Inter-tester reliability of selected clinical tests for long-lasting temporomandibular disorders

Elisabeth Heggem Julsvoll MSc[1, 2 *], RPT, MT, Nina Køpke Vøllestad PhD[1], Gro Opseth MSc[1, 2], RPT, MT, Hilde Stendal Robinson PhD[1], RPT, MT

[1]Department of Health Sciences, Institute of Health and Society, University of Oslo, Norway

[2]Current work-address: Hans and Olaf Physiotherapy Clinic, Oslo, Norway

The work should be attributed to: Department of Health Sciences, Institute of Health and Society, University of Oslo, P.O. Box 1089, Blindern, N-0318 Oslo, Norway

* Corresponding author: Elisabeth Heggem Julsvoll heggemjulsvoll@gmail.com

Tel.: + 47 95941262

E-mail addresses: Nina Køpke Vøllestad n.k.vollestad@medisin.uio.no

Gro Opseth gro.opseth@gmail.com

Hilde Stendal Robinson h.s.robinson@medisin.uio.no

Biographical note

Elisabeth Heggem Julsvoll

Julsvoll is a Norwegian physical therapist with over 30 years of clinical experience. She completed her postgraduate education in manual therapy in 1991 and Master of Health Science at the University of Oslo, Norway in 2013. Julsvoll received the Norwegian Musculosceletal Research Award 2014 for her research on temporomandibular disorders and have presented her research results both at international and national congresses. Julsvoll works at Hans & Olaf Physical therapy clinic, Oslo. She collaborates with other physical therapists and with dentists, oral surgeons and maxillo-facial radiologists. She is also responsible for specialization courses for physiotherapists, manual therapists and chiropractors in Norway and has given support to the Norwegian guidelines for patients with TMD, published by the Norwegian Directorate for Health.

Nina Køpke Vøllestad

Vøllestad is Professor and Head of the Institute of Health and Society at the University of Oslo. Her scientific interest is primarily related to musculoskeletal disorders; their underlying mechanisms and how they are related to functioning. Over the last years she has become increasingly interested in the underlying mechanisms of recovery through rehabilitation and treatment. Her research spans from physiological mechanisms, through studies of effectiveness and health economics, to studies of methodological properties. Currently, she leads a large research project (FYSIOPRIM) investigating several aspects of musculoskeletal disorders and treatment in primary care, including clinical and health service topics.

Gro Opseth

Opseth is a physical therapist working at Hans & Olaf Physical therapy clinic in Oslo Norway. She has more than 25 years of clinical experience, and completed her postgraduate education in manual therapy in 1998. She finished her Master of Health Science at the University of Oslo, Norway in 2013. Her scientific interest is patients with long-term musculoskeletal disorders, and she has presented her research results both at international and national congresses.

Hilde Stendal Robinson

Robinson is Associate Professor at the Institute of Health and Society at University of Oslo and combines this with a clinical position as a physical therapist in primary health care, specialized in Manual Therapy. She has broad experience as a researcher in interdisciplinary projects and her main interests are on musculoskeletal disorders and women's health. Her research spans from methodological studies, through clinical and mixed method studies to epidemiological studies.

Introduction

Temporomandibular disorder (TMD) is a collective term used to describe a number of conditions involving the temporomandibular joint (TMJ), the masticatory muscles and associated structures [1]. Clinically TMD is characterized by signs and symptoms, such as joint sounds (click and crepitus), dysfunctional movement patterns and pain in the jaw area. The prevalence of TMD in Norway is unknown, but 4-7 percent (75% are women) of the Norwegian population is reported to seek treatment for their TMD during their lifetime [2].

According to Bermejo-Fernoll (2010), the TMJ is considered as one of the most complicated joints in the human body [3]. Each joint is divided into two compartments by a fibrous disc and has three axes of movement. These axes shift during displacement of the condyles and result in infinite movement axes. Activities of the mandible involve movement in both TMJs simultaneously. Hence, a disorder in one joint, such as a displacement of the fibrous disc, will affect the other and vice versa. This makes it difficult to examine the TMJ area.

In clinical practice there is a need for both valid and reliable tests to diagnose correctly and thereby give suitable treatments. Several tests, such as joint-sound tests, functional tests and pain provocation tests are used when examining patients with TMD [4,5]. However, the commonly used tests differ in their reliability. According to John and Zwijnenberg (2001) and Scmitter et al (2005) the reliability is good to excellent for tests used to measure the range of movements and according to John and Zwijnenberg (2001) and de Wijer (1995) the reliability is moderate for tests used to evaluate joint sounds [6-8]. Additionally the reliability of other tests used in clinical practice, for example joint play tests and different pain provocation tests, have not yet been studied. Furthermore, different pain provocation tests and joint mobility tests commonly used by physical therapists when examining TMD patients, are not included in the Diagnostic Criteria for Temporomandibular Disorders (DC/TMD) [5]. Hence, there is a need to explore their reliability [9,10].

In clinical practice, decisions are usually based on the response to several tests and not to one single test. According to Julsvoll et al (2015) a cluster of seven tests (the dental stick test, the isometric test, the joint provocation test, the joint sound test, the deviation test, the laterotrusion test and the joint mobility test) is sensitive to reveal anterior disc displacement without reduction (ADDWOR) [11]. However, the reliability of this ADDWOR-cluster needs to be explored. For therapists to be able to reproduce each other's findings, it is important to use reliable single tests and cluster of tests.

The main objective of the present study was to explore inter-tester reliability for clinical tests used by physical therapists examining patients with long-lasting painful TMD. A secondary objective was to explore the reliability for the identified ADDWOR-cluster.

Methods

Design and participants

Forty participants, 36 women and 4 men, mean (SD) age 44 (13) years with pain in or around one (unilateral) or both TMJ´s (bilateral) were included in this cross-sectional study. The participants reported a history of persistent pain for more than one year and could, according to "The classification of chronic pain for *ICD-11*" [12], be classified as chronic or long-lasting. Rheumatoid arthritis, osteoarthritis, previous injury and surgery in the jaw area, as well as known pathology of the teeth and/or oral cavity, were defined as exclusion criteria. The participants, mainly from the Oslo area, were recruited consecutively in 2012 by different clinicians, who were treating them for their TMD problems. The number of people invited to the study that declined participation was not recorded.

Procedure

Each participant was examined by two assessors on the same day with approximately 30 minutes break between examinations. The assessors were physical therapists with postgraduate education in manual therapy; one with 22 years of experience treating mainly TMD patients, while the other with 15 years of experience with mixed musculoskeletal practice including TMD. The sequence of the assessor's examinations was randomly allocated. The tests were performed in the same order on each side separately by both assessors. They registered the participants' responses on each test, and since the purpose was to examine reliability, no diagnoses were set during examinations. Furthermore, the assessors were blinded for all information about the participants, including the results of the other examination. The clinical tests were standardized and consensus on the procedures was made before the study started. The study also included MRI scans of the participants TMJ area for validity purposes to confirm if ADDWOR was present or not [11]. The MRIs (both coronal and sagittal planes) were all taken at the same radiology center. The participants were examined clinically within 24 hours before or after the MRI. The assessors were blinded for the MRI results until all participants were examined.

The study was carried out according to the Helsinki Declaration [13]. The Norwegian Social Science Data Services and Local Ethical committee at the Department of Health Sciences, University of Oslo, approved the study. Written informed consent was obtained from all participants before the examinations.

Descriptive characteristic of the participants

On the examination day, the participants filled out a questionnaire including demographic data, pain location and pain intensity. The Patient-Specific Functional Scale (PSFS) was used to assess activity limitations [14]. Pain at present and worst pain ever in the TMJ area were measured by Visual Analog Scales (VAS, 0-100, 100 being the worst).

Clinical tests

Three types of clinical tests were included; joint-sound, functional and pain provocation tests. *Joint-sounds* ("click" and "crepitus") were tested on mouth-opening, -closing, protrusion and bilateral laterotrusion using both a stethoscope and digital palpation (Figure 1). The following *functional tests* were used: range of motion (ROM) (mouth-opening, protrusion and bilateral laterotrusion), movement quality (mouth-opening), joint-mobility and anterior glide (end-feel) (Figure 2-5). Different *pain provocation tests* were included; the dental stick test, isometric test (hold on mouth-opening, protrusion and bilateral laterotrusion), joint provocation test, distraction test and pain-provocation during anterior glide (Figure 6-10). Also, whenever movements (mouth-opening, protrusion and bilateral laterotrusion) provoked pain, were recorded (Figure 11.1-11.3). (See appendix for description of the clinical tests).

Data analysis

Statistical analyses were performed using SPSS statistical package version 22.0 (IBM Corp., New York, NY). Demographic data, mouth-opening and pain characteristics are described by mean with standard deviation (SD) and median with range, while activity limitations and additional complaints are described by frequencies and percentages.

The responses on each clinical test were registered and analyzed separately for the right and the left side except for mouth-opening since this movement involves both joints. Joint-sounds were coded “Yes” for presence and “No” for absence of sound. It was distinguished between “click” and “crepitus” and each was coded separately. Based on definitions by Bermejo-Fenoll (2010) and Hylander (2006) the cut-off values for reduced jaw movements were defined as follows; mouth-opening < 40 mm, protrusion ≤ 7 mm and laterotrusion to either side ≤ 9 mm [3,15]. The total ability to open the mouth was of interest in this study, hence mouth-opening on the two sides separately was not measured, solely one measurement between the front teeth. Reduced and normal mobility were coded “Yes” and “No”, respectively, for analyzing purposes. To examine the movement quality of the jaw during mouth-opening it was registered if the mandible moved straight (“Yes”/”No”), if it deviated with a correction (“Yes”/”No”) or without correction (“Yes”/”No”) to either side.

The movement end feel was tested coding hard and firm end-feel as “Yes” and soft and empty end-feel as “No”. Whenever tests provoked pain, it was coded “Yes” and registered on a numeric pain rating scale (NPRS, 0-10, 10 being the worst pain ever). For the analyses, no pain or unfamiliar (discordant) pain was coded “No” (0) and familiar (concordant) pain (1-10) was coded “Yes”.

Inter-tester reliability for the tests with categorical outcome variables was calculated by percentage agreement and the kappa agreement coefficient (*k*) with 95% confidence interval (CI) [16]. The strength of agreements of the results was interpreted as follows: $k \leq 0.40$: poor, $0.40 < k < 0.75$: fair to good, $0.75 \leq k < 1.00$: excellent [17,18]. The result from each assessor on each test is given by frequencies (positive/negative).

No order differences were found between assessor 1 and 2 on tests with continuous outcomes (A1 and A2). The relative inter-tester reliability was assessed by the intra-class-correlation-coefficient ($ICC_{3,1}$), presented with 95% CI [16,19,20]. The results were categorized as follows: ICC ≥ 0.75: excellent reliability, 0.40 < ICC < 0.75: good reliability, and ICC ≤ 0.40: poor reliability [16,18]. Absolute reliability was calculated by the smallest detectable change (SDC) by using standard error of measurement;

$SEM = SD_{difference} / \sqrt{2}$ and $SDC = 1.96 \times \sqrt{2} \times SEM$.[16,20]

The reliability of the previously identified ADDWOR-cluster was calculated by the sum score for A1 and A2 on each side separately. Maximum score was 7 (seven positive tests). Agreement on the case definition was based on at least four/five positive tests.

Results

Forty participants with 65 symptomatic (33 right/ 32 left) and 15 asymptomatic joints were included (Table 1). Ninety percent were women. Median pain intensity (range) on the examination day was 38 (4-90), while the median worst pain ever in the TMJ area was 85 (33-99). Mean (SD) mouth-opening on the examination day was 36 (5) mm. The main activity limitations reported in PSFS were mouth-opening (90%), chewing hard food (88%) and yawning (85%) (Table1).

[Table 1 near here]

The dental-stick test and the joint-sound tests (click and crepitus on mouth-opening and mouth-closing and crepitus on laterotrusion) had the highest percentage agreement (95-100%) with excellent kappa values (0.80 -1.0) (Table 2).

[Table 2 near here]

For the tests evaluating quality of mouth-opening the kappa values (95% CI) for opening straight and opening with deviation to either left or right side without correction ranged from 0.81(0.53-1.0) to 0.94 (0.79-1.0). Right deviation with correction showed excellent reliability (*k* = 0.88) while deviation to the left with correction showed poor reliability (*k* = 0.38) (Table 3).

[Table 3 near here]

The relative reliability, $ICC_{3,1}$ (95% CI), for the range of mouth-opening was 0.97 (0.95-0.98) while the absolute reliability, SDC, was 4 mm (Table 3). The relative reliability for the protrusion and laterotrusion tests (ROM) varied from 0.90 to 0.94 with SDCs from 2 to 3 mm.

MRI revealed ADDWOR in 22 joints, anterior disc displacement without reduction (ADDwR) in 24 joints and negative findings in 34 joints. When using this categorization on the participants and calculating kappa for the same tests used for each group, there were no significant differences in the inter-tester reliability of the tests, but there were somewhat broader confidence intervals (data not shown).

Only small differences in kappa were found for the cluster with case definition of four or five positive tests; *k* = 0.72 (0.42-0.94) and 0.76 (0.40-1.0), respectively (Table 4). When four positive tests out of seven were required, agreement reached 36 and 35 cases for the right and left side, respectively, and when five of seven tests were positive, agreement was found on 37 and 36 cases, respectively.

[Table 4 near here]

Discussion

The following single tests had best reliability in the present study: the joint-sound tests (click and crepitus on mouth-opening and closing and crepitus on laterotrusion), range of motion test (mouth-opening) and the dental stick test. Furthermore, the reliability for the ADDWOR-cluster was good with case definition of both four and five positive tests.

Joint sound

The reliability of joint-sound tests (click and crepitus) was excellent on vertical movements (opening and closing of the mouth) and varied from fair to excellent on horizontal movements (protrusion and laterotrusion to either side). These results are better than the results from John and Zwijnenburg (2001) [6]. The authors found only moderate reliability for joint-sound tests. However, the difference in methodology might have influenced the results. They included participants without TMD and tested joint-sounds only during vertical movement. Dworkin and co-workers (1990) reported better kappa values for joint-sounds on mouth-opening than on protrusion and laterotrusion, which support the results of the present study [21]. Since joint-sounds can either be weak and hidden or easily heard by both the patient and the examiner, the authors recommended using a stethoscope [21]. The use of both palpation and a stethoscope in the present study might have improved the results, since the use of stethoscope increases the ability to hear sounds. Even though the agreement is fairly high for the joint-sound tests, some of the kappa values are low with broad confidence intervals, indicating some uncertainty in the results. According to Altman (1991) kappa is influenced by a low number of participants with positive response on the test [17]. This might be the case for the test of click-sound on the right side on laterotrusion to the left when only six (A1) and five

(A2) participants, respectively, had a positive test. The results should thus be interpreted with caution.

Functional tests

Measurement of TMJ ROM is dependent on a reference tooth from which to measure. With vertical movement (Figure 2.1), the reference tooth is determined in advance and easier to identify than for the horizontal movements where the assessors have to select the reference tooth themselves (Figure 2.2 and 2.3). The selection of the reference tooth on horizontal movement might then be a source of error. This can influence the reliability especially if different teeth are chosen and the results for horizontal movements could then be biased. Moreover, measurement of vertical movements can be easier to learn for less experienced therapists because of the preselected reference tooth, and may influence the reliability positively. In the present study excellent reliability was found for both tests (measuring vertical and horizontal movements), however best agreement was found on vertical movement. The absolute reliability (SDC) was 4 mm for mouth-opening, 3 and 4 mm for laterotrusion to the right and left side, respectively, and 2 mm for protrusion on both left and right side. These findings of SDC imply that larger differences after treatment are needed to ascertain true improvements rather than measurements errors.

As mentioned above; activities of the mandible involve movements in both TMJs simultaneously. This is important when evaluating the quality of mouth-opening (Figure 3a-3b). Hence, it is possible that the difference in kappa on right (0.88) and left (0.38) deviation with correction could be caused by instability on one side and thereby an ADDwR or a transient locking [15]. Another plausible explanation is that the low number of participants with positive response on the test have influenced the kappa scores [17]. Because of the low number of participants testing positive for jaw deviation with correction (quality of motion), conclusion on the reliability on this test is uncertain. Future studies could explore the reliability of this test further by including a larger sample size with jaw deviating with correction during mouth-opening.

The joint-mobility test agreement (Figure 4) was fair. Tests of joint-mobility are considered an important tool in physical therapy and manual therapy, but it seems difficult to obtain good inter-examiner agreement, most probably because the tests are based on palpation. The results of the tests for TMJ-mobility in the present study are slightly better than what have been reported on palpation tests for other joints/structures [22-25]. Robinson and co-workers (2009) found poor to moderate inter-tester reliability on palpation tests for identifying the spinous processes of C7 and L5. The same group also found poor reliability for the joint-play test for the sacroiliac joint [23]. Generally, joint-mobility tests involving palpation seem to have lower reliability than pain provocation tests, and it has been hypothesized that this might depend on the experience and skills of the assessors [25,26]. Furthermore, the quality of joint play is very difficult both to define and to quantify. Despite the difference between the assessors in years of experience with TMD patients, the inter-tester reliability of the test for joint mobility and anterior glide was acceptable in the present study.

Pain-provocation tests

The best reliability among the pain-provocation tests used in the study was found for the dental stick test showing excellent agreement [18]. The DC/TMD does not include the pain-provocation tests used in the present study. However, different pain-provocation tests such as bite-tests, static and dynamic tests, “end-feel” tests and joint-play-test such as compression and distraction are mentioned as adjunctive tests that one can include in an examination. When performing the pain provocation tests in the present study, the aim was to reproduce the patient`s actual pain. However, pain provoked during the first assessment could have resulted in increased pain on the second assessment. To reduce this effect the participants were allowed a 30-minute break between the assessments.

Categorization of the participants

There were no significant differences in the reliability of the tests calculated for the different groups of participants based on MRI results (ADDWOR, ADDwR and negative MRI). Hence, the tests seem quite reliable and, independent of TMD diagnosis, the clinicians agreed on test responses. However, since the CIs were broader, the estimates have some uncertainties. Additionally, there were few participants in each group and this might have influenced kappa. More research is needed to establish validity for the tests for each of the TMD diagnoses.

Cluster of tests

The reliability was excellent for the previously defined ADDWOR-cluster both when four and five of seven tests were positive. When case definition was set to six of seven tests, the reliability was fair (data not shown). The excellent kappa results, when four and five of seven tests were positive, should be linked to both the case definition and the number of cases; there were agreement in a high number of cases in the present study (between 88 and 93%). One weakness of clusters could be that the assessors did not agree on exactly the same tests. However, since the reliability of the ADDWOR-cluster was excellent, its use in clinical practice for diagnosing ADDWOR could be recommended.

Methodology

To minimize systematic faults, randomizing procedures were used and the tests were performed in equal order by both assessors. All measurements were taken on the second attempt, to ensure that the conditions were as similar as possible. Furthermore, to reduce the learning effect, each test was repeated only twice in both examinations with one exception, the joint-sound tests were repeated three times. By allowing a 30-minute break between assessments, the conditions were assumed to be stable.

In the present study, patients with long-lasting painful TMD were studied. Hence, the results cannot be generalized to patients with acute TMD. Moreover, since 90% of the participants in the study displayed limited mouth-opening, the results cannot necessarily be generalized to patients having TMD with normal or increased mouth-opening. The fact that the study included people with both uni- and bilateral pain and both symptomatic and asymptomatic joints (65/15) might also have influenced the reliability on some tests.

The skills of the assessors might have influenced the tests results. In the present study the assessors were experienced manual therapists working in the same setting. We attempted to optimize agreement by practicing the skills of required procedures. Thus, less agreement can be expected in different clinical settings or between various medical specialists.

The sample size was relatively small for a methodological study. However, according to Cosmins checklist, 40 subjects with 65 symptomatic joints can represent a patient group adequately [27].

Conclusion

The study found good to excellent reliability among experienced therapists for selected clinical tests and the ADDWOR-cluster (four or five tests positive out of seven) for patients with long-lasting painful TMD. The best reliability scores among single tests were found for the joint sound tests, range of mouth-opening and the dental stick test. More than 4 mm difference in mouth-opening after treatment is needed to identify a real improvement. The tests require no advanced equipment, are easy to perform and suitable for use in clinical settings. Future studies should include a larger sample and subjects with acute TMD. Furthermore, validity of the tests for diagnosing various TMDs conditions needs to be established.

References

1. Okeson JP, de Kanter RJ. Temporomandibular disorders in the medical practice. J Fam Pract. 1996;43(4):347-56.
2. Høvik H, Ninkov P. National academic guidelines for examination and treatment of TMD. The Norwegian Directorate for Health. Nasjonal faglig retningslinje for utredning og behandling av TMD (tyggemuskulatur- og kjeveleddsplager). Helsedir. Oslo 2016.
3. Bermejo-Fenoll A. Anatomy of the Temporomandibular Joint and Masticatory Muscles. Current concepts on temporomandibular disorders. Anatomy of the Temporomandibular Joint and masticatory Muscles. London: Quintessence; 2010. p. 3-21.
4. Shaffer SM, Brismee JM, Sizer PS, Courtney CA. Temporomandibular disorders. Part 1: anatomy and examination/diagnosis. J Man Manip Ther 2014;22(1):2-12.
5. Schiffman E, Ohrbach R, Truelove E, Look J, Anderson G, Goulet JP, et al. Diagnostic Criteria for Temporomandibular Disorders (DC/TMD) for Clinical and Research Applications: recommendations of the International RDC/TMD Consortium Network* and Orofacial Pain Special Interest Groupdagger. J Orofac Pain 2014;28(1):6-27.
6. John MT, Zwijnenburg AJ. Interobserver variability in assessment of signs of TMD. Int J Prosthodont. 2001;14(3):265-70.
7. Schmitter M, Ohlmann B, John MT, Hirsch C, Rammelsberg P. Research diagnostic criteria for temporomandibular disorders: a calibration and reliability study. Cranio. 2005;23(3):212-8.
8. de Wijer A, Lobbezoo-Scholte AM, Steenks MH, Bosman F. Reliability of clinical findings in temporomandibular disorders. J Orofac Pain 1995;9(2):181-91.

9. Lobbezoo-Scholte AM, de Wijer A, Steenks MH, Bosman F. Interexaminer reliability of six orthopaedic tests in diagnostic subgroups of craniomandibular disorders. J Oral Rehabil. 1994;21(3):273-85.
10. Schmitter M, Kress B, Leckel M, Henschel V, Ohlmann B, Rammelsberg P. Validity of temporomandibular disorder examination procedures for assessment of temporomandibular joint status. Am J Orthod Dentofacial Orthop: official publication of the American Association of Orthodontists, its constituent societies, and the American Board of Orthodontics. 2008;133(6):796-803.
11. Julsvoll EH, Vollestad NK, Robinson HS. Validation of clinical tests for patients with long-lasting painful temporomandibular disorders with anterior disc displacement without reduction. Man Ther. 2016;21:109-119.
12. Treede RD, Rief W, Barke A, Aziz Q, Bennett MI, Benoliel R, et al. A classification of chronic pain for ICD-11. Pain. 2015;156(6):1003-7.
13. Declaration of Helsinki (1964) - Ethical Principles for Medical Research Involving Human Subjects, World Medical Association Declaration of Helsinki, General Assembly, Helsinki, Finland.
14. Stratford P GC, Westaeay M, Binkley J. Assesing Disability and Change on Individual patients: A Report of Patient Specific Measure. Phys Ther. 1995;47:258 - 63.
15. Hylander WL. Functional Anatomy and Biomechanics of the Masticatory Apparatus. TMDs - An Evidence-Based Approach to Diagnosis and Treatment: Quintessence; 2006. p. 3-34.
16. de Vet HCW. Measurement in medicine: a practical guide. Cambridge: Cambridge University Press; 2011.
17. Altman DG. Practical statistics for medical research. London: Chapman and Hall; 1991.

18. Fleiss JL. The design and analysis of clinical experiments. New York: Wiley; 1999.
19. Rousson V, Gasser T, Seifert B. Assessing intrarater, interrater and test-retest reliability of continuous measurements. Statistics in medicine. 2002;21(22):3431-46.
20. Overend T, Anderson C, Sawant A, Perryman B, Locking-Cusolito H. Relative and absolute reliability of physical function measures in people with end-stage renal disease. Physiother Can. 2010;62(2):122-8.
21. Dworkin SF, LeResche L, DeRouen T, Von Korff M. Assessing clinical signs of temporomandibular disorders: reliability of clinical examiners. J Prosthet Dent. 1990;63(5):574-9.
22. Billis EV, Foster NE, Wright CC. Reproducibility and repeatability: errors of three groups of physiotherapists in locating spinal levels by palpation. Man Ther. 2003;8(4):223-32.
23. Seffinger MA, Najm WI, Mishra SI, Adams A, Dickerson VM, Murphy LS, et al. Reliability of spinal palpation for diagnosis of back and neck pain: a systematic review of the literature. Spine. 2004;29(19):E413-25.
24. Robinson HS, Brox JI, Robinson R, Bjelland E, Solem S, Telje T. The reliability of selected motion- and pain provocation tests for the sacroiliac joint. Man Ther. 2007;12(1):72-9.
25. Robinson R, Robinson HS, Bjorke G, Kvale A. Reliability and validity of a palpation technique for identifying the spinous processes of C7 and L5. Man Ther. 2009;14(4):409-14.
26. Harlick JC, Milosavljevic S, Milburn PD. Palpation identification of spinous processes in the lumbar spine. Man Ther. 2007;12(1):56-62.
27. Carter RE, Lubinsky J, Domholdt E. Rehabilitation research. St. Louis, Miss.: Elsevier Saunders; 2011. VIII, 503 s. : ill. p.

28. Kaltenborn FM, Evjenth O. Manual mobilization of the joints. Oslo: Norli; 2011. XVI, 319 s. : ill. p.

Table 1

Descriptive characteristics of the participants.

n = 40	Frequency (%)	Mean (SD)	Median	Min - Max
Women	36 (90)			
Men	4 (10)			
Age		44 (13)		18 – 66
Years of education		15 (3)		9 – 25
ROM, mm				
Mouth-opening		36 (5)		21 – 50
Protrusion; right side		6 (2)		2 – 11
Protrusion; left side		6 (2)		2 – 10
Laterotrusion right		9 (3)		3 – 14
Laterotrusion left		10 (4)		3 – 20
Duration of pain, years		10 (2)		1 – 40
Pain at present, VAS			38	4 – 90
Worst pain, VAS			85	33 – 99
Activity limitation (PSFS)				
Mouth-opening	36 (90)			
Chewing hard food	35 (88)			
Yawning	34 (85)			
Additional complaints				
Neck pain	33 (83)			
Headache	31 (78)			
Vertigo	0 (50)			
Tinnitus	16 (40)			
Bruxism	27 (68)			

n; number of persons, SD; standard deviation, Min; minimum, Max; maximum, mm; millimeter, ROM; range of motion, VAS; Visual Analog Scale. Pain at present and worst pain ever were measured by VAS with the questions; On a 100 millimeter scale where 0 represent absence of pain and 100 represent the worst pain ever: 1. How will you grade your pain in the jaw today? 2. How will you grade the worst pain you ever have had in the jaw?

PSFS; Patients rate their ability to complete an activity on an 11-point scale; 0 represents “unable to perform the activity” and 10 represents “able to perform the activity at prior level”.

Table 2

Inter-tester reliability for clinical tests (right and left side) for persons with temporomandibular disorders (n = 40) presented with rater frequencies, kappa and percentage agreement.

	Right side				Left side			
Clinical tests	A1	A2	Kappa (95 % CI)	% Agreement	A1	A2	Kappa (95 % CI)	% Agreement
	pos/neg	pos/neg			pos/neg	pos/neg		
Joint sound tests								
Opening the mouth								
Click	14/26	13/27	0.94 (0.82-1.0)	98	17/23	16/24	0.95 (0.83-1.0)	98
Crepitus	15/25	14/26	0.95 (0.82-1.0)	98	13/27	14/26	0.83 (0.62-1.0)	93
Closing the mouth								
Click	11/29	9/31	0.87 (0.64-1.0)	95	10/30	9/31	0.93 (0.76-1.0)	98
Crepitus	12/28	10/30	0.88 (0.66-1.0)	95	10/30	9/31	0.93 (0.77-1.0)	98
Protrusion								
Click	8/32	6/34	0.66 (0.26-0.93)	90	8/32	6/34	0.66 (0.25-0.93)	90
Crepitus	12/28	9/31	0.81 (0.58-1.0)	93	9/31	8/32	0.78 (0.47-1.0)	93
Laterotrusion to the right								
Click	10/30	10/30	0.87 (0.63-1.0)	95	9/31	6/34	0.59 (0.23-0.90)	88
Crepitus	7/33	8/32	0.92 (0.72-1.0)	98	6/34	6/34	1.0 (1.0-1.0)	100
Laterotrusion to the left								
Click	6/34	5/35	0.47 (0.04-0.90)	88	6/34	4/36	0.77 (0.38-1.0)	95
Crepitus	6/34	6/34	0.80 (0.47-1.0)	95	5/35	4/36	0.88 (0.48-1.0)	98
Functional tests								
Joint mobility	19/21	10/30	0.54 (0.31-0.78)	78	18/22	10/30	0.55 (0.31-0.80)	75
Anterior glide (end-feel)	16/24	16/24	0.79 (0.55-0.95)	90	17/23	16/24	0.95 (0.83-1.0)	98
Pain provocation tests								
Dental stick test	13/27	11/29	0.88 (0.69-1.0)	95	15/25	15/25	1.0 (1.0 - 1.0)	100
Isometric test	15/25	11/29	0.78 (0.55-0.95)	90	14/26	17/23	0.75 (0.50-0.92)	88
Joint provocation test	30/10	28/12	0.81 (0.55-1.0)	93	29/11	27/13	0.88 (0.68-1.0)	95
Distraction test								
Pain relief	10/30	10/30	0.73 (0.42-0.94)	90	10/30	8/32	0.86 (0.59-1.0)	95
Pain provocation	9/31	15/25	0.54 (0.23-0.78)	80	8/32	14/26	0.63 (0.36-0.87)	85
Anterior glide (pain)	23/17	23/17	0.80 (0.60-0.95)	90	19/21	21/19	0.80 (0.58-0.95)	88
Movement-pain on								
Mouth- opening	24/16	24/16	0.79 (0.55-0.95)	90	27/13	25/15	0.89 (0.70-1.0)	95
Protrusion	24/16	21/19	0.44 (0.16-0.70)	73	28/12	24/16	0.57 (0.30-0.81)	80
Right laterotrusion	17/23	18/22	0.75 (0.52-0.95)	88	23/17	21/19	0.60 (0.33-0.80)	80
Left laterotrusion	25/15	22/18	0.64 (0.38-0.87)	83	23/17	17/23	0.71 (0.49-0.90)	85

Kappa and % agreement assessed for categorical variables, n; number of persons, A; Assessor, CI; Confidence interval.

Table 3

Inter-tester reliability for the functional tests **movement quality** (how the jaw moves during mouth-opening) presented with rater frequencies, kappa and percentage agreement. **Range of motion** (ROM; protrusion and laterotrusion (right and left side) and mouth-opening) presented with intraclass correlation coefficient and smallest detectable change. n = 40 persons with temporomandibular disorders.

Movement quality	A1	A2		
	pos/neg	pos/neg	Kappa (95 % CI)	% Agreement
Straight	12/28	10/30	0.88 (0.67 -1.0)	95
Right deviation without correction	12/28	11/29	0.94 (0.79 -1.0)	98
Left deviation without correction	11/29	10/30	0.81 (0.53 -1.0)	93
Right deviation with correction	4/36	5/35	0.88 (0.48-1.0)	98
Left deviation with correction	1/39	4/36	0.38 (0 -1.0)	93

	Right side		Left side			
ROM, mm	ICC $_{(3.1)}$ (95 % CI)	SDC (mm)	ICC $_{(3.1)}$ (95 % CI)	SDC (mm)	ICC $_{(3.1)}$ (95 % CI)	SDC (mm)
Protrusion	0.90 (0.81-0.95)	2.4	0.92 (0.86-0.96)	2.0		
Laterotrusion	0.94 (0.89-0.97)	2.5	0.93 (0.88-0.97)	3.4		
Mouth-opening					0.97 (0.95 - 0.98)	3.8

Kappa and % agreement assessed for the categorical variables of the different types of mouth-opening, n; number of persons, A; Assessor, CI; Confidence interval**,** Intra-class correlation coefficient (ICC) and smallest detectable change (SDC) assessed for continuous variables.

Table 4

Cluster reliability results (n=40)

	Case definition	% Agreement	Kappa	95% CI
Cluster right side	4 of 7	90	0.76	0.51-0.95
Cluster left side	4 of 7	88	0.74	0.49-0.94
Cluster right side	5 of 7	93	0.76	0.40-1.0
Cluster left side	5 of 7	90	0.72	0.42-0.94

Cluster; (the dental stick test, the isometric test, the joint provocation test, the joint sound test, the deviation test, the laterotrusion test and the joint mobility test, n; number of persons, CI; Confidence interval.