disability index in patients with neck pain. *Spine (Phila Pa 1976)* 2013;**38**:609–15. doi:10.1097/BRS.0b013e31828b2d09

26 Takeshita K, Hosono N, Kawaguchi Y, *et al.* Validity, reliability and responsiveness of the Japanese version of the Neck Disability Index. *J Orthop Sci* 2013;**18**:14–21. doi:10.1007/s00776-012-0304-y

27 Trouli MN, Vernon HT, Kakavelakis KN, *et al.* Translation of the Neck Disability Index and validation of the Greek version in a sample of neck pain patients. *BMC Musculoskelet Disord* 2008;**9**:1–8. doi:10.1186/1471-2474-9-106

28 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. *Aust J Physiother* 2006;**52**:281–5. doi:10.1016/S0004-9514(06)70008-3

29 Ngo Trung, Stupar Maja, Coˆte´ 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

30 Björklund M, Wiitavaara B, Heiden M. Responsiveness and minimal important change for the ProFitMap-neck questionnaire and the Neck Disability Index in women with neck–shoulder pain.

Qual Life Res 2017;**26**:161–70. doi:10.1007/s11136-016-1373-8

31 Evans R, Bronfort G, Maiers M, *et al.* ‘“ I know it ” s changed ’’: a mixed-methods study of the meaning of Global Perceived Effect in chronic neck pain patients. 2014;:888–97. doi:10.1007/s00586-013-3149-y

32 Jorritsma W, Dijkstra PU, De Vries GE, *et al.* Detecting relevant changes and responsiveness of Neck Pain and Disability Scale and Neck Disability Index. *Eur Spine J* 2012;**21**:2550–7. doi:10.1007/s00586-012-2407-8

33 Monticone M, Frigau L, Vernon H, *et al.* Responsiveness and minimal important change of the NeckPix© in subjects with chronic neck pain undergoing rehabilitation. *Eur Spine J* 2018;**27**:1324–31. doi:10.1007/s00586-017-5343-9

34 Monticone M, Ambrosini E, Vernon H, *et al.* Responsiveness and minimal important changes for the Neck Disability Index and the Neck Pain Disability Scale in Italian subjects with chronic neck pain. *Eur Spine J* 2015;**24**:2821–7. doi:10.1007/s00586-015-3785-5

35 Young BA, Walker MJ, Strunce JB, *et al.* Responsiveness of the Neck Disability Index in patients with mechanical neck disorders. *Spine J* 2009;**9**:802–8. doi:10.1016/j.spinee.2009.06.002

36 Farooq MN, Mohseni-Bandpei MA, Gilani SA, *et al.* Urdu version of the neck disability index: A reliability and validity study. *BMC Musculoskelet Disord* 2017;**18**:1–11. doi:10.1186/s12891-017-1469-5

37 Williams GN, Gangel TJ, Arciero RA, *et al.* Comparison of the single assessment numeric evaluation method and two shoulder rating scales. Outcomes measures after shoulder surgery. *Am J Sports Med* 1999;**27**:214–21. doi:10.1177/03635465990270021701

38 Schmitt J, Abbott JH. Global Ratings of Change Do Not Accurately Reflect Functional Change Over Time in Clinical Practice. *J Orthop Sport Phys Ther* 2015;**45**:106–11. doi:10.2519/jospt.2015.5247

Table 1. Study Characteristics

Study	Population	Setting	Sample Size	Properties Evaluated	GROC evaluated	Interval
Bjorklund et al (2017)	Women with non-specific neck-shoulder pain	Not specified	104	Validity (correlation) Between NDI and GRoC	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 scale administered only after intervention at one time point (1 week)
Cleland et al (2006)	Patients with cervical radiculopathy	Hospital	38	Validity (correlation) Between NDI and GRoC Between PSFS and GRoC	GRoC 15-points -7 (a very great deal worse) to zero (about the same) to +7 (a very great deal better)	GRoC was completed at follow up. Within a week over the period of 7 weeks.
Cleland et al. (2008)	Patients with neck pain only	5 Outpatient physical therapy clinics	137	Validity (correlation) Between NDI and GRoC Between NPRS and GRoC	GRoC 15-points -7 (a very great deal worse) to zero (about the same) to +7 (a very great deal better)	GRoC was completed at follow up. Within a week
Cook et al (2014)	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-points -7 (a very great deal worse) to zero (about the same) to +7 (a very great deal better)	Baseline and at follow up 48- and 96-hours post baseline
Farooq et al. (2017)	Patients with neck pain	Physical therapy clinics	106	Validity (correlation) Between NDI-U and GRoC	GRoC 15-points -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
Guzy et al. (2013)	Patients with neck pain	Outpatient rehabilitation clinic	95	Validity (correlation) Between NDI-P and GRoC	GRoC 7-points 'complete recovery'' over ''no change'' to ''my complaints are worse than ever''	GRoC scale was completed at 2 weeks and at 4 weeks
Jorritsma et al. (2012)	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-points 3 (completely recovered) to zero (no change) to -3 (worse than ever)	After completion of the program varying from 3 to 5 months, patients filled the GPE
Kamper et al. (2010)	Patients with any whiplash-associated disorder.	Physical therapy clinics	134	Test-retest reliability	GPE 11-points -5 (vastly worse) to zero (unchanged) to +5 (completely recovered)	Baseline, 6 weeks, and 12 months
Monticone et al. 2017	Patients with chronic neck pain	Outpatient Rehabilitation Unit	153	Validity (correlation) Between NeckPix and GPE	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)	At the end of treatment (8 weeks) and one year before follow-up
Monticone et al. 2015	Patients with chronic neck pain	Outpatient Rehabilitation Unit	200	Validity (correlation) Between NDI and GPE Between NPDS and GPE	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)	At the end of treatment 8 weeks

Ngo et al. (2010)	Patients with WAD. Most participants (69.6%) had grade II WAD.	Interviewed by person or by telephone in Ontario	46	Test-retest reliability	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	3-5 days
Shaheen et al. (2015)	Patients with neck pain lasting more than 3 months	3 primary health centers	70	Validity (correlation) Between NDI-Ar and GRoC	GRoC 15-points -7 (a very great deal worse) to zero (about the same) to +7 (a very great deal better)	1 week
Takeshita et al. (2014)	Patients with neck pain, cervical radiculopathy and/or cervical myelopathy	Variety of clinics and hospital settings	130	Validity (correlation) Between NDI-J and GRoC	PGIC 7-points much better, better, slightly better, unchanged, slightly worse, worse and much worse	Over 8 weeks
Trouli et al. (2008)	Patients with neck pain	Primary healthcare clinic	68	Validity (correlation) Between NDI-Gr and GRoC	GRoC 15-points -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 months, but 1 week for test-retest
Tuttle et al. (2006)	Patients with neck pain for more than 2 weeks	Private physiotherapy clinics	29	Validity (correlation) Between NDI and GPE Between PSFS and GPE Between VAS and GPE Between ROM and GPE	GPE 11-points –5 is vastly worse and +5 is completely recovered	6 weeks
Young et al. (2009)	Patients presenting with mechanical neck pain	Outpatient physical therapy clinics.	91	Validity (correlation)	GRoC 15-points -7 (‘‘a very great deal worse’’) to 0 (‘‘about the same’’) to +7 (‘‘a very great deal better’’)	3 weeks

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

Study	Item Evaluation Criteria*													
	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

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*§ (Criteria)**	**Quality of Studies** (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).

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.*

Total score = (sum of subtotals ÷ 24 × 100). If for a specific paper an item is deemed NA (Not Applicable), then, Total score = (sum of subtotals ÷ (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

Bjorklund et al (2017)	Spearman's correlation between the change scores of
	GRoC and ProFitMap-neck
	GRoC and NDI
		rho = 0.47, (p<0.05)
		rho = 0.59, (p<0.05)
			Very Good	Excellent

Cleland et al. (2006)	Correlations (Pearson r) between change scores
	NDI and GRoC
	PSFS and GRoC
		r = 0.19
		r = 0.82
			Very Good	Excellent

Cleland et al. (2008)	Correlations (Pearson r) between change scores
	NDI and GRoC
	NRS and GRoC
		r = 0.58
		r = 0.57
			Very Good	Excellent

Cook et al. (2014)	Receiver operator characteristics (ROC) Within-session change Between-session change Between session change of Pain and GROC Sensitivity Specificity	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) 79.2% (62.2, 91.1)	Very Good	Excellent
Farooq et al. (2017)	Correlations (Pearson r) NDI-U	r = 0.50	Very Good	Excellent
Guzy et al. (2013)	Correlations (Pearson r) NDI vs GROC	Two- week interval (r = -0.73) Four-week interval (r = -0.56)	Very Good	Very good
Jorritsma et al. (2012)	Correlation between change scores of NPAD and GPE	r = 0.49 (95 % CI 0.30–0.64)	Very Good	Excellent
Monticone et al. (2017)	Correlations (Spearman) between change scores of the NeckPix© and GPE	rho = 0.69–0.82	Very Good	Excellent

Monticone et al. (2015)	Correlation (Spearman) between change scores
	NDI-I and GPE
	NDPS and GPE
		rho = 0.71, p<0.01
		rho = 0.59, p<0.01
			Very Good	Excellent

Shaheen et al. (2013)	Correlations (Spearman's) NDI-Ar and GROC	rho = 0.81, p<0.001	Very Good	Excellent

Takeshita et al. (2014)	Correlations
	NDI and PGIC
	NDI-J and PGIC
		Spearman (rho)
		rho = 0.47, p<o.oo1
		rho = 0.59, p<o.oo1
			Very Good	Very good

Trouli et al. (2008)	Correlation (Spearman's) GROC vs Gr-NDI	rho = 0.30, p=0.02	Very Good	

TABLE 4. SUMMARY OF RELIABILITY PROPERTIES OF GROC SCALES

y	Type of Reliability	Reliability Estimates	COSMIN	Qu
)	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	
	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'' 0.80 (0.62–0.99) when defined as ''very much better'' or ''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	

Tuttle et al. (2006)	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)
	PSFS vs GPE (post 1, minus pre-1)
	PSFS vs GPE (post 2, minus pre-1)
	PSFS vs GPE (post 2, minus pre-2)
	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)
	Total ROM (post 2, minus pre-1)
	Total ROM (post 2, minus pre-2)
		rho = 0.17
		rho = 0.01
		rho = 0.03
		rho = 0.06
		rho = 0.03
		rho = 0.03
		rho = 0.00
		rho = 0.05
		rho = 0.01
		rho = 0.03
		rho = 0.01
		rho = 0.00
			Very Good	Excellent

Young et al. (2009)	Correlations (Pearson's) between change scores NDI and GRoC	r =0.52 (p<0.01)	Very Good	Excellent

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

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).

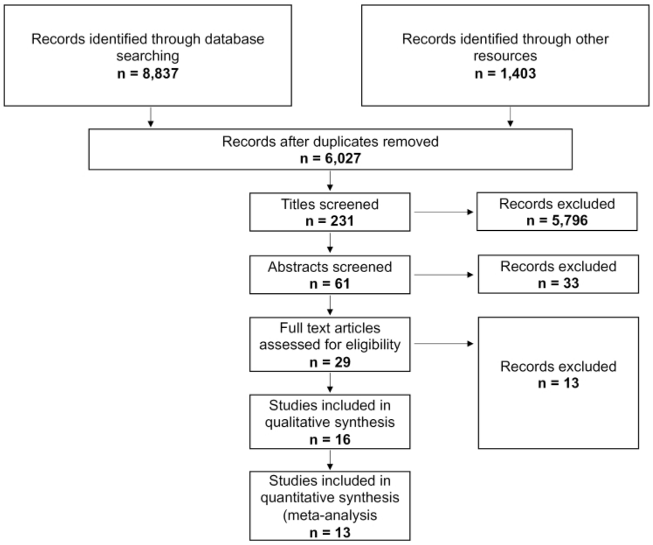
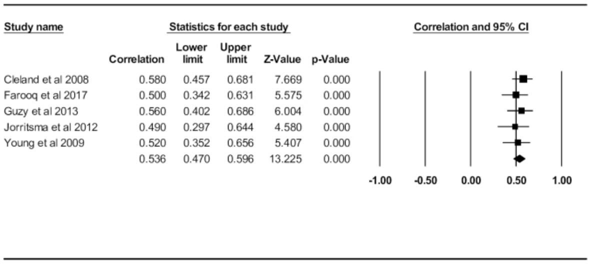

Records identified through database searching
n = 8,837
Records identified through other resources
n = 1,403
Records after duplicates removed
n = 6,027
Titles screened
n = 231
Records excluded
n = 5,796
Abstracts screened
n = 61
Records excluded
n = 33
Full text articles assessed for eligibility
n = 29
Records excluded
n = 13
Studies included in qualitative synthesis
n = 16
Studies included in quantitative synthesis (meta-analysis
n = 13

Figure 1. Flow diagram of included studies

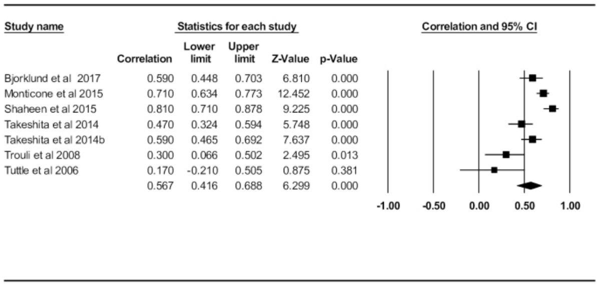
338x190mm (300 x 300 DPI)



Study name	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

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)



Study name	Statistics for each study					Correlation and 95% CI
	Correlation	Lower limit	Upper limit	Z-Value	p-Value	
Bjorklund et al 2017	0.590	0.448	0.703	6.810	0.000	
Monticone et al 2015	0.710	0.634	0.773	12.452	0.000	
Shaheen et al 2015	0.810	0.710	0.878	9.225	0.000	
Takeshita et al 2014	0.470	0.324	0.594	5.748	0.000	
Takeshita et al 2014b	0.590	0.465	0.692	7.637	0.000	
Trouli et al 2008	0.300	0.066	0.502	2.495	0.013	
Tuttle et al 2006	0.170	-0.210	0.505	0.875	0.381	
	0.567	0.416	0.688	6.299	0.000	

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.

215x279mm (300 x 300 DPI)

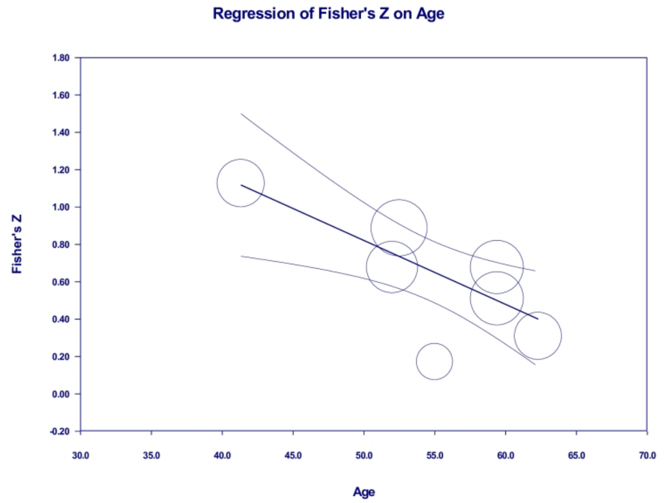


Figure 4. Random effects univariate meta-regression between age and the Fisher’s Z estimates. Each circle represents a study and the size of the circle indicates the influence of that study on the model. The regression prediction is illustrated by the straight line and the curved lines represent the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

215x279mm (300 x 300 DPI)

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.

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/
54. genital disease*.mp.
55. or/53-54
56. 52 not 55
57. 51 or 56
58. neck/
59. neck muscles/
60. exp cervical plexus/
61. exp cervical vertebrae/
62. atlanto-axial joint/
63. 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/
69. Thoracic Vertebrae/
70. cervical vertebrae.mp.
71. cervical plexus.mp.
72. cervical spine.mp.
73. (neck adj3 muscles).mp.
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/
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.
130. 128 or 129
131. exp genital diseases, female/

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

178. (collaborative research or collaborative review* or collaborative overview*).tw.
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

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	3-5
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	4-5
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	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
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	6
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	Appendix1
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	6
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	6-7
Data items	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
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	8-9

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	8-9
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	8=9
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	9
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	9-10
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	10
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	10-12
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	13
Risk of bias across studies	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-regression [see Item 16]).	13
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	14-15
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	16
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	14-15
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	18

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

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

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

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

Pavlos Bobos[1], Joy C MacDermid[2], Goris Nazari[3], Rochelle Furtado[4] and CATWAD co-authors[5]

[1]Pavlos Bobos PT, PhD(c), (corresponding author) Doctoral Candidate, Western's Bone and Joint Institute, Department of Health and Rehabilitation Sciences, Western University, Elborn College, 1201 Western Road, N6G 1H1, London, Ontario, Dalla Lana School of Public Health, Institute of Health Policy Management and Evaluation, Department of Clinical Epidemiology and Health Care Research, University of Toronto, Canada, (pbobos@uwo.ca), tel: +1 519 661 2111 x88912

[2]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 Health Centre, London, Ontario; Professor Rehabilitation Science McMaster University, Hamilton, ON, Canada (jmacderm@uwo.ca)

[3]Goris Nazari PT, PhD(c) Doctoral Candidate, Western's Bone and Joint Institute, School of Physical Therapy, Department of Health and Rehabilitation Sciences, Western University, London, Ontario, Canada, (gnazari@uwo.ca)

[4]Rochelle Furtado MSc Western's Bone and Joint Institute, School of Physical Therapy, Department of Health and Rehabilitation Sciences, Western University, London, Ontario, Canada, (rfurtad5@uwo.ca)

[5]**CATWAD:** Michele Sterling m.sterling@uq.edu.au, Anne Söderlund anne.soderlund@mdh.se, Michele Curatolo, curatolo@uw.edu, James M Elliott j-elliott@northwestern.edu, David Walton dwalton5@uwo.ca, Helge Kasch helgkasc@rm.dk, Linda Carroll linda.carroll@ualberta.ca, Hans Westergren Hans.Westergren@skane.se, Gwendolen Jull g.jull@uq.edu.au, Eva-Maj Malmström eva-maj.malmstrom@med.lu.se, Luke B Connelly l.connelly@uq.edu.au, Joy C MacDermid jmacderm@uwo.ca, Mandy Nielsen mandy.nielsen@griffith.edu.au, Pierre Côté pierre.cote@uoit.ca, Tonny Elmose Andersen tandersen@health.sdu.dk, Trudy Rebbeck trudy.rebbeck@sydney.edu.au, Annick Maujean a.maujean@uq.edu.au, Sarah Robins s.robins1@uq.edu.au, Kenneth Chen k.chen8@uq.edu.au, Julia Treleaven j.treleaven@uq.edu.au

Kewords: neck pain, global assessment, psychometric properties, systematic review

Word count: 3908

ABSTRACT

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

Data sources: A search was performed in 4 databases (MEDLINE, EMBASE, CINAHL, SCOPUS) until February 2019.

Data extraction and synthesis: Eligible articles were appraised using Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) checklist and the Quality Appraisal for Clinical Measurement Research Reports Evaluation Form.

Results: The search obtained 16 eligible studies and included in total 1533 patients with neck pain. 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: 0.47 to 0.59). Pooled data of Spearman's correlations indicated that GROC scores were moderately 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 reliability of GPE for patients with Whiplash Associated Disorders. Evidence from very good-to-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 4th 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 measurement properties of GROC scales themselves in patients with neck disorders. Therefore, this systematic review aims to critically appraise and synthesize the psychometric properties of the GROC scales in patients with neck disorders.

METHODS

Patient and Public Involvement

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

Study Design and Protocol Registration

We conducted a systematic review to evaluate the psychometric properties of GROC scales in patients with neck disorders. The protocol was registered in PROSPERO register database with registration number: CRD 42018117874

Eligibility Criteria

We included studies in this systematic review if the following criteria were met [10–12]:

- Design: psychometric testing, randomized/ cohort studies
- Participants: > 50% of the study's patient population with neck conditions/disorders,
- Intervention/Comparison: studies that reported on the psychometric properties (reliability, validity, responsiveness) of GROC, Global Perceived Effect (GPE) and Patient Global Impression of Change (PGIC),
- Outcomes: GROC, GPE and PGIC
- Articles were written in English language only

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 (≥ 0.20 to < 0.50), moderate (≥ 0.50 to < 0.80) or large (≥ 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^2 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.**

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.**

Reliability Measures

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 (β = -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%, 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-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 (I^2>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. 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

Not applicable

Availability of data and material

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

Funding Statement

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

Competing Interest Statement

None to report

References

1 Murray CJL, Abraham J, Ali MK, *et al.* The State of US health, 1990-2010: Burden of diseases, injuries, and risk factors. *JAMA - J Am Med Assoc* Published Online First: 2013. doi:10.1001/jama.2013.13805

2 Fejer R, Kyvik KO, Hartvigsen J. The prevalence of neck pain in the world population: A systematic critical review of the literature. Eur. Spine J. 2006. doi:10.1007/s00586-004-0864-4

3 Hogg-Johnson S, van der Velde G, Carroll LJ, *et al.* The Burden and Determinants of Neck Pain in the General Population. Results of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. *J Manipulative Physiol Ther* Published Online First: 2009. doi:10.1016/j.jmpt.2008.11.010

4 Bobos P, Nazari G, Palimeris S, *et al.* The contribution of health and psychological factors in patients with chronic neck pain and disability: A cross-sectional study. *J Clin Diagnostic Res* 2018;**12**:YC04–7. doi:10.7860/JCDR/2018/31284.11203

5 MacDermid JC, Walton DM, Bobos P, *et al.* A Qualitative Description of Chronic Neck Pain has Implications for Outcome Assessment and Classification. *Open Orthop J* 2017;**10**:746–56. doi:10.2174/1874325001610010746

6 Treleaven J. Sensorimotor disturbances in neck disorders affecting postural stability, head and eye movement control-Part 2: Case studies. *Man Ther* 2008;**13**:266–75. doi:10.1016/j.math.2007.11.002

7 Cohen SP. Epidemiology, diagnosis, and treatment of neck pain. *Mayo Clin Proc* 2015;**90**:284–99. doi:10.1016/j.mayocp.2014.09.008

8 Bobos P, Macdermid JC, Walton DM, *et al.* Patient-reported outcome measures used for neck disorders: An overview of systematic reviews. J. Orthop. Sports Phys. Ther. 2018;**48**:775–88. doi:10.2519/jospt.2018.8131

9 Kamper SJ, Maher CG, Mackay G. Global Rating of Change Scales: A Review of Strengths and Weaknesses and Considerations for Design. *J Man Manip Ther* 2009;**17**:163–70. doi:10.1002/mus.21062

10 Nazari G, Bobos P, MacDermid JC, *et al.* The Effectiveness of Instrument-Assisted Soft Tissue Mobilization in Athletes, Participants Without Extremity or Spinal Conditions, and Individuals with Upper Extremity, Lower Extremity, and Spinal Conditions: A Systematic Review. *Arch Phys Med Rehabil* Published Online First: February 2019. doi:10.1016/j.apmr.2019.01.017

11 Bobos P, Nazari G, Szekeres M, *et al.* The effectiveness of joint-protection programs on pain, hand function, and grip strength levels in patients with hand arthritis: A systematic review and meta-analysis. *J Hand Ther* 2018;**32**:194–211. doi:10.1016/j.jht.2018.09.012

12 Nazari G, Bobos P, MacDermid JC, *et al.* Psychometric properties of the Zephyr bioharness device: A systematic review. *BMC Sports Sci Med Rehabil* 2018;**10**. doi:10.1186/s13102-018-0094-4

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

14 Roy JS, Desmeules F, MacDermid JC. Psychometric properties of presenteeism scales for musculoskeletal disorders: A systematic review. *J Rehabil Med* Published Online First: 2011. doi:10.2340/16501977-0643

15 Sańchez J. Rosner, B.: Fundamentals of Biostatistics, third edition. PWS-Kent, Boston 1990, xv, 655 pp., ISBN 0-534-91973-1. *Biometrical J* 1993;**35**:150. doi:10.1002/bimj.4710350205

16 Wuensch KL, Evans JD. Straightforward Statistics for the Behavioral Sciences. *J Am Stat Assoc* Published Online First: 2006. doi:10.2307/2291607

17 Cohen J. Statistical power analysis for the behavioral sciences. Stat. Power Anal. Behav. Sci.

1988. doi:10.1234/12345678

18 Mokkink LB, de Vet HCW, Prinsen CAC, *et al.* COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. *Qual Life Res* Published Online First: 2018. doi:10.1007/s11136-017-1765-4

19 Terwee CB, Mokkink LB, Knol DL, *et al.* Rating the methodological quality in systematic reviews of studies on measurement properties : a scoring system for the COSMIN checklist. 2012;:651–7. doi:10.1007/s11136-011-9960-1

20 Viechtbauer W. Conducting meta-analisys in R with metafor package. *J Stat Softw* 2010;**36**:1–48.

21 Higgins JPT, Thompson SG, Deeks JJ, *et al.* Measuring inconsistency in meta-analyses. *BMJ* Published Online First: 2003. doi:10.1136/bmj.327.7414.557

22 Cleland J, Fritz J, Whitman J, *et al.* The reliability and construct validity of the Neck Disability Index and Patient Specific Functional Scale. *Spine (Phila Pa 1976)* 2006;**31**:598–602.

23 Cleland JA, Childs JD, Whitman JM. Psychometric Properties of the Neck Disability Index and Numeric Pain Rating Scale in Patients With Mechanical Neck Pain. *Arch Phys Med Rehabil* 2008;**89**:69–74. doi:10.1016/j.apmr.2007.08.126

24 Cook C, Lawrence J, Michalak K, *et al.* Is there preliminary value to a within- and/or between-session change for determining short-term outcomes of manual therapy on mechanical neck pain? *J Man Manip Ther* 2014;**22**:173–80. doi:10.1179/2042618614y.0000000071

25 Guzy G, Vernon H, Polczyk R, *et al.* Psychometric validation of the authorized Polish version of the Neck Disability Index. *Disabil Rehabil* 2013;**35**:2132–7. doi:10.3109/09638288.2013.771706

26 Kamper SJ, Ostelo RWJG, Knol DL, *et al.* Global Perceived Effect scales provided reliable assessments of health transition in people with musculoskeletal disorders, but ratings are strongly influenced by current status. *J Clin Epidemiol* 2010;**63**:760-766.e1. doi:10.1016/j.jclinepi.2009.09.009

27 Shaheen AAM, Omar MTA, Vernon H. Cross-cultural adaptation, reliability, and validity of the arabic version of neck disability index in patients with neck pain. *Spine (Phila Pa 1976)*

2013;**38**:609–15. doi:10.1097/BRS.0b013e31828b2d09

28 Takeshita K, Hosono N, Kawaguchi Y, *et al.* Validity, reliability and responsiveness of the Japanese version of the Neck Disability Index. *J Orthop Sci* 2013;**18**:14–21. doi:10.1007/s00776-012-0304-y

29 Trouli MN, Vernon HT, Kakavelakis KN, *et al.* Translation of the Neck Disability Index and validation of the Greek version in a sample of neck pain patients. *BMC Musculoskelet Disord* 2008;**9**:1–8. doi:10.1186/1471-2474-9-106

30 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. *Aust J Physiother* 2006;**52**:281–5. doi:10.1016/S0004-9514(06)70008-3

31 Ngo Trung, Stupar Maja, Coˆte´ 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

32 Björklund M, Wiitavaara B, Heiden M. Responsiveness and minimal important change for the ProFitMap-neck questionnaire and the Neck Disability Index in women with neck–shoulder pain. *Qual Life Res* 2017;**26**:161–70. doi:10.1007/s11136-016-1373-8

33 Evans R, Bronfort G, Maiers M, *et al.* ‘“ I know it ” s changed ’’: a mixed-methods study of the meaning of Global Perceived Effect in chronic neck pain patients. 2014;:888–97. doi:10.1007/s00586-013-3149-y

34 Jorritsma W, Dijkstra PU, De Vries GE, *et al.* Detecting relevant changes and responsiveness of Neck Pain and Disability Scale and Neck Disability Index. *Eur Spine J* 2012;**21**:2550–7. doi:10.1007/s00586-012-2407-8

35 Monticone M, Frigau L, Vernon H, *et al.* Responsiveness and minimal important change of the NeckPix© in subjects with chronic neck pain undergoing rehabilitation. *Eur Spine J* 2018;**27**:1324–31. doi:10.1007/s00586-017-5343-9

36 Monticone M, Ambrosini E, Vernon H, *et al.* Responsiveness and minimal important changes for the Neck Disability Index and the Neck Pain Disability Scale in Italian subjects with chronic neck pain. *Eur Spine J* 2015;**24**:2821–7. doi:10.1007/s00586-015-3785-5

37 Young BA, Walker MJ, Strunce JB, *et al.* Responsiveness of the Neck Disability Index in patients with mechanical neck disorders. *Spine J* 2009;**9**:802–8. doi:10.1016/j.spinee.2009.06.002

38 Farooq MN, Mohseni-Bandpei MA, Gilani SA, *et al.* Urdu version of the neck disability index: A reliability and validity study. *BMC Musculoskelet Disord* 2017;**18**:1–11. doi:10.1186/s12891-017-1469-5

39 Williams GN, Gangel TJ, Arciero RA, *et al.* Comparison of the single assessment numeric evaluation method and two shoulder rating scales. Outcomes measures after shoulder surgery. *Am J Sports Med* 1999;**27**:214–21. doi:10.1177/03635465990270021701

40 Schmitt J, Abbott JH. Global Ratings of Change Do Not Accurately Reflect Functional Change Over Time in Clinical Practice. *J Orthop Sport Phys Ther* 2015;**45**:106–11. doi:10.2519/jospt.2015.5247

41 Chiarotto A, Ostelo RW, Boers M, *et al.* A systematic review highlights the need to investigate the content validity of patient-reported outcome measures for physical functioning in patients with low back pain. *J Clin Epidemiol* 2018;**95**:73–93. doi:10.1016/j.jclinepi.2017.11.005

42 Ailliet L, Knol DL, Rubinstein SM, *et al.* Definition of the construct to be measured is a prerequisite for the assessment of validity. the Neck Disability Index as an example. *J Clin Epidemiol* 2013;**66**:775-782.e2. doi:10.1016/j.jclinepi.2013.02.005

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)

Table 1. Study Characteristics

Study	Population	Setting	Sample Size	Properties Evaluated	GROC evaluated	Interval
Bjorklund et al (2017)	Women with non-specific neck-shoulder pain	Not specified	104	Validity (correlation) Between NDI and GRoC	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 scale administered only after intervention at one time point (1 week)
Cleland et al (2006)	Patients with cervical radiculopathy	Hospital	38	Validity (correlation) Between NDI and GRoC Between PSFS and GRoC	GRoC 15-points -7 (a very great deal worse) to zero (about the same) to +7 (a very great deal better)	GRoC was completed at follow up. Within a week over the period of 7 weeks.
Cleland et al. (2008)	Patients with neck pain only	5 Outpatient physical therapy clinics	137	Validity (correlation) Between NDI and GRoC Between NPRS and GRoC	GRoC 15-points -7 (a very great deal worse) to zero (about the same) to +7 (a very great deal better)	GRoC was completed at follow up. Within a week
Cook et al (2014)	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-points -7 (a very great deal worse) to zero (about the same) to +7 (a very great deal better)	Baseline and at follow up 48- and 96-hours post baseline
Farooq et al. (2017)	Patients with neck pain	Physical therapy clinics	106	Validity (correlation) Between NDI-U and GRoC	GRoC 15-points -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
Guzy et al. (2013)	Patients with neck pain	Outpatient rehabilitation clinic	95	Validity (correlation) Between NDI-P and GRoC	GRoC 7-points 'complete recovery'' over ''no change'' to ''my complaints are worse than ever''	GRoC scale was completed at 2 weeks and at 4 weeks
Jorritsma et al. (2012)	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-points 3 (completely recovered) to zero (no change) to -3 (worse than ever)	After completion of the program varying from 3 to 5 months patients filled the GPE
Kamper et al. (2010)	Patients with any whiplash-associated disorder.	Physical therapy clinics	134	Test-retest reliability	GPE 11-points -5 (vastly worse) to zero (unchanged) to +5 (completely recovered)	Baseline, 6 weeks and 12 months
Monticone et al. 2017	Patients with chronic neck pain	Outpatient Rehabilitation Unit	153	Validity (correlation) Between NeckPix and GPE	GPE 5-points (helped a lot = 1, helped = 2), one no change level (helped only a little = 3), and two worsening	At the end of treatment (8 weeks) and one year before follow-up

					levels (did not help = 4, made things worse = 5)	
Monticone et al. 2015	Patients with chronic neck pain	Outpatient Rehabilitation Unit	200	Validity (correlation) Between NDI and GPE Between NPDS and GPE	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)	At the end of treatment 8 weeks
Ngo et al. (2010)	Patients with WAD. Most participants (69.6%) had grade II WAD.	Interviewed by person or by telephone in Ontario	46	Test-retest reliability	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	3-5 days
Shaheen et al. (2015)	Patients with neck pain lasting more than 3 months	3 primary health centers	70	Validity (correlation) Between NDI-Ar and GRoC	GRoC 15-points -7 (a very great deal worse) to zero (about the same) to +7 (a very great deal better)	1 week
Takeshita et al. (2014)	Patients with neck pain, cervical radiculopathy and/or cervical myelopathy	Variety of clinics and hospital settings	130	Validity (correlation) Between NDI-J and GRoC	PGIC 7-points much better, better, slightly better, unchanged, slightly worse, worse and much worse	Over 8 weeks
Trouli et al. (2008)	Patients with neck pain	Primary healthcare clinic	68	Validity (correlation) Between NDI-Gr and GRoC	GRoC 15-points -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 months, but 1 week for test-retest
Tuttle et al. (2006)	Patients with neck pain for more than 2 weeks	Private physiotherapy clinics	29	Validity (correlation) Between NDI and GPE Between PSFS and GPE Between VAS and GPE Between ROM and GPE	GPE 11-points –5 is vastly worse and +5 is completely recovered	6 weeks
Young et al. (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

clinics.	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*§ (Criteria)	Quality of Studies** (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

Study	Item Evaluation Criteria*													
	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.*

Total score = (sum of subtotals ÷ 24 × 100). If for a specific paper an item is deemed NA (Not Applicable), then, Total score = (sum of subtotals ÷ (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 4. Summary of reliability 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’’ 0.80 (0.62–0.99) when defined as ‘‘very much better’’ or ‘‘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

TABLE 5. Summary of validity properties of GRoC scales

Study	Type of Reliability	Validity Estimates	COSMIN	Quality of Studies

Bjorklund et al (2017)	Spearman's correlation between the change scores of GRoC and ProFitMap-neck
	GRoC and NDI
		rho = 0.47, (p<0.05)
		rho = 0.59, (p<0.05)
			Very Good	Excellent

Cleland et al. (2006)	Correlations (Pearson r) between change scores
	NDI and GRoC
	PSFS and GRoC
		r = 0.19
		r = 0.82
			Very Good	Excellent

Cleland et al. (2008)	Correlations (Pearson r) between change scores
	NDI and GRoC
	NRS and GRoC
		r = 0.58
		r = 0.57
			Very Good	Excellent

Cook et al. (2014)	Receiver operator characteristics (ROC) Within-session change Between-session change Between session change of Pain and GROC Sensitivity Specificity	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) 79.2% (62.2, 91.1)	Very Good	Excellent
Farooq et al. (2017)	Correlations (Pearson r) NDI-U	r =0.50	Very Good	Excellent
Guzy et al. (2013)	Correlations (Pearson r) NDI vs GROC	Two- week interval (r = -0.73) Four-week interval (r = -0.56)	Very Good	Very good
Jorritsma et al. (2012)	Correlation between change scores of NPAD and GPE	r = 0.49 (95 % CI 0.30–0.64)	Very Good	Excellent
Monticone et al. (2017)	Correlations (Spearman) between change scores of the NeckPix© and GPE	rho = 0.69–0.82	Very Good	Excellent

Monticone et al. (2015)	Correlation (Spearman) between change scores
	NDI-I and GPE
	NDPS and GPE
		rho = 0.71, p<0.01
		rho = 0.59, p<0.01
			Very Good	Excellent

Shaheen et al. (2013)	Correlations (Spearman's) NDI-Ar and GROC	rho = 0.81, p<o.oo1	Very Good	Excellent

Takeshita et al. (2014)	Correlations
	NDI and PGIC
	NDI-J and PGIC
		Spearman (rho)
		rho = 0.47, p<o.oo1
		rho = 0.59, p<o.oo1
			Very Good	Very good

Trouli et al. (2008)	Correlation (Spearman's) GROC vs Gr-NDI	rho = 0.30, p=0.02	Very Good	Excellent

Tuttle et al. (2006)	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)
	
	PSFS vs GPE (post 1, minus pre-1)
	PSFS vs GPE (post 2, minus pre-1)
	PSFS vs GPE (post 2, minus pre-2)
		rho = 0.17
		rho = 0.01
		rho = 0.03
		
		rho = 0.06
		rho = 0.03
		rho = 0.03
			Very Good	Excellent

	Pain Intensity (post 1, minus pre-1)
	Pain Intensity (post 2, minus pre-1)
	Pain Intensity (post 2, minus pre-2)
		rho = 0.00
		rho = 0.05
		rho = 0.01		

	Total ROM (post 1, minus pre-1)
	Total ROM (post 2, minus pre-1)
	Total ROM (post 2, minus pre-2)
		rho = 0.03
		rho = 0.01
		rho = 0.00		

Young et al. (2009)	Correlations (Pearson's) between change scores NDI and GRoC	r =0.52 (p<0.01)	Very Good	Excellent

Monticone et al. (2015)	Correlation (Spearman) between change scores
	NDI-I and GPE
	NDPS and GPE
		rho = 0.71, p<0.01
		rho = 0.59, p<0.01
			Very Good	Excellent

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?’’*

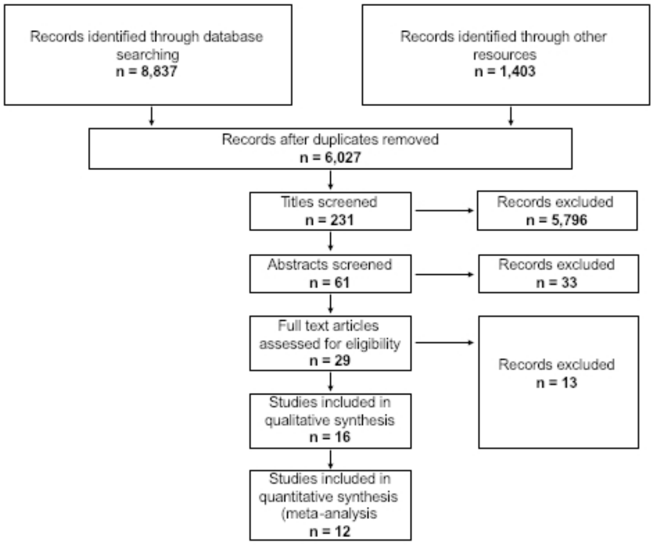

Records identified through database searching
n = 8,837
Records identified through other resources
n = 1,403
Records after duplicates removed
n = 6,027
Titles screened
n = 231
Records excluded
n = 5,796
Abstracts screened
n = 61
Records excluded
n = 33
Full text articles assessed for eligibility
n = 29
Records excluded
n = 13
Studies included in qualitative synthesis
n = 16
Studies included in quantitative synthesis (meta-analysis
n = 12

Figure 1. Flow diagram of included studies

60x34mm (300 x 300 DPI)

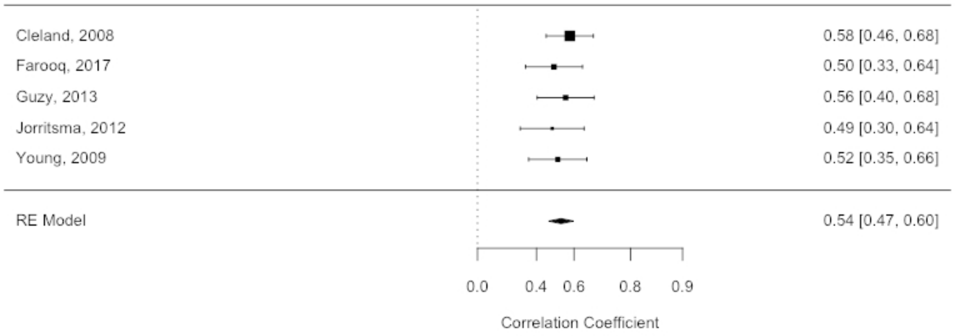


Figure 2. Meta-analysis of Pearson's correlation coefficients between neck disability change scores and GROC scores in patients with neck disorders based on 5 very good to excellent quality studies.

67x34mm (300 x 300 DPI)

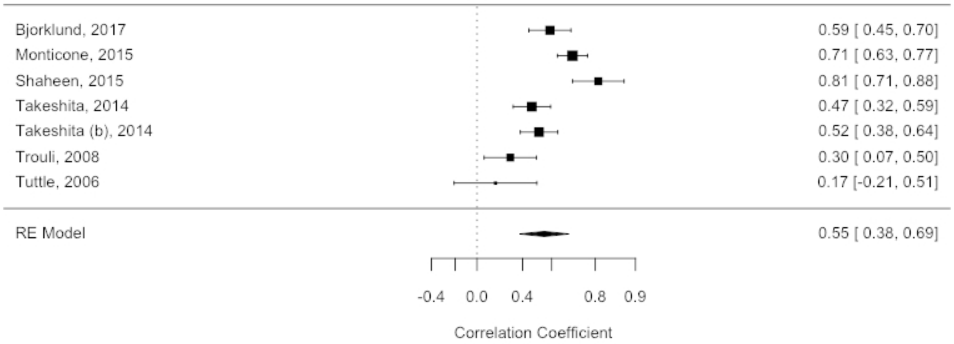


Figure 3. Meta-analysis of Spearman's correlation coefficients between neck disability change scores and GROC scores in patients with neck disorders based on 6 very good to excellent quality studies.

67x34mm (300 x 300 DPI)

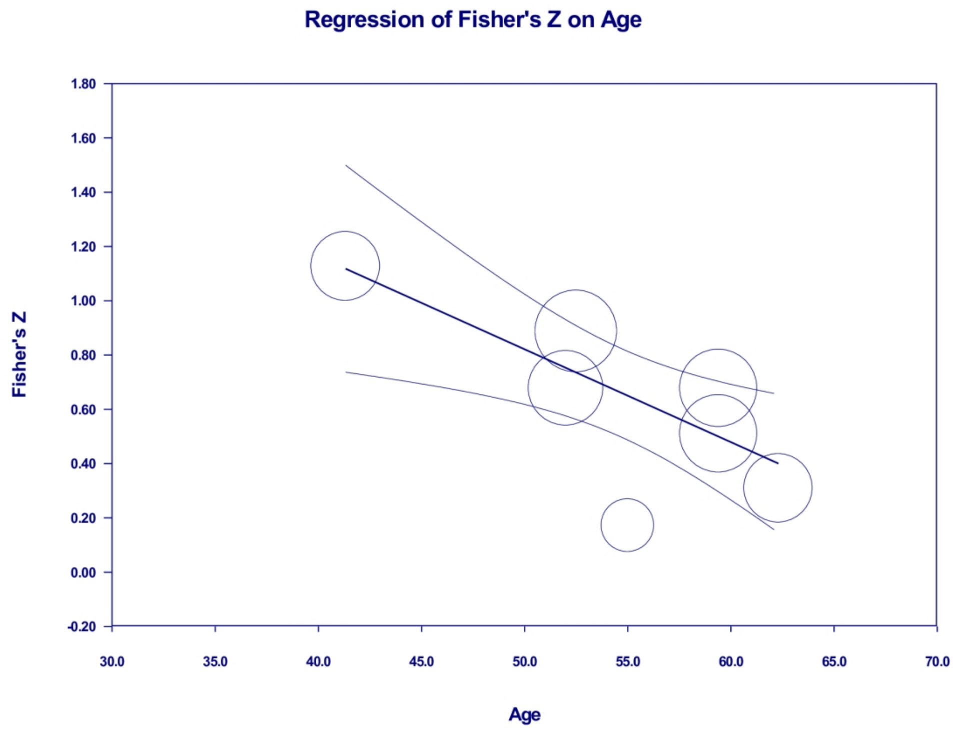


Figure 4. Random effects univariate meta-regression between age and the Fisher's Z estimates. Each circle represents a study and the size of the circle indicates the influence of that study on the model. The regression prediction is illustrated by the straight line and the curved lines represent the 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.

40. brachial neuralgia.mp.
41. neck pain.mp.
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/
54. genital disease*.mp.
55. or/53-54
56. 52 not 55
57. 51 or 56
58. neck/
59. neck muscles/
60. exp cervical plexus/
61. exp cervical vertebrae/
62. atlanto-axial joint/
63. 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/
69. Thoracic Vertebrae/
70. cervical vertebrae.mp.
71. cervical plexus.mp.
72. cervical spine.mp.
73. (neck adj3 muscles).mp.
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/

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.

130. 128 or 129
131. exp genital diseases, female/
132. genital disease*.mp.
133. exp *Uterus/
134. or/131-133
135. 130 not 134
136. (thoracic adj3 spine).mp.
137. cervical spine.mp.
138. 123 or 124 or 125 or 126 or 127 or 135 or 136 or 137
139. Intervertebral Disk/
140. (disc or discs).mp.
141. (disk or disks).mp.
142. 139 or 140 or 141
143. 138 and 142
144. herniat*.mp.
145. slipped.mp.
146. prolapse*.mp.
147. displace*.mp.
148. degenerat*.mp.
149. (bulge or bulged or bulging).mp.
150. 144 or 145 or 146 or 147 or 148 or 149
151. 143 and 150
152. intervertebral disk degeneration/ or intervertebral disk displacement/
153. intervertebral disk displacement.mp.
154. intervertebral disc displacement.mp.
155. intervertebral disk degeneration.mp.
156. intervertebral disc degeneration.mp.
157. 152 or 153 or 154 or 155 or 156
158. 138 and 157
159. 57 or 99 or 122 or 151 or 158
160. animals/ not (animals/ and humans/)
161. 159 not 160
162. exp *neoplasms/
163. exp *wounds, penetrating/
164. 162 or 163
165. 161 not 164
166. 29 and 165
167. guidelines as topic/
168. practice guidelines as topic/
169. guideline.pt.
170. practice guideline.pt.
171. (guideline? or guidance or recommendations).ti.
172. consensus.ti.
173. or/167-172
174. meta-analysis/
175. exp meta-analysis as topic/

176. (meta analy* or metaanaly* or met analy* or metanaly*).tw.
177. review literature as topic/
178. (collaborative research or collaborative review* or collaborative overview*).tw.
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

Quality Appraisal for Clinical Measurement Research Reports

Evaluation Form

Authors: ______________________________ Year: ____________________ Rater: ____

Use this form to rate the quality of a clinical measurement study. To decide which score to provide for each item on your quality checklist, pick the descriptor that sounds most *like what was reported in the study you are evaluating. Items rank descriptors are provided in the guide. (Forms and guides to extract study data for evidence synthesis are available from developer at macderj@mcmaster.ca)*

Evaluation criteria	Score		
Study question	2	1	0
1. Was the relevant background work cited to define what is currently known about the measurement properties of measures under study, and the potential contributions of the current research question to informing that knowledge base?			
Study Design			
2. Were appropriate inclusion/exclusion criteria defined?			
3. Were specific clinical measurement questions/hypotheses identified?			
4. Was an appropriate scope of measurement properties considered?			
5. Was an appropriate sample size used?			
6. Was appropriate retention/follow-up obtained? (for studies involving retesting; otherwise n/a)			
Measurements			
7. Were specific descriptions provided of the measure under study and the method(s) used to administer it?			
8. Were standardized procedures used to administer all study measures in a manner that minimized potential sources of error/bias (including the study measure and its comparators)?			
Analyses			

9. Were analyses conducted for each specific hypothesis or purpose?			
10. Were appropriate statistical tests performed to obtain point estimates of the measurement properties?			
11. Were appropriate ancillary analyses done to quantify the confidence in the estimates of the clinical measurement property (Precision/Confidence intervals; benchmark comparisons/ROC curves, alternate forms of analysis like SEM/MID, etc.)?			
Recommendations			
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?			
Subtotals (of columns 1 and 2)			
Total score (sum of subtotals/24*100); if for a specific paper or topic an item is deemed inappropriate then you can sum of items/2*number of items *100			

Quality Appraisal of a Clinical Measurement Study

Interpretation Guide

To decide which score to provide for each item on your quality checklist, read the following descriptors. Pick the descriptor that sounds *most* like the study you were evaluating with respect to a given item. If there is no documentation about any specific aspect of an item; then you must evaluate assuming that it was not done. Given the diversity in clinical measurement properties and design options, the evaluator has to make judgments using the criteria below and extend the principles to specific aspects that may not be covered in these brief exemplars. In many cases, the study will not look exactly like the descriptor so there will be some interpretation as to which level of optimal methods for clinical measurement studies have been achieved. In such cases, the evaluator can use the general approach that if this study research design and conduct is consistent with best practice (score=2); is acceptable but suboptimal (score=1); is not done/documented, substantially inadequate or inappropriate (score=0).

		Descriptors
Study question		
Score		
1	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.
	1	All of the above criteria were not fulfilled, but a sound rationale was provided for the research question.
	0	A foundation for the current research question was not clear; and the rationale was not founded on previous literature.
Study design		

2	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).
3	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 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).
4	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 – intra-rater, 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 rich view on measurement properties.
	1	Two or more clinical measurement properties were evaluated, however, scope was narrow and did not meet above criteria. (e.g., internal consistency and one other indicator of validity or reliability).
	0	The scope of clinical measurement properties was very narrow as indicated by a narrow evaluation of only one form of reliability or validity.

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 not 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.
6	2	90% or more of the patients enrolled for study were re-evaluated.
	1	70% or more of the enrolled patients were re-evaluated.
	0	Less than 70% of the patients enrolled in the study were re-evaluated
Measurements		
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 if 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.
	1	The test(s) and its administration procedures are referenced; but there is inadequate description of the test procedures.
	0	Minimal description of test procedures without appropriate references.

8	2	This item addresses the overall study procedures for administering all study measures (study measure and its comparators) in an unbiased way. Test procedures should not introduce systematic errors in the estimation of the clinical measurement properties. This includes standardized procedures for who completed or administered the measures. For self-report, this includes order of presentation, who completed at what time interval; handling of missing items. If relevant, then the paper should include how cultural literacy issues were handled (e.g., exclusion, assisted or surrogate completion). For impairment measures, procedures would include calibration of any equipment; use of consistent measurement tools and scoring, a priori exclusion of any participants likely to give invalid results/unable to complete testing (not exclusion of after enrollment); use of standardized instructions and test procedures. This can include order of administration of test and quality checking of scores. For reliability testing, the appropriate retest interval will depend on the nature of the condition; but for acute conditions it may require retesting within 48 hours; whereas chronic/stable conditions are commonly retested within 4-14 days. For estimation of clinical change, retest intervals should be ones during which a meaningful clinical change would have occurred (and from an intervention with known effectiveness). The evaluator decides overall whether this has sufficiently been addressed by the methods described.
	1	No obvious sources of bias in the study test protocol or how tests were performed/administered is apparent; but there were suboptimal procedures or an inadequate description of the measurement protocol to be insured control of bias or that procedures were standardized.
	0	No description of the overall procedures for administering study tests; OR an obvious source of bias in data collection methods.
Analyses		
9	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 results under specific subheadings or by demarcating which analyses addressed specific clinical measurement properties. Data was presented for each hypothesis/research question posed.
	1	Data was presented that addressed each of the measurement questions posed, but authors did not link specific analyses to specific research questions or hypotheses.
	0	Data was not presented for every hypothesis or clinical measurement property outlined in the purposes or methods.

10	2	Tests selected - 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 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	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 of 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.

	1	Either precision definition (confidence intervals) or appropriate benchmark comparison were used - NOT both. OR Some analyses were associated with indicators of precision or alternate form of analysis -but not all key indicators.
	0	Inappropriate use of benchmarks or confidence intervals; or indicators of precision or alternate form are absent
Recommendations		
12	2	Authors made specific conclusions and clinical measurement recommendations that were clearly related to each hypotheses/question posed in the study and that were supported by the data presented. Ideal 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 must be specific; and conclusions cannot overstate the clinical measurement properties observed the study; nor ignore suboptimal measurement properties found.
	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 of the findings. (The measure is "reliable and valid ") OR authors made specific clinical measurement recommendations; but for only some of the study hypotheses.
	0	Authors did not make conclusions about clinical measurement; OR made recommendations that were in contradiction to the actual data presented

List with excluded studies with reasons

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. Chien et al 2015	No properties for GRoC scales
6. Cruz et al. 2015	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. Sharma et al 2017	Ineligible population
12. Stevens et al 2019	Ineligible population
13. Meyer et al 2014	Ineligible population

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	3-5
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	4-5
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	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
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	6
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	Appendix1
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	6
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	6-7
Data items	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
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	8-9

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	8-9
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	8=9
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	9
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	9-10
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	10
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	10-12
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	13
Risk of bias across studies	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-regression [see Item 16]).	13
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	14-15
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	16
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	14-15
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	18

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

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

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

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

Pavlos Bobos[1], Joy C MacDermid[2], Goris Nazari[3], Rochelle Furtado[4] and CATWAD co-authors[5]

[1]Pavlos Bobos PT, PhD(c), (corresponding author) Doctoral Candidate, Western's Bone and Joint Institute, Department of Health and Rehabilitation Sciences, Western University, Elborn College, 1201 Western Road, N6G 1H1, London, Ontario, Dalla Lana School of Public Health, Institute of Health Policy Management and Evaluation, Department of Clinical Epidemiology and Health Care Research, University of Toronto, Canada, (pbobos@uwo.ca), tel: +1 519 661 2111 x88912

[2]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 Health Centre, London, Ontario; Professor Rehabilitation Science McMaster University, Hamilton, ON, Canada (jmacderm@uwo.ca)

[3]Goris Nazari PT, PhD(c) Doctoral Candidate, Western's Bone and Joint Institute, School of Physical Therapy, Department of Health and Rehabilitation Sciences, Western University, London, Ontario, Canada, (gnazari@uwo.ca)

[4]Rochelle Furtado MSc Western's Bone and Joint Institute, School of Physical Therapy, Department of Health and Rehabilitation Sciences, Western University, London, Ontario, Canada, (rfurtad5@uwo.ca)

[5]**CATWAD:** Michele Sterling m.sterling@uq.edu.au, Anne Söderlund anne.soderlund@mdh.se, Michele Curatolo, curatolo@uw.edu, James M Elliott j-elliott@northwestern.edu, David Walton dwalton5@uwo.ca, Helge Kasch helgkasc@rm.dk, Linda Carroll linda.carroll@ualberta.ca, Hans Westergren Hans.Westergren@skane.se, Gwendolen Jull g.jull@uq.edu.au, Eva-Maj Malmström eva-maj.malmstrom@med.lu.se, Luke B Connelly l.connelly@uq.edu.au, Joy C MacDermid jmacderm@uwo.ca, Mandy Nielsen mandy.nielsen@griffith.edu.au, Pierre Côté pierre.cote@uoit.ca, Tonny Elmose Andersen tandersen@health.sdu.dk, Trudy Rebbeck trudy.rebbeck@sydney.edu.au, Annick Maujean a.maujean@uq.edu.au, Sarah Robins s.robins1@uq.edu.au, Kenneth Chen k.chen8@uq.edu.au, Julia Treleaven j.treleaven@uq.edu.au

Kewords: neck pain, global assessment, psychometric properties, systematic review

Word count: 3908

ABSTRACT

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

Data sources: A search was performed in 4 databases (MEDLINE, EMBASE, CINAHL, SCOPUS) until February 2019.

Data extraction and synthesis: Eligible articles were appraised using Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) checklist and the Quality Appraisal for Clinical Measurement Research Reports Evaluation Form.

Results: The search obtained 16 eligible studies and included in total 1533 patients with neck pain. 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: 0.47 to 0.59). Pooled data of Spearman's correlations indicated that GROC scores were moderately 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 reliability of GPE for patients with Whiplash Associated Disorders. Evidence from very good-to-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 4th 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 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 appraise and synthesize the psychometric properties of the GROC scales in patients with neck disorders.

METHODS

Patient and Public Involvement

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

Study Design and Protocol Registration

We conducted a systematic review to evaluate the psychometric properties of GROC scales in patients with neck disorders. The protocol was registered in PROSPERO register database with registration number: CRD 42018117874

Eligibility Criteria

We included studies in this systematic review if the following criteria were met [10–12]:

- Design: psychometric testing, randomized/ cohort studies
- Participants: > 50% of the study's patient population with neck conditions/disorders,
- Intervention/Comparison: studies that reported on the psychometric properties (reliability, validity, responsiveness) of GROC, Global Perceived Effect (GPE) and Patient Global Impression of Change (PGIC),
- Outcomes: GROC, GPE and PGIC
- Articles were written in English language only

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 (≥ 0.20 to < 0.50), moderate (≥ 0.50 to < 0.80) or large (≥ 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^2 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.**

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.**

Reliability Measures

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 (β = -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%, 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 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 (I^2>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 current study

Funding Statement

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

Competing Interest Statement

None to report

References

1 Murray CJL, Abraham J, Ali MK, *et al.* The State of US health, 1990-2010: Burden of diseases, injuries, and risk factors. *JAMA - J Am Med Assoc* Published Online First: 2013. doi:10.1001/jama.2013.13805

2 Fejer R, Kyvik KO, Hartvigsen J. The prevalence of neck pain in the world population: A systematic critical review of the literature. Eur. Spine J. 2006. doi:10.1007/s00586-004-0864-4

3 Hogg-Johnson S, van der Velde G, Carroll LJ, *et al.* The Burden and Determinants of Neck Pain in the General Population. Results of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. *J Manipulative Physiol Ther* Published Online First: 2009. doi:10.1016/j.jmpt.2008.11.010

4 Bobos P, Nazari G, Palimeris S, *et al.* The contribution of health and psychological factors in patients with chronic neck pain and disability: A cross-sectional study. *J Clin Diagnostic Res* 2018;**12**:YC04–7. doi:10.7860/JCDR/2018/31284.11203

5 MacDermid JC, Walton DM, Bobos P, *et al.* A Qualitative Description of Chronic Neck Pain has Implications for Outcome Assessment and Classification. *Open Orthop J* 2017;**10**:746–56. doi:10.2174/1874325001610010746

6 Treleaven J. Sensorimotor disturbances in neck disorders affecting postural stability, head and eye

movement control-Part 2: Case studies. *Man Ther* 2008;**13**:266–75. doi:10.1016/j.math.2007.11.002

7 Cohen SP. Epidemiology, diagnosis, and treatment of neck pain. *Mayo Clin Proc* 2015;**90**:284–99. doi:10.1016/j.mayocp.2014.09.008

8 Bobos P, Macdermid JC, Walton DM, *et al.* Patient-reported outcome measures used for neck disorders: An overview of systematic reviews. J. Orthop. Sports Phys. Ther. 2018;**48**:775–88. doi:10.2519/jospt.2018.8131

9 Kamper SJ, Maher CG, Mackay G. Global Rating of Change Scales: A Review of Strengths and Weaknesses and Considerations for Design. *J Man Manip Ther* 2009;**17**:163–70. doi:10.1002/mus.21062

10 Nazari G, Bobos P, MacDermid JC, *et al.* The Effectiveness of Instrument-Assisted Soft Tissue Mobilization in Athletes, Participants Without Extremity or Spinal Conditions, and Individuals with Upper Extremity, Lower Extremity, and Spinal Conditions: A Systematic Review. *Arch Phys Med Rehabil* Published Online First: February 2019. doi:10.1016/j.apmr.2019.01.017

11 Bobos P, Nazari G, Szekeres M, *et al.* The effectiveness of joint-protection programs on pain, hand function, and grip strength levels in patients with hand arthritis: A systematic review and meta-analysis. *J Hand Ther* 2018;**32**:194–211. doi:10.1016/j.jht.2018.09.012

12 Nazari G, Bobos P, MacDermid JC, *et al.* Psychometric properties of the Zephyr bioharness device: A systematic review. *BMC Sports Sci Med Rehabil* 2018;**10**. doi:10.1186/s13102-018-0094-4

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

14 Roy JS, Desmeules F, MacDermid JC. Psychometric properties of presenteeism scales for musculoskeletal disorders: A systematic review. *J Rehabil Med* Published Online First: 2011. doi:10.2340/16501977-0643

15 Sańchez J. Rosner, B.: Fundamentals of Biostatistics, third edition. PWS-Kent, Boston 1990, xv,

655 pp., ISBN 0-534-91973-1. *Biometrical J* 1993;**35**:150. doi:10.1002/bimj.4710350205

16 Wuensch KL, Evans JD. Straightforward Statistics for the Behavioral Sciences. *J Am Stat Assoc* Published Online First: 2006. doi:10.2307/2291607

17 Cohen J. Statistical power analysis for the behavioral sciences. Stat. Power Anal. Behav. Sci. 1988. doi:10.1234/12345678

18 Mokkink LB, de Vet HCW, Prinsen CAC, *et al.* COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. *Qual Life Res* Published Online First: 2018. doi:10.1007/s11136-017-1765-4

19 Terwee CB, Mokkink LB, Knol DL, *et al.* Rating the methodological quality in systematic reviews of studies on measurement properties : a scoring system for the COSMIN checklist. 2012;:651–7. doi:10.1007/s11136-011-9960-1

20 Viechtbauer W. Conducting meta-analisys in R with metafor package. *J Stat Softw* 2010;**36**:1–48.

21 Higgins JPT, Thompson SG, Deeks JJ, *et al.* Measuring inconsistency in meta-analyses. *BMJ* Published Online First: 2003. doi:10.1136/bmj.327.7414.557

22 Cleland J, Fritz J, Whitman J, *et al.* The reliability and construct validity of the Neck Disability Index and Patient Specific Functional Scale. *Spine (Phila Pa 1976)* 2006;**31**:598–602.

23 Cleland JA, Childs JD, Whitman JM. Psychometric Properties of the Neck Disability Index and Numeric Pain Rating Scale in Patients With Mechanical Neck Pain. *Arch Phys Med Rehabil* 2008;**89**:69–74. doi:10.1016/j.apmr.2007.08.126

24 Cook C, Lawrence J, Michalak K, *et al.* Is there preliminary value to a within- and/or between-session change for determining short-term outcomes of manual therapy on mechanical neck pain? *J Man Manip Ther* 2014;**22**:173–80. doi:10.1179/2042618614y.0000000071

25 Guzy G, Vernon H, Polczyk R, *et al.* Psychometric validation of the authorized Polish version of the Neck Disability Index. *Disabil Rehabil* 2013;**35**:2132–7. doi:10.3109/09638288.2013.771706

26 Kamper SJ, Ostelo RWJG, Knol DL, *et al.* Global Perceived Effect scales provided reliable assessments of health transition in people with musculoskeletal disorders, but ratings are strongly

influenced by current status. *J Clin Epidemiol* 2010;**63**:760-766.e1. doi:10.1016/j.jclinepi.2009.09.009

27 Shaheen AAM, Omar MTA, Vernon H. Cross-cultural adaptation, reliability, and validity of the arabic version of neck disability index in patients with neck pain. *Spine (Phila Pa 1976)* 2013;**38**:609–15. doi:10.1097/BRS.0b013e31828b2d09

28 Takeshita K, Hosono N, Kawaguchi Y, *et al.* Validity, reliability and responsiveness of the Japanese version of the Neck Disability Index. *J Orthop Sci* 2013;**18**:14–21. doi:10.1007/s00776-012-0304-y

29 Trouli MN, Vernon HT, Kakavelakis KN, *et al.* Translation of the Neck Disability Index and validation of the Greek version in a sample of neck pain patients. *BMC Musculoskelet Disord* 2008;**9**:1–8. doi:10.1186/1471-2474-9-106

30 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. *Aust J Physiother* 2006;**52**:281–5. doi:10.1016/S0004-9514(06)70008-3

31 Ngo Trung, Stupar Maja, Coˆte´ 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

32 Björklund M, Wiitavaara B, Heiden M. Responsiveness and minimal important change for the ProFitMap-neck questionnaire and the Neck Disability Index in women with neck–shoulder pain. *Qual Life Res* 2017;**26**:161–70. doi:10.1007/s11136-016-1373-8

33 Evans R, Bronfort G, Maiers M, *et al.* '" I know it " s changed '': a mixed-methods study of the meaning of Global Perceived Effect in chronic neck pain patients. 2014;:888–97. doi:10.1007/s00586-013-3149-y

34 Jorritsma W, Dijkstra PU, De Vries GE, *et al.* Detecting relevant changes and responsiveness of Neck Pain and Disability Scale and Neck Disability Index. *Eur Spine J* 2012;**21**:2550–7.

doi:10.1007/s00586-012-2407-8

35 Monticone M, Frigau L, Vernon H, *et al.* Responsiveness and minimal important change of the NeckPix© in subjects with chronic neck pain undergoing rehabilitation. *Eur Spine J* 2018;**27**:1324–31. doi:10.1007/s00586-017-5343-9

36 Monticone M, Ambrosini E, Vernon H, *et al.* Responsiveness and minimal important changes for the Neck Disability Index and the Neck Pain Disability Scale in Italian subjects with chronic neck pain. *Eur Spine J* 2015;**24**:2821–7. doi:10.1007/s00586-015-3785-5

37 Young BA, Walker MJ, Strunce JB, *et al.* Responsiveness of the Neck Disability Index in patients with mechanical neck disorders. *Spine J* 2009;**9**:802–8. doi:10.1016/j.spinee.2009.06.002

38 Farooq MN, Mohseni-Bandpei MA, Gilani SA, *et al.* Urdu version of the neck disability index: A reliability and validity study. *BMC Musculoskelet Disord* 2017;**18**:1–11. doi:10.1186/s12891-017-1469-5

39 Williams GN, Gangel TJ, Arciero RA, *et al.* Comparison of the single assessment numeric evaluation method and two shoulder rating scales. Outcomes measures after shoulder surgery. *Am J Sports Med* 1999;**27**:214–21. doi:10.1177/03635465990270021701

40 Schmitt J, Abbott JH. Global Ratings of Change Do Not Accurately Reflect Functional Change Over Time in Clinical Practice. *J Orthop Sport Phys Ther* 2015;**45**:106–11. doi:10.2519/jospt.2015.5247

41 Chiarotto A, Ostelo RW, Boers M, *et al.* A systematic review highlights the need to investigate the content validity of patient-reported outcome measures for physical functioning in patients with low back pain. *J Clin Epidemiol* 2018;**95**:73–93. doi:10.1016/j.jclinepi.2017.11.005

42 Ailliet L, Knol DL, Rubinstein SM, *et al.* Definition of the construct to be measured is a prerequisite for the assessment of validity. the Neck Disability Index as an example. *J Clin Epidemiol* 2013;**66**:775-782.e2. doi:10.1016/j.jclinepi.2013.02.005

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)

Table 1. Study Characteristics

Study	Population	Setting	Sample Size	Properties Evaluated	GROC evaluated (ranked categories)	Interval
Bjorklund et al (2017)	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. Very much improved.	GRoC scale administered only after intervention at one time point (1 week)
Cleland et al (2006)	Patients with cervical radiculopathy	Hospital	38	Validity (correlation) Between NDI and GRoC Between PSFS 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. Within a week over the period of 7 weeks.
Cleland et al. (2008)	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. Within a week
Cook et al (2014)	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 follow up 48- and 96-hours post baseline
Farooq et al. (2017)	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
Guzy et al. (2013)	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 completed at 2 weeks and at 4 weeks
Jorritsma et al. (2012)	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 the program varying from 3 to 5 months patients filled the GPE

Kamper et al. (2010)	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
Monticone et al. 2017	Patients with chronic neck pain	Outpatient Rehabilitation Unit	153	Validity (correlation) Between NeckPix and GPE	GPE (5) (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
Monticone et al. 2015	Patients with chronic neck pain	Outpatient Rehabilitation Unit	200	Validity (correlation) Between NDI and GPE Between NPDS and GPE	GPE (5) (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
Ngo et al. (2010)	Patients with WAD. Most participants (69.6%) had grade II WAD.	Interviewed by person or by telephone in Ontario	46	Test-retest reliability	GPE (7) 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	3-5 days
Shaheen et al. (2015)	Patients with neck pain lasting more than 3 months	3 primary health centers	70	Validity (correlation) Between NDI-Ar and GRoC	GRoC (15) -7 (a very great deal worse) to zero (about the same) to +7 (a very great deal better)	1 week
Takeshita et al. (2014)	Patients with neck pain, cervical radiculopathy and/or cervical myelopathy	Variety of clinics and hospital settings	130	Validity (correlation) Between NDI-J and GRoC	PGIC (7) much better, better, slightly better, unchanged, slightly worse, worse and much worse	Over 8 weeks
Trouli et al. (2008)	Patients with neck pain	Primary healthcare clinic	68	Validity (correlation) Between NDI-Gr and GRoC	GRoC (15) -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 months, but 1 week for test-retest

Tuttle et al. (2006)	Patients with neck pain for more than 2 weeks	Private physiotherapy clinics	29	Validity (correlation) Between NDI and GPE Between PSFS and GPE Between VAS and GPE Between ROM and GPE	GPE (11) –5 is vastly worse and +5 is completely recovered	6 weeks
Young et al. (2009)	Patients presenting with mechanical neck pain	Outpatient physical therapy clinics.	91	Validity (correlation)	GRoC (15) -7 (‘‘a very great deal worse’’) to 0 (‘‘about the same’’) to +7 (‘‘a very great deal better’’)	3 weeks

NDI = Neck Disaiblity Index, NPRS=Numeric Pain Rating Scale, PSFS= Patient Specific Functional Scale, ROC= Receiver Operator Characteristic, VAS=Visual Analog Scale, NPAD=Neck Pain and Disability Scale, AUC= Area Under the Curve, ROM=Range of Motion

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

Study	Item Evaluation Criteria*												Total (%)	Quality Summary
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**		
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.*

Total score = (sum of subtotals ÷ 24 × 100). If for a specific paper an item is deemed NA (Not Applicable), then, Total score = (sum of subtotals ÷ (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 4. Summary of reliability 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’’ 0.80 (0.62–0.99) when defined as ‘‘very much better’’ or ‘‘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

TABLE 5. Summary of validity properties of GRoC scales

Study	Type of Reliability	Validity Estimates	COSMIN	Quality of Studies

Bjorklund et al (2017)	Spearman's correlation between the change scores of GRoC and ProFitMap-neck
	GRoC and NDI
		rho = 0.47, (p<0.05)
		rho = 0.59, (p<0.05)
			Very Good	Excellent

Cleland et al. (2006)	Correlations (Pearson r) between change scores
	NDI and GRoC
	PSFS and GRoC
		r = 0.19
		r = 0.82
			Very Good	Excellent

Cleland et al. (2008)	Correlations (Pearson r) between change scores
	NDI and GRoC
	NRS and GRoC
		r = 0.58
		r = 0.57
			Very Good	Excellent

Cook et al. (2014)	Receiver operator characteristics (ROC) Within-session change Between-session change Between session change of Pain and GROC Sensitivity Specificity	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) 79.2% (62.2, 91.1)	Very Good	Excellent
Farooq et al. (2017)	Correlations (Pearson r) NDI-U	r =0.50	Very Good	Excellent
Guzy et al. (2013)	Correlations (Pearson r) NDI vs GROC	Two- week interval (r = -0.73) Four-week interval (r = -0.56)	Very Good	Very good
Jorritsma et al. (2012)	Correlation between change scores of NPAD and GPE	r = 0.49 (95 % CI 0.30–0.64)	Very Good	Excellent
Monticone et al. (2017)	Correlations (Spearman) between change scores of the NeckPix© and GPE	rho = 0.69–0.82	Very Good	Excellent

Monticone et al. (2015)	Correlation (Spearman) between change scores
	NDI-I and GPE
	NDPS and GPE
		rho = 0.71, p<0.01
		rho = 0.59, p<0.01
			Very Good	Excellent

Shaheen et al. (2013)	Correlations (Spearman's) NDI-Ar and GROC	rho = 0.81, p<o.oo1	Very Good	Excellent

Takeshita et al. (2014)	Correlations
	NDI and PGIC
	NDI-J and PGIC
		Spearman (rho)
		rho = 0.47, p<o.oo1
		rho = 0.59, p<o.oo1
			Very Good	Very good

Trouli et al. (2008)	Correlation (Spearman's) GROC vs Gr-NDI	rho = 0.30, p=0.02	Very Good	Excellent

Tuttle et al. (2006)	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)
	PSFS vs GPE (post 1, minus pre-1)
	PSFS vs GPE (post 2, minus pre-1)
	PSFS vs GPE (post 2, minus pre-2)
		rho = 0.17
		rho = 0.01
		rho = 0.03
		rho = 0.06
		rho = 0.03
		rho = 0.03
			Very Good	Excellent

	Pain Intensity (post 1, minus pre-1)
	Pain Intensity (post 2, minus pre-1)
	Pain Intensity (post 2, minus pre-2)
		rho = 0.00
		rho = 0.05
		rho = 0.01		

	Total ROM (post 1, minus pre-1)
	Total ROM (post 2, minus pre-1)
	Total ROM (post 2, minus pre-2)
		rho = 0.03
		rho = 0.01
		rho = 0.00		

Young et al. (2009)	Correlations (Pearson's) between change scores NDI and GRoC	r =0.52 (p<0.01)	Very Good	Excellent

Monticone et al. (2015)	Correlation (Spearman) between change scores
	NDI-I and GPE
	NDPS and GPE
		rho = 0.71, p<0.01
		rho = 0.59, p<0.01
			Very Good	Excellent

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

Author	GROC (ranked categories)	Patients with neck disorders were asked:
Bjorklund et al. (2017)	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 regarding my neck–shoulder problem is now’’*
Evans et al (2014)	GPE (9)	*‘‘Overall, how much has your neck pain changed since you started treatment in the study?’’*
Kamper et al. (2010)	GPE (11)	*“With respect to your whiplash injury how would you describe yourself now compared to immediately after your accident”*
Monticone et al. (2017)	GPE (5)	*“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)	*‘‘Overall, how much did the treatment you received help your neck problem?’’*
Ngo et al. (2010)	GPE (7)	*‘‘How well do you feel you are recovering from your injuries?’’* *‘‘How do you feel your neck pain has changed since the injury?’’*

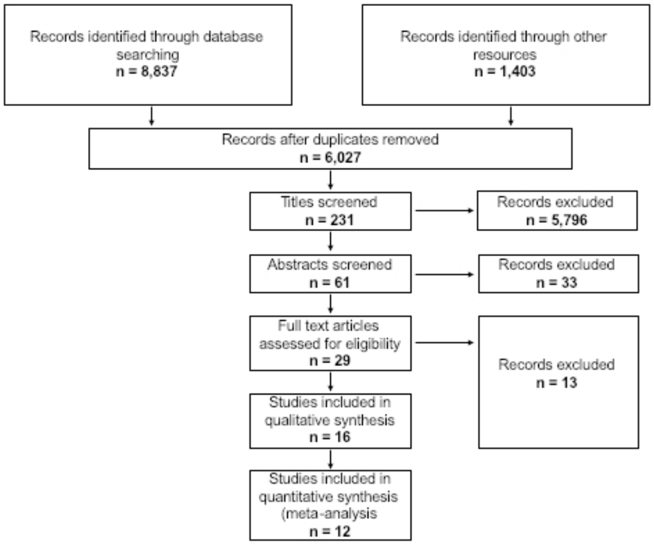

Records identified through database searching
n = 8,837
Records identified through other resources
n = 1,403
Records after duplicates removed
n = 6,027
Titles screened
n = 231
Records excluded
n = 5,796
Abstracts screened
n = 61
Records excluded
n = 33
Full text articles assessed for eligibility
n = 29
Records excluded
n = 13
Studies included in qualitative synthesis
n = 16
Studies included in quantitative synthesis (meta-analysis
n = 12

Figure 1. Flow diagram of included studies

60x34mm (300 x 300 DPI)

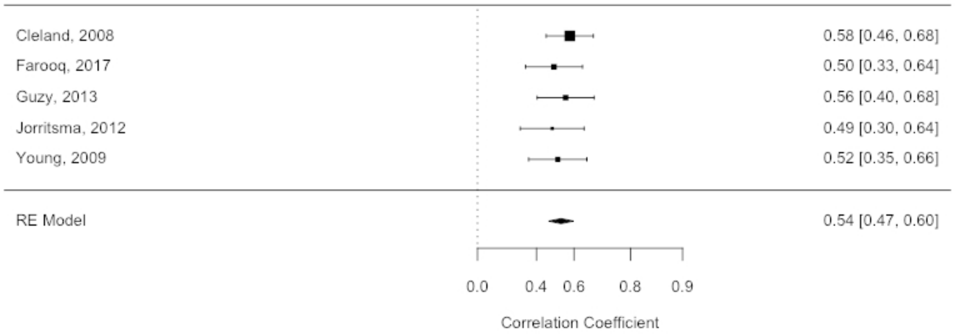


Figure 2. Meta-analysis of Pearson’s correlation coefficients between neck disability change scores and GROC scores in patients with neck disorders based on 5 very good to excellent quality studies.

67x34mm (300 x 300 DPI)

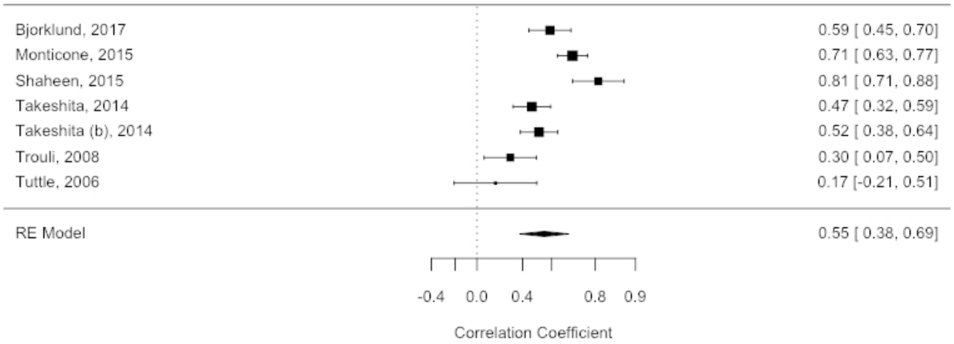


Figure 3. Meta-analysis of Spearman’s correlation coefficients between neck disability change scores and GROC scores in patients with neck disorders based on 6 very good to excellent quality studies.

67x34mm (300 x 300 DPI)

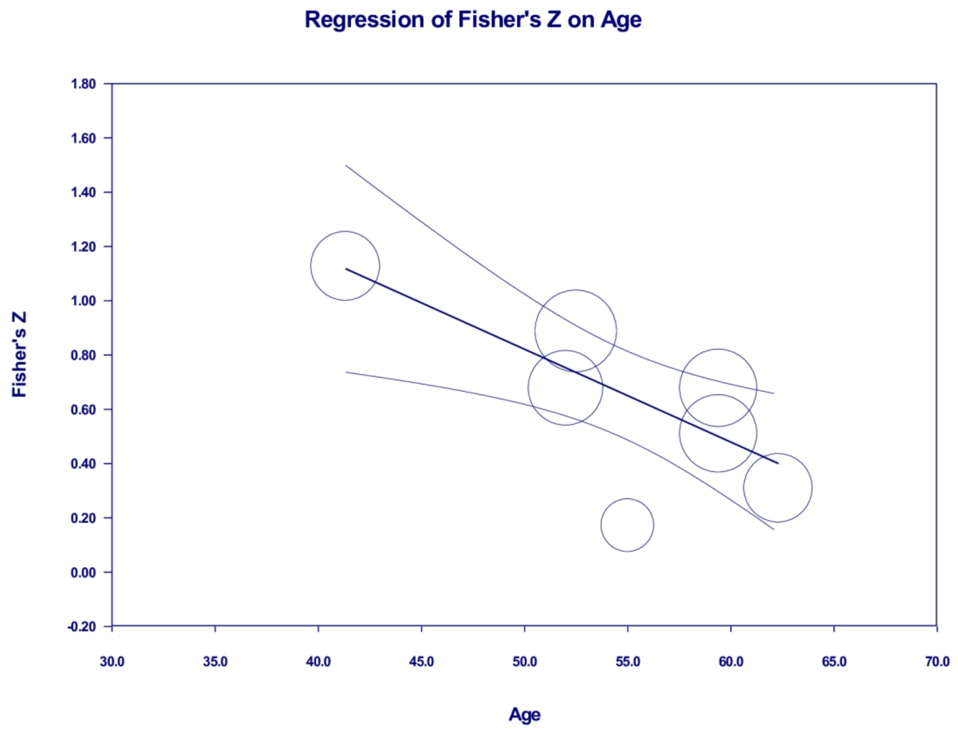


Figure 4. Random effects univariate meta-regression between age and the Fisher’s Z estimates. Each circle represents a study and the size of the circle indicates the influence of that study on the model. The regression prediction is illustrated by the straight line and the curved lines represent the 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.

40. brachial neuralgia.mp.
41. neck pain.mp.
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/
54. genital disease*.mp.
55. or/53-54
56. 52 not 55
57. 51 or 56
58. neck/
59. neck muscles/
60. exp cervical plexus/
61. exp cervical vertebrae/
62. atlanto-axial joint/
63. 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/
69. Thoracic Vertebrae/
70. cervical vertebrae.mp.
71. cervical plexus.mp.
72. cervical spine.mp.
73. (neck adj3 muscles).mp.
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/

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.

130. 128 or 129
131. exp genital diseases, female/
132. genital disease*.mp.
133. exp *Uterus/
134. or/131-133
135. 130 not 134
136. (thoracic adj3 spine).mp.
137. cervical spine.mp.
138. 123 or 124 or 125 or 126 or 127 or 135 or 136 or 137
139. Intervertebral Disk/
140. (disc or discs).mp.
141. (disk or disks).mp.
142. 139 or 140 or 141
143. 138 and 142
144. herniat*.mp.
145. slipped.mp.
146. prolapse*.mp.
147. displace*.mp.
148. degenerat*.mp.
149. (bulge or bulged or bulging).mp.
150. 144 or 145 or 146 or 147 or 148 or 149
151. 143 and 150
152. intervertebral disk degeneration/ or intervertebral disk displacement/
153. intervertebral disk displacement.mp.
154. intervertebral disc displacement.mp.
155. intervertebral disk degeneration.mp.
156. intervertebral disc degeneration.mp.
157. 152 or 153 or 154 or 155 or 156
158. 138 and 157
159. 57 or 99 or 122 or 151 or 158
160. animals/ not (animals/ and humans/)
161. 159 not 160
162. exp *neoplasms/
163. exp *wounds, penetrating/
164. 162 or 163
165. 161 not 164
166. 29 and 165
167. guidelines as topic/
168. practice guidelines as topic/
169. guideline.pt.
170. practice guideline.pt.
171. (guideline? or guidance or recommendations).ti.
172. consensus.ti.
173. or/167-172
174. meta-analysis/
175. exp meta-analysis as topic/

176. (meta analy* or metaanaly* or met analy* or metanaly*).tw.
177. review literature as topic/
178. (collaborative research or collaborative review* or collaborative overview*).tw.
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

Quality Appraisal for Clinical Measurement Research Reports

Evaluation Form

Authors: ________________________________ Year: _____________________ Rater: _____

Use this form to rate the quality of a clinical measurement study. To decide which score to provide for each item on your quality checklist, pick the descriptor that sounds most like what was reported in the study you are evaluating. Items rank descriptors are provided in the guide. (Forms and guides to extract study data for evidence synthesis are available from developer at macderj@mcmaster.ca)

Evaluation criteria	**Score**		
Study question	2	1	0
1. Was the relevant background work cited to define what is currently known about the measurement properties of measures under study, and the potential contributions of the current research question to informing that knowledge base?			
Study Design			
2. Were appropriate inclusion/exclusion criteria defined?			
3. Were specific clinical measurement questions/hypotheses identified?			
4. Was an appropriate scope of measurement properties considered?			
5. Was an appropriate sample size used?			
6. Was appropriate retention/follow-up obtained? (for studies involving retesting; otherwise n/a)			
Measurements			
7. Were specific descriptions provided of the measure under study and the method(s) used to administer it?			
8. Were standardized procedures used to administer all study measures in a manner that minimized potential sources of error/bias (including the study measure and its comparators)?			
Analyses			

9. Were analyses conducted for each specific hypothesis or purpose?			
10. Were appropriate statistical tests performed to obtain point estimates of the measurement properties?			
11. Were appropriate ancillary analyses done to quantify the confidence in the estimates of the clinical measurement property (Precision/Confidence intervals; benchmark comparisons/ROC curves, alternate forms of analysis like SEM/MID, etc.)?			
Recommendations			
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?			
Subtotals (of columns 1 and 2)			
Total score (sum of subtotals/24*100); if for a specific paper or topic an item is deemed inappropriate then you can sum of items/2*number of items *100			

© MacDermid 2011

Quality Appraisal of a Clinical Measurement Study

Interpretation Guide

To decide which score to provide for each item on your quality checklist, read the following descriptors. Pick the descriptor that sounds *most* like the study you were evaluating with respect to a given item. If there is no documentation about any specific aspect of an item; then you must evaluate assuming that it was not done. Given the diversity in clinical measurement properties and design options, the evaluator has to make judgments using the criteria below and extend the principles to specific aspects that may not be covered in these brief exemplars. In many cases, the study will not look exactly like the descriptor so there will be some interpretation as to which level of optimal methods for clinical measurement studies have been achieved. In such cases, the evaluator can use the general approach that if this study research design and conduct is consistent with best practice (score=2); is acceptable but suboptimal (score=1); is not done/documented, substantially inadequate or inappropriate (score=0).

		Descriptors
Study question		
Score		
1	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.
	1	All of the above criteria were not fulfilled, but a sound rationale was provided for the research question.
	0	A foundation for the current research question was not clear; and the rationale was not founded on previous literature.
Study design		

2	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).
3	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 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).
4	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 – intra-rater, 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 rich view on measurement properties.
	1	Two or more clinical measurement properties were evaluated, however, scope was narrow and did not meet above criteria. (e.g., internal consistency and one other indicator of validity or reliability).
	0	The scope of clinical measurement properties was very narrow as indicated by a narrow evaluation of only one form of reliability or validity.

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 not 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.
6	2	90% or more of the patients enrolled for study were re-evaluated.
	1	70% or more of the enrolled patients were re-evaluated.
	0	Less than 70% of the patients enrolled in the study were re-evaluated
Measurements		
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 if 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.
	1	The test(s) and its administration procedures are referenced; but there is inadequate description of the test procedures.
	0	Minimal description of test procedures without appropriate references.

8	2	This item addresses the overall study procedures for administering all study measures (study measure and its comparators) in an unbiased way. Test procedures should not introduce systematic errors in the estimation of the clinical measurement properties. This includes standardized procedures for who completed or administered the measures. For self-report, this includes order of presentation, who completed at what time interval; handling of missing items. If relevant, then the paper should include how cultural literacy issues were handled (e.g., exclusion, assisted or surrogate completion). For impairment measures, procedures would include calibration of any equipment; use of consistent measurement tools and scoring, a priori exclusion of any participants likely to give invalid results/unable to complete testing (not exclusion of after enrollment); use of standardized instructions and test procedures. This can include order of administration of test and quality checking of scores. For reliability testing, the appropriate retest interval will depend on the nature of the condition; but for acute conditions it may require retesting within 48 hours; whereas chronic/stable conditions are commonly retested within 4-14 days. For estimation of clinical change, retest intervals should be ones during which a meaningful clinical change would have occurred (and from an intervention with known effectiveness). The evaluator decides overall whether this has sufficiently been addressed by the methods described.
	1	No obvious sources of bias in the study test protocol or how tests were performed/administered is apparent; but there were suboptimal procedures or an inadequate description of the measurement protocol to be insured control of bias or that procedures were standardized.
	0	No description of the overall procedures for administering study tests; OR an obvious source of bias in data collection methods.
Analyses		
9	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 results under specific subheadings or by demarcating which analyses addressed specific clinical measurement properties. Data was presented for each hypothesis/research question posed.
	1	Data was presented that addressed each of the measurement questions posed, but authors did not link specific analyses to specific research questions or hypotheses.
	0	Data was not presented for every hypothesis or clinical measurement property outlined in the purposes or methods.

10	2	Tests selected - 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 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	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 of 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.

	1	Either precision definition (confidence intervals) or appropriate benchmark comparison were used - NOT both. OR Some analyses were associated with indicators of precision or alternate form of analysis -but not all key indicators.
	0	Inappropriate use of benchmarks or confidence intervals; or indicators of precision or alternate form are absent
Recommendations		
12	2	Authors made specific conclusions and clinical measurement recommendations that were clearly related to each hypotheses/question posed in the study and that were supported by the data presented. Ideal 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 must be specific; and conclusions cannot overstate the clinical measurement properties observed the study; nor ignore suboptimal measurement properties found.
	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 of the findings. (The measure is “reliable and valid ”) OR authors made specific clinical measurement recommendations; but for only some of the study hypotheses.
	0	Authors did not make conclusions about clinical measurement; OR made recommendations that were in contradiction to the actual data presented

List with excluded studies with reasons

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. Chien et al 2015	No properties for GRoC scales
6. Cruz et al. 2015	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. Sharma et al 2017	Ineligible population
12. Stevens et al 2019	Ineligible population
13. Meyer et al 2014	Ineligible population

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	3-5
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	4-5
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	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
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	6
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	Appendix1
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	6
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	6-7
Data items	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
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	8-9

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	8-9
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	8=9
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	9
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	9-10
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	10
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	10-12
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	13
Risk of bias across studies	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-regression [see Item 16]).	13
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	14-15
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	16
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	14-15
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	18

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

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