PTJ: Physical Therapy & Rehabilitation Journal | *Physical Therapy*, 2022;102:1–12
https://doi.org/10.1093/ptj/pzac054
Advance access publication date May 13, 2022
Review

Reliability, Validity, and Responsiveness of the Craniocervical Flexion Test in People Who Are Asymptomatic and Patients With Nonspecific Neck Pain: A Systematic Review and Meta-Analysis

Antonio Romeo, PT, MSc, OMPT[1,*], Marco Baccini, PT, MSc[2], Giulia Carreras, PhD[3], Marco Sagripanti, PT, MSc, OMPT[1], Martina Ruggeri, PT, OMPT[1], Paolo Pillastrini, PT, MSc[1,4], Mauro Di Bari[5]

[1]Department of Biomedical and Neuromotor Sciences (DIBINEM), Alma Mater Studiorum, University of Bologna, Bologna, Italy
[2]PROMISE Lab, IRCCS Fondazione Don Gnocchi Firenze, Florence, Italy
[3]Research Unit of Medicine of Aging, Department of Experimental and Clinical Medicine, University of Florence, Florence, Italy
[4]Division of Occupational Medicine, IRCCS Azienda Ospedaliero-Universitaria di Bologna, Bologna, Italy
[5]Department of Experimental and Clinical Medicine, University of Florence and Azienda Ospedaliero-Universitaria Careggi, Florence, Italy

*Address all correspondence to Antonio Romeo at: anton.rom63@gmail.com

Abstract

Objective. Among the tests designed to evaluate neck neuromuscular function, the craniocervical flexion test (CCFT) assesses the function of the deep cervical flexor muscles (DCFs). The purpose of this study was to conduct a review and meta-analysis of published articles about all measurement properties of the different CCFT versions (CCFT Activation Score [CCFT-AS], CCFT Performance Index [CCFT-PI], CCFT Cumulative Performance Index [CCFT-CPI], and CCFT alternative procedures for measuring activation level (CCFT1) or endurance (CCFT2) in people who were asymptomatic and people with nonspecific neck pain.

Methods. PubMed Central, MEDLINE, CINAHL, Scopus, Web of Science, and Google Scholar were searched from inception to June 30, 2020. Studies were selected if they reported data on reliability, validity, and/or responsiveness of the CCFT in adults who were asymptomatic or who had nonspecific neck pain. Two reviewers independently selected the studies, conducted quality assessment, and extracted the results. All meta-analyses used a random-effects model.

Results. Twenty-one studies met the inclusion criteria. The rating of interrater reliability (assessed for CCFT-AS and CCFT-CPI) was positive only for using the test at a group level. The same rating was ascribed to the intrarater reliability of CCFT-AS, CCFT1, and CCFT2, whereas CCFT-PI and CCFT-CPI showed positive intrarater reliability for assessment of individuals as well. CCFT validity was rated as positive for expressly assessing DCF action when measuring DCF activation through electromyography—not through ultrasonography—or craniocervical flexion motion as well as for differentiating patients who were asymptomatic and patients who had nonspecific neck pain (only the AS version). CCFT validity was rated as negative for investigating the CCFT performance correlation with the severity of nonspecific neck pain. CCFT responsiveness was rated as negative.

Conclusions. The CCFT is a potentially useful tool for detecting impairment in DCF control and identifying patients who have nonspecific neck pain and who would benefit from a targeted intervention. However, the limited reliability affects its suitability for that purpose. Further research on the reliability of different CCFT versions in which the raters are thoroughly trained is strongly recommended.

Impact. The CCFT might help to detect impairment in DCF control and identify patients who have nonspecific neck pain and who would benefit from a targeted intervention. However, the poor reliability of most versions of the test greatly limits its application in clinical practice. Only CCFT-PI and CCFT-CPI seem reliable enough to help in clinical decision-making at the individual level.

Keywords: Craniocervical Flexion Test, Deep Cervical Flexor Muscle, Neck Pain, Reliability, Responsiveness, Validity

Received: July 7, 2021. **Revised:** December 8, 2021. **Accepted:** June 6, 2022
© The Author(s) 2022. Published by Oxford University Press on behalf of the American Physical Therapy Association. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Downloaded from academic.oup.com/ptj/article/102/7/pzac054/6585158 by Biblioteca di Scienze, Università degli studi di Firenze user on 27 October 2026

Introduction

Neck pain is a highly prevalent condition that leads to pain, disability, and financial expense, with high rates of recurrence and chronicity.[1] Recurrence may at least partly depend on inadequate recovery post injury of cervical muscle function, which contributes approximately 80% to cervical spine stability.[2] Evidence exists of an association between mechanical neck pain and changes in cervical muscle structure and function.[3–7]

Among the tests designed to evaluate neck neuromuscular function, the craniocervical flexion test (CCFT) assesses the function of the deep cervical flexor muscles (DCFs).[8] The CCFT is performed with the patient in a supine crook-lying position with the neck in a neutral position. An air-filled pressure biofeedback sensor is placed behind the neck and inflated to 20 mmHg to fill the space under the neck without causing lordosis. A pressure sensor monitors the slight flattening of the cervical lordosis that occurs when DCFs are activated. The individual performs a gentle head-nodding action of craniocervical flexion (CCF) for 5 incremental stages of increasing range (2 mm Hg each stage). An ideal DCF control should be able to increase pressure to 30 mmHg. A clinician observes the movement quality, discouraging compensatory movements (eg, retraction) or muscle strategies (eg, excessive use of the superficial cervical flexor muscles [SCFs]).[8,9]

There are different ways to perform and score the test. The CCFT Activation Score (CCFT-AS) is the pressure an individual can achieve and hold in a steady manner for 10 seconds. The CCFT Performance Index (CCFT-PI) is calculated as the number of times an individual can hold for 10 seconds at the highest-pressure increment achieved. For example, if the individual can achieve the second level (24 mm Hg; ie, 4-mm Hg increment), performing six 10-second holds with the correct CCF, then the CCFT-PI is $4 \times 6 = 24$.[8] The highest CCFT-AS and CCFT-PI are 30 and 100, respectively. The CCFT Cumulative Performance Index (CCFT-CPI; maximum score = 300) is computed by adding the preceding score to the CCFT-PI to reflect the entire test, not just the position at which it terminates.[10]

Two other versions of the test have been described.[9] In the first one, each level is maintained for 3 seconds with the correct chin position, without palpable activity of SCFs (CCFT1). The second variation (CCFT2) is performed as described but with requesting three 10-second-hold repetitions at each incremental level. The test ends when participants are unable to maintain the position for the requested time or use inappropriate strategies to achieve or maintain that level of pressure.

A 2008 systematic review rated the reliability of CCFT as poor, whereas other properties were not reported.[11] However, the number of publications evaluating CCFT measurement properties has subsequently increased, leading to 2 new systematic reviews published in 2020. One[12] examined all metric properties but did not aggregate data from different studies and did not provide any quantitative synthesis, merely rating each property as positive, negative, or indeterminate. The second review[13] considered only reliability and validity and computed meta-analytic estimates but failed to include many reliability studies and, actually, included no studies on the CCFT validity. Moreover, both reviews also selected studies on persons with traumatic neck pain or cervicogenic headache, conditions considered as separate entities.[1,14] The results of the previous reviews are conflicting, but none of them evaluated separately the different versions and scores of the CCFT. In our opinion, this is a severe limitation, because different CCFT versions and scores quantify either different abilities (level of CCF action achieved or endurance) or the same ability with different procedures (eg, endurance capacity requesting 10 or three 10-second holds).

The purpose of this review is to critically appraise, summarize, and meta-analyze the published articles about all measurement properties (reliability, validity, and responsiveness) of CCFT in individuals who are asymptomatic or are affected by nonspecific neck pain.

Methods

We conducted this review according to Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) recommendations.

Data Sources and Searches

The following databases were searched to identify studies evaluating measurement properties of CCFT in adults with nonspecific neck pain or adults who were asymptomatic: PubMed Central, MEDLINE, CINAHL, Scopus, Web of Science, and Google Scholar, from inception to December 20, 2020. Details about the search strategies are provided in Supplementary Appendix 1. Reference lists of articles retrieved were also evaluated for relevant publications.

Two authors (M.S. and M.R.) independently conducted the study selection. After removing duplicates, titles, abstracts, and full texts were thoroughly analyzed for eligibility. Two experts (A.R. and M.B.) were consulted in case of disagreement.

Study Selection

We included studies reporting data on reliability, validity, and responsiveness of the CCFT without restrictions of publication data or language. Only studies using the instrumentation expressly adapted for the CCFT were considered.

Participants were men and women who were at least 18 years old and who were asymptomatic or had nonspecific neck pain in either the acute or the persistent form. Also, studies including patients with neck pain due to specific (eg, cervical radiculopathy) or traumatic (eg, whiplash) causes or other defined neurologic, rheumatic, or systemic disorders were considered only when data on individuals with nonspecific neck pain or individuals who were asymptomatic could be extracted.

We included studies addressing the following metric properties as described in the Consensus-Based Standards for the Selection of Health Measurement Instruments (COSMIN) taxonomy[15]: reliability, defined as "the extent to which scores for patients who have not changed are the same for repeated measurement under several conditions," including reliability and measurement error (test–retest, interrater, and intrarater); validity, defined as "the degree to which the instrument measures the construct(s) it purports to measure," including construct validity ("the degree to which the scores of the instrument are consistent with hypotheses based on the assumption that the instrument validly measures the construct to be measured"); and responsiveness, defined as "the ability of the instrument to detect change over time in the construct to be measured."[15]

Downloaded from academic.oup.com/ptj/article/102/7/pzac054/6585158 by Biblioteca di Scienze, Università degli studi di Firenze user on 27 October 2026

Data Extraction

General information (eg, authors, publication year, journal), characteristics of enrolled samples (eg, sample size, age, sex, inclusion/exclusion criteria), property (eg, reliability, including measurement error, validity, responsiveness), and CCFT versions under estimation were collected.

For each property, 2 reviewers (G.C. and M.B.) independently extracted the following measures (from graphs when necessary): reliability, convergent validity, discriminant validity, and responsiveness.

Reliability

Reliability included intraclass correlation coefficients (ICCs, standard error of measurement (SEM), and minimal detectable change (MDC).

Convergent Validity

Convergent validity included data showing the association of CCFT pressure levels with the variation of activity, measured by electromyography (EMG) or ultrasonography (US), of SCFs (sternocleidomastoid and anterior scalene muscles) and DCFs, and with the variation of the range of motion (ROM) of CCF (CCF-ROM) in participants who were asymptomatic. Convergent validity also included data showing the correlation of CCFT with neck pain severity measured by pain or disability scales.

Discriminant Validity

Discriminant validity included data showing the association of CCFT pressure levels with the variation of SCF or DCF activity and with CCF-ROM in participants who were asymptomatic or who had nonspecific neck pain. Discriminant validity also included data comparing the CCFT performance in participants who were asymptomatic or who had nonspecific neck pain.

Responsiveness

Responsiveness included any index computed using a distribution- or anchor-based approach.

Quality Assessment

Two reviewers (M.S. and M.R.) rated the risk of bias independently. Disagreements were resolved by a third reviewer (M.B.). For reliability studies, we used the Quality Appraisal for Reliability (QAREL) checklist,[16] which comprises 11 items answered with "yes" (good quality), "no" (poor quality), "unclear" (insufficient information), or possibly "not applicable." Items are related to participant (item 1) or rater (item 2) representativeness, rater masking (items 3–7), examination order (item 8), time interval among measurement suitability (item 9), appropriate test application/interpretation (item 10), and correct statistics (item 11). Nine QAREL items have good or moderate reliability, and 2 items have fair reliability.[17] A study was considered to be of moderate to high quality when scoring "yes" to ≥50% of applicable items.

The COSMIN risk-of-bias checklist[18] was used for studies on validity (box 9 concerning hypotheses testing for construct validity, 7 items related to convergent/discriminant validity) and responsiveness (box 10 with 13 items concerning anchor- and distribution-based methods). Each item was scored, whenever applicable, as "very good," "adequate," "doubtful," or "inadequate."[19] The lowest rating of any item within a box was used to determine the risk of bias of the study for that domain ("the worst score counts" principle).[20]

Although COSMIN comprises boxes for assessing reliability studies, QAREL was preferred for this property because it itemizes possible sources of bias more thoroughly than COSMIN, which was mainly developed for Patient-Reported Outcome Measures.

On the basis of COSMIN-updated criteria (Suppl. Appendix 2), the quality of each measurement property was rated as positive, indeterminate, or negative.[19] With regard to reliability, however, a distinction was made, and different standards[21] were set for using the test in groups (ICC > 0.70; ie, the COSMIN criterion) or in individuals (ICC ≥ 0.90). When assessing groups, in fact, increase of sample size can compensate for relatively low reliability, but the random variability around the individual's true score is wide if the reliability coefficient is below the recommended level.[22]

Statistics

All meta-analyses were based on random-effects models to account for the heterogeneity of estimates.[23] Heterogeneity and inconsistency were assessed using the Q statistic and the I^2 statistic, respectively.[24]

The inverse-variance pooling method was used to combine coefficients from different studies into 1 pooled estimate.[25] Pooled estimates were computed by CCFT version or by grouping CCFT versions, which measure the level of CCF action achieved (CCFT-AS/CCFT1) or holding capacity (CCFT2/CCFT-PI/CCFT-CPI).

When missing, SEM and MDC were computed provided that authors reported the necessary data (ICC and SD/SE of the measure or rough data); otherwise, SEM was imputed using the prognostic method, which uses the SEs from other studies[26]; when the study reported different reliability estimates (eg, kappa statistics), data were reanalyzed computing the ICCs, provided that rough data were reported or could be extracted from graphs.

Pooled estimates of SEM and MDC for interrater and intrarater reliability were obtained by first pooling the SDs of measurements and then by applying the formulas $SEM = SD\sqrt{1 - ICC}$ and $MDC = 1.96 \bullet SEM \bullet \sqrt{2}$ to pooled SD and ICC.

Convergent validity was assessed through the analysis of variations of ROM and SCF or DCF activity in individuals who were asymptomatic and through correlations of CCFT with pain/disability scales.

The Bonett-Wright method[27] was used to estimate SEs on correlation of CCFT with pain/disability. All pooled correlations were obtained after Fisher z transformation to normalize their distribution.[25]

Discriminant validity was assessed by comparing individuals with nonspecific neck pain and individuals who were asymptomatic in terms of CCFT values and ROM, SCF, or DCF activity variation by increasing pressure levels. The difference in CCFT values was evaluated through the Cohen effect size based on the standardized mean difference, and results were pooled with standard meta-analytical methods.

Pooled estimates of ROM and muscle activity variation at each pressure level were obtained using the conventional random-effects model for meta-analysis after logit transformation of each study's proportion estimate for better approximation to normal distribution. Pooled estimates were

Downloaded from academic.oup.com/ptj/article/102/7/pzac054/6585158 by Biblioteca di Scienze, Università degli studi di Firenze user on 27 October 2026

obtained, both overall and stratified by method of muscle activity assessment.

Trends in pooled variations of ROM and muscle activities by increasing pressure levels for patients who were asymptomatic and patients with nonspecific neck pain were estimated using a linear regression model. All analyses were conducted using R-software version 3.4.1.

Results

A total of 3217 titles were retrieved through database searching. After removing duplicates and screening titles and abstracts, 35 full-text articles were assessed for eligibility and 21 studies (Tab. 1), published from 1999 to 2018, were eventually included (Fig. 1).

Seven studies included only participants who were asymptomatic,[10,28–33] 3 included only patients with nonspecific neck pain,[34–36] and 7 included both populations.[37–43] From 1 study,[38] we excluded data on individuals with a previous history of nonspecific neck pain who were asymptomatic at the time of the investigation. Four studies also included patients with other neck pain conditions[8,44–46] but reported separately data regarding individuals who were asymptomatic or patients with nonspecific neck pain. The CCFT versions assessed were CCFT-AS,[8,31,35,37,39,41] CCFT-PI,[8,36] CCFT-CPI,[10,33] CCFT1,[43] and CCFT2.[32,43]

Reliability and Measurement Error

Eight studies evaluated reliability. The interrater reliability was assessed in CCFT-AS[31,39] and CCFT-CPI,[33] and the intrarater reliability was assessed in CCFT-AS,[8,37,39] CCFT-PI,[8] CCFT-CPI,[10,33] CCFT1,[43] and CCFT2.[32,43] All studies but 2[8,37] also evaluated measurement error. In 1 study[33] reporting clearly incorrect data, the interrater ICC, SEM, and MDC were recomputed by extracting rough data from graphs, and the study was not included in intrarater reliability analyses. Only 2 of 5 studies on interrater reliability[31,39] and 1 of 7 studies on intrarater[39] reliability obtained a QAREL score ≥50% (Suppl. Appendix 3A).

For interrater reliability, 2 studies on CCFT-AS[31,39] were subjected to meta-analysis (Tab. 2; Suppl. Appendix 4A). For intrarater reliability, we performed 3 meta-analyses, pooling the data of 3 studies on CCFT-AS,[8,37,39] 2 studies on CCFT2,[32,43] and 1 study on CCFT1 with different assessment sessions and different populations (Tab. 3; Suppl. Appendix 4B).[43]

We classified the overall rating of interrater reliability of CCFT-AS and CCFT-CPI as positive for using the test at a group level, but negative for making clinical decisions at the individual level (level of evidence: low for CCFT-CPI, moderate for CCFT-AS).

For intrarater reliability, we classified the overall rating of CCFT-AS, CCFT1, and CCFT2 as positive or negative for using the test at a group or individual level, respectively (level of evidence: low); we also classified the overall rating of CCFT-PI and CCFT-CPI as positive for the assessment of individuals (level of evidence: very low).

Validity and Responsiveness

Quality ratings of included studies are reported in (Suppl. Appendix 3B).

Convergent validity was assessed in individuals who were asymptomatic by correlating CCFT stages with EMG activity of DCFs[29,45] and SCFs[29,38,42,44–46] or with thickness variation of DCFs and SCFs measured with US.[30–40] We extracted the data of 3 studies from graphs.[38,42,46]

The quality was doubtful in 6 studies[29,30,38,40,44,46] and adequate in 2.[42,46] The meta-analytic estimates included all studies (Tab. 4; Suppl. Appendix 5) and showed progressively increasing DCF activity, measured with either EMG (trend = 10.96; 95% CI = 8.87 to 13.04) or US (trend = 4.4; 95% CI = 3.21 to 5.59). SCFs showed increasing activation as well, but the trend was much lower compared with DCFs when measured with EMG (trend = 6.58; 95% CI = 4.66 to 8.5) and slightly higher when measured with US (6.23; 95% CI = 4.9 to 7.56).

Convergent validity was also investigated in 3 studies (quality rating: adequate) by correlating in individuals who were asymptomatic the CCFT stages with CCF-ROM measured with a digital imaging technique.[28,29,45] The meta-analytic estimate (all studies; Suppl. Appendix 6) showed that CCF-ROM increased with incremental pressure targets (trend = 13.63; 95% CI = 10.83 to 16.44).

We classified the overall rating of CCFT convergent validity for expressly assessing DCF activation as positive or negative when measuring EMG activity or US thickness variation, respectively, and as positive when measuring CCF-ROM (level of evidence: low).

Convergent validity was assessed in 3 studies (quality rating: very good) by correlating CCFT-AS,[41] CCFT1 or CCFT2,[43] or CCFT-PI[36] with pain or disability in patients with nonspecific neck pain. Pooled correlation coefficients were mostly not significant and ranged from −0.08 to −0.24 (pain) and from 0.20 to −0.26 (disability) (Suppl. Appendix 7). We classified the overall rating of CCFT convergent validity with neck pain severity as negative (level of evidence: CCFT1 or CCFT2, low; CCFT-PI, moderate).

Discriminant validity was assessed by comparing the CCFT performance of patients with nonspecific neck pain and people who were asymptomatic using CCFT-AS,[37,39,41] CCFT1, or CCFT2.[43] Quality was rated as very good,[41,43] adequate,[37] or doubtful.[39] The effect size estimates (Fig. 2) were 1.01 (95% CI = 0.52 to 1.50; CCFT-AS; 3 studies pooled[11,14,15]), 0.03 (95% CI = −0.45 to 0.50; CCFT1), and 0.18 (95% CI = −0.30 to 0.67; CCFT2).

We classified the overall rating of CCFT discriminant validity in differentiating participants who were asymptomatic and patients with nonspecific neck pain as positive for CCFT-AS and negative for CCFT1 or CCFT2 (level of evidence: high and moderate, respectively).

Discriminant validity was also assessed by comparing participants who were asymptomatic and individuals with nonspecific neck pain with regard to EMG activity of DCFs[29,34,45] and SCFs,[29,34,38,42,44–46] US thickness variation of DCFs[30,40] and SCFs,[30,40] and CCF-ROM[28,29,34,45] at the 5 incremental CCFT stages. Quality was rated as inadequate,[40] doubtful,[29,30,34,38,42,44] or adequate.[45,46] Pooled estimates of both EMG and US studies consistently showed that patients with nonspecific neck pain performed the 5 CCFT stages with a higher increase of SCFs and lower increase of DCFs than individuals who were asymptomatic (Tab. 4; Suppl. Appendix 5). They also showed lesser CCF-ROM at each level, although the trend in ROM increment by level was similar (Suppl. Appendix 6).

Downloaded from academic.oup.com/ptj/article/102/7/pzac054/6585158 by Biblioteca di Scienze, Università degli studi di Firenze user on 27 October 2026

Table 1. Characteristics of the Selected Studies[a]

Study	Participants	Age, Mean (SD)	No. of Men/No. of Women	Measurement Property Estimated From Data Presented in Study	Was Measurement Property Estimate a Primary Study Aim?	Outcome Measure(s)
Jull et al[8] (1999)	15 asymptomatic	NR	NR	Intrarater reliability (data from 50 asymptomatic participants, imputing SD for ICC and computing MDC using SD of 15 asymptomatic participants	No	CCFT-AS; CCFT-PI
	15 with cervical headache	NR	NR			
	50 asymptomatic (for reliability)	NR	NR			
James and Doe[10] (2010)	19 asymptomatic	24.9 (3.5)	12/7	Intrarater reliability	Yes	CCFT-CPI
Falla et al[28] (2003)	20 asymptomatic	28.9 (8.2)	12/8	Discriminant validity; convergent validity (data from 17 participants)	Yes (convergent validity)	CCF-ROM
Falla et al[29] (2003)	10 asymptomatic	30.7 (10.3)	4/6	Discriminant validity; convergent validity (ROM, SCM, anterior scalene muscle, DCFs)	No	CCF-ROM; EMG
Jesus et al[30] (2008)	10 asymptomatic	26.0 (6.7)	6/4	Discriminant validity; convergent validity (SCM, DCFs)	Yes (convergent validity)	US
Arumugam et al[31] (2011)	30 asymptomatic	33.7 (8.55)	15/15	Interrater reliability (recomputed from rough data)	Yes	CCFT-AS
Dimitriadis et al[32] (2013)	20 asymptomatic	34.1 (12.3)	8/12	Intrarater reliability	Yes	CCFT2
Kotwani et al[33] (2018)	60 asymptomatic	NR	NR	Interrater reliability (recomputed from rough data)	Yes	CCFT-CPI
Jull and Falla[34] (2016)	32 with neck pain	38.0 (11.6)	0/32	Discriminant validity (SCM, anterior scalene muscle, DCFs)	No	EMG
Jørgensen et al[35] (2017)	164 with neck pain	45.1 (11.7)	36/128	Responsiveness	Yes	CCFT-AS
Subbarayalu and Ameer[36] (2017)	84 with neck pain	NR	57/27	Convergent validity	No	CCFT-PI
Chiu et al[37] (2005)	20 asymptomatic	40.6 (11.0)	8/12	Intrarater reliability (data from 10 asymptomatic participants); discriminant validity (both measures computed from rough data)	Yes (discriminant validity)	CCFT-AS
	20 with neck pain	43.4 (10.9)	7/13			
Johnston et al[38] (2008)	22 asymptomatic	37.4 (10.4)	NR	Discriminant validity (SCM, anterior scalene muscle; only data from asymptomatic participants and participants with moderate or mild neck pain)	Yes	EMG
	14 with moderate neck pain (NDI ≥ 30)	45.4 (10.3)	NR			
	38 with mild neck pain (NDI = 9–29)	43.8 (9.4)	NR			
	33 with no pain (NDI ≤ 8)	43 (10.6)	NR			
Juul et al[39] (2013)	30 asymptomatic	34 (14.7)	13/17	Inter- and intrarater reliability; discriminant validity (estimates from 2 examiners and 2 assessments)	Yes (reliability)	CCFT-AS
	33 with neck pain	54 (15.1)	8/25			
Jun and Kim[40] (2013)	20 asymptomatic	24.8 (0.7)	11/9	Discriminant validity; convergent validity (SCM, DCFs)	Yes	US
	20 with neck pain	23.8 (0.3)	8/12			
Elsig et al[41] (2014)	30 asymptomatic	37.2 (13.5)	5/25	Discriminant validity; convergent validity	Yes	CCFT-AS
	30 with neck pain	36.9 (13.6)	5/25			
Steinmetz et al[42] (2016)	21 asymptomatic (violinists)	28.2 (11.9)	6/15	Discriminant validity (SCM)	Yes	EMG
	21 asymptomatic (nonmusicians)	31.8 (9.8)	6/15			
	12 with neck pain	29.3 (13.3)	3/9			
Martins et al[43] (2018)	33 asymptomatic (26 in session 2)	36.2 (2.3)	5/28	Intrarater reliability (estimates from 2 examiners, within 2 sessions and between sessions); discriminant validity (estimates from 1 session); convergent validity (estimates from 2 sessions)	Yes	CCFT1; CCFT2
	33 with neck pain (31 in session 2)	36.8 (2.4)	5/28			
Jull[44] (2000)	12 with whiplash	37.8 (11.6)	NR	Discriminant validity (SCM; only data from 7 asymptomatic participants)	Yes	EMG
	12 asymptomatic	NR	NR			
Falla et al[45] (2004)	10 asymptomatic	26.4 (5.8)	3/7	Discriminant validity (ROM, SCM, anterior scalene muscle, DCFs; only data from 10 asymptomatic participants)	Yes	CCF-ROM; EMG
	10 with neck pain (3 with whiplash)	32.2 (9.1)	4/6			
Jull et al[46] (2004)	25 asymptomatic	39.3 (14.0)	10/15	Discriminant validity (SCM; only data from asymptomatic participants and participants with neck pain)	Yes	EMG
	25 with neck pain	40.3 (9.2)	5/20			
	25 with whiplash	36.3 (10.2)	8/17			

[a]CCF-ROM = range of motion of craniocervical flexion; CCFT = craniocervical flexion test; CCFT1 = level of craniocervical action, scored with the modified version of the test; CCFT2 = level of isometric endurance, scored with the modified version of the test; CCFT-AS = CCFT Activation Score; CCFT-CPI = CCFT Cumulative Performance Index; CCFT-PI = CCFT Performance Index; DCFs = deep cervical flexor muscles; EMG = variation of muscle activity measured with electromyography; MDC = minimal detectable change; NDI = Neck Disability Index; NR = not reported; ROM = range of motion; SCM = sternocleidomastoid muscle; US = variation of muscle thickness measured with ultrasonography.

Downloaded from academic.oup.com/ptj/article/102/7/pzac054/6585158 by Biblioteca di Scienze, Università degli studi di Firenze user on 27 October 2026

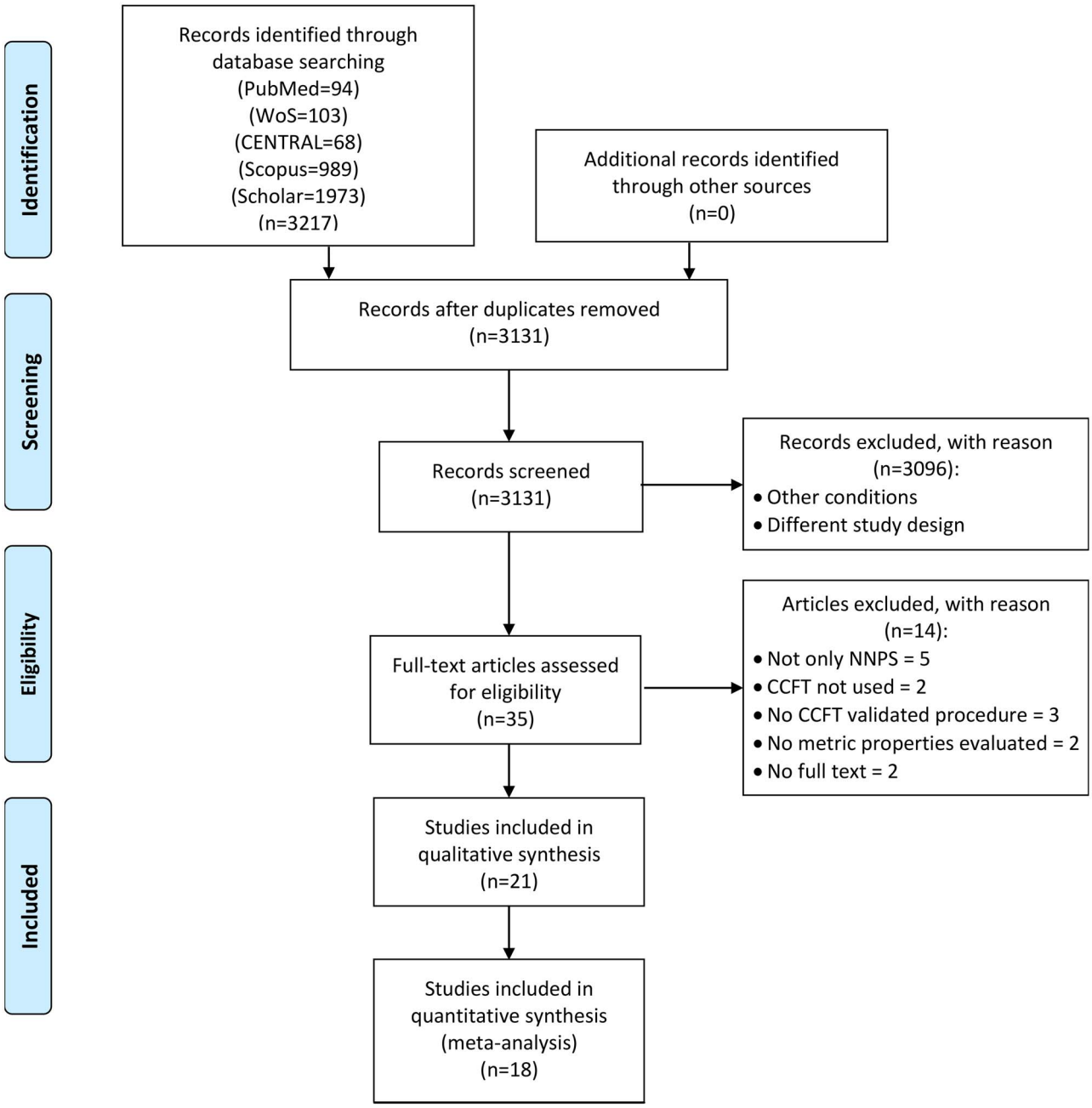


Figure 1. Flowchart of the process of article selection. CCFT = craniocervical flexion test; NNP = individuals with nonspecific neck pain; WoS = Web of Science.

We classified the overall rating of CCFT discriminant validity in differentiating participants who were asymptomatic and individuals with nonspecific neck pain based on muscle activity or CCFT-ROM as positive (level of evidence: low).

One study of doubtful quality[35] evaluated the responsiveness of CCFT-AS in patients with neck pain who completed a 4-month follow-up after rehabilitation, using the Neck Disability Index (NDI) as anchor and the receiver operating characteristic curve approach. A cut score >7 NDI change points was chosen for differentiating patients with improvement (n = 20) and patients without improvement (n = 108). The area under the curve was 0.54 (95% CI = 0.42 to 0.67), with poor sensitivity (50.0) and specificity (54.2). The CCFT-AS change score barely correlated with the NDI change score ($r = 0.19$; 95% CI = 0.03 to 0.35) and was, on average, similar between individuals showing improvement and those not showing improvement. We classified the overall rating of CCFT responsiveness as negative (level of evidence: low).

Discussion

Our study aimed to investigate the metric properties of the CCFT for assessing DCF control in asymptomatic and nonspecific neck pain populations, but, contrary to existing systematic reviews, we evaluated separately different CCFT versions and scores. We will first discuss results about CCFT reliability because they help in understanding the conflicting findings about validity.

Downloaded from academic.oup.com/ptj/article/102/7/pzac054/6585158 by Biblioteca di Scienze, Università degli studi di Firenze user on 27 October 2026

Table 2. Interrater Reliability of Each CCFT Version[a]

Study	CCFT version	Participant Group (No. of Participants)	Assessment Session	No. of Raters	ICC (95% CI)	SEM	MDC	Meta-analytical estimates		
								ICC (95% CI) [Heterogeneity]	SEM	MDC
Arumugam et al (2011)[31]	AS	Asympt (30)	1	2	0.84 (0.7–0.92)	0.68	1.88	0.86 (0.82–0.89) [$I^2 = 0\%$]	1.27	3.52
Juul et al (2013)[39]	AS	NNP (63)	1	2	0.85 (0.76–0.91)	1.64	4.54			
		Asympt (63)	2		0.86 (0.81–0.93)	1.55	4.30			
Kotwani et al (2018)[33]	CPI	Asympt (60)	1	2	0.83 (0.72–.09)[b]	4.41[b]	12.23[b]			

[a]AS = activation score; Asympt = asymptomatic; CCFT = craniocervical flexion test; CPI = Cumulative Performance Index; NNP = nonspecific neck pain. [b]Recomputed from rough data.

Table 3. Intrarater Reliability of Each CCFT Version and Pooled Estimates[a]

Study	CCFT Version	Participant Group (No. of Participants)	Rater Who Performed Assessment	No. of Raters	ICC (95% CI)	SEM	MDC	Meta-Analytical Estimates			
								ICC (95% CI) [Heterogeneity]		SEM	MDC
Jull et al (1999)[8]	AS	Asympt (50)	1	1	0.81 (0.68–0.89)[b]	0.65[c]	1.81[c]	0.78 (0.71–0.83) [$I^2 = 5.8\%$]	0.78 (0.74–0.82) [$I^2 = 0\%$]	1.26	3.50
Chiu et al (2005)[37]	AS	Asympt (10)	1	1	0.84 (0.47–0.96)[d]	1.02[d]	2.82[d]				
Juul et al (2013)[39]	AS	NNP and Asympt (63)	1	2	0.69 (0.53–0.80)	1.84	5.10				
			2		0.81 (0.70–0.88)	1.48	4.10				
Martins et al (2018)[43]	CCFT1	NNP (33)	1 (w-s 1)	2	0.83 (0.73–0.91)	1.55	4.30	0.79 (0.73–0.83) [$I^2 = 0\%$]		1.55	4.30
		NNP (31)	1 (w-s 2)		0.81 (0.69–0.9)	1.62	4.49				
		NNP (31)	1 (b-s)		0.7 (0.39–0.86)	1.79	4.96				
		Asympt (33)	2 (w-s 1)		0.73 (0.58–0.85)	1.66	4.60				
		Asympt (26)	2 (w-s 2)		0.82 (0.7–0.9)	1.48	4.10				
		Asympt (26)	2 (b-s)		0.77 (0.52–0.89)	1.45	4.02				
Dimitriadis et al (2013)[32]	CCFT2	Asympt (20)	1	1	0.81 (0.59–0.92)	0.97	2.69	0.84 (0.78–0.88) [$I^2 = 37.8\%$]	0.88 (0.81–0.93) [$I^2 = 78.8\%$]	1.27	3.51
Martins et al (2018)[43]	CCFT2	NNP (33)	1 (w-s 1)	2	0.89 (0.81–0.94)	1.13	3.13				
		NNP (31)	1 (w-s 2)		0.70 (0.54–0.83)	1.55	4.30				
		NNP (31)	1 (b-s)		0.90 (0.80–0.95)	0.9	2.49				
		Asympt (33)	2 (w-s 1)		0.86 (0.77–0.93)	1.16	3.22				
		Asympt (26)	2 (w-s 2)		0.84 (0.73–0.91)	1.33	3.69				
		Asympt (26)	2 (b-s)		0.81 (0.6–0.91)	1.29	3.58				
Jull et al (1999)[8]	PI	Asympt (50)	1	1	0.93 (0.88–0.96)	7.57	20.97				
James and Doe (2010)[10]	CPI	Asympt (19)	1	1	0.98 (0.96–0.99)	8.94	24.78				

[a]AS = Activation Score; Asympt = asymptomatic; b-s = between sessions; CCFT = craniocervical flexion test; CCFT1 = level of activation of deep cervical flexor muscles with the new CCFT protocol; CCFT2 = level of isometric endurance of deep cervical flexor muscles with the new CCFT protocol; CPI = Cumulative Performance Index; MDC = minimal detectable change; NNP = nonspecific neck pain; PI = Performance Index; SEM = standard error of measurement; w-s = within session. [b]CI imputed. [c]Computed from ICC and imputed SD. [d]Recomputed from rough data.

Downloaded from academic.oup.com/ptj/article/102/7/pzac054/6585158 by Biblioteca di Scienze, Università degli studi di Firenze user on 27 October 2026

Table 4. Trends and 95% CIs in Variation of Muscle Activity for 2-mmHg Pressure Increase by Muscle, Individuals, and Method of Assessment[a]

Muscle	Trend (95% CI) for:					
	Participants Who Were Asymptomatic			Participants Who Had Neck Pain		
	All Measures	Only EMG	Only US	All Measures	Only EMG	Only US
Sternocleidomastoid	6.36 (4.86–7.87)	6.58 (4.66–8.5)	6.23 (4.9–7.56)	10.58 (10.27–10.9)	11.02 (10.09–11.95)	9.1 (7.16–11.04)
Anterior scalene	5.16 (2.5–7.82)	5.16 (2.5–7.82)		14.4 (12.33–16.47)	14.4 (12.33–16.47)	
Deep cervical flexor	6.36 (4.6–8.12)	10.96 (8.87–13.04)	4.4 (3.21–5.59)	3.16 (2.31–4.01)	5.29 (2.74–7.84)	1.5 (1.18–1.82)

[a]EMG = electromyography; US = ultrasonography.

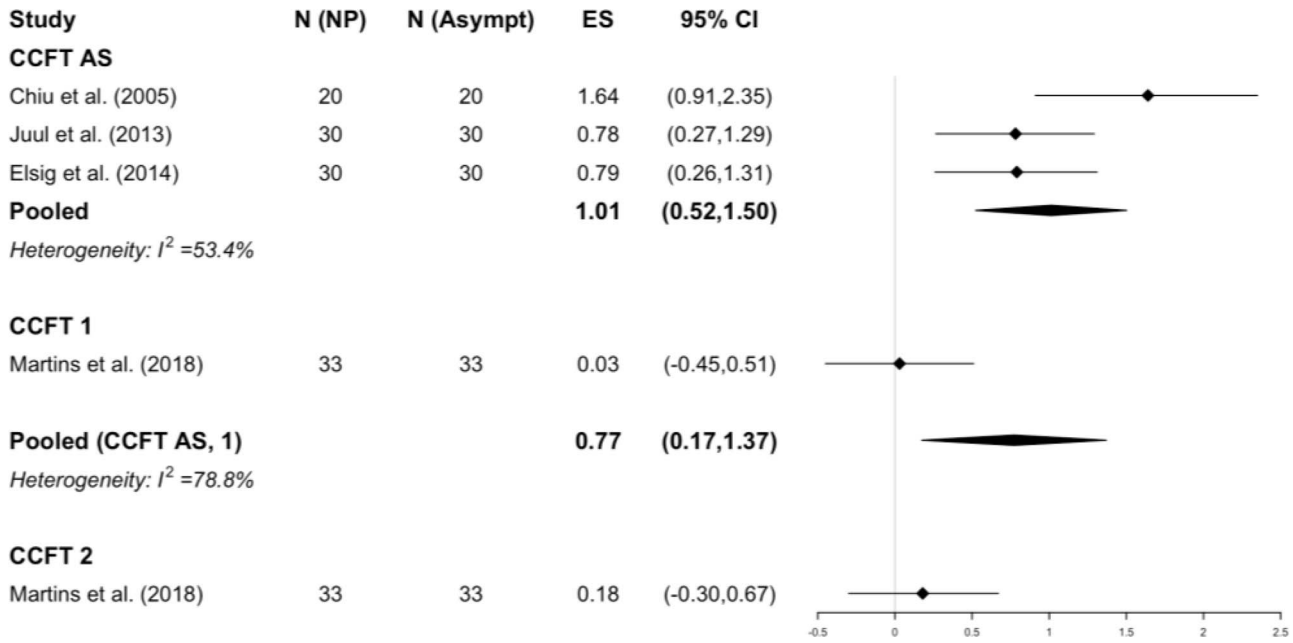


Figure 2. Forest plot of discriminant validity of the craniocervical flexion test (CCFT), measured as the Cohen effect size (ES) between participants with neck pain (NP) and participants who were asymptomatic (Asympt), along with the corresponding sample size (N) and 95% CI. AS = activation score.

Reliability

Our review demonstrated that CCFT versions are mostly reliable only at a group level. CCFT-PI and CCFT-CPI showed sufficient reliability for making clinical decisions on individuals as well, but only when the test is applied by the same rater. We rated the level of evidence on reliability from low to moderate due to the overall quality of the studies and/or the limited sample sizes. Previous systematic reviews classified CCFT reliability as negative,[11] positive with moderate level of evidence,[12] and good but needing further investigations.[13] Measurement error was not rated because it was not described in the included literature,[11] rated as indeterminate with unknown level of evidence,[12] or not considered at all.[13] Several important differences among these reviews and ours can explain these conflicting conclusions.

First, as outlined in the introduction, none of the previous reviews conducted separate analyses for different CCFT versions that evaluate different or the same ability with different testing procedures. Two reviews[11,12] also included studies investigating the reliability of EMG activity measured during the test; because the test does not consider the appraisal of EMG activity, this provides no information about CCFT reliability.

Second, different reviews included different numbers of studies and adopted different inclusion criteria. We included 8 articles, all providing data on CCFT reliability in participants who were asymptomatic and/or participants with nonspecific neck pain. The previous reviews included 4,[11] 12,[12] and 5[13] articles, but when studies assessing reliability in patients with specific neck pain or reliability of EMG data—or not using the original biofeedback pressure unit—were excluded, the number of studies evaluated in the previous reviews dropped to 2,[11] 7,[12] and 3.[13]

Third, interpretation of reliability indexes greatly affects authors' conclusions. In the present review, we used the criteria suggested by Fitzpatrick et al,[21] where reliability coefficients >0.70 are considered as acceptable when studying a group but coefficients of at least 0.90 are required for making decisions about an individual depending on their score. This interpretation of reliability coefficients (ICCs) is consistent with our finding of large or moderately large MDC for most of the CCFT measures: an MDC of 2 target levels (4.0 mmHg) should undoubtedly be considered as excessive for measures with only 5 target levels (CCFT-AS, CCFT1, and CCFT2). Conversely, an MDC of approximately 25 is acceptable for a measure (CCFT-CPI) whose value may range from 0 to 300.

Downloaded from academic.oup.com/ptj/article/102/7/pzac054/6585158 by Biblioteca di Scienze, Università degli studi di Firenze user on 27 October 2026

De Konig et al[11] adopted restrictive criteria (ICC ≥ 0.85) without differentiating about test application (group/individual) and rated as negative CCFT-AS reliability (all studies included used this version). Adopting the COSMIN cutoff (sufficient reliability when ICC/kappa is >0.70), De Araujo et al[12] rated CCFT reliability as positive and measurement error as indeterminate. Indeed, this property should have been rated as negative because an MCID of 2.0 has been reported for CCFT-AS by 1 study,[35] also included in their review, and this value is inferior to our meta-analytic estimate of CCFT-AS MDC. However, authors did not subject the data to a meta-analysis. Selistre et al[13] interpreted reliability as good or excellent when ICCs were 0.75 to 0.90 or > 0.90, respectively. This interpretation[47] is similar to our criteria (good and excellent reliability might be equate to sufficient reliability at group or individual level, respectively). However, pooling data of different CCFT versions, the authors failed to detect differences in the intrarater reliability of activation and endurance scores and rated CCFT reliability as good.

Selistre et al[13] calculated meta-analytic estimates of CCFT reliability separately in participants who were asymptomatic, individuals with neck pain, and a mixed population. They found the highest intrarater reliability in participants who were asymptomatic, but this finding might be biased by the limited number of studies and because different CCFT versions were pooled. Whereas the unique study on participants with neck pain used CCFT1, the 3 studies on individuals who were asymptomatic used CCFT-AS, CCFT-CPI, or CCFT1. Nevertheless, distinguishing results in different populations is an important issue. Conceivably, persons with neck pain might more frequently use compensatory strategies, and this fact would require higher rater ability in detecting deviations from the correct execution. Although we did not differentiate results between participants who were asymptomatic and participants with nonspecific neck pain, because few data are available on the latter, reliability indexes seem better when calculated in samples of participants who were asymptomatic than in mixed populations (asymptomatic + nonspecific neck pain). For example, the MDC of CCFT-AS is almost double in the latter. This might also be due to a relatively higher performance variability with individuals (likely with nonspecific neck pain) unable to achieve even the first pressure increments and individuals (likely asymptomatic) who complete all the levels. Further studies on large samples of patients with nonspecific neck pain are strongly recommended.

A final issue is related to quality assessment. De Konig et al[11] developed a checklist where quality assessment (simply rated as adequate or doubtful, without details) was not separated by the analysis of reliability values (ICC, kappa, SEM) found in the study; thus, they rated reliability as positive when methods were adequate, and indexes met the required standard but made no inferences about the level of evidence. Selistre et al[13] adopted 2 checklists, including QAREL, to assess quality. Their rating of the 3 studies also included in the present review is similar to our rating, but they made no clear statements about evidence levels. De Araujo et al[12] used the COSMIN checklist, which does not list important sources of bias (eg, rater masking to other raters' findings) that need to be considered under the item asking to note "any other important flaws in the design or statistical methods of the study." Quality assessment of a reliability study may lead to dubious conclusions if these important features are not considered. For example, authors rated as adequate the quality of 2 studies[8,33] that met only 1 of 9 and 2 of 9, respectively, of the applicable QAREL items.

Validity and Responsiveness

We found that CCFT seems suitable to investigate individuals' ability to activate DCFs and discriminate between participants who were asymptomatic and individuals with nonspecific neck pain, the latter achieving on average lower CCFT-AS without compensations. This finding confirms the hypothesis that they may experience impaired DCF control.[9] However, their performance seems poorly related to neck pain severity, but this finding might be biased by the poor reliability of the CCFT when applied to scoring the individual performance. The insufficient reliability might also explain the poor responsiveness of the test. We rated the level of evidence as generally low or moderate, except for discriminant validity of CCFT-AS between participants who were asymptomatic and individuals with nonspecific neck pain, whose level of evidence is high.

Previously, CCFT validity was reviewed by De Araujo et al.[12] and Selistre et al,[13] the latter including only 1 study that investigated the concurrent validity of using a sphygmomanometer cuff compared with the biofeedback unit designed for the test[48]; actually, this study provides no information about CCFT validity.

As far as CCFT validity, 3 aspects will be considered, linked to each other. First, does CCFT quantify DCF control (convergent validity: correlations between CCFT and DCF activity or CCF-ROM)? Second, can CCFT identify people assumed to partly lack such control (discriminant validity: differences in CCFT performance between people who are asymptomatic and people with nonspecific neck pain)? Third, does CCFT correlate with neck pain severity (convergent validity: correlations between CCFT and pain/disability scales; discriminant validity: differences in CCFT performance between subgroups with mild/moderate/severe nonspecific neck pain)?

De Araujo et al[12] rated CCFT convergent validity as positive based on studies of participants who were asymptomatic, correlating pressure levels to SCF and/or DCF EMG activity (first aspect of validity, as outlined above). Our results confirm and extend this finding because we also analyzed studies where DCF activity was measured with US, and we computed meta-analytic estimates of muscle activity variation with pressure increase. We also provided evidence of a positive association of CCFT pressure levels and CCF-ROM, which was not considered by the previous review.[12] Data on ROM and muscle activity demonstrate that, when correctly executed, CCFT requires a progressive increase of CCF and DCF activation. We classified the level of evidence as low because of the small cumulative sample size and insufficient data on metric properties of US and surface EMG for measuring DCF activity. Limited evidence of surface EMG reliability[29] and validity[49] has been provided in very small samples of individuals who were asymptomatic, whereas published data on US reliability and validity are conflicting.[50–52]

The finding that DCF activity was increased by CCFT pressure levels is somewhat limited by the finding that SCFs also are progressively more activated. The trend is almost double in DCF when activity is recorded by EMG but slightly higher in SCFs when measured with US. Pooling EMG and US data, trends for SCFs and DCFs are almost identical, also confirming that individuals who are asymptomatic may lack enough DCF control to correctly perform the test. Indeed,

Downloaded from academic.oup.com/ptj/article/102/7/pzac054/6585158 by Biblioteca di Scienze, Università degli studi di Firenze user on 27 October 2026

1 of the included studies[44] found that only 7 of 85 asymptomatic volunteers enrolled were able to complete CCFT with minimal SCF activity, as measured by EMG. However, when muscle activity was compared in participants who were asymptomatic and individuals with nonspecific neck pain, our meta-analysis demonstrated that the latter showed a greater increase in SCFs and a lower increase in DCFs than participants who were asymptomatic. This addresses the discriminant validity of the test.

De Araujo et al[12] concluded that CCFT has good discriminant validity in differentiating participants who were asymptomatic and individuals with neck pain, mixing studies in heterogeneous conditions (nonspecific neck pain, whiplash, headache) and mixing data of all CCFT versions and EMG activity. Pooling data separately for different variables, we demonstrated that when the test is applied measuring muscle activity and/or CCF-ROM, discouraging compensations but without stopping the test when excessive SCF activation is detected, individuals with nonspecific neck pain showed a larger increase of SCFs than individuals who were asymptomatic, mostly in the last pressure increments and irrespective of the measurement instrument (EMG/US), and performed the test with approximately 20% less CCF-ROM at any pressure level. These findings do not provide evidence of CCFT discriminant validity but further stress the need of careful qualitative monitoring when it is applied. CCFT does not require muscle activity or CCF-ROM recording, and the use of compensatory strategies needs to be detected by the rater through observation/palpation.[8,9] A rater's ability is a critical requirement of the test, which would be useless otherwise.

We found a good ability of CCFT-AS, but not of CCFT1 and CCFT2, to discriminate between individuals who were asymptomatic and individuals with neck pain. All of these versions have sufficient reliability at a group level. CCFT-AS and CCFT1 assess individuals' ability to perform a correct CCF action lasting 10 seconds and 2 to 3 seconds, respectively. The latter might not be enough for an accurate evaluation. CCFT2 evaluates endurance of DCFs, requesting three 10-second repetitions at each pressure level until failure. Because impaired DCF endurance was demonstrated in neck pain,[53] three 10-second repetitions might not be enough to assess this ability. Unfortunately, no studies investigated the discriminative ability of CCFT-PI or CCFT-CPI, which measure endurance as the ability to perform ten 10-second repetitions.

We found weak and mostly non-significant correlations between different CCFT versions and neck pain severity. This aspect of CCFT convergent validity was not considered in the previous review. The lack of correlation was unexpected because of our results on differences between participants who were asymptomatic and individuals with nonspecific neck pain in both muscle activity pattern and CCFT score, and because a CCF training was found effective in reducing pain and disability in persons with neck pain.[54] It is likely biased by the poor reliability of most CCFT versions for use at the individual level.

Responsiveness has been investigated only for CCFT-AS and was found poor.[35] This finding must be treated with caution for both the sample size enrolled and the great imbalance between individuals with no change (n = 118) and those with improvement (n = 20). Nevertheless, it is consistent with the poor correlation between CCFT and neck pain severity and is likely also due to the insufficient reliability of CCFT-AS to measure the individual performance. An MCID of 2.0 (equal to 1 pressure level) is high for a measure that may assume 5 levels. In our opinion, further studies on this property would be recommended only using the CCFT versions with sufficient reliability (CCFT-PI/CCFT-CPI). For the other versions, reliability should be improved before using the test as an outcome measure at the individual level. CCFT was able to detect longitudinal changes in groups of patients with neck pain[55] and differences in change scores between groups in clinical trials.[56–58] This stresses again the need of distinguishing whether results are presented for groups (small changes may be interpreted as important) or for individuals (larger changes are needed to be confident that a meaningful change occurred).[59]

Limitations

Our results should be interpreted with caution because they are based on studies mostly of poor quality and some meta-analytical estimates have significant heterogeneity. The exclusion of studies where results were reported without differentiating between patients with nonspecific neck pain or patients with traumatic neck pain, even when the latter were a small minority, might have led to omitting relevant information. Moreover, most meta-analyzed studies assessed reliability and measurement error without differentiating individuals who were asymptomatic and individuals who were symptomatic.

Conclusions and Need for Further Research

CCFT is a potentially useful tool for detecting impairment in DCF control and identifying patients with nonspecific neck pain who would benefit from a targeted intervention. However, the poor reliability of most of CCFT versions for assessing the individual performance greatly limits its suitability for that purpose in clinical practice. A crucial requirement is raters' ability to detect compensations, but this detection is also complicated because in individuals who are asymptomatic, a growing activation of SCFs occurs with increasing pressure. Thus, one may wonder what should be considered "excessive activity" of SCFs. Proper rater training could improve reliability, but unfortunately almost none of the included studies addressed this issue, and often no information about raters were provided. We stress the need for further research on the reliability of different CCFT versions where raters are thoroughly trained. The use of EMG biofeedback to measure SCM muscle activity during the test might improve reliability and the other properties of the CCFT as well but would require a substantial modification of the procedure that would limit its clinical application. The evidence provided by this study indicates that, when applying the test in clinical practice with only the pressure biofeedback apparatus and no other instrumentations, it is strongly recommended to use the CCFT-PI or the CCFT-CPI as the outcome measure. Although far more time-consuming than the AS, these scores are the only measures found to be reliable enough to help make a clinical decision at the individual level. Whether they are also capable of detecting longitudinal changes is yet to be determined in future research.

Author Contributions

Concept/idea/research design: A. Romeo, M. Baccini, P. Pillastrini, M. Di Bari
Writing: A. Romeo, M. Baccini, M. Sagripanti, M. Ruggeri
Data collection: A. Romeo, M. Baccini, M. Sagripanti, M. Ruggeri

Downloaded from academic.oup.com/ptj/article/102/7/pzac054/6585158 by Biblioteca di Scienze, Università degli studi di Firenze user on 27 October 2026

Data analysis: A. Romeo, M. Baccini, G. Carreras, M. Sagripanti, M. Ruggeri
Project management: P. Pillastrini, M. Di Bari
Consultation (including review of manuscript before submitting): P. Pillastrini, M. Di Bari

Funding

There are no funders to report for this study.

Systematic Review Registration

This study was registered in the International Prospective Register of Systematic Reviews (PROSPERO) (CRD42020200155).

Disclosures

Paolo Pillastrini is a member of *PTJ*'s Editorial Board. The authors completed the ICMJE Form for Disclosure of Potential Conflicts of Interest and reported no other conflicts of interest.

References

1. Blanpied PR, Gross AR, Elliott JM, et al. Neck pain: revision 2017. *J Orthop Sports Phys Ther*. 2017;47:A1–A83.
2. Panjabi MM, Cholewicki J, Nibu K, Grauer J, Babat LB, Dvorak J. Critical load of the human cervical spine: an in vitro experimental study. *Clin Biomech (Bristol, Avon)*. 1998;13:11–17.
3. De Pauw R, Coppieters I, Kregel J, De Meulemeester K, Danneels L, Cagnie B. Does muscle morphology change in chronic neck pain patients? - A systematic review. *Man Ther*. 2016;22:42–49.
4. De Marco M, Vanti C, Pillastrini P, Romeo A. Structural and functional changes of cervical neuromuscular system associated with insidious onset mechanical neck pain: a literature review. *Int J Phys Ther Rehab*. 2015;1:103.
5. Miranda IF, Wagner Neto ES, Dhein W, Brodt GA, Loss JF. Individuals with chronic neck pain have lower neck strength than healthy controls: a systematic review with meta-analysis. *J Manip Physiol Ther*. 2019;4:608–622.
6. Hesby BB, Hartvigsen J, Rasmussen H, Kjaer P. Electronic measures of movement impairment, repositioning, and posture in people with and without neck pain-a systematic review. *Syst Rev*. 2019;8:220.
7. Stanton TR, Leake HB, Chalmers KJ, Moseley GL. Evidence of impaired proprioception in chronic, idiopathic neck pain: systematic review and meta-analysis. *Phys Ther*. 2016;96:876–887.
8. Jull G, Barrett C, Magee R, Ho P. Further clinical clarification of the muscle dysfunction in cervical headache. *Cephalalgia*. 1999;19:179–185.
9. Jull GA, O'Leary SP, Falla DL. Clinical assessment of the deep cervical flexor muscles: the craniocervical flexion test. *J Manip Physiol Ther*. 2008;31:525–533.
10. James G, Doe T. The craniocervical flexion test: intra-tester reliability in asymptomatic subjects. *Physiother Res Int*. 2010;15:144–149.
11. de Koning CH, van den Heuvel SP, Staal JB, Smits-Engelsman BC, Hendriks EJ. Clinimetric evaluation of methods to measure muscle functioning in patients with non-specific neck pain: a systematic review. *BMC Musculoskelet Disord*. 2008;9:142.
12. Araujo FX, Ferreira GE, Scholl Schell M, Castro MP, Ribeiro DC, Silva MF. Measurement properties of the craniocervical flexion test: a systematic review. *Phys Ther*. 2020;100:1094–1117.
13. Selistre LFA, Melo CS, Noronha MA. Reliability and validity of clinical tests for measuring strength or endurance of cervical muscles: a systematic review and meta-analysis. *Arch Phys Med Rehabil*. 2021;102:1210–1227.
14. Coppieters I, De Pauw R, Kregel J, et al. Differences between women with traumatic and idiopathic chronic neck pain and women without neck pain: interrelationships among disability, cognitive deficits, and central sensitization. *Phys Ther*. 2017;97:338–353.
15. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol*. 2010;63:737–745.
16. Lucas NP, Macaskill P, Irwig L, Bogduk N. The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). *J Clin Epidemiol*. 2010;63:854–861.
17. Lucas N, Macaskill P, Irwig L, et al. The reliability of a quality appraisal tool for studies of diagnostic reliability (QAREL). *BMC Med Res Methodol*. 2013;13:111.
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*. 2018;27:1171–1179.
19. Prinsen CAC, Mokkink LB, Bouter LM, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. *Qual Life Res*. 2018;27:1147–1157.
20. Terwee CB, Mokkink LB, Knol DL, Ostelo RWJG, Bouter LM, De Vet HCW. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Qual Life Res*. 2012;21:651–657.
21. Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluating patient-based outcome measures for use in clinical trials. *Health Technol Assess*. 1998;2:1–74.
22. Williams JI, Naylor CD. How should health status measures be assessed? Cautionary notes on procrustean frameworks. *J Clin Epidemiol*. 1992;45:1347–1351.
23. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials*. 1986;7:177–188.
24. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med*. 2002;21:1539–1558.
25. Fisher RA. 014: On the "probable error" of a coefficient of correlation deduced from a small sample. *Metro*. 1921;1:3–32.
26. Ma J, Liu W, Hunter A, Zhang W. Performing meta-analysis with incomplete statistical information in clinical trials. *BMC Med Res Methodol*. 2008;8:56.
27. Bonett DG, Wright TA. Sample size requirements for estimating Pearson, Kendall and Spearman correlations. *Psychometrika*. 2000;65:23–28.
28. Falla DL, Campbell CD, Fagan AE, Thompson DC, Jull GA. Relationship between cranio-cervical flexion range of motion and pressure change during the cranio-cervical flexion test. *Man Ther*. 2003;8:92–96.
29. Falla D, Jull G, Dall'Alba P, Rainoldi A, Merletti R. An electromyographic analysis of the deep cervical flexor muscles in performance of craniocervical flexion. *Phys Ther*. 2003;83:899–906.
30. Jesus FM, Ferreira PH, Ferreira ML. Ultrasonographic measurement of neck muscle recruitment: a preliminary investigation. *J Man Manip Ther*. 2008;16:89–92.
31. Arumugam A, Mani R, Raja K. Interrater reliability of the craniocervical flexion test in asymptomatic individuals–a cross-sectional study. *J Manip Physiol Ther*. 2011;34:247–253.
32. Dimitriadis Z, Strimpakos N, Kapreli E, Oldham J. Test-retest reliability of craniocervical flexion test for assessing the endurance of deep neck flexors in healthy individuals. *Physiotherapy Issues (article in Greek)*. 2013;9:11–20.
33. Kotwani S, Bid DN, Ghatamaneni D, Alahmari KA, Ramalingam T, Paul SS. Determining the reliability of craniocervical flexion test in asymptomatic individuals. *Hong Kong Physiother J*. 2018;38:33–40.
34. Jull G, Falla D. Does increased superficial neck flexor activity in the craniocervical flexion test reflect reduced deep flexor activity in people with neck pain? *Man Ther*. 2016;25:43–47.
35. Jørgensen R, Ris I, Juhl C, Falla D, Juul-Kristensen B. Responsiveness of clinical tests for people with neck pain. *BMC Musculoskelet Disord*. 2017;18:548.
36. Subbarayalu AV, Ameer MA. Relationships among head posture, pain intensity, disability and deep cervical flexor muscle

Downloaded from academic.oup.com/ptj/article/102/7/pzac054/6585158 by Biblioteca di Scienze, Università degli studi di Firenze user on 27 October 2026

performance in subjects with postural neck pain. *J Taibah Univ Med Sci*. 2017;12:541–547.
37. Chiu TT, Law EY, Chiu TH. Performance of the craniocervical flexion test in subjects with and without chronic neck pain. *J Orthop Sports Phys Ther*. 2005;35:567–571.
38. Johnston V, Jull G, Souvlis T, Jimmieson NL. Neck movement and muscle activity characteristics in female office workers with neck pain. *Spine (Phila Pa 1976)*. 2008;33:555–563.
39. Juul T, Langberg H, Enoch F, Søgaard K. The intra- and inter-rater reliability of five clinical muscle performance tests in patients with and without neck pain. *BMC Musculoskelet Disord*. 2013;14:339.
40. Jun I, Kim K. A comparison of the deep cervical flexor muscle thicknesses in subjects with and without neck pain during craniocervical flexion exercises. *J Phys Ther Sci*. 2013;25:1373–1375.
41. Elsig S, Luomajoki H, Sattelmayer M, Taeymans J, Tal-Akabi A, Hilfiker R. Sensorimotor tests, such as movement control and laterality judgment accuracy, in persons with recurrent neck pain and controls - a case-control study. *Man Ther*. 2014;19:555–561.
42. Steinmetz A, Claus A, Hodges PW, Jull GA. Neck muscle function in violinists/violists with and without neck pain. *Clin Rheumatol*. 2016;35:1045–1051.
43. Martins F, Bento A, Silva AG. Within-session and between-session reliability, construct validity, and comparison between individuals with and without neck pain of four neck muscle tests. *PM R*. 2018;10:183–193.
44. Jull GA. Deep cervical flexor muscle dysfunction in whiplash. *J Musculoskelet Pain*. 2000;8:143–154.
45. Falla DL, Jull GA, Hodges PW. Patients with neck pain demonstrate reduced electromyographic activity of the deep cervical flexor muscles during performance of the craniocervical flexion test. *Spine (Phila Pa 1976)*. 2004;29:2108–2114.
46. Jull G, Kristjansson E, Dall'Alba P. Impairment in the cervical flexors: a comparison of whiplash and insidious onset neck pain patients. *Man Ther*. 2004;9:89–94.
47. Portney LG, Watkins MP. *Foundations of Clinical Research: Applications to Practice*. Upper Saddle River, NJ: Pearson/Prentice Hall; 2015.
48. Mahashabde R, Fernandez R, Sabnis S. Validity and reliability of the aneroid sphygmomanometer using a paediatric size cuff for craniocervical flexion test. *Int J Evid Based Healthc*. 2013;11:285–290.
49. Falla D, Jull G, O'Leary S, Dall'Alba P. Further evaluation of an EMG technique for assessment of the deep cervical flexor muscles. *J Electromyogr Kinesiol*. 2006;16:621–628.
50. Cagnie B, Derese E, Vandamme L, Verstraete K, Cambier D, Danneels L. Validity and reliability of ultrasonography for the longus colli in asymptomatic subjects. *Man Ther*. 2009;14:421–426.
51. Javanshir K, Rezasoltani A, Mohseni-Bandpei MA, Amiri M, Ortega-Santiago R, Fernández-de-Las-Peñas C. Ultrasound assessment of bilateral longus colli muscles in subjects with chronic bilateral neck pain. *Am J Phys Med Rehabil*. 2011;90:293–301.
52. Øverås CK, Myhrvold BL, Røsok G, Magnesen E. Musculoskeletal diagnostic ultrasound imaging for thickness measurement of four principal muscles of the cervical spine -a reliability and agreement study. *Chiropr Man Therap*. 2017;25:2.
53. O'Leary S, Jull G, Kim M, Vicenzino B. Cranio-cervical flexor muscle impairment at maximal, moderate, and low loads is a feature of neck pain. *Man Ther*. 2007;12:34–39.
54. Martin-Gomez C, Sestelo-Diaz R, Carrillo-Sanjuan V, Navarro-Santana MJ, Bardon-Romero J, Plaza-Manzano G. Motor control using cranio-cervical flexion exercises versus other treatments for non-specific chronic neck pain: a systematic review and meta-analysis. *Musculoskelet Sci Pract*. 2019;42:52–59.
55. Gallego Izquierdo T, Pecos-Martin D, Lluch Girbés E, et al. Comparison of cranio-cervical flexion training versus cervical proprioception training in patients with chronic neck pain: a randomized controlled clinical trial. *J Rehabil Med*. 2016;48:48–55.
56. Dunning JR, Cleland JA, Waldrop MA, et al. Upper cervical and upper thoracic thrust manipulation versus nonthrust mobilization in patients with mechanical neck pain: a multicenter randomized clinical trial. *J Orthop Sports Phys Ther*. 2012;42:5–18.
57. Ris I, Søgaard K, Gram B, Agerbo K, Boyle E, Juul-Kristensen B. Does a combination of physical training, specific exercises and pain education improve health-related quality of life in patients with chronic neck pain? A randomised control trial with a 4-month follow up. *Man Ther*. 2016;26:132–140.
58. Lytras D, Sykaras E, Christoulas K, Myrogiannis I, Kellis E. Effects of an integrated neuromuscular inhibition technique program on neck muscle strength and endurance in individuals with chronic mechanical neck pain. *J Bodyw Mov Ther*. 2019;23:643–651.
59. Beaton DE, Bombardier C, Katz JN, et al. Looking for important change/differences in studies of responsiveness. OMERACT MCID working group. Outcome measures in rheumatology. Minimal clinically important difference. *J Rheumatol*. 2001;28:400–405.

Downloaded from academic.oup.com/ptj/article/102/7/pzac054/6585158 by Biblioteca di Scienze, Università degli studi di Firenze user on 27 October 2026