Acta Odontologica Scandinavica, 2006; 64: 146–152

Taylor & Francis
Taylor & Francis Group

ORIGINAL ARTICLE

Recalibration improves inter-examiner reliability of TMD examination

THOMAS LIST[1], MIKE T. JOHN[2,3], SAMUEL F. DWORKIN[3] & PETER SVENSSON[4,5]

[1]*Department of Stomatognathic Physiology, Malmö University, Malmö, Sweden,* [2]*Department of Prosthodontics and Material Science, University of Leipzig, Leipzig, Germany,* [3]*Department of Oral Medicine, University of Washington, Seattle, Wash., USA,* [4]*Department of Clinical Oral Physiology, School of Dentistry, University of Aarhus, Aarhus, Denmark,* [5]*Department of Oral and Maxillofacial Surgery, Aarhus University Hospital, Aarhus, Denmark*

Abstract

Objective. The purpose of this study was to assess whether recalibration of examiners would improve the reliability of gathering clinical findings and related diagnoses of temporomandibular disorders (TMD) in accordance with the Research Diagnostic Criteria for TMD (RDC/TMD). ***Material and Methods***. Two clinicians independently examined a total of 48 symptomatic and asymptomatic subjects according to the RDC/TMD on two occasions: examination 1 (E1). Aarhus, Denmark ($n=24$; 18 female, ages 18–59 years); examination 2 (E2). Malmö, Sweden ($n=24$; 18 female, ages 18–86 years). The clinicians were calibrated in the use of the RDC/TMD Axis-I examination on the day before E1. Six months later, they were recalibrated on the day before E2. Intra-class correlation coefficients (ICCs) were used to examine the inter-examiner reliability of the two clinicians on the two occasions (E1, E2). ***Results***. The intra-class correlation coefficients of vertical range of jaw motion differed little between E1 and E2. At E2, all other examination components consistently improved in reliability relative to E1. Similar improvements were seen for the frequently occurring RDC/TMD clinical diagnoses: Ia. Myofascial pain [ICC = 0.83 (E1) and 1.00 (E2)], IIa. Disk displacement with reduction [ICC = 0.26 (E1) and 0.64 (E2)], and IIIa. Arthralgia [ICC = 0.16 (E1) and 0.73 (E2)]. ***Conclusion***. Recalibration considerably improved inter-examiner reliability for assessing RDC/TMD clinical variables and diagnoses, which are critically dependent on reliable assessment of clinical signs; improvement was most marked when initial inter-examiner reliability was low. Final inter-examiner reliabilities after recalibration were all associated with acceptable to excellent levels.

Key Words: *Calibration, classification, reliability, temporomandibular disorders*

Introduction

The consequences of measurement error are serious in research settings, where outcome relationships can be biased or attenuated due to unreliable data-gathering and in clinical practice where poor reliability of clinical data-gathering is a clear risk factor for misdiagnosis and faulty clinical decision-making with regard to treatment. To improve quality of dental research and, ultimately, of the public's dental health, the World Health Organization has recommended that the reliability of clinical measures be part of oral health reports [1].

In the field of temporomandibular disorders (TMD), indeed for most aspects of clinical measurement in medicine and dentistry, measurement reliability for clinical signs is problematic in the absence of standardized calibration of clinical examiners, so what factors determine clinical reliability [2–14]? Variability of clinical examination reliability is found repeatedly, and the reasons include lack of clear specification of examination procedures; number of patients examined, distribution of clinical symptoms among the group; experience of the examiner, the fluctuating nature of the observed symptoms and, most relevant to the present study, failure to calibrate clinical examiners to a standardized set of examination procedures and criteria for identifying clinical signs. Guidelines were therefore proposed by Dworkin et al. [6] for how reliability studies should be conducted. The most important factors cited for improving reliability include definitions of variables and specifications for examination procedures to enhance a priori the possibility to achieve sufficient reliability. The operational criteria are already available for the most widely used diagnostic and

Correspondence: Thomas List, DDS, Odont dr., Orofacial Pain Unit, Department of Stomatognathic Physiology, Malmö University, SE-214 21 Malmö, Sweden. Tel: + 46 40 6658424. Fax: +46 40 6658420. E-mail: thomas.list@od.mah.se

(Received 8 June 2005; accepted 16 November 2005)

ISSN 0001-6357 print/ISSN 1502-3850 online © 2006 Taylor & Francis
DOI: 10.1080/00016350500483137

classification system for TMD, i.e. the RDC/TMD [15]. Another factor that potentially could be used to improve reliability is training and recalibration of examiners. Ideally, examiners should undergo iterations of training and calibration until acceptable levels of reliability are attained.

There is evidence that training and recalibration do improve reliability. Duinkerke et al. [3] reported satisfactory temporomandibular joint (TMJ) and muscle palpation reliability for both experienced and non-experienced examiners, indicating that training is as important as clinical experience. Dahlström et al. [5] found that even though training tended to improve reliability, several clinical signs had poor reliability. In a series of reliability studies, Dworkin et al. [5] reported that trained examiners were more consistent than untrained examiners, and that retraining improved reliability. Although the insight into repeated reliability assessment was interesting, these studies did not use statistical tests to distinguish observed improvements in reliability from sampling variability. Hence, chance as a cause of reliability improvement cannot be excluded.

The aim of this study was therefore to demonstrate the influence of recalibration on the reliability of clinical TMD signs measured by the RDC/TMD and the inevitable increase in reliability of clinical diagnoses based on clinical measurements obtained by improving the reliability of clinical examiners. The research hypothesis for the studies reported here is that recalibration has no effect on TMD reliability when measured with a set of reliability coefficients for clinical TMD signs that were evaluated using the RDC/TMD.

Material and methods

Subjects

On two occasions, two clinicians independently examined a total of 48 symptomatic and asymptomatic subjects according to the RDC/TMD. Examination 1 (E1): Aarhus, Denmark, where after an initial calibration session the two clinicians examined 24 subjects (6 male, 18 female, mean age 35.6 $\pm$ 10.1, range 18–59 years); Examination 2 (E2) was conducted in Malmö, Sweden, where, again after a recalibration session, the two clinicians examined 24 subjects (6 male, 18 female, mean age 42.8 $\pm$ 18.7, range 18–86 years). The TMD patients had all been referred to respective university clinics because of chronic orofacial pain. The participants in these reliability trials were selected from the catchment area of each university so that a broad variety of symptoms would be represented in the studies. Four asymptomatic, healthy volunteers were also selected from each of the university communities. All participants were reimbursed USD 30 for expenses. Subject recruitment and the study protocol were approved by the local Ethics Committee at Lund University, Lund, Sweden and informed consent was obtained from all participants.

Clinical variables

Clinical TMD signs were selected to represent core concepts of the RDC/TMD. Specifications for the examination and algorithms for the diagnostic criteria have been presented by Dworkin & Le Resche [15]. The following variables were selected:

Mandibular range of motion variables (unassisted opening without pain, maximum unassisted opening, maximum assisted opening, lateral and protrusive jaw excursions).
TMJ clicking sounds (joint clicking during opening, during closing, during contralateral movement, during ipsilateral movement, protrusion, reciprocal clicking eliminated in opening from protruded position).
Masticatory muscle and TMJ palpation variables (Temporalis posterior, middle, anterior; masseter origin, body, insertion; retromandibular region; submandibular region; lateral pterygoid area; tendon of temporalis; lateral part, posterior part of TMJ). Palpation tenderness was graded on a 4-point scale according to the RDC/TMD: 0 = no pain, 1 = slight pain, 2 = moderate pain, and 3 = severe pain but used as pain/no pain in the analysis.
TMD composite measures (8 RDC/TMD diagnoses, 2 summary palpation scores – the sum of 20 yes/no muscle palpation pain sites and the sum of 4 yes/no TMJ palpation pain sites).

For variables measured on both face sites, data of both sites were combined into one variable for analytical purposes and treated as independent observations, e.g. the variable *click during opening* contained the 24 data of the left joint and the 24 data of the right joint.

Design

The examiners were two experienced clinical TMD specialists. An incomplete Latin square design was used to ensure that each subject was seen by each examiner in a randomized sequence so that the influence of the examination order on the responses would be balanced. It was ensured that the number of subjects seen first by one examiner was equal to the number for the other examiner. The examiners had not seen any of the patients before and were blind to the results of the previous examination. Each examination lasted about 10 min.

Calibration

Information was sent to the participants prior to the examination so that they would be familiar with the specific RDC/TMD examination methods and protocols. An experienced TMD dentist (SFD) held an 8-h calibration session on the day preceding the inter-examiner reliability study at both sites. The session comprised the following steps: Day One (E1), a video with the clinical examination was observed, and the clinical measures were discussed. A scale was used to ensure that the examiners applied the same pressure during digital palpation of the muscle sites and the joints (0.45 kg and 0.9 kg, respectively). The examiners investigated each other and compared their findings at the same site with those of the instructor (the "gold standard"). Thereafter, patients were examined blind, and the results of the examiners being calibrated were compared and discussed with the results of the gold standard examiner. If the results of those being calibrated differed from those of the gold standard examiner, further calibration took place. On Day Two (E2), 6 months later, recalibration comprised a reliability assessment. The two examiners who were recalibrated were both experienced TMD specialists (TL, PS).

Statistical methods

Reliability was determined by computing intra-class correlation coefficients (ICCs) for both continuous and dichotomous measurement scales [14]. ICCs based on random effects analysis of variance [ANOVA], which treats subjects and raters as random factors, were calculated [16,17]. If the prevalence of a variable was low ($<5\%$), an ICC was not calculated. If the ICC was <0.4, reliability was considered poor; 0.4–0.75 was considered fair to good; and >0.75 was considered excellent according to the guidelines [17].

Differences between ICCs on the two occasions were calculated as ICC_{Malmoe}–ICC_{Aarhus} for each variable. The mean and the 95% confidence interval (CI) for differences were calculated for single clinical TMD variables of the following groups: the range of motion variables ($n=5$), joint clicking ($n=6$), TMJ and muscle palpation sites ($n=11$, M. temporalis ICC was excluded because of low variable prevalence). In a second analysis, the mean and the 95% confidence intervals for differences in ICCs were calculated for TMD composite measures: Four RDC/TMD diagnoses (four diagnoses out of the available eight diagnoses were excluded because of low variable prevalence) and two palpation summary measures.

A paired *t*-test was used to investigate whether observed differences between the two occasions were statistically significant for the three groups of clinical variables (range of mandibular motion, joint clicking, muscle and joint palpation) and the composite measures (RDC/TMD diagnoses including the two palpation summary scores). $P<0.05$ was considered statistically significant. All analyses were carried out using the statistical software package STATA, Release 7.0 (StataCorp. 1999, Stata Statistical Software, College Station, Tx., USA).

Results

Range of motion

The reliability of measurements for range-of-motion clinical signs is given in Table I. Vertical range of motion measured in millimeters was associated with extremely high reliability levels and, for all practical (and statistical) purposes, was equivalent at E1 and E2 [$ICC_{mean}=0.93$ (E1), 0.90 (E2)]. Assessments of extent in millimeters of lateral and protrusive excursions were associated with fair to good reliability and showed similar stability from E1 to E2. The ICC mean difference of –0.02 between E2 and E1 for all 5 mandibular range of motion variables, taken together with the observed 95% CI (–0.12–0.07) was not statistically significant ($p=0.55$).

TMJ sounds

The joint sounds assessed were clicking during different mandibular movements. Reliability ranged from poor reliability on E1 to fair to good reliability on E2, after retraining for agreement in detecting joint sounds associated with vertical opening

Table I. Reliability (ICC), mean, and SD for measurements of range of motion

	Initial calibration		Recalibration	
	Mean (SD)	ICC	Mean (SD)	ICC
Vertical movements (measured in mm)				
Unassisted opening without pain	47.1 (9.7)	0.90	43.8 (8.3)	0.88
Maximum unassisted opening	53.1 (8.1)	0.96	51.0 (8.0)	0.90
Maximum assisted opening	55.0 (7.7)	0.93	52.9 (7.8)	0.93
Horizontal movements (measured in mm)				
Lateral excursion	10.7 (2.4)	0.66	10.5 (2.9)	0.75
Protrusion	5.5 (2.9)	0.70	6.8 (2.7)	0.59

Table II. Reliability (ICC) and prevalence of clicking of the TMJ

	Initial calibration		Recalibration	
	Prevalence (%)	ICC	Prevalence (%)	ICC
Click during opening (present/absent)	20	0.42	20	0.54
Click during closing (present/absent)	22	0.33	22	0.54
Click during contralateral movement	9	0.39	13	0.63
Click during ipsilateral movement	7	0.24	9	0.88
Click during protrusion	8	0.46	18	0.79
Click eliminated by opening from a protruded opening	10	0.35	16	0.46

(Table II). Reliability of clicking eliminated in protrusive opening was initially poor but similarly exhibited fair reliability after re-training. Detection of TMJ click during laterotrusion and protrusive movements also improved from poor to good reliability. The mean of the ICC difference between E1 and E2 was 0.27 (95% CI: 0.07–0.48), i.e. reliability increased substantially after recalibration ($p = 0.02$).

Masticatory muscles and TMJ palpation pain

The ICC statistics associated with inter-rater reliability for measuring whether individual masticatory muscles were painful to standardized digital palpation are summarized in Table III. Several of the muscle palpation sites were associated with fair reliability for measurement of pain, while some also exhibited poor reliability, such as the posterior mandibular region, the origin and belly of the masseter, and the tendon of the temporalis on occasion I. For muscle and TMJ examinations, the ICC improved consistently from E1 to E2. All palpation sites improved from fair to good reliability, with the exception of the intra-oral palpation sites of the lateral pterygoid muscle and temporalis tendon after retraining. The reliability of TMJ palpation improved to good on E2. The mean of the ICC difference between E1 and E2 was 0.23 (95% CI: 0.08 to 0.38), i.e. reliability increased substantially after recalibration ($p = 0.01$).

Composite measures

Reliability between the examiners increased when combinations of the TMJ pain palpation scores were computed. The reliability of summary scores for detecting the presence of pain in response to palpation of extra- and intramuscular sites (20 sites) and TMJ sites (4 sites) is given in Table V.

RDC/TMD diagnosis

The prevalence of the RDC/TMD diagnoses for muscle disorders, disk displacements, and arthralgia did not differ substantially between E1 and E2.

When signs and symptoms measured during the clinical examination were combined according to the RDC/TMD algorithms for diagnosing TMD – using the clinical measurement from each examiner – improvements in reliability of diagnosing TMD according to RDC/TMD criteria were seen for all

Table III. Reliability (ICC) and prevalence of muscle and TMJ painful upon palpation

	Initial calibration		Recalibration	
	Prevalence (%)	ICC	Prevalence (%)	ICC
Extra-oral muscle pain (present/absent)				
Temporalis posterior	2	–*	8	0.73
Temporalis middle	15	0.36	27	0.59
Temporalis anterior	36	0.61	39	0.53
Masseter origin	43	0.28	50	0.59
Masseter body	61	0.30	53	0.50
Masseter insertion	49	0.55	54	0.60
Retromandibular region	25	0.45	28	0.64
Submandibular region	13	0.29	26	0.73
Intra-oral muscle pain (present/absent)				
Lateral pterygoid area	78	0.46	82	0.37
Tendon of Temporalis	57	0.13	69	0.25
TMJ pain (present/absent)				
Lateral part	36	0.08	36	0.69
Posterior part	6	0.29	15	0.83

*If prevalence was low ($<5\%$) the ICC was not calculated.

Table IV. Reliability (ICC) and prevalence of RDC/TMD diagnoses and summary measures

	Initial calibration		Recalibration	
	Prevalence (%)	ICC	Prevalence (%)	ICC
Group I (present/absent)				
Myofascial pain	63	0.83	54	1.00
Myofascial pain with limitation	21	0.76	21	1.00
Group II (present/absent)				
Disk displacement with reduction	12	0.26	19	0.64
Disk displacement without reduction (acute form)	1	–*	2	–*
Disk displacement without reduction (chronic form)	1	–*	0	–*
Group III (present/absent)				
Arthralgia	23	0.16	33	0.73
Osteoarthritis	0	–*	0	–*
Osteoarthrosis	2	–*	2	–*
Palpation pain summary measures	Mean		Mean	
All muscles sites (range 0–20)	7.6	0.78	8.7	0.89
All TMJ sites (range 0–4)	0.9	0.29	1.0	0.80

*If prevalence was low (<5%) the ICC was not calculated.

diagnostic categories examined: The relevant findings for all frequently occurring RDC/TMD clinical diagnoses are given in Table IV. The mean of the differences between E1 and E2 for TMD composite measures ($n=6$, 4 RDC/TMD diagnoses plus 2 palpation summary scores) was 0.33 (95% CI: 0.13 to 0.52), i.e. reliability increased substantially after recalibration ($p=0.01$).

Discussion

In other studies assessing the reliability of clinical signs, the composition of participants varied from only TMD patients [4] or only healthy volunteers [3] to a mixture of both [5,11,18]. The main finding in this study was that re-training clinical TMD examiners improved reliability in a majority of the clinical measures and led to an overall better reproducibility of the diagnosis of myofascial pain, disk displacement with reduction, and arthralgia.

In this study, patients with the most common TMD conditions as well as asymptomatic healthy controls were included to ensure that a broad spectrum ranging from none to severe findings was present. On both occasions (E1 and E2), a good representation of the target TMD population was sought and no major statistical difference was seen between the two groups in prevalence of clinical signs. The latter is an important point because substantial differences in prevalence are known to influence reliability.

The number of participants examined in previous reliability studies varies between 12 and 100 [5,19]. The number of participants in our studies was based on numerous previous reliability studies which indicated that 24 individuals were sufficient to capture the most common TMD diagnoses such as myofascial pain, disk displacements, and arthralgia [6]. The reliability coefficient is dependent on the prevalence of the symptom being measured; some diagnoses were therefore not computed because of low prevalence or because they were absent in the sample.

In our study, the examiner was blinded to whether the participant was a patient or a healthy control. This was to avoid bias during examination; in only a minority of reliability studies has a blinding of the examiner been reported [6]. The methodology has been emphasized to strengthen the evidence of the results [20].

Several studies evaluating the reliability of clinical findings have pointed out that the experience and calibration of the examiners are crucial for the result [3,6]. In our study, experienced examiners were used. In a group of examiners, Dahlström et al. [5] were able to show that agreement between previously calibrated examiners was better than between newly calibrated examiners, even though all had experience in TMD. Duinkerke et al. [3] employed inexperienced and experienced examiners who were then calibrated to investigate the reliability of a palpation test. The results from both groups were satisfactory, indicating that calibration is as important as clinical experience.

It is possible that the increase in the reliability of the RDC/TMD examination between E1 and E2 is due to a learning effect over the 6-month period between the sessions rather than the calibration per se. Data from another study, however, suggest an alternative explanation [6]. Dworkin et al. [6], who employed experienced examiners and compared untrained with trained examiners, reported that without calibration the reliability of experienced clinicians was low compared with that of calibrated clinicians. It is likely that not only the calibration per se but also the amount of time spent in calibration is

important. In several of the studies with higher reliability scores for diagnoses of TMD, there was a tendency to spend more time in calibration, 40 h, prior to the reliability trial [11,18,21]. In two of these studies, inexperienced dentists were calibrated, and after 1 week of calibration high reliability values were achieved [11,21]. In our study, experienced examiners were used. In summary, devoting adequate time to the calibration process seems to override professional experience, per se, as prior studies have shown; even non-dentists (dental hygienists and medical nurses) with adequate calibration experience achieved reliability in the RDC/TMD examination comparable to, or exceeding reliability levels yielded by non- or minimally calibrated experienced dentist–clinicians.

Vertical mandibular movements measured by a millimeter ruler as maximum unassisted opening without pain, maximum unassisted opening, and maximum assisted opening were all found to be highly reliable on both occasions. This is in line with results from most other studies [2,9,11,18,22]. Comparisons of TMD patients with asymptomatic controls have revealed a significant difference [11,18]. Given that these measures on the first occasion had excellent reliability, retraining did not improve results.

TMJ sounds may occur as a single click or pop or may consist of multiple sounds or crepitus. To improve reliability in our study, the reciprocal click had to occur on two of three consecutive trials, which eliminates indistinct or temporary clicking sounds. Studies evaluating the presence (or absence) and type of TMJ sounds have reported acceptable reliability, which was also found in this study [10–12,18]. Observer reliability was improved by retraining the examiners, which is in line with the findings of others [6,7]. Stethoscopes have been employed in clinical settings to detect joint sounds, and reliability reported to be similar to that found with finger palpation [18] or lower [12]. TMJ sounds vary widely from one assessment to the next in the same individual, and the description can vary from no joint sound to crepitus to clicking. TMJ sounds are a questionable indicator of disease, since the prevalence of TMJ sounds in an asymptomatic population of adults and adolescents has been reported to be approximately 30% and 14%, respectively [23,24].

In two studies using trained examiners, the reliability of muscle palpation in both symptomatic and asymptomatic populations was investigated [11,18]. Similar findings were reported in both studies: intra-oral muscle reliability estimates were found to be lower than extra-oral muscle reliability. This is in agreement with our results where extra-oral sites exhibited higher ICC values than the intra-oral sites. Dworkin et al. [6] reported that retraining examiners improved reliability from acceptable to good levels for extra-oral and intra-oral muscles, while reliability for TMJ palpation improved to acceptable levels. In our study, palpation of extra-oral muscles and the TMJs improved to good reliability levels following retraining, whereas although intra-oral muscle sites improved they still exhibited poor reliability levels. The reliability of the summary scores of muscle and TMJ palpation sites improved to a level near those found in other studies [11,18,22].

Many studies require the examiner to rate the patient's pain response to palpation rather than having the patient rate the pain. Only one study has examined this issue, and a high reliability of patient's ratings with trained examiners compared to untrained examiners was found. One explanation for this might be that the variability in palpation pressure applied by the calibrated examiners was less than that applied by the untrained examiners. Similar findings have also been reported by others [9].

One of the difficulties with estimating the reliability of muscle and TMJ palpation is the stability of the phenomenon being measured over time. Since muscle and TMJ palpation responses can vary from one examination to the next during the same day and from one day to the next, the difficulty with obtaining high reliability scores is apparent. To overcome this problem, some investigators have created a single composite score for muscle palpation that produces higher reliability scores. Cut-off scores have also been created for diagnoses; for example, 3 or more of 20 painful sites are necessary to obtain a diagnosis of myofascial pain.

Reliability improved considerably for most diagnostic groups, which is consistent with the observed improvement in reliability of measuring most clinical signs which associated with recalibration. It is axiomatic that poor reliability of assessing clinical signs is associated with poorer reliability for diagnoses based on those clinical signs and, most critically, is related to poor validity of the clinical diagnoses – validity of clinical findings is statistically limited by reliability of clinical measurement. The improvement in clinical diagnoses from E1 to E2 was most evident in arthralgia, which is likely related not only to recalibration but also to the increased prevalence of signs related to the condition. However, overall, all diagnostic groups exhibited good reliability after the examiners had been recalibrated. The reliability values are similar to those reported in other studies that have used the RDC/TMD [11,21]. Reliabilities for uncommon diagnoses such as osteoarthrosis were not calculated owing to their low prevalence in this study. Consequently, if the reliability of all TMD diagnostic groups, including the rare conditions, has to be estimated, special attention to the number of participants of each diagnostic group needs to be considered so that a sufficiently

large group is included in the study to be able to calculate the reliabilities of even the rarest conditions.

Conclusion

Recalibration considerably improved inter-examiner reliability for assessing RDC/TMD clinical variables and diagnoses, especially where initial inter-examiner reliability was low. Finally, inter-examiner reliability in all areas was associated with acceptable to excellent levels.

Acknowledgments

The staff at the Dental Schools in Malmö and Aarhus are acknowledged for providing help during the calibration studies.

References

[1] World Health Organization. Oral health surveys – basic methods. Geneva: WHO; 1971.
[2] Carlsson GE, Egermark-Eriksson I, Magnusson T. Intra- and inter-observer variation in functional examination of the masticatory system. Swed Dent J 1980;4:187–94.
[3] Duinkerke AS, Luteijn F, Bouman TK, de Jong HP. Reproducibility of a palpation test for the stomatognathic system. Community Dent Oral Epidemiol 1986;14:80–5.
[4] de Wijer A, Lobbezoo-Scholte AM, Steenks MH, Bosman F. Reliability of clinical findings in temporomandibular disorders. J Orofac Pain 1995;9:181–91.
[5] Dahlstrom L, Keeling SD, Fricton JR, Galloway Hilsenbeck S, Clark GM, Rugh JD. Evaluation of a training program intended to calibrate examiners of temporomandibular disorders. Acta Odontol Scand 1994;52:250–4.
[6] Dworkin SF, LeResche L, DeRouen T. Reliability of clinical measurement in temporomandibular disorders. Clin J Pain 1988;4:89–99.
[7] Goulet JP, Clark GT, Flack VF, Liu C. The reproducibility of muscle and joint tenderness detection methods and maximum mandibular movement measurement for the temporomandibular system. J Orofac Pain 1998;12:17–26.
[8] John MT, Zwijnenburg AJ. Interobserver variability in assessment of signs of TMD. Int J Prosthodont 2001;14: 265–70.
[9] Kopp S, Wenneberg B. Intra- and interobserver variability in the assessment of signs of disorder in the stomatognathic system. Swed Dent J 1983;7:239–46.
[10] Westling L, Helkimo E, Mattiasson A. Observer variation in functional examination of the temporomandibular joint. J Craniomandib Disord 1992;6:202–7.
[11] Wahlund K, List T, Dworkin SF. Temporomandibular disorders in children and adolescents: reliability of a questionnaire, clinical examination, and diagnosis. J Orofac Pain 1998;12:42–51.
[12] Wabeke KB, Spruijt RJ, van der Zaag J. The reliability of clinical methods for recording temporomandibular joint sounds. J Dent Res 1994;73:1157–62.
[13] Stockstill JW, Gross AJ, McCall WD Jr. Interrater reliability in masticatory muscle palpation. J Craniomandib Disord 1989;3:143–6.
[14] John MT, Dworkin SF, Mancl LA Reliability of clinical temporomandibular disorder diagnoses. Pain 2005.
[15] Dworkin SF, LeResche L. Research diagnostic criteria for temporomandibular disorders: review, criteria, examinations and specifications, critique. J Craniomandib Disord 1992;6: 301–55.
[16] Shrout PR, Fleiss JL. Intraclass correlation: uses in assessing rater reliability. Psychol Bull 1979;86:420–8.
[17] Fleiss JL. The design and analysis of clinical experiments. New York: Wiley; 1986.
[18] Dworkin SF, LeResche L, DeRouen T, Von Korff M. Assessing clinical signs of temporomandibular disorders: reliability of clinical examiners. J Prosthet Dent 1990;63: 574–9.
[19] Smith JP. Observer variation in the clinical diagnosis of mandibular pain dysfunction syndrome. Community Dent Oral Epidemiol 1977;5:91–3.
[20] Altman D. Practical statistics for medical researchers. London: Chapman & Hall; 1991.
[21] Marcusson A, List T, Paulin G, Dworkin S. Temporomandibular disorders in adults with repaired cleft lip and palate: a comparison with controls. Eur J Orthod 2001;23:193–204.
[22] Fricton JR, Schiffman EL. Reliability of a craniomandibular index. J Dent Res 1986;65:1359–64.
[23] Dworkin SF, Huggins KH, LeResche L, Von Korff M, Howard J, Truelove E, et al. Epidemiology of signs and symptoms in temporomandibular disorders: clinical signs in cases and controls. J Am Dent Assoc 1990;120:273–81.
[24] List T, Wahlund K, Wenneberg B, Dworkin SF. TMD in children and adolescents: prevalence of pain, gender differences, and perceived treatment need. J Orofac Pain 1999;13: 9–20.