Interobserver Variability in Assessment of Signs of TMD

Mike T. John, Dr Med Dent[a]
Arco J. Zwijnenburg, PhD[b]

Purpose: The aim of this report was to study the ability of examiners to measure reliably the clinical signs of temporomandibular disorders (TMD). Four examiners participated in this study of 11 TMD patients and 25 nonpatients. **Materials and Methods:** Vertical and lateral excursions of the jaw were measured using a millimeter ruler. Joint sounds during vertical jaw movements were assessed using digital palpation. The reliability of delivering appropriate degrees of digital pressure to assess masticatory muscle pain was assessed using a manometer after training examiners to exert specified pressures. **Results:** Intraclass correlation coefficients for the measurement of vertical and protrusive jaw movements were ≥ 0.87, which was considered excellent. The intraclass correlation coefficient for measurements of left and right lateral jaw excursions varied between 0.73 and 0.85, which was considered acceptable. The interobserver agreement for detecting the joint sounds showed overall agreement across examiners of 78%. Kappa for every possible pair of examiners varied between .52 and .86 (median .75, interquartile range .18). Reliability for diagnostic categories from the Helkimo index and Research Diagnostic Criteria for Temporomandibular Disorders involving joint noises showed modest reliability. **Conclusion:** Point estimates and measures of spread for reliability measures of single clinical TMD signs as well as combinations of signs into diagnostic categories from the Helkimo index and Research Diagnostic Criteria for Temporomandibular Disorders involving joint noises were sufficient in a group of four examiners. *Int J Prosthodont 2001;14:265–270.*

It is generally agreed that clinical data are collected with some degree of uncertainty and inaccuracy in that no diagnostic procedure is completely free of measurement error. Reliability, or reproducibility of clinical data, is fundamental to arriving at a valid diagnosis. When various examiners perform clinical measurements, the results can only be interpreted with confidence if examiners are able to measure clinical signs and symptoms with consistency.[1] The World Health Organization has recommended that reliability information should be a routine part of all oral health survey final reports.[2]

Several studies assessing the reliability of clinical signs and symptoms of temporomandibular disorders (TMD) have measured reliability of examiners for assessing continuous variables using intraclass correlation coefficients (ICC) and examined reliability of pairs of examiners for clinical categoric variables using the kappa statistic (Cohen's kappa)[1,3–8] and sometimes, but rarely, Scott's pi.[9]

Most studies present only mean ICC and kappa when more than two examiners are involved. The assumption that the examiners are interchangeable is essential to pooling clinical measurements across all examiners. This is particularly important for large epidemiologic and clinical multicenter studies with many investigators, when information about the reliability for all examiners is warranted. But to be confident that such pooled data result from reliable

[a]*Assistant Professor, Department of Prosthodontics, Martin Luther University Halle-Wittenberg, Halle/Saale, Germany.*
[b]*Assistant Professor, Department of Prosthodontics and Material Sciences, University of Leipzig, Germany.*

***Reprint requests:** Dr M. John, Department of Prosthodontics, Martin Luther University Halle-Wittenberg, Große Steinstraße 19, 06108 Halle (Saale), Germany. Fax: + 49 345-5573779. e-mail: mike.john@medizin.uni-halle.de*

measurements, it would be useful to report, in addition to mean kappa and ICC, measures representing the spread of kappa and ICC among the clinical examiners being tested for reliability.

Most reliability studies investigate isolated clinical TMD signs. However, some subjects may present multiple signs and symptoms. They can be combined into indices describing the severity of TMD for the individual. A common index is the Helkimo index.[10] Inclusion and exclusion criteria group signs and symptoms to define specific clinical disorders.[11,12] These alternatives for characterizing TMD are also influenced by measurement variability. Guidelines for assessing the confidence one can have in reliability of clinical measurements exist.[13,14]

In preparation of an epidemiologic research project and a multicenter TMD study, this report describes the results of a reliability study involving the calibration of four trained TMD examiners and one TMD study coordinator. The aim of this study was to assess the ability of examiners to reliably measure signs of TMD. Methodologic issues involved with measuring the different clinical signs, including palpation pain of masticatory muscles and temporomandibular joints (TMJ), joint noises, and range of mandibular motion, are discussed.

Materials and Methods

Four examiners (three dentists and one examiner with a PhD in oral physiology) trained in TMD examinations from their clinical experience in the diagnosis and treatment of TMD participated in the study, which was executed in two sessions.

This study was performed according to the guidelines of Dworkin and Whitney.[15] Criteria for defining the variables assessed in this study were provided by the specification contained in the Research Diagnostic Criteria for Temporomandibular Disorders (RDC/TMD).[12] The measurement definition of the variables is summarized in Table 1.

Examination Specifications

Joint sounds for both the left and the right TMJs were determined using digital palpation. Sounds were classified according to the categories "no sound," "clicking," and "occasional clicking." To assess joint sounds, the examiners judged three consecutive opening and closing movements. A clicking sound was reported when the clicking was perceived in at least two of three movements. When the clicking was perceived in only one movement, the category occasional clicking was selected.

Measurements of mandibular range of motion were performed to the nearest millimeter using a ruler. Measurements of range of motion were repeated three times, and the highest value was recorded.

The vertical opening patterns were determined using a millimeter ruler held vertically between the maxillary central incisal embrasures. Only deviations or deflections larger than 2 mm were noted (Table 1).

Calibration of Examiners

Previous reliability studies have reported marginal to good intraobserver and interobserver reproducibility for assessment of muscle and joint tenderness using finger palpation.[4,7–9]

Reported pain provoked by palpation tests is subjective. The pressure and the method of palpation (location of muscle and joint points, technique of finger palpation) determine the pain response from the patient. However, repeatedly performed palpation on the same subject during the training process could distort the results and mislead the trainee.[1] Therefore, for initial training and calibration purposes, determination of muscle and joint tenderness was not performed directly on subjects (patients or controls). Instead, the examiners were given a manual with specifications of the palpation points and the method. The location of the exact palpation points was demonstrated on volunteers by the study coordinator, and the examiners determined the palpation location on the same volunteers and on each other.

According to the RDC/TMD, palpation should be done with a pressure equivalent to 0.91 kg for the extraoral muscles and 0.45 kg for the joints and the intraoral muscles.[12] In our study, the examiners were trained to exert fingertip pressure equivalent to a range between 0.73 and 1.09 kg for extraoral palpation and between 0.36 and 0.54 kg for intraoral palpation in at least four of five trials. This was performed with a manometer (Haag-Streit).

Selecting Appropriate Clinical Samples

To examine a wide spectrum of TMD signs,[16] TMD cases (persons with signs and symptoms of TMD) and symptom-free persons not included in the calibration sessions were recruited; 11 patients, 19 students, and six staff members from the Martin Luther University School of Dentistry were included in the investigation. In the first session of the reliability study, four examiners independently studied 13 subjects. In the second session, the remaining 23 subjects were studied. The subjects varied in age from 17 to 71 years, and their mean age was 29.3 ± 12.3 years. Independent diagnosis by the study coordinator prior

Table 1 Definition of Clinical Measurements

Variable	Measurement definition
Range of motion	
Maximum active opening	Distance (in mm) between the edge of the most vertically oriented maxillary incisor and the labioincisal edge of the opposing mandibular incisor during maximum active opening + vertical incisal overlap at this line
Lateral excursions	Distance (in mm) between maxillary and mandibular labioincisal embrasures between the maxillary central incisors and mandibular central incisors during maximum lateral mandibular movement (corrected by the midline deviation) in slight jaw opening
Protrusive excursions	Distance (in mm) from maxillary labial to mandibular labial incisor surfaces during maximum protrusive mandibular movement in slight jaw opening
Joint sounds during open-close movement	Click = a short, distinct sound of brief and very limited duration, with a clear beginning and end; usually sounds like a "click" Crepitation = a sound that is continuous over a longer period of jaw movement; the joint may make overlapping, continuous noises
Opening pattern	Straight pattern = no perceptible deviation or deflection (< 2 mm) upon opening Right or left deflection = perceptible deflection (≥ 2 mm) to the right or left at maximum opening; no correction to the midline before reaching maximum unassisted opening Right or left deviation = perceptible deviation (≥ 2 mm) to the right or left; corrected to the midline before or upon maximum unassisted opening

to the calibration sessions revealed that seven subjects presented bilateral joint clicking, 11 subjects had unilateral clicking, and 18 subjects had no joint clicking.

The examinations of the reliability trial were performed at four dental units (measurement stations) in two examination rooms simultaneously. At each station, the measurement results were recorded by research assistants. All subjects were examined in a random order by each examiner, all of whom were blind to subject status (case or noncase).

Reliability of continuous variables was determined using ICC. Values of ICC ≥ 0.75 were taken as an indication of acceptable interexaminer agreement.[5]

The reliability of categoric variables was determined using percent agreement between the examiners and the kappa statistic (Cohen's kappa), which was determined for every possible pair of examiners. Median $\tilde{\chi}$ and interquartile range (IQR) of the kappa statistic for each pair of examiners were calculated. Kappa controls for agreement that would occur simply by chance.[17] Kappa was interpreted according to Landis and Koch.[18] Thus, $\kappa \geq .60$ was taken as indicative of acceptable interexaminer agreement.[15]

Results

Regarding range of motion, the ICC of the measurement of maximum opening as well as protrusion was 0.87 or higher (Table 2). For left and right lateral excursions, the ICC varied between 0.73 and 0.85.

The interobserver agreement of joint sound assessment with categories "clicking present" and "no clicking" showed an overall agreement of 78% across all examiners. Kappa for every possible pair of examiners varied between .52 and .86 ($\tilde{\chi}$ .75, IQR .18; Table 3).

Table 2 Reliability of Mandibular Range of Motion Using Intraclass Correlation Coefficients*

Variable	First session	Second session
Active maximum opening	0.93 (0.85, 0.98)	0.94 (0.86, 0.98)
Protrusion	0.91 (0.80, 0.97)	0.87 (0.73, 0.95)
Left lateral excursion	0.79 (0.60, 0.93)	0.85 (0.70, 0.95)
Right lateral excursion	0.73 (0.50, 0.90)	0.79 (0.59, 0.92)

*95% confidence interval values are given in parentheses.

Crepitation was assessed on five clinical TMD patients. Each case was discussed by the group to reach an acceptable agreement on the phenomenon.

The procedure for determining pain on palpation recommended by Goulet et al[6] was used to ensure acceptable reliability. All examiners were able to locate the specified palpation locations and to exert fingertip pressure equivalent to a range between 0.73 and 1.09 kg for extraoral palpation and between 0.36 and 0.54 kg for intraoral palpation in at least four of five trials. Each examiner subsequently performed six series of five trials at both target pressures. The mean values for the series of five trials ranged between 0.75 and 0.96 kg, with standard deviations between 0.04 and 0.14 kg for extraoral palpation. Mean values for intraoral palpation were between 0.40 and 0.52 kg, with standard deviations between 0.04 and 0.07 kg.

When TMJ dysfunction was calculated as part of the Helkimo index,[10] the index varied between 0 and 2

Table 3 Reliabiliy of Joint Clicking Using Kappa*

	Examiner 1	Examiner 2	Examiner 3
First session			
Examiner 2	.61 (81)		
Examiner 3	.72 (88)	.52 (77)	
Examiner 4	.74 (88)	.52 (77)	.81 (92)
Second session			
Examiner 2	.86 (93)		
Examiner 3	.81 (91)	.86 (93)	
Examiner 4	.76 (89)	.63 (86)	.77 (89)

*% agreement values are given in parentheses.

Table 4 Kappa and Examiner Agreement for Joint Dysfunction According to the Helkimo Index and Disc Displacement with Reduction According to the RDC/TMD (Median; Interquartile Range)

	Joint dysfunction		Disc displacement with reduction	
	First session	Second session	First session	Second session
% agreement	69; 17	83; 4	88; 5	87; 9
Kappa	.41; .32	.52; .15	.51; .13	.32; .33

and the agreement varied between 52% and 77% ($\tilde{\chi}$ 69%, IQR 17%) in the first session. Kappa ranged from –.05 to .57 ($\tilde{\chi}$ .41, IQR .32). In the second session, the examiner agreement improved, to a range from 74% to 91% ($\tilde{\chi}$ 83%, IQR 4%), as did kappa, which changed to .33 to .78 ($\tilde{\chi}$ .52, IQR .15).

To determine the ability of the examiners to diagnose TMD patients, the joints were classified as disc displacement with reduction versus no disc displacement according to the RDC/TMD.[12] The agreement between observers was similar in both sessions (Table 4). In the first session, kappa varied between .33 and .71 ($\tilde{\chi}$ .51, IQR .13) and in the second session between .10 and .73 ($\tilde{\chi}$ .32, IQR .33).

Discussion

The results of this study were consistent with other studies.[4,5] Certain signs of TMD are less reliable than others. They fluctuate because of examination influences and the transient nature of the phenomenon itself. The variability of clinical signs can be assigned to three different sources: *(1)* error in measurement (ie, instruments are unreliable); *(2)* unreliability of examiners; or *(3)* instability or changing characteristics of the underlying structure being measured.

The measurement reliability of maximum jaw opening in this study was excellent. This is in agreement with other studies.[1,6,8,9] The lower but adequate ICC values for determination of the lateral excursive movement range are also shown in other studies.

Determination of joint sounds was marginally reliable. This finding is supported by other studies.[1,5] Dworkin and coworkers[1,4,12] believe that training of examiners is essential when examination results have to be compared. The results of the present study underline this recommendation. When each of the four examiners were compared with the other three examiners, the kappa values for five of the possible six combinations of observers were higher during the second session compared to the first (Table 3). Although the examiners were trained in the diagnosis of TMD, between both sessions the reliability improved, probably because of the training effect of the first calibration session.

A majority of the variability of the joint noises is likely because of the variability of the joint noise itself. An estimate of the amount of the true intraobserver variability (corrected for the phenomenon joint clicking) can be approximated by having two observers assess the same joint simultaneously. This is possible by means of recording the joint sounds on tape and letting different observers analyze these sounds.[19] A drawback of this procedure is that the observer cannot see the jaw movements and is thus not able to interpret the sounds adequately. This disadvantage can be overcome by using a stereo stethoscope, ie, a stethoscope with two sets of earpieces that allow two observers to listen to the same TMJ at the same time during excursions of the jaw.[1]

Dworkin and coworkers found that successive pairs of examiners listening in turn to the same TMJ will

disagree with the sound detected by the previous pair of examiners roughly 50% of the time. They state that "poor inter-examiner variability . . . may more importantly reflect highly variable patterns of joint noises than inadequacies of clinical examiners." Joint sounds exist in a wide range of intensity. Results from previous studies show an increased reliability when joint sounds that are audible to the subjects are recorded.[20]

Although it is generally accepted to present the agreement between various pairs of observers by the average kappa,[1,4] the results of this study showed that an average kappa does not adequately represent reliability of measurement because the spread of all pairs of observers is substantial. This indicates that perhaps some examiners are less able to become calibrated or need more training, and hence are less able to acquire satisfactory reliability. It is therefore concluded that the agreement between more than two observers can be better represented by median and interquartile range of kappa.

Unquestionably, pain is the most striking symptom of TMD. One of the major issues is believing in the patient's complaints.[13] The variation in palpation pressure may influence the pain report. Therefore, it is important to use the same palpation pressure for all patients. The results of the manometer palpation pressures in this study showed that the examiners were able to exert a relatively constant fingertip pressure during palpation. It is a limitation of our study that we used the application of palpation pressure (although it was precise) as a proxy for the reliability of palpation assessed on patients.

The mean kappa for joint dysfunction in our study was comparable with the result of Kopp and Wenneberg[8] for the clinical dysfunction index (Scott's pi = .30, 63% agreement). This finding underlines the well-known rule that the reliability of a diagnosis cannot be better than the reliability of the assessment of component clinical signs and symptoms.

According to the RDC/TMD,[12] the diagnosis of disc displacement with reduction can be assigned when reciprocal clicking in the TMJ that is reproducible on two of three consecutive trials, or a reproducible click during opening or closing and a reproducible click during lateral excursion and/or protrusion, is reported. In this study, however, joint sounds were only determined during normal open-close jaw movements, so the diagnosis disc displacement with reduction is probably underestimated. When the patients in this study were diagnosed according to the RDC/TMD, the agreement between the pairs of examiners varied between 80% and 96%, with a corresponding kappa between .10 and .73. Wahlund et al[21] reported higher kappa values between .85 and .90 for disc displacements. This difference may be because of more reliable examiners or a better-balanced study population in the previous study. The paradoxic result, a relatively good agreement and low kappa, has been addressed by Feinstein and Cicchetti.[22,23] These authors argue that the basis for this paradox is the dependence of the kappa statistic on the prevalence of the identifiable trait in the sample population. Lantz and Nebenzahl,[24] for example, showed that with an observer agreement of 90%, the kappa can vary between –.05 and .80 according to the configuration of the data, ie, the prevalence of the identifiable trait in the sample population. Similar possibilities were reported by Dworkin and Whitney.[15]

Numerous studies, including the present report, have demonstrated that reliability of TMD signs varies from excellent to poor. Certain signs of TMD are more reliable than others. Assessment of range of motion is very consistent. Muscle and joint palpation as well as joint noises are less reproducible, but measurement can achieve acceptable levels of reliability with training, calibration, and selection of examiners. Unstable phenomena and the influence of the examination procedure itself might be responsible for the less-reliable TMD variables, especially assessment of TMJ sounds.

Impact of results from calibration studies for epidemiologic investigations can be profound for at least two reasons: *(1)* reported prevalence of TMD signs and symptoms is known to vary widely, and it seems reasonable to assume that a portion of this variability can be related to the use of uncalibrated examiners using nonstandardized examination methods; and *(2)* unreliable data, no matter the source of inconsistency of measurement, cannot yield valid clinical diagnoses—reliability represents a mathematic constraint on the maximum validity one can expect from a set of clinical measurements. Prevalences range from 0% to 93% for clinically assessed signs.[25] The majority of the variability in prevalence is supposed to be because of methodologic differences.[26] Therefore, lack of reliability, beyond a lack of validity, may contribute to these methodologic differences.

When reporting TMD research results, reproducibility of examiners should be aimed for and reported, because as shown, training of examiners is essential to maximize reliability of clinical examination data.

Acknowledgments

The authors thank Dr S. F. Dworkin, University of Washington, Seattle for his valuable comments. This study was supported by Forschungsverbund Public Health Sachsen, grant No. 01EG9532/0 and by Deutsche Akademie der Naturforscher Leopoldina grant No. BMBF-LPD 9901/8-4.

References

1. Dworkin SF, LeResche L, DeRouen T. Reliability of clinical measurements in temporomandibular disorders. Clin J Pain 1988;49: 89–99.
2. World Health Organization. Oral Health Surveys Basic Methods, ed 4. Geneva: WHO, 1998.
3. Duinkerke AS, Luteijn F, Bouman TK, de Jong HP. Reproducibility of a palpation test for the stomatognathic system. Community Dent Oral Epidemiol 1986;14:80–85.
4. Dworkin SF, LeResche L, DeRouen T, Von Korff M. Assessing clinical signs of temporomandibular disorders: Reliability of clinical examiners. J Prosthet Dent 1990;63:574–579.
5. Goulet JP, Clark GT. Clinical TMJ examination methods. J Calif Dent Assoc 1990;18:25–33.
6. Goulet JP, Clark GT, Flack VF. Reproducibility of examiner performance for muscle and joint palpation in the temporomandibular system following training and calibration. Community Dent Oral Epidemiol 1993;21:72–77.
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. Kopp S, Wenneberg B. Intra- and interobserver variability in the assessment of signs of disorder in the stomatognathic system. Swed Dent J 1983;7:239–246.
9. Carlsson GE, Egermark-Eriksson I, Magnusson T. Intra- and interobserver variation in functional examination of the masticatory system. Swed Dent J 1980;4:187–194.
10. Helkimo M. Studies on function and dysfunction of the masticatory system. II. Index for anamnestic and clinical dysfunction and occlusal state. Swed Dent J 1974;67:101–121.
11. Okeson JP. Orofacial Pain: Guidelines for Assessment, Diagnosis, and Management. Chicago: Quintessence, 1996.
12. Dworkin SF, LeResche L. Research diagnostic criteria for temporomandibular disorders: Review, criteria, examinations and specifications, critique. J Craniomandib Disord Facial Oral Pain 1992;6:301–355.
13. Turk DC, Melzack R. Trends and future directions in human pain assessment. In: Turk DC, Melzack R (eds). Handbook of Pain Assessment. New York: Guilford, 1992:473–479.
14. Rudy TE, Turk DC, Brody MC. Quantification of biomedical findings and chronic pain: Problems and solution. In: Turk DC, Melzack R (eds). Handbook of Pain Assessment. New York: Guilford, 1992:447–469.
15. Dworkin SF, Whitney CW. Relying on objective and subjective measures of chronic pain: Guidelines for use and interpretation. In: Turk DC, Melzack R (eds). Handbook of Pain Assessment. New York: Guilford, 1992:429–446.
16. Fletcher RH, Fletcher SW, Wagner EH. Clinical Epidemiology: The Essentials, ed 3. Baltimore: Williams & Wilkins, 1996.
17. Maclure M, Willett WC. Misinterpretation and misuse of the kappa statistic. Am J Epidemiol 1987;126:161–169.
18. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–174.
19. Eriksson L, Westesson PL, Sjöberg H. Observer performance in describing temporomandibular joint sounds. Cranio 1987;5:32–35.
20. Fricton JR, Schiffman EL. Reliability of a craniomandibular index. J Dent Res 1986;65:1359–1364.
21. 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.
22. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol 1990;43:543–549.
23. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol 1990;43:551–558.
24. Lantz CA, Nebenzahl E. Behavior and interpretation of the kappa statistic: Resolution of the two paradoxes. J Clin Epidemiol 1996;49:431–434.
25. de Kanter RJAM, Truin GJ, Burgersdijk RC, Van't Hof MA, Battistuzzi PG, Kalsbeek H, Kayser AF. Prevalence in the Dutch adult population and a meta-analysis of signs and symptoms of temporomandibular disorder. J Dent Res 1993;72:1509–1518.
26. Carlsson GE, LeResche L. Epidemiology of temporomandibular disorders. In: Sessle BJ, Bryant PS, Dionne RA (eds). Temporomandibular Disorders and Related Pain Conditions. Seattle: IASP, 1995:211–226.