Is There a Difference in the Reliable Measurement of Temporomandibular Disorder Signs Between Experienced and Inexperienced Examiners?

Anna Leher, Dr Med Dent, MPH
Researcher
Department of Medical Informatics, Biometry and Epidemiology
Friedrich-Alexander University
Erlangen-Nuremberg, Germany

Kathrin Graf, Dr Med Dent
Researcher

Jean-Marc PhoDuc, Dr Med Dent
Researcher
Department of Prosthodontics
Ludwig-Maximilian University
Münich, Germany

Peter Rammelsberg, Prof, Dr Med Dent
Professor and Department Chairman
Department of Prosthodontics
Ruprecht-Karl University
Heidelberg, Germany

Correspondence to:
Dr Anna Leher
Alt-Wuerttemberg-Allee 33
71638 Ludwigsburg
Germany
Fax: +49 7141 144 9512
E-mail: anna.leher@gmx.de

Aims: *To determine whether there is a difference in terms of reliability between experienced examiners and inexperienced examiners in the measurement of signs of temporomandibular disorders (TMD).* ***Methods:*** *A total of 27 patients seen for treatment of TMD were rated blindly and in random sequence by 2 experienced and 2 inexperienced examiners. The examiners participated in a 4-hour calibration session on the day preceding the reliability study. Both experienced and inexperienced examiners participated in the calibration session to reduce the effect of examiner subjectivity and allow the study focus to be on the effect of experience. The rating followed the Research Diagnostic Criteria for Temporomandibular Disorders and included mandibular movements, joint sounds, and digital palpation of muscles and joints. Intraclass correlation coefficients and kappa statistics were calculated to estimate interrater reliability. The Wilcoxon signed rank test was performed to test for differences between experienced and inexperienced examiners' results, and the Friedman test was used for differences between all 6 examiner combinations.* ***Results:*** *Excellent overall reliability was found for vertical mandibular motions, acceptable reliability was found for the summed muscle palpation pain sites, and moderate to poor reliability was found for excursive movements, joint sounds, and single muscle palpation pain sites. No significant differences in the measurement results could be found between the experienced examiners and the inexperienced examiners.* ***Conclusion:*** *Examiner calibration rather than professional experience seems to be the most important factor for reliable measurement of TMD symptoms.* J OROFAC PAIN 2005;19:58–64

Key words calibration, reliability, temporomandibular disorders

Dependable and valid diagnoses are a prerequisite for the proper treatment of disease. Reliable diagnoses are also important to ensure that epidemiologic research studies generate accurate results. In the first case, a wrong diagnosis may have a negative effect on the individual; in the latter case, it may bias prevalence rates on which guidelines, public health measures, and research focus are based. It is important for researchers preparing prevalence studies to ensure that the data to be collected are representative of the target population and that diagnostic measurements are carried out with due precision. There are many well-known methods to enhance data representativeness, such as randomization.

However, establishing study validity can be more problematic. One of the main challenges in establishing validity is that the

observed symptoms may vary. This variability can usually be attributed to 3 causes. First, the variability can be due to the changing characteristics of the observed symptoms. Symptoms can vary naturally or because of the influence of examination. It is well known that joint sounds have fluctuating characteristics and the intraindividual variability is very high.[1–7] Second, variability of observed symptoms can be due to the use of unreliable diagnostic instruments. This can be controlled for by checking the reliability of the diagnostic tools employed. For example, the validity and reliability of the Research Diagnostic Criteria for Temporomandibular Disorders (RDC/TMD),[8] which have gained wide acceptance as a way to classify temporomandibular disorders (TMD), have been proven several times.[1,2,7,9] Third, low agreement between observers measuring the same item can lead to different conclusions about the same patient. This variability can be attributed to individual characteristics of the examiner, eg, his or her experience level and whether he or she underwent previous calibration. Many earlier studies have dealt with reliability,[1–20] and in most instances, examiner calibration has been considered crucial. For example, Dahlström et al[10] were able to show in a group of examiners that previously calibrated examiners had better agreement than newly calibrated examiners, although all of the examiners had experience with TMD patients, as did those used by Dworkin et al,[1,3] who employed experienced examiners and compared those who were trained with those who were untrained. Duinkerke et al[12] did utilize some inexperienced examiners, but their intention was to evaluate the reproducibility of a palpation test.

Since there has been no research focus on the influence of experience, the aim of the present investigation was to determine whether the reliability of measurements of TMD symptoms would vary with the experience of the examiners. The results of this study may have an impact on how examiners are chosen and trained for epidemiologic investigations of TMD.

Materials and Methods

Study Population

A total of 22 women and 5 men participated in this study. All were TMD patients who sought treatment for jaw pain or dysfunction at the prosthodontic department of the Ludwig-Maximilian University in Münich, Germany. Participants were carefully prepared for examination. They were well informed about the study by an independent clinician, and all provided written consent. Potential candidates were asked to participate if they were between 18 and 75 years old and had reported TMJ pain or sounds (Table 1). Only a few of those asked refused because of the time required.

Table 1 Criteria for Inclusion and Exclusion of Subjects

Inclusion	Exclusion
•18 to 75 years old •Report of TMD pain or joint sounds	•Presence of sinus disease or ear disorders •Systemic inflammatory polyarthritis •Cervical pain or dysfunction •Episodic headaches •Dental infection

The study sample was limited to TMD patients because a study design incorporating both TMD patients and healthy controls was not considered necessary or feasible. A single symptom (eg, tenderness of a muscular palpation point) has a low prevalence even among TMD patients, and the inclusion of healthy individuals would have reduced the prevalence of the symptoms studied in the patient population and also the possibility of achieving acceptable reliability values. Furthermore, not every TMD patient experiences every symptom, so controls for each symptom could be expected within the group. It was estimated that a sample size of at least 30 patients ($\alpha = 0.05$, power = 80%) would be required to detect a significant difference in the reliability results of the examiners. However, of the 33 patients who initially consented, 6 were absent on the day of examination.

Clinical Examination

All subjects were clinically examined by 4 examiners. Two of the examiners (A and B) were experienced clinicians who had been examining TMD patients for several years each. The other 2 examiners (C and D) were undergraduate students in their last year of dental school who had no practical experience in TMD examination.

The clinical examination was performed according to the guidelines described in the RDC/TMD diagnostic criteria.[8] Range of motion was measured in millimeters with a ruler. Tenderness to palpation was recorded as described in the guidelines. However, following the RDC/TMD guidelines, since only the presence or absence of tenderness at the site of muscle palpation is important for a diagnostic decision, muscle tenderness was

Table 2 Clinical Signs

Pain on palpation*
1. Masseter (origin, body, and insertion), right side
2. Masseter (origin, body, and insertion), left side
3. Temporalis (posterior, middle, anterior), right side
4. Temporalis (posterior, middle, anterior), left side
5. Retromandibular region, right side
6. Retromandibular region, left side
7. Submandibular region, right side
8. Submandibular region, left side
9. Lateral pterygoid area, right side
10. Lateral pterygoid area, left side
11. Tendon of temporalis, right side
12. Tendon of temporalis, left side
13. Lateral pole and posterior attachment of TMJ, right side
14. Lateral pole and posterior attachment of TMJ, left side
TMJ sounds
15. Sounds with vertical opening, right side
16. Sounds with vertical opening, left side
Range of motion
1. Unassisted opening, no pain
2. Maximum unassisted opening
3. Maximum assisted opening
4. Lateral excursions, right side
5. Lateral excursions, left side
6. Horizontal overbite
7. Vertical overlap

*Items 1 to 8, 13, and 14 are extraoral palpation sites; items 9 to 12 are intraoral palpation sites.

simply counted as present or absent for the purpose of determining reliability. Extraoral sites were palpated with a force of 0.9 kp and intraoral sites with a 0.45 kp force. In addition, joint sounds were not categorized as crepitation or clicking; examiners were only required to assess the presence or absence of sounds in the temporomandibular joint (TMJ). This was necessary because the low prevalence of the different sounds in the study population would have prevented the calculation of meaningful kappa values. All examined variables are described in Table 2.

Calibration

All examiners participated in a calibration session to reduce the effect of examiner subjectivity and allow the study focus to be on the effect of experience. A 4-hour session was held on the day preceding the reliability study by an independent dentist experienced in TMD examination. The 4 examiners watched a training video and practiced on each other, on the trainer, and on 3 additional patients not included in the reliability study. At the end of the session they discussed all measuring problems that came up during the session. A scale was used to learn how much pressure to apply for the digital palpation of muscle sites and TMJs (0.45 and 0.9 kp, respectively).

Reliability Measurement

Examiners saw patients in a random sequence to control for possible effects of the passage of time and for repeated measurements. The number of subjects seen first by an examiner was equal for all examiners. The examiners had never seen any of the patients before and were blinded to the results of the previous examinations. Each examination lasted 10 minutes, and all examinations were performed in 1 afternoon to make sure that the time period between calibration and investigation was identical for all the examiners.

Statistical Analyses

Reliability was analyzed by 2 statistical methods. For data measured on continuous scales (eg, millimeter rulers), the intraclass correlation coefficient (ICC) was calculated.[21] The ICC values were considered acceptable if they were greater than 0.7. For categorical variables (eg, response to muscle palpation), Cohen's κ was calculated. It adjusts for the likelihood of agreement by chance[20,22] and can be averaged if more than 2 raters are involved.[2] Kappa values of $0.4 < \kappa \leq 0.6$ were considered to indicate moderate interrater agreement, values of $0.6 < \kappa \leq 0.8$ were interpreted as acceptable interrater agreement, and values of $0.8 < \kappa \leq 1.0$ were considered to demonstrate almost perfect interrater agreement.[22] Statistics were performed with SAS Software version 8.0 (SAS Institute).[23] There were no missing or implausible values. Reliability was calculated for each combination of 2 examiners and for the mean of the 6 examiner combinations (AB, CD, and the 4 experienced/inexperienced couples: AC, AD, BC, and BD). Differences between the experienced (A, B) and inexperienced (C, D) observers were tested by the Wilcoxon signed rank test ($\alpha = 0.05$). The Friedman test ($\alpha = 0.05$) was carried out on κ and ICC values for the 6 examiner combinations to test for statistical differences.

Results

The mean age of the subjects was 40 years (standard deviation 18.5 years; range 19 to 71 years). A total of 16 participants (59%) had been treated for TMD before, with a mean time interval of 2 years (range 1 to 12 years) since the last treatment.

Reasons for visiting the dental clinic were jaw pain (52%), TMJ sounds (30%), and other reasons (18%).

The reliability values of measurement for range of motion for the experienced and inexperienced examiner pairs and the overall results are given in Table 3. The results for the vertical movements were highly reliable, and acceptably reliable results were found for measures of overbite and vertical overlap. The results for the excursive movements were moderately reliable.

Results for pain reported on palpation of muscle sites and joints as well as for joint sounds are listed in Table 4. Some palpation sites had very good reliability, eg, the temporalis muscle, and some had moderate reliability, eg, the condyle. The pain prevalences at various palpation sites (eg, the posterior and anterior temporalis) were too low to obtain interpretable results (data not shown). The authors therefore looked at the sum of those sites and counted the muscle as painful if at least 1 palpation point of the muscle was painful on digital palpation (Table 4). The diagnosis of "myofascial pain," requiring the presence of at least 3 muscle sites tender to palpation, in addition to a patient report of pain, demonstrated overall acceptable reliability ($\kappa = 0.71$). The reliability of joint sounds measured by palpation was moderate to poor, even though crepitus and clicking sounds were grouped together.

There were no significant differences between the AB and CD κ values ($P = .4$) and ICC values ($P = .3$). No single pair of examiners produced more reliable results than the other pairs. Indeed, as shown in Figs 1 and 2, the kappa and ICC values across the 6-pair combinations were widely scattered. They did not show any trend or pattern and did not vary with statistical significance (ICC: $P = 0.3$; κ: $P = .98$).

Discussion

Given the variability of symptoms associated with TMD, their reliable measurement is particularly important. Firstly, for the affected individual, adequate treatment depends on correct diagnosis. Secondly, reliability between different examiners is essential for conducting valid epidemiologic investigations. Since clinical signs of disease may change spontaneously over time and thereby make it difficult to find the same sign on successive examinations, other sources of diagnostic uncertainty (eg, unreliable instruments and observers' subjectiveness) must be minimized. This study investigated the reliability between 4 TMD examiners to determine whether there is a measurement difference between experienced and inexperienced examiners after undergoing the same calibration procedure. The practical question addressed was whether it is more important for prevalence studies to calibrate examiners stringently or to have only examiners with previous experience.

The values for the reliability of experienced examiners in the present study were similar to those found in previously published studies.[1–3,5–7,13–18] Metric variables showed excellent

Table 3 Reliability of Mandibular Movements on the Basis of ICC

No.		AB	CD	Overall
1	Unassisted opening, no pain	0.91	0.78	0.83
2	Maximum unassisted opening	0.87	0.89	0.89
3	Maximum assisted opening	0.92	0.93	0.93
4	Lateral excursion, right side	0.06	0.54	0.41
5	Lateral excursion, left side	0.31	0.47	0.40
6	Horizontal overbite	0.81	0.67	0.79
7	Vertical overlap	0.52	0.89	0.70

Table 4 Reliability of Palpation Measurements on the Basis of Mean Kappa Values

No.		AB	CD	Overall
1/2	Masseter*	0.62/0.46	1.00/0.67	0.78/0.56
3/4	Temporalis*	1.00/1.00	1.00/1.00	0.87/0.91
5/6	Retromandibular region	0.61/0.87	0.37/0.12	0.56/0.50
7/8	Submandibular region	0.78/1.00	0.78/0.46	0.73/0.68
9/10	Lateral pterygoid area	0.50/0.37	0.56/0.40	0.50/0.37
11/12	Tendon of temporalis	0.44/0.62	0.48/0.60	0.53/0.48
13/14	Lateral pole and posterior attachment of TMJ*	0.35/0.44	0.36/0.22	0.43/0.46
15/16	TMJ sounds	0.59/0.05	0.33/0.09	0.52/0.25
17	Myofascial pain**	0.77	0.70	0.71

*Pain reported on digital palpation at at least 1 palpation site.
**Pain reported on digital palpation at at least 3 of 20 muscle sites.

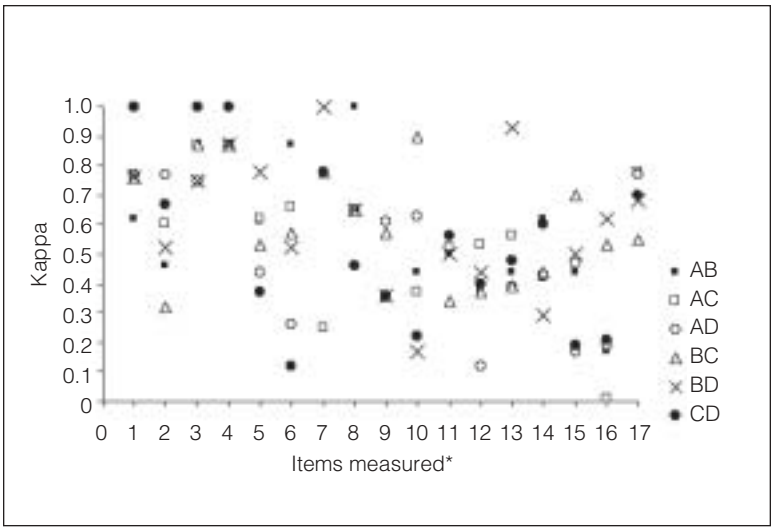

Fig 1 Plot showing widely scattered distribution of the κ values of the 17 measured dichotome variables for all 6 examiner combinations. *Items are listed by number in Tables 2 and 4.

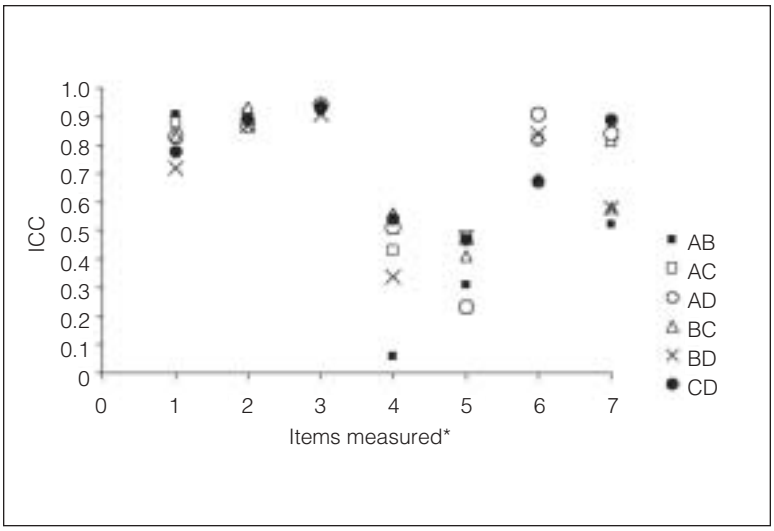

Fig 2 Plot showing widely scattered distribution of the ICC values of the 7 metric variables for all 6 examiner combinations. *Items are listed by number in Tables 2 and 3.

agreement, and and the agreements found for the combined palpation sites were satisfactory. The finding of myofascial pain as 1 of the primary treatment indications showed good agreement for the single examiner combinations and overall. For the other RDC diagnoses, prevalence of combined symptoms was too low to calculate any reliability coefficients. The approach of summing several muscle palpation sites has been previously described as reliable, although different criteria were used.[2,11,17] In accordance with previous studies[1,3,5,6,11,15,19] moderate to poor reliability for joint sounds was found in the present study, even with simplification of the diagnosis of joint sounds.

Although the criteria used for measurement in previous studies were not identical, it can be concluded that the low reliability of measuring joint sounds was not solely because of examiner error. Two studies[4,6] reported fair to good reliability for palpation and auscultation of joint sounds but high reliability for classifying recorded joint sounds by the same observer. This indicates that joint sounds vary over time, and further research should focus on other instruments to properly assess them. The very low reliability coefficients for joint sounds in the left TMJ may be due to a combination of the changing character of the symptom and the low prevalence in the study population. The κ values were low for all 6 examiner pairs but were consistent with the conclusion that there is no difference between experienced and inexperienced examiners.

Most of the previous studies focused on the difference between trained and less-trained or untrained examiners, but they all employed experienced examiners. Duinkerke et al[12] found no dif-

ference between experienced and inexperienced examiners, but their study design differed from the one used here. Their study sample consisted of healthy volunteers with no TMD symptoms, and they based their findings on percent agreement and correlation coefficients, which do not take into account chance agreement. Contrary to the observations of Dahlström et al,[10] the data presented here do not confirm that differences in muscle and joint palpation results between experienced and less experienced examiners are due to palpation pressure, because no significant differences were found in palpation results between the more and less experienced pairs of examiners. There was better reliability for combined muscle sites than single sites, which could be because of differences in finding the same muscular palpation point. However, this difference is more likely due to variation between individual examiners than to experience. In accordance with the presumption of Dahlström et al,[10] the results of the present study suggest that experience with TMD examinations per se does not increase reliability. No significant difference could be found even between very experienced examiners and very inexperienced examiners in terms of ability to find signs of TMD. The present data suggest that calibration has more influence on diagnostic consistency than does experience.

Although this investigation generated some important findings, it was limited by its small sample size, which affects the precision of the estimates. The small number of single symptoms reduced the power of the study. Studies with small sample sizes and low symptom prevalence can also result in misleading κ values.[21,24,25] The failure of Dahlström et al[10] to demonstrate reliability in all categories (using the Cranio-Mandibular Index) may be also due to their small sample size (n = 12) and low symptom prevalence resulting in low κ values; with only 6 subjects suffering from TMD, the power might have been insufficient to produce reliable results. Feinstein[21] showed that the best possible κ results can be obtained when the measured symptom is distributed equally in the study population. Therefore, healthy individuals were not recruited for the study. Because not every TMD patient has all of the symptoms, it was expected that enough controls would be found within the TMD group. Studies examining reliability should make sure that symptom prevalence in their sample is high enough to obtain interpretable reliability values.

Variability between examiners is relevant to the precision and power of statistical analysis of survey data. If misclassification occurred because of observer differences, precision and power would be weakened, which would make the distribution of population parameters more difficult to identify. Our findings suggest that it is not as important that future epidemiologic investigations be limited to experienced investigators as it is to have a thorough calibration lesson prior to the investigation. Also, investigating the reliability of examiners is only possible when symptoms occur frequently enough to be observed.

References

1. 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.
2. 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.
3. Dworkin SF, LeResche L, DeRouen T. Reliability of clinical measurement in temporomandibular disorders. Clin J Pain 1988;4:89–99.
4. Wabeke KB, Spruijt RJ, van der Zaag J. The reliability of clinical methods for recording joint sounds. J Dent Res 1994;73:1157–1162.
5. 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.
6. Westling L, Helkimo E, Mattiasson A. Observer variation in functional examination of the temporomandibular joint. J Craniomandib Disord 1992;6:202–207.
7. John MT, Zwijnenburg AJ. Interobserver variability in assessment of signs of TMD. Int J Prosthodont 2001;14: 265–270.
8. Dworkin SF, LeResche L. Research Diagnostic Criteria for Temporomandibular Disorders: Review, criteria, examinations and specifications, critique. J Craniomandib Disord 1992;6:301–355.
9. List T, Dworkin SF. Comparing TMD diagnoses and clinical findings at Swedish and US TMD centers using Research Diagnostic Criteria for Temporomandibular Disorders. J Orofac Pain 1996;10:240–253.
10. Dahlström 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;52:250–254.
11. 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;21:273–285.
12. 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.
13. Smith JP. Observer variation in the clinical diagnosis of mandibular pain dysfunction syndrome. Community Dent Oral Epidemiol 1977;5:91–93.
14. John M. Prävalenz von kraniomandibulären Dysfunktionen. Dt Zahnaerztliche Zeitschrift 1999;5:302–309.

15. de Wijer A, Lobbezoo-Scholte AM, Steenks MH, Bosman F. Reliability of clinical findings in temporomandibular disorders. J Orofac Pain 1995;9:181–191.
16. Stockstill JW, Gross AJ, McCall WD Jr. Interrater reliability in masticatory muscle palpation. J Craniomandib Disord 1989;3:143–146.
17. Fricton JR, Schiffman EL. Reliability of a craniomandibular index. J Dent Res 1986;65:1359–1364.
18. Carlsson GE, Egermark-Eriksson I, Magnusson T. Intra- and inter-observer variation in functional examination of the masticatory system. Swed Dent J 1980;4:187–194.
19. Hunt RJ. Percent agreement, Pearson's correlation and kappa as measures of inter-examiner reliability. J Dent Res 1986;65:128–130.
20. 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.
21. Feinstein AR. Principles of medical statistics. Boca Raton: Chapman & Hall/CRC, 2002.
22. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–174.
23. SAS Institute. SAS/STAT User's Guide, v 8.0. Cary, NC: SAS Institute, 1999.
24. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol 1990;43:543–549.
25. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol 1990;43:551–558.