Journal of Orthopaedics 21 (2020) 40–48

Contents lists available at ScienceDirect

Journal of Orthopaedics

journal homepage: www.elsevier.com/locate/jor

Original Article

Psychometric properties of the global rating of change scales in patients with low back pain, upper and lower extremity disorders. A systematic review with meta-analysis

Pavlos Bobos[a,b,d,*], Christina Ziebart[a,b], Rochelle Furtado[a,b], Ze Lu[b,c], Joy C. MacDermid[a,b,c]

[a] Western's Bone and Joint Institute, School of Physical Therapy, Department of Health and Rehabilitation Sciences, Western University, London, Ontario, Canada
[b] Roth McFarlane Hand and Upper Limb Centre, St. Joseph's Hospital, London, ON, Canada
[c] School of Rehabilitation Science, McMaster University, Hamilton, Ontario, Canada
[d] 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

ARTICLE INFO

Keywords:
Global rating of change
Psychometrics
Low back pain
Shoulder pain
Musculoskeletal disorders

ABSTRACT

Objective: The purpose of this systematic review was to critically appraise and synthesize the psychometric properties of the Global Rating of Change (GRoC) scales on the assessment of patients with low back pain (LBP), upper extremity and lower extremity disorders.
Methods: A search was performed in 4 databases (MEDLINE, EMBASE, CINAHL, SCOPUS) until February 2019. 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 8 eligible studies included participants with orthopedic lumbar spine impairments (n = 52,767), patients with work-related musculoskeletal disorders (n = 1944), patients with low back pain (n = 183) and individuals with upper extremity disorders (n = 151). Risk of bias was ranging from "adequate" to "very good" and quality was found excellent for all studies. Based on pooled data, test-retest reliability of 11-item GRoC for patients with low back pain was found excellent ICC = 0.84, 95% CI: 0.65 to 0.94. Test-retest reliability in patients with shoulder pain was found fair to good ICC of 0.62 in a 15-point GRoC scale. Seven studies (n = 7) examined the convergent validity between GRoC and another outcome measure. Minimum important change on the Portuguese version of Global Perceived Effect (GPE) for patients with LBP was 2.5 points out of 11 points.
Conclusions: The current pool of clinical measurement studies indicates that the GRoC has excellent test-retest reliability for patients with low back pain, shoulder pain and with lumbar spine disorders. However, the validity of it as a reference standard in responsiveness studies or as an accurate overall assessment of change has been questioned. While future studies might provide more insight into its measurement properties, this limitation is unlikely to change. Therefore, we suggest that future responsiveness in the studies that want a global indicator measure need to use an additional measure to mitigate recall bias.
Prospero registration number: CRD 42020149122.

1. Introduction

A change in a patient's health condition can be evaluated by a patient-reported outcome measure (PRO), which allows patients to provide a subjective report of their perceived change.[1] Researchers are directed towards evaluating these PROs and performance-based functional tests, as they can provide an assessment of the health conditions change over time.[2,29] The measurement property of responsiveness indicates whether a patient has undergone meaningful change over a period of time. Therefore, the responsiveness of a PRO must be strong, as they are a critical component in clinical decision-making.[2,29]

The Global Rating of Change (GRoC) is a scale that assesses whether the patient condition has gotten worse, better, or stayed the same and to quantify the magnitude of that change, typically following treatment.[3] GRoC scales, can be administered for assessing interventions or as a reference standard when evaluating the responsiveness of another

* Corresponding author. Western University School of Physical Therapy Elborn College, London, ON, N6G 1H1, Canada.
E-mail address: pbobos@uwo.ca (P. Bobos).

https://doi.org/10.1016/j.jor.2020.01.047
Received 15 November 2019; Accepted 31 January 2020
Available online 10 February 2020
0972-978X/

PRO.[4] The GRoC requires the individual to recall their initial status post-injury and compare it to their current health status.[1,4] GRoC scales are presented in a variety of formats with different anchors and with variations in names such as “Global Perceived Effect”, “Patient Global Impression of Change”, “Transition Ratings”, and “Global Scale.[4] In the middle of the response scale is a “0” indicating no improvement or no change, the negative values towards the left indicate worsening symptoms or a deterioration in status, and positive values towards the right indicate improvement in the health status.[5–8] The GRoC is frequently used in clinical practice for musculoskeletal conditions to assess the effectiveness of interventions.[1,4,9] It reduces administration burden, it is easy for patients to understand, and it is easy for clinicians to interpret the results.[1,4] Since is it not disease specific, the GRoC is appropriate for use in multiple different conditions.[5,7,10]

While the GRoC is a common tool used to assess the effectiveness of interventions, it is a retrospective tool that requires patients to recall information from the past, making it vulnerable to recall bias.[1,4,7] This has led researchers to believe that this retrospective tool would not provide an accurate measure of functional change over time, as there is a possibility of a stronger correlation with the current health status of the patient.[4,7,8,10] Although the psychometric properties of GRoC scales have been critically appraised and synthesized for patients with neck pain[11,26] the psychometric properties of GRoC for other conditions have yet to be synthesized. The purpose of this study was to assess and synthesize the psychometric properties of the GRoC scales in patients with low back pain, upper extremity disorders and lower extremity disorders.

2. Methods

2.1. Patient and public involvement

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

2.2. 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 42020149122.

2.3. Eligibility criteria

We included studies in this systematic review if the following criteria were met[12–14]:

- Design: psychometric testing, randomized/cohort studies
- Participants: > 50% of the study's patient population with low back pain, upper and lower extremity 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.

2.4. 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 the following keywords: Reliability OR consistency OR validity OR responsiveness OR calibration OR validation OR agreement OR minimal detectable change OR clinically important difference OR psychometric properties OR measurement properties AND hip OR knee OR ankle/foot pathologies OR lower extremity OR lower limb conditions/disorders OR upper extremity disorders OR low back pain AND Global Perceived Effect OR Patient Global Impression of Change OR Global Rating of Change. Furthermore, we identified additional studies by examining the reference list of each of the selected studies.

2.5. Study selection

Two investigators (PB and CZ) 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.

2.6. Data extraction

The third 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).[15,16] For reliability estimates, Standard Error of Measurement (SEM), Intra-class Correlation Coefficient (ICC), Minimal Detectable Change (MDC) and 95% confidence intervals were extracted.[15,16] The ICC interpretation of ICC < 0.40 indicating poor, $0.40 \leq$ ICC < 0.75 indicating fair-to-good and ICC ≥ 0.75 indicating excellent reliability were used as a common benchmark. The strength of agreement was used as “None” - (0.00–0.20), “Minimal” - (0.21–0.39), “Weak” - (0.40–0.59), “Moderate” - (0.60–0.79), “Strong” - (0.80–0.90) and “Almost Perfect” - (> 0.90).[17] For validity estimates, correlation coefficient (Pearson's/Spearman) and the 95% confidence intervals were extracted.[15,16] 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”.[18] 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.[15,16] 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.[19] When insufficient data were presented, PB contacted the authors by email and requested further data.

2.7. 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.[20] 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.[21] 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 (?).[20]

2.7.1. 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.[15,16] 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.[15,16] 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%.[15,16] 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.[15,16]

2.7.2. Synthesis of results

A qualitative synthesis was conducted to report findings of the included articles based on the condition, reported psychometric estimate and the study quality ratings. A meta-analysis of test-retest reliability (intra-class correlation coefficient) was performed with R (metafor package) software.[22] The meta-analysis was conducted using a random effects model (RE). Heterogeneity was deemed substantial if I^2 values were more than 50%.[23]

3. Results

3.1. 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, 8 studies were considered eligible.[1,5–9,24,25] The flow of the study selection process is presented in Fig. 1.

3.2. Study characteristics

The 8 eligible studies were conducted between 2005 and 2019 and included 52,767 participants with orthopedic lumbar spine impairments,[9] 1944 patients with work-related musculoskeletal disorders,[24] 183 patients with low back pain (acute and chronic),[5,25] 151 individuals with upper extremity disorders (shoulder impingement and shoulder pain). Study size ranged from 52 to 52767 participants. A summary description of all the studies included is displayed in Table 1. Validity as determined by correlation was evaluated in 6 studies[1,5,6,8,9,24] by comparing the Medrisk Instrument for Measuring Patient Satisfaction with Physical Therapy Care (MR-12), Functional Rating Index (FRI), Roland-Morris Disability Questionnaire (RDMQ), Patient-Specific Functional Scale (PSFS), American Shoulder and Elbow Surgeon's Scale (ASES), Disabilities of the Arm, Shoulder and Hand (DASH), Short-Form 12 (SF-12), The Shoulder Pain and Disability Index (SPADI), Patient-Rated Wrist Evaluation (PRWE), Functional Status (FS), and FSCH to the GRoC. Three studies[6,9,25] examined test-retest reliability. One study[5] evaluated reliability through reproducibility, and one study[25] evaluated convergent validity.

3.3. COSMIN risk of bias rating and quality appraisal of the included studies

Regarding the risk of bias, the rating was ranging from “adequate” to “very good”. The risk of bias and criteria of good measurement properties are summarized in Table 2. The quality of the studies ranged from 92% to 100% (Table 3). The most common flaws were 1) lack of/inadequate sample size calculations, and 2) missing data (i.e. inadequate follow up).

3.4. Reported GROC scales

The most commonly reported GRoC scale (n = 4 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)”. An 11-point scale was reported in 3 studies, 29- and 9- point scales were reported in one study. The anchors in those scales varied greatly and are presented in Table 1. Five out of the 9 studies[5,6,9,24,25] reported full detail regarding the specific questions asked when a GRoC scale was administered. Those questions that were reported are presented in Table 6.

3.5. Meta-analysis of test-retest reliability for patients with low back pain

Based on pooled data (n = 183 participants) from two studies,[5,25] test-retest reliability of 11-item GRoC for patients with low back pain was found excellent ICC = 0.84, 95% CI: 0.65 to 0.94, I^2 = 85% (Fig. 2).

3.6. Reliability measures for shoulder pain and lumbar spine disorders

Two studies were included that examined test-retest reliability of GPE for patients with upper and lower extremity disorders. Moore-Reed et al. (2017) examined the test-retest reliability of an 15-point GRoC scale in 99 patients with shoulder pain, and reported an ICC of 0.62, but the 95% CI was not reported. Wang et al. (2018) reported test-retest reliability using a 15-point GRoC scale in 52,767 patients with orthopaedic lumbar spine impairments, and reported an ICC of 0.61, but did not report a 95% CI (Table 4).

3.7. Inter-rater reliability and agreement between physician and patient

Moore-Reed et al. (2017)[6] examined the degree of patient–physician discordance in the assessment of the shoulder pain change in status with a 15-item GRoC scale 6 weeks after physical therapy intervention. ICC, Pearson's r and weighted Kappa between patient and physician were found with moderate agreement (0.62, 0.63 and 0.62 respectively). Wang et al. (2018)[9] examined the agreement of patient GRoCp rating and therapists GRoCt rating. Using the entire data, the ICC absolute agreement was 0.14 (“none”) while using patient data that GRoC was > 0 the ICC increased to 0.61 (“moderate”).

3.8. Validity measures

Eight studies examined the validity measures between GRoC and another PRO (Table 5). Beattie et al. (2011)[24] examined the relationship between perceived clinical change with 9-item GRoC scale and measurement of overall patient satisfaction with MR-12 in 1944 patients with work-related musculoskeletal disorders after a 4 week intervention. A weak correlation was found between GRoC and MR-12 (Pearson r = −0.30). Costa et al. (2008)[5] examined the correlations of the Brazilian-Portuguese versions of GPE and the disability scores of FRI (very weak correlation, Pearson r = 0.11, p = 0.30), the disability scores of RMDQ (very weak correlation, Pearson r = 0.14, p = 0.18) and the functional ability scores of PSFS (weak correlation, Pearson r = 0.34, p < 0.01), in 99 patients with acute low back pain. Freitas

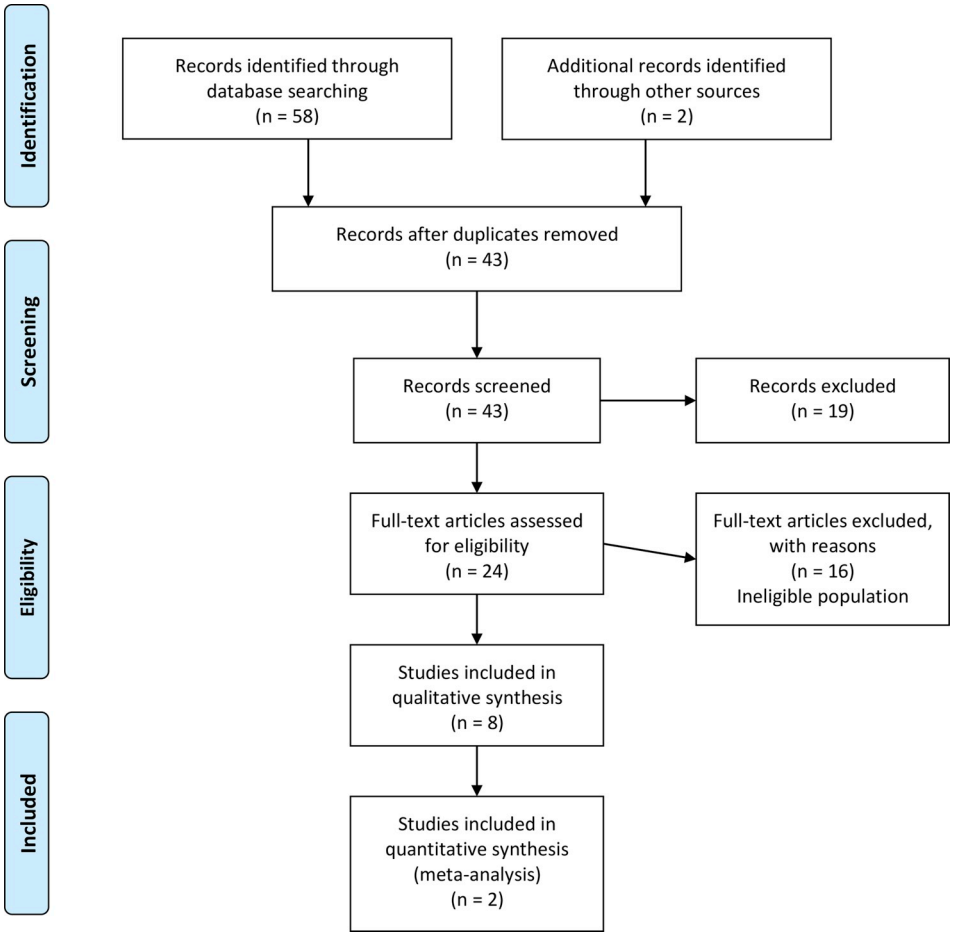

Fig. 1. PRISMA Flow diagram for article inclusion process.

et al. (2019)[25] examined the convergent validity between Portuguese version of an 11-item Global Perceived Effect Scale GPE and PGIC. A strong correlation rho = 0.68 was found between the GPE and PGIC after 6 weeks of intervention in 84 patients with chronic lower back pain. Garrison et al. (2012)[1] examined the correlation on a weekly basis between GRoC and ASES over time from week 1–8 in patients with shoulder impingement syndrome. The correlations were ranging from −0.05 to 0.31 (very weak to weak) and only up to 3 weeks the GRoC was only correlated to functional measures ($p < 0.05$). Specific correlations for each week are presented in Table 5. Schmidt et al. (2005) examined Spearman's Rho correlation between a 15-item GRoC scale and DASH, SF-12, SPADI and PRWE at baseline and at 3- months, in 211 patients with upper limb extremity musculoskeletal problems. The correlations ranged from 0.16 to −0.59 (very weak to moderate), but specific relationships are presented in Table 5. Schmidt et al. (2015) examined the correlation between FS change scores and a 15-item GROC scale. The Pearson's r correlation ranged from −0.12 to 0.53 (very weak to moderate) between GRoC and initial score of the FS between zero days and 180 days, in 7341 patients with disorders in the hip, foot or ankle (Table 5). Wang et al.[9] examined the relationship between the patient GRoCp patient and the therapist GRoCt and the correlation was weak (r = 0.21). GRoCp consistently did not correlate well with FS at intake (r ranged from 0.02 to 0.12 – very weak).

3.9. Responsiveness

The study of Freitas et al.[25] tested the responsiveness of Portuguese version of GPE through the relationship of the following change scores: PGIC scores, the Numeric Pain Rating Scale scores and the Quebec Back Pain Disability Scale scores using the Spearman rank order correlation after 6 weeks of intervention in patients with LBP. The GPE change scores were strongly correlated with PGIC (rho = 0.60, $p < 0.01$) and moderately correlated with Quebec Back Pain Disability Scale and Numeric Pain Rating Scale change scores (0.45, $p < 0.01$ and 0.45, $p < 0.01$, respectively). The ROC curve revealed an absolute optimal cutoff (minimally important change - MIC) of 2.5 points out of 11 points on the Portuguese version of GPE for patients with LBP.

4. Discussion

This review has synthesized the current evidence from 8 studies aimed to evaluate the psychometric properties of GRoC scales, and suggest clinical recommendations associated with its application for patients with low back pain, upper and lower extremity disorders. Risk of bias (COSMIN) was ranging from “adequate” to “very good” and it was downgrade it mostly because no hypothesis was provided or partly confirmed by the analysis.[1,6–9] Quality was found excellent in all studies however, the most common flaws were the lack of sample size calculations, and the inadequate follow up. While risk of bias and quality overall scores performed really well across the included studies, we believe there is room for improvement in terms of sample size calculations and proper hypothesis testing.

Based on 2 excellent quality studies,[5,25] pooled estimates of test-retest reliability of 11-item GRoC scale and for patients with low back pain were found excellent (ICC = 0.84, 95% CI: 0.65 to 0.94). Based on an excellent quality study,[9] test-retest reliability of 15-item GRoC was found fair-to-good (ICC = 0.61) for patients with lumbar spine disorders. The strength of the agreement between patient and physician

Table 1
Study characteristics.

Study	Population	Setting	Sample Size	Properties Evaluated	GRoC evaluated	Interval
Beattie et al. (2011)	Patients with work-related musculoskeletal disorders	Physical therapy clinic	1944	Validity (correlation) MR-12 vs GRoC	GRoC 9 points 1 = very much better, 5 = no change and 9 = very much worse	Completed once after the 4-week intervention
Costa et al. (2008)	Patients with acute lower back pain	Physiotherapy clinics	99	Reliability (reproducibility) Validity (construct and correlation for external responsiveness) GPE vs FRI, RMDQ, PSFS (Brazilian-Portuguese versions)	GPE 11-points −5 (“vastly worse”) through 0 (“no change”) to +5 (completely recovered”)	GPE was completed at time points of: baseline, 24 h and then 2 weeks
Freitas et al. (2019)	Patients with chronic lower back pain	Physiotherapy clinic	84	Reliability (test-retest) Validity (convergent) GPE vs PGIC Responsiveness (correlations) of GPE	GPES 11- points −5 (“vastly worse”) through 0 (“no change”) to +5 (completely recovered”)	GPES-PT was completed at baseline, after 48 h and six weeks after intervention.
Garrison et al. (2012)	Patients with shoulder impingement	Physical therapy clinics	52	Validity (correlation) GRoC vs ASES	GRoC 15-point −7 (worse) to +7 (better)	GRoC was completed each week for a time period of 8 weeks
Moore-Reed et al. (2017)	Patients with shoulder pain	Sports medical centre	99	Reliability (inter-rater/agreement)	GRoC 15-points −7 (‘‘a very great deal worse’’) to 0 (‘‘about the same’’) to +7 (‘‘a very great deal better’’)	GRoC was measured at baseline and 24 h
Schmidt et al. (2005)	Patients with upper limb extremity musculoskeletal problems	Physical or occupational therapy outpatient clinics	211	Validity (correlation) GRoC vs DASH, SF-12 PCS, SPADI, PRWE	Retrospective GRC 29-points −1 to −14 (deterioration), 0 for no change and +1 to +14 for improvement	GRoC was completed at baseline and 3 months
Schmidt et al. (2015)	Patients with disorders in the hip, foot or ankle	Physical therapy outpatient clinics	7341	Validity (correlation) GRoC vs FS	GRoC 15-point −7 (worse), 0 (no change) and +7 (better)	GRoC was completed at 5 different time points over 180 days
Wang et al. (2018)	Patients with orthopaedic lumbar spine impairments	Outpatient Rehabilitation clinics	52767	Reliability (test-retest) Validity (correlation) GRoCt vs GRoCP GRoCp vs FSCH GRoCt vs FSCH	GRoC 15-point −7 (worse), 0 (no change) and +7 (better)	GRoC was completed at both intake and again at discharge

GRoC = Global Rating of Change, GPE = Global Perceived Effect, PGIC = Patient Global Impression of Change, MR- 12 = Measuring Patient Satisfaction with Physical Therapy Care, FRI = Functional Rating Index, RMDQ = Roland-Morris Disability Questionnaire, PSFS = Patient Specific Functional Scale, ASES = American Shoulder and Elbow Surgeon's Scale, SF- 12 = Short Form 12, SPADI = The Shoulder Pain and Disability Index, PRWE = Patient Rated Wrist Evaluation, FS = Functional Status, GPES-PT = Portuguese version of GPE, Functional status change score (FSCH) was defined by subtracting the FS score at intake from the FS score at discharge (FSCH = discharge FS–intake FS), GRoCt = GRoC completed from the treated physician, GRoCP = GRoC completed by patient.

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[a] (QACMRR)
Beattie et al. (2011)	Validity (convergent)	Very good		?	Excellent
Costa et al. (2008)	Reliability (test-retest)	Adequate		+	Excellent
	Validity (correlation)	Very good		?	
Freitas et al. (2019)	Reliability (test-retest)	Very good		+	Excellent
	Validity (convergent)	Very good		+	
	Responsiveness (correlation)	Very good		+	
Garrison et al. (2012)	Validity (convergent)	Very good		+	Excellent
Moore-Reed et al. (2017)	Reliability (inter-rater)	Adequate		–	Excellent
	Validity (convergent)	Adequate		+	
Schmidt et al. (2005)	Validity (convergent)	Very good	?		Excellent
Schmidt et al. (2015)	Validity (convergent)	Very good		+	Excellent
Wang et al. (2018)	Reliability (inter-rater)	Adequate		–	Excellent
	Validity (convergent)	Very good		?	

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.
[a] 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[a]												Total (%)	Quality Summary
	1	2	3	4	5	6	7	8	9	10	11	12		
Costa et al. (2008)	2	2	2	2	2	2	2	2	2	2	2	2	100	Excellent
Moore-Reed et al. (2017)	2	2	2	2	1	2	2	2	2	2	2	2	96	Excellent
Schmidt et al. (2005)	2	2	2	2	2	1	2	2	2	2	2	2	96	Excellent
Wang et al. (2018)	2	2	2	2	1	2	2	2	2	2	2	2	96	Excellent
Beattie et al. (2011)	2	2	1	2	1	2	2	2	2	2	2	2	92	Excellent
Freitas et al. (2019)	2	2	2	2	1	1	2	2	2	2	2	2	92	Excellent
Garrison et al. (2012)	2	2	2	2	1	1	2	2	2	2	2	2	92	Excellent
Schmidt et al. (2015)	2	2	2	2	1	1	2	2	2	2	2	2	92	Excellent

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%).
[a] 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.

Table 4
Summary of reliability properties of GRoC scales.

Study	Type of Reliability	Reliability Estimates	COSMIN	Quality of Studies
Costa et al. (2008)	Test-retest (LBP)	Intra-class coefficients (ICC) 0.90 (0.84–0.93)	Adequate	Excellent
Freitas et al. (2019)	Test-retest (LBP)	Intra-class coefficients (ICC) 0.76 (0.69–0.85)	Very good	Excellent
Moore-Reed et al. (2017)	Inter-rater/Agreement (SP)	Intra-class coefficients (ICC) 0.62 Weighted Kappa 0.62 Pearson's r 0.63	Adequate	Excellent
Wang et al. (2018)	Test –retest (LSD)	Intra-class coefficients (ICC) 0.61	Adequate	Excellent

LBP = Low Back Pain, SP = Shoulder Pain, LSD = Lumbar Spine Disorders.

was found as "moderate" in 2 different studies which may support the hypothesis that patient-reported and physician-reported GRoC appears to represent the same perceived of change by each group.

Evidence from a number of individual studies indicated that the GRoC validity measures ranged from very weak to weak when compared to another PRO. In the Beattie et al. (2011)[24] (excellent quality; indeterminate properties) GRoC with MR-12 scores yielded weak correlations between the $r = -0.19$ and $r = -0.30$. This finding may suggest that there is generally weak relationship between perceived satisfaction and perceived change for individuals with work-related musculoskeletal problems. Other findings indicated that the majority of associations between GRoC scales and functional ability (FRI), PSFS and ASES confirm that GRoC scales do not adequately or consistently correlate with functional change. On the other hand, GRoC was correlated much higher with the current status of the patient for DASH, PRWE and SPADI. This is consistent with other studies which have indicated that

Table 5
Summary of validity properties of GRoC scales.

Study	Type of Reliability	Validity Estimates	COSMIN	Quality of Studies
Beattie et al. (2011)	Correlations (Pearson r)	−0.30, p < 0.01	Very good	Excellent
	GRoC vs MR-12 Overall Satisfaction			
Costa et al. (2008)	Correlations (Pearson r)	−0.37, p < 0.01	Very good	Excellent
	At Baseline	−0.42, p < 0.01		
	GPE vs FRI	0.33, p < 0.01		
	GPE vs RMDQ	0.11, p = 0.30		
	GPE vs PSFS	0.14, p = 0.18		
	Correlations At Discharge	0.34, p < 0.01		
	GPE vs FRI			
	GPE vs RMDQ			
	GPE vs PSFS			
Freitas et al. (2019)	Convergent (Spearman's Rho)	rho = 0.68, p < 0.01	Very good	Excellent
	GPE vs PGIC	rho = 0.60, p < 0.01		
	Responsiveness	rho = 0.45, p < 0.01		
	GPE vs PGIC	rho = 0.45, p < 0.01		
	GPE vs NRS			
	GPE vs QBPDS			
Garrison et al. (2012)	Correlations (Pearson r)	0.31, p < 0.01	Very good	Excellent
	GRoC vs ASES week 1	0.29, p < 0.05		
	GRoC vs ASES week 2	0.27, p < 0.05		
	GRoC vs ASES week 3	0.23		
	GRoC vs ASES week 4	0.13		
	GRoC vs ASES week 5	0.31		
	GRoC vs ASES week 6	−0.05		
	GRoC vs ASES week 7	0.9		
	GRoC vs ASES week 8			
Schmidt et al. (2005)	Correlations (Spearman's Rho)	Rho = 0.16	Very good	Excellent
	GRoC vs DASH (initial)	Rho = −0.54		
	GRoC vs DASH (3-month)	Rho = 0.17		
	GRoC vs SF-12 PCS (initial)	Rho = −0.42		
	GRoC vs SF-12 PCS (3-month)	Rho = 0.05		
	GRoC vs SPADI (initial)	Rho = −0.59		
	GRoC vs SPADI (3-month)	Rho = 0.10		
	GRoC vs PRWE (initial)	Rho = −0.52		
	GRoC vs PRWE (3-month)			
Schmidt et al. (2015)	Correlation (Pearson)	0.16 (0.10,0.22)	Very good	Excellent
	GRoC vs FS 0–30 days (initial)	0.53(0.48,0.57)		
	GRoC vs FS 0–30 days (initial)	0.00 (−0.06,0.06)		
	GRoC vs FS 31–60 days (initial)	0.39(0.34,0.44)		
	GRoC vs FS 31–60 days (initial)	0.06(-0.04,0.16)		
	GRoC vs FS 61–90 days (initial)	0.46(0.38,0.54)		
	GRoC vs FS 61–90 days (initial)	−0.12 (−0.24, 0.00)		
	GRoC vs FS 91–180 days (initial)	0.30(0.18,0.41)		
	GRoC vs FS 91–180 days (initial)	−0.09(-0.37,0.20)		
	GRoC vs FS 180 < days (initial)	0.42(0.15,0.63)		
	GRoC vs FS 180 < days (initial)			
Wang et al. (2008)	Correlation (Pearson)	0.15	Very good	Excellent
	GRoCt vs FS intake −7 to +7	0.02		
	GRoCp vs FS intake −7 to +7	0.51		
	GRoCt vs FS discharge −7 to +7	0.19		
	GRoCp vs FS discharge −7 to +7	0.39		
	GRoCt vs FSCH -7 to +7	0.17		
	GRoCp vs FSCH -7 to +7	0.15		
	GRoCt vs FS intake ≥ −3	0.11		
	GRoCp vs FS intake ≥ −3	0.51		
	GRoCt vs FS discharge ≥ −3	0.53		
	GRoCp vs FS discharge ≥ −3	0.39		
	GRoCt vs FSCH ≥ −3	0.45		
	GRoCp vs FSCH ≥ −3	0.15		
	GRoCt vs FS intake ≥ 0	0.12		
	GRoCp vs FS intake ≥ 0	0.50		
	GRoCt vs FS discharge ≥ 0	0.56		
	GRoCp vs FS discharge ≥ 0	0.38		
	GRoCt vs FSCH ≥ 0	0.47		
	GRoCp vs FSCH ≥ 0			

GRoC = Global Rating of Change, GPE = Global Perceived Effect, Functional status change score (FSCH) was defined by subtracting the FS score at intake from the FS score at discharge (FSCH = discharge FS–intake FS), GRoCt = GRoC completed from the treated physician, GRoCP = GRoC completed by patient.

GRoC is correlated more with current status than actual change.[26] Only one study Freitas et al. (Risk of bias: very good; sufficient properties; excellent quality) provided responsiveness of the Portuguese version of GPE by using an anchor-based method and found a minimally important change (MIC) of 2.5 points out of 11 in patients with LBP. While this MIC was calculated based on a Portuguese population with LBP the 2.5 points are consistent with other findings[4] in the literature and therefore, can have clinical applications.

Given that recall bias is a substantial concern, further measurement studies are unlikely to resolve this issue. There are a number of ways

Table 6
Questions of Global Rating of Change scales.

Author	GROC item- scale	Patients with neck disorders were asked:
Beattie et al., 2011	GRoC 9 points	*" How does your current condition compare to how it was before you started physical therapy treatment?"*
Costa et al., 2008	GPE 11-points	*Compared to when this episode first started, how would you describe your back these days?"*
Freitas et al., 2019	GPE 11- points	*The participants had the opportunity to choose between three options of anchor questions:* (1) "Compared to when this episode first started, how would you describe your back at this moment?"; (2) "Compared to the day on which physiotherapy was arranged/referred, how would you describe your back at this moment?"; (3) Compared to the beginning of treatment, how would you describe your back at this moment?"
Moore-Reed et al., 2017	GRoC 15- points	*Subjects were instructed to select the statement that best represented their perceived change in functional status subsequent to the initial evaluation*
Wang et al., 2018	GRoC 15- points	*"Rate the overall change during the treatment for this condition.* *Use the center '0' as the overall level of the condition at the beginning of treatments in this facility"*

GRoC = Global Rating of Change, GPE = Global Perceived Effect.

where global status can be measured by a single item such as visual analog scales that ask about overall health such as quality of life on the EQ-5D or the assessment of how normal a patient judge their current status with Single Assessment Numerical Evaluation (SANE).[27] Administering such global measures on two occasions would mitigate the problems with recall bias. One might argue what the best global indicator single item might be. However, methodologists are reviewers who suggest that global rating of change is necessary and dismiss other criteria and measures are contributing to flawed assessments of PRO. We suggest that researchers should make a case for how they chose to assess their global rating of change, and how they mitigated sources of bias by this choice. While a lot of research studies used GRoC scales especially for calculating the minimum clinically important difference (MCID) of another PRO the number of published papers addressing the measurement properties of GRoC scales is quite limited. As a result, the number of the included studies was relatively small (n = 8) and indicates the current state of the literature. Our included studies used different populations, interventions, and time intervals when addressing the specific psychometric properties of the GRoC, which may have contributed to the differences in final results and findings. Furthermore, we used 2 different approaches to evaluate the risk of bias and the quality of the individual studies. Various critical appraisal tools exist, and the perspectives and ratings may differ across instruments.

5. Conclusions

Very limited evidence has evaluated the measurement properties of the GRoC scales. The current pool of clinical measurement studies indicates that the GRoC has excellent test-retest retest reliability for patients with low back pain, shoulder pain and with lumbar spine disorders. However, the validity of it as a reference standard in responsiveness studies or as an accurate overall assessment of change has been questioned. While future studies might provide more insight into its measurement properties, this limitation is unlikely to change. Therefore, we suggest that future responsiveness in the studies that want a global indicator measure need to use another measure to mitigate recall bias.

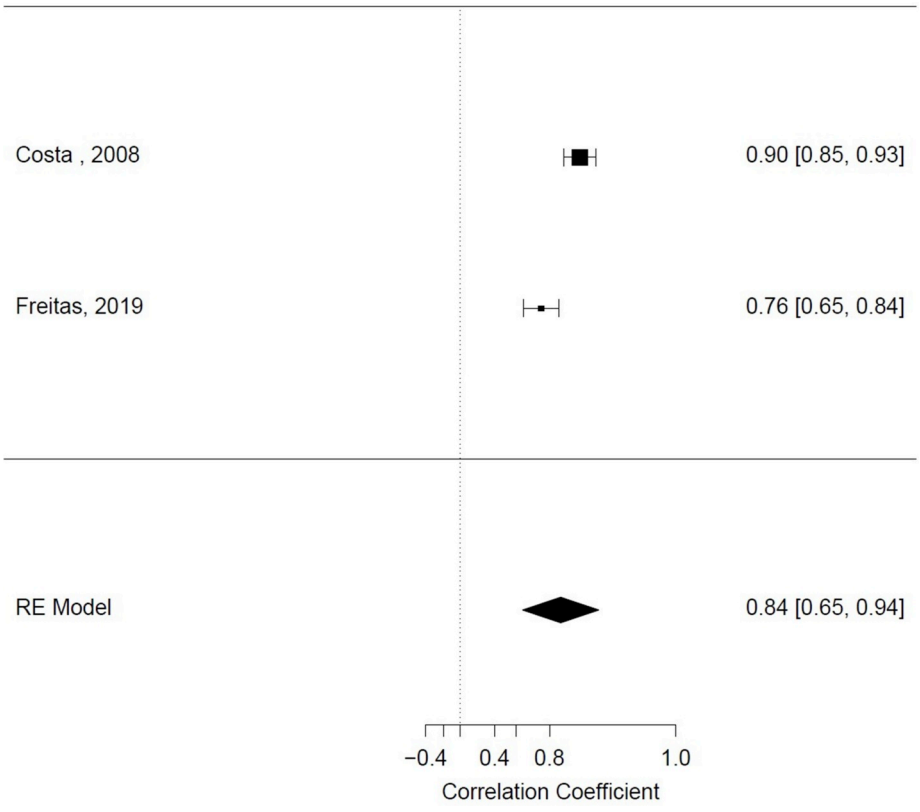

Fig. 2. Forest plot of test-retest pooled estimates for low back pain patients with 95% confidence intervals.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The authors have none to report.

References

1. Garrison C, Cook C. Clinimetrics corner: the global rating of change score (GRoC) poorly correlates with functional measures and is not temporally stable. *J Man Manip Ther.* 2012;20(4):178–181. https://doi.org/10.1179/1066981712Z.00000000022.
2. McGee S, Sipos T, Allin T, et al. CATWAD. Systematic review of the measurement properties of performance-based functional tests in patients with neck disorders. *BMJ Open.* 2019;9(11):e031242. https://doi.org/10.1136/bmjopen-2019-031242.
3. Schmitt J, Abbott JH. Global ratings of change do not accurately reflect functional change over time in clinical practice. *J Orthop Sports Phys Ther.* 2015;45(2):106–111. https://doi.org/10.2519/jospt.2015.5247.
4. 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(3):163–170. https://doi.org/10.1002/mus.21062.
5. Oliveira L, Costa P, Maher CG, et al. Clinimetric testing of three self-report outcome measures for low back pain patients in Brazil which one is the Best ? *Spine.* 2008;33(22):2459–2463.
6. Moore-Reed SD, Kibler W Ben, Bush H, Uhl TL. Level of patient–physician agreement in assessment of change following conservative rehabilitation for shoulder pain. *Shoulder Elbow.* 2017;9(2):127–132. https://doi.org/10.1177/1758573216658799.
7. Schmitt J, Abbott JH. Global ratings of change do not accurately reflect functional change over time in clinical practice. *J Orthop Sports Phys Ther.* 2015;45(2):106–111. https://doi.org/10.2519/jospt.2015.5247.
8. Schmitt J, Di Fabio RP. The validity of prospective and retrospective global change criterion measures. *Arch Phys Med Rehabil.* 2005;86(12):2270–2276. https://doi.org/10.1016/j.apmr.2005.07.290.
9. Wang YC, Sindhu BS, Kapellusch J, Yen SC, Lehman L. Global rating of change: perspectives of patients with lumbar impairments and of their physical therapists. *Physiother Theory Pract.* 2019;35(9):851–859. https://doi.org/10.1080/09593985.2018.1458930.
10. Kamper SJ, Ostelo RWJG, Knol DL, Maher CG, de Vet HCW, Hancock MJ. 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(7):760–766.e1. https://doi.org/10.1016/j.jclinepi.2009.09.009.
11. Bobos P, Macdermid JC, Walton DM, Gross A, Santaguida PL. Patient-reported outcome measures used for neck disorders: an overview of systematic reviews. *J Orthop Sports Phys Ther.* 2018;48(10):775–788. https://doi.org/10.2519/jospt.2018.8131.
12. Nazari G, Bobos P, MacDermid JC, Birmingham T. 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.* 2019. https://doi.org/10.1016/j.apmr.2019.01.017 (c).
13. Bobos P, Nazari G, Szekeres M, Lalone EA, Ferreira L, MacDermid JC. 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(2):194–211. https://doi.org/10.1016/j.jht.2018.09.012.
14. Nazari G, Bobos P, Lu Z, MacDermid JC. Measurement Properties of the Hand Grip Strength Assessment: A Systematic Review With Meta-analysis. *Arch Phys Med Rehabil.* 2019. https://doi.org/10.1016/j.apmr.2019.10.183.
15. Law MC, MacDermid J. *Evidence-Based Rehabilitation : A Guide to Practice.* Thorofare, NJ: Slack Incorporated; 2014.
16. Roy JS, Desmeules F, MacDermid JC. Psychometric properties of presenteeism scales for musculoskeletal disorders: a systematic review. *J Rehabil Med.* 2011. https://doi.org/10.2340/16501977-0643.
17. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med.* 2012;22(3):276–282. https://doi.org/10.11613/bm.2012.031.
18. Wuensch KL, Evans JD. Straightforward statistics for the behavioral sciences. *J Am Stat Assoc.* 2006. https://doi.org/10.2307/2291607.
19. Cohen J. Statistical power analysis for the behavioral sciences. *Stat Power Anal Behav Sci.* 1988. https://doi.org/10.1234/12345678.
20. 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.* 2018. https://doi.org/10.1007/s11136-017-1765-4.
21. Terwee CB, Mokkink LB, Knol DL, Ostelo RWJG. *Rating the methodological quality in systematic reviews of studies on measurement properties : a scoring system for the COSMIN checklist.* 2012; 2012:651–657. https://doi.org/10.1007/s11136-011-9960-1.
22. Viechtbauer W. Conducting meta-analysis in R with metafor package. *J Stat Software.* 2010;36(3):1–48.
23. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ.* 2003. https://doi.org/10.1136/bmj.327.7414.557.
24. Beattie PF, Nelson RM, Heintzelman M. The relationship between patient satisfaction with physical therapy care and global rating of change reported by patients receiving worker's compensation. *Physiother Theory Pract.* 2011;27(4):310–318. https://doi.org/10.3109/09593985.2010.490575.
25. Freitas P, Pires D, Nunes C, Cruz EB. Cross-cultural adaptation and psychometric properties of the European Portuguese version of the Global Perceived Effect Scale in patients with chronic low back pain. *Disabil Rehabil.* 2019:1–7. https://doi.org/10.1080/09638288.2019.1648568 0(0).
26. Bobos P, MacDermid J, Nazari G, Furtado R. Psychometric properties of the global rating of change scales in patients with neck disorders: a systematic review with meta-analysis and meta-regression. *BMJ Open.* 2019;9(11) https://doi.org/10.1136/bmjopen-2019-033909 e033909.
27. Nazari G, Bobos P, Lu Z, et al. Psychometric properties of the single assessment numeric evaluation in patients with lower extremity pathologies. A systematic review. *Disabil Rehabil.* 2019. https://doi.org/10.1080/09638288.2019.1693641.
29. Hunter SW, Bobos P, Somerville L, Howard J, Vasarhelyi E, Lanting B. Comparison of functional and patient-reported outcomes between direct anterior and lateral surgical approach one-year after total hip arthroplasty in Canadian population: a cross-sectional study. *J Orthop.* 2019. https://doi.org/10.1016/j.jor.2019.11.004.