Original Article

Reproducibility of Characteristics Assessing the Occlusion of Young Adults

Anna-Liisa Svedström-Oristo, L Odont[a]; Hans Helenius, M Sc[b]; Terttu Pietilä, L Odont, D Odont[c]; Ilpo Pietilä, L Odont[d]; Pentti Alanen, L Odont, D Odont, D Soc Sci[e]; Juha Varrela, L Odont, D Odont[f]

Abstract: The aim of the present investigation was to analyze the reproducibility in the assessment of six morphological and three functional characteristics included in a new method evaluating the occlusion in young adults. These characteristics comprised coincidence of midlines, overjet, overbite, canine relationship, crossbite, scissors bite, recurrent deviation on opening, guided lateral excursions, and discrepancy between the centric relation and the intercuspal position. The study was conducted in three stages: (1) five observers assessed the occlusions of five volunteers, (2) seven observers assessed nine volunteers, and (3) five observers assessed nine volunteers. Two calibrated orthodontists were used as references. For numerical variables, the nonparametric method for repeated measurements (Friedman's test) was used to test the significance of differences, while the proportion of agreement was calculated for categorical assessments. The results were analyzed using two precision levels: within a measurement unit/the same category and an acceptable/nonacceptable dichotomy. The magnitude of systematic differences was small and of minor clinical importance except in measurements of recurrent deviation on opening. The proportional agreement for acceptance was good in the assessment of overjet, coincidence of midlines, crossbite, scissors bite, open bite, and discrepancy between the centric relation and the intercuspal position. Moderate agreement was achieved in the assessment of overbite, canine relationship, recurrent deviation on opening, and guided lateral excursions. Among the nonacceptable cases, the agreement ranged from poor to good. The results indicated that noncalibrated observers assess categorical characteristics inconsistently. (*Angle Orthod* 2002; 72:310–315.)

Key Words: Orthodontics; Clinical assessment; Interobserver variation; Proportional agreement

INTRODUCTION

Occlusal classifications are descriptive tools used by orthodontists and craniofacial biologists for clinical and research purposes. The usefulness of these classifications has, however, been questioned, mainly because they are found to give inconsistent results.[1–3]

The reproducibility of classifications has been tested both in clinical settings[4–8] and using patient records, such as facial and dental photographs, radiographs, or dental casts.[2,9–13] In some studies, clinical data have been combined with data obtained from study models.[14,15] Examiners have variously comprised orthodontists,[2,3,11–14] orthodontists and other specialists,[10] TMD specialists and auxiliary personnel,[7,8,16,17] or general practitioners.[4,9] In general, the results have shown low consistency in assessments of the tested characteristics,[2,4,5,7,11–14,17] but there are findings indicating that specialists can reach an acceptable level of agreement in the assessment of morphological characteristics.[3,10] Of functional assessments, on the other hand, only maximal mouth opening has frequently shown high reproducibility.[5,6,8,16–20] While some investigators have reported that training and calibration of the examiners results in a high level of agreement,[3,7,8,16,17,21] others are sceptical and suggest that these will have only a minor impact on reproducibility.[10,14]

In Finland, free dental care, including orthodontics, is

[a] Specialist in Orthodontics, Department of Oral Development and Orthodontics, Institute of Dentistry, University of Turku, Turku, Finland.

[b] Consulting Biostatistician, Department of Biostatistics, University of Turku, Turku, Finland.

[c] Specialist in Orthodontics, The Health Authority of Pori, Pori, Finland.

[d] Chief Dental Officer, The Health Authority of Pori, Pori, Finland.

[e] Professor of Community Dentistry, University of Turku, Turku, Finland.

[f] Professor of Oral Development and Orthodontics, Institute of Dentistry, University of Turku, Turku, Finland.

Corresponding author: Anna-Liisa Svedström-Oristo, Institute of Dentistry, University of Turku, Lemminkäisenkatu 2, FIN 20520 Turku, Finland
(e-mail: anlisve@utu.fi).

Accepted: February 2002. Submitted: September 2001.

provided on a population basis up to 18 years of age. The health care system is showing increasing interest in the effectiveness, quality, and efficiency of orthodontic treatment, but there are no satisfactory tools that could be applied in occlusal evaluations. Our research group has been developing a method that could be used to assess the occlusions of young adults when studying the targeting and outcome of orthodontic care. A group of specialists in orthodontics and stomatognathic physiology has selected a set of morphological and functional characteristics that would meet the requirements of the health care system and orthodontic professionals in Finland.[22] The aim of the present study was to analyze the reproducibility of the assessment of the selected characteristics.

MATERIALS AND METHODS

The investigation was conducted in three stages. In the first stage, five orthodontists examined five orthodontically treated volunteers. In the second stage, seven observers (three orthodontists and three orthodontically experienced and one inexperienced general practitioner) examined nine orthodontically treated volunteers. In the third stage, five observers (three orthodontists and one experienced and one inexperienced general practitioner) examined a group of nine volunteers including both orthodontically treated and untreated individuals. The examinations were carried out during routine orthodontic follow-up visits or annual dental examinations. In all stages, the volunteers were rated in a random sequence and informed consent was obtained from all of them.

The reproducibility of the assessment of six morphological and three functional characteristics was evaluated. These characteristics were selected using a modified Delphi process. For each characteristic, a group of specialists in orthodontics and stomatognathic physiology had defined a demarcation line for an acceptable–nonacceptable dichotomy. Overbite, canine relationship, crossbite, scissors bite, and guided lateral excursions were assessed categorically, while numerical measurements were taken for the coincidence of the facial midline and midline of the upper dental arch, overjet, recurrent deviation on opening, and discrepancy between the centric relation (CR) and the intercuspal position (ICP) (Table 1). The CR was defined according to Dawson[23] as ''the relationship of the mandible to the maxilla when the properly aligned condyle–disk assemblies are in the most superior position against the eminentia, irrespective of tooth position or vertical dimension.'' Before each stage, all assessment procedures were demonstrated, and detailed instructions were given to the observers. To achieve the CR, a bimanual manipulation technique of the mandible[23] was used during the demonstration. However, the use of this technique was not insisted on; the observers were allowed to use their own methods.

Two orthodontists, who participated in all stages of the study, were calibrated for the assessment of the chosen criteria. During a training session, they independently evaluated 20 dental casts. In case of a disagreement, the cast was reevaluated and the source of disagreement was discussed. Thereafter, both observers clinically assessed the occlusions of 20 randomly selected adolescents. The first five adolescents were assessed together and their recordings were excluded from the analyses. Calibration of other observers was not performed.

Statistical analyses

For the numerical variables, the disagreement between the observers concerning each volunteer was quantified by calculating the average of absolute values of the differences between every pair of observers. The percentage of pairs in which the absolute value of the difference was not more than 1 mm was also calculated. Because the comparisons concerned five to seven observers at the same time and because it was not found appropriate to assume that the distributions of the measurements were normal distributions, the nonparametric method for repeated measurements (Friedman's test) was used to test the significance of differences.[24] *P*-values of less than .05 were interpreted as statistically significant.

For categorical assessments, the proportion of agreement was used to avoid the pitfalls inherent in the intraclass correlation and the kappa coefficient.[21,25,26] Clinically, it is often relevant to be aware of the agreement for both the acceptable and the nonacceptable classifications, especially if there is a low number of observations in one of the categories. Statistical computing was performed using the SAS System for Windows, release 8.1/2000.

RESULTS

At the dichotomous level, the proportion of agreement for acceptance among all observers ranged from moderate to good, while that for the nonacceptable category varied between poor and perfect (Tables 2 through 5). In the acceptable category, the orthodontists achieved a good level of agreement for all numerical variables (Tables 2 and 4).

Although systematic differences were found in numerical measurements among both orthodontists and general practitioners, these differences were of minor clinical importance except in measurements of recurrent deviation on opening. According to this criterion, only 0–22% of volunteers were found to be within one measurement unit (1 mm) by all observers. Further, the mean of the average differences (calculated from the absolute values of differences) was more than twice that of the other criteria (Table 4). Even the measurements made by the calibrated orthodontists indicated a systematic difference at the level of calibration ($P = .04$). Their measurements fell within one measurement unit in 55% of all examined volunteers ($n = 38$).

TABLE 1. Assessments Used in the Reproducibility Study; Numerical Measurements Taken to the Nearest Millimeter With a Ruler [a]

	Acceptability	Reference	Assessment	Conventions
MORPHOLOGY				
1. Coincidence of facial midline and midline of the upper dental arch	Max 3 mm deviation	Frontal plane	Numerical	
2. Overjet	Max 6 mm	CR	Numerical	From the labial surface of d 41 to the labial surface of d 11
3. Overbite	Occlusal contact incisal to the gingival third of the palatal surface of the upper incisors	CR	3 categories Incisal Middle Gingival	Marked with articulating paper
	Open bite only in laterals		Open bite listed in tooth pairs	
4. Canine relationship	Class I Class II in case of missing upper incisors	CR	4 categories Class I Class II Class III Cusp to cusp	
5. Scissors bite	Not accepted	CR	2 categories Present Absent	
6. Crossbite	One tooth pair/side if no interference or slide between CR and ICP	CR	3 categories Absent Present, no slide Present, slide	
FUNCTION				
7. Recurrent deviation on opening	Max 4 mm	Frontal plane	Numerical	Recorded using a toothpick between the lower central incisors; deviation read from a transparent scale paper; repeated at least three times
8. Lateral excursions	Canine protection/group contact	CR	5 categories Canine protection Group contact Contact in incisors, premolars, and molars Contact distal to the canine Other	Guided lateral gliding until upper and lower canines at the same transversal level
9. Discrepancy between CR and ICP		CR	Numerical	Measured from pencil markings in one pair of premolars and incisors
Sagittally and vertically	Max 2 mm			
Laterally	Not accepted			

[a] CR, centric relation [23]; ICP, intercuspal position.

DISCUSSION

In many Finnish health centers, general practitioners, under the supervision of an orthodontist, carry out screening of malocclusions and simple treatment procedures.[27] In these cases, a satisfactory level of agreement between the orthodontists and general practitioners is of importance. All orthodontists participating in our study were familiar with the assessments, and their agreement level was considered to represent the level that could be achieved through training. The accuracy of measuring was set to 1 mm, which was considered adequate for measurements taken directly from the mouth. For a number of reasons, the study was conducted in several stages, with relatively few observers participating in each stage. As the assessment took about 6–7 minutes/observer, we suspected that a larger number of repeated examinations could have affected the volunteers' functional status and distorted the results. Furthermore, the time available for the assessment was limited because it took place during an orthodontic follow-up visit or an annual dental examination. This design made it possible to study samples of both orthodontically treated and untreated occlusions and enabled the inclusion of observers with varying orthodontic backgrounds.

Of all assessments, the widest variability was found in

TABLE 2. Reproducibility of Numerical Morphological Variables Among All Observers; Results Are Shown for All Observers or Separately for Orthodontists and General Practitioners (GPs)

	Average Absolute Difference		Systematic Differences		All Observers			Orthodontists		
	All Observers		Orthodontists	GP		Proportional Agreement			Proportional Agreement	
	Mean (mm)	SD	P[a]	P[a]	±1 mm[b] (%)	Acceptable[c]	Nonacceptable[c]	±1 mm[b] (%)	Acceptable[c]	Nonacceptable[c]
Overjet	0.20–0.51	0.15–0.35	.047*–.663	<.001***–.564	78–100	.93–1.00	.14–1.00	100	.93–1.00	.00–1.00
Coincidence of midlines	0.60–0.91	0.23–0.36	.018*–.594	.232–.655	33–80	.90–1.00	.25–1.00	80–89	.93–1.00	.00–1.00

[a] Friedman's test; * $P < .05$, ** $P < .01$, *** $P < .001$.
[b] Among all subjects and all observers, the percentage of pairs differing not more than 1 mm.
[c] The categories of acceptable and nonacceptable defined as in Table 1; <.40 = poor, .40–.75 = moderate, >.75 = good agreement.

TABLE 3. Reproducibility of Categorical Morphological Variables Among All Observers; Results Are Shown for All Observers or Separately for Orthodontists

	All Observers			Orthodontists		
	Same category[a] (%)	Proportional Agreement		Same category[a] (%)	Proportional Agreement	
		Acceptable[b]	Nonacceptable[b]		Acceptable[b]	Nonacceptable[b]
Scissors bite	78–89	.92–.96	.00–.09	80–100	.92–1.00	.00–1.00
Crossbite	78–89	.91–.97	.00–.78	80–89	.90–.92	.00–.78
Open bite	67–100	.87–1.00	.00–1.00	78–100	.85–1.00	.00–1.00
Overbite	20–33	.73–1.00	.56–1.00	20–89	.70–1.00	.54–1.00
Right canine relationship	22–80	.47–.85	.52–.65	56–80	.56–.85	.53–.65
Left canine relationship	22–80	.45–.84	.41–.85	33–80	.33–.84	.43–.75

[a] Among all subjects and all observers, the percentage of pairs classified in the same category.
[b] The categories of acceptable and nonacceptable defined as in Table 1; <.40 = poor, .40–.75 = moderate, >.75 = good agreement.

TABLE 4. Reproducibility of Numerical Functional Variables Among All Observers: Results Are Shown for All Observers or Separately for Orthodontists and General Practitioners (GPs)

	Average Absolute Difference		Systematic Differences		All Observers			Orthodontists		
	All Observers		Orthodontists	GP		Proportional Agreement			Proportional Agreement	
	Mean (mm)	SD	P[a]	P[a]	±1 mm[b] (%)	Acceptable[c]	Nonacceptable[c]	±1 mm[b] (%)	Acceptable[c]	Nonacceptable[c]
Slide sagittally	0.29–0.58	0.21–0.42	.021*–1.000	.014*–.223	89–100	1.00	1.00	100	1.00	1.00
Slide vertically	0.20–0.58	0.19–0.27	.006**–.788	.112–.564	89–100	1.00	1.00	100	1.00	1.00
Slide laterally	0.00–0.38	0.00–0.23	.406–1.000	1.00	100	.57–1.00	.24–1.00	100	.77–1.00	.14–1.00
Recurrent deviation	1.69–1.96	0.36–0.68	.003**–.121	.005**–.013*	0–22	.68–.92	.00–.05	0–67	.85–1.00	.00–1.00

[a] Friedman's test: * $P < .05$, ** $P < .01$, *** $P < .001$.
[b] Among all subjects and all observers, the percentage of pairs differing not more than 1 mm.
[c] The categories of acceptable and nonacceptable defined as in Table 1; <.40 = poor, .40–.75 = moderate, >.75 = good agreement.

measurements of recurrent deviation on opening. This finding is in line with earlier studies, in which the reproducibility of categorically assessed jaw opening patterns has ranged from poor to good.[16–18,20,28] It is possible, however, that the high variation in recurrent deviation on opening does not reflect differences in technical management but rather exemplifies the instability of the characteristic.[8,17,18,28]

As in earlier studies,[2,3,11,13,16,17] the classification of canine relationship was found to be ambiguous. Given that the sagittal measurements were reproduced with high precision, it is unlikely that the observed discrepancies in canine classification could be assigned to variation in mandibular position. Instead, it is possible that not all observers were familiar with applying the Angle's classification to canines. It is also possible that the observers did not use the same viewing angle when assessing the buccal segment occlu-

TABLE 5. Reproducibility of Categorical Functional Variables Among All Observers; Results Are Shown for All Observers or Separately for Orthodontists

	All Observers			Orthodontists		
	Same category[b] (%)	Proportional Agreement		Same category[b] (%)	Proportional Agreement	
		Acceptable[c]	Nonacceptable[c]		Acceptable[c]	Nonacceptable[c]
Ltr[a]/ right	0–40	.60–1.00	.11–1.00	33–44	.56–1.00	.00–1.00
Ltr[a]/ left	0–40	.67–.90	.18–.71	40–56	.69–.90	.11–.71

[a] Ltr, laterotrusion.
[b] Among all subjects and all observers, the percentage of pairs classified in the same category.
[c] The categories of acceptable and nonacceptable defined as in Table 1; $<.40$ = poor, .40–.75 = moderate, $>.75$ = good agreement.

sion,[29,30] which might explain some of the observed variation. In borderline cases, the differences may have arisen from judgmental variation[31] based on differing interpretations of Angle's classes. Practical training, together with clear instructions and well-defined demarcation lines, would probably increase the reproducibility of the classification of this characteristic.

When measured in millimeters, overbite has been shown to have good reproducibility.[17] In line with the present results, the agreement in categorical assessments has varied between moderate and good.[3,10,13] However, in our study, the percentages of exact agreement (within the same category) indicated a wider variability than was found by Keeling et al.[3]

CONCLUSIONS

The agreement among all observers concerning the acceptable category was good in the assessment of overjet, coincidence of midlines, crossbite, scissors bite, open bite, and discrepancy between the CR and the ICP. Moderate agreement was achieved in the assessment of overbite, canine relationship, and guided lateral excursions.

In the nonacceptable category, the variability in agreement may partly reflect the low number of observations in this group.

Exact agreement in categorical assessments was highly variable.

The reproducibility of measurements of recurrent deviation on opening was poor, as described by the relatively high mean of the average absolute differences and by the low percentage of pairs within one measurement unit.

ACKNOWLEDGMENTS

The authors wish to thank the chief dental officers and their staff in the municipal health centers of Pori and Rauma and the staff at the Department of Oral Development and Orthodontics in the Institute of Dentistry, University of Turku, for their cooperation and assistance in conducting the study. We are grateful to Mr Heikki Hiekkanen for performing the statistical analyses, and we warmly thank all the participating volunteers. This study was supported by a grant from the Emil Aaltonen Foundation.

REFERENCES

1. Rinchuse DJ, Rinchuse DJ. Ambiguities of Angle's classification. *Angle Orthod.* 1989;59:295–298.
2. Katz MI. Angle classification revisited 1: is current use reliable? *Am J Orthod Dentofac Orthop.* 1992;102:173–179.
3. Keeling SD, McGorray S, Wheeler TT, King GJ. Imprecision in orthodontic diagnosis: reliability of clinical measures of malocclusion. *Angle Orthod.* 1996;66:381–392.
4. 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.
5. 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.
6. Nielsen L, Melsen B, Terp S. Clinical classification of 14–16-year-old Danish children according to functional status of the masticatory system. *Commun Dent Oral Epidemiol.* 1988;16:47–51.
7. 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.
8. de Wijer A, Lobbezoo-Scholte AM, Steenks MH, Bosman F. Reliability of clinical findings in temporomandibular disorders. *J Orofacial Pain.* 1995;181–191.
9. Lewis EA, Albino JE, Cunat JJ, Tedesco LA. Reliability and validity of clinical assessments of malocclusion. *Am J Orthod.* 1982; 81:473–477.
10. Phillips C, Bailey LT, Sieber RP. Level of agreement in clinicians' perceptions of class II malocclusions. *J Oral Maxillofac Surg.* 1994;52:565–571.
11. Baumrind S, Korn EL, Boyd RL, Maxwell R. The decision to extract: part 1—interclinician agreement. *Am J Orthod Dentofacial Orthop.* 1996;109:297–309.
12. Du SQ, Rinchuse DJ, Zullo TG, Rinchuse DJ. Reliability of three methods of occlusion classification. *Am J Orthod Dentofacial Orthop.* 1998;113:463–470.
13. Luke LS, Atchison KA, White SC. Consistency of patient classification in orthodontic diagnosis and treatment planning. *Angle Orthod.* 1998;68:513–520.
14. Gravely JF, Johnson DB. Angle's classification of malocclusion: an assessment of reliability. *Br J Orthod.* 1974;1:79–86.
15. Buchanan IB, Downing A, Stirrups DR. A comparison of the Index of Orthodontic Treatment Need applied clinically and to diagnostic records. *Br J Orthod.* 1994;21:185–188.
16. Dworkin SF, LeResche L, DeRouen T. Reliability of clinical measurement in temporomandibular disorders. *Clin J Pain.* 1988;4: 89–99.
17. 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.
18. Kopp S. Constancy of clinical signs in patients with mandibular dysfunction. *Commun Dent Oral Epidemiol.* 1977;5:94–98.
19. Westling L, Helkimo E, Mattiasson A. Observer variation in functional examination of the temporomandibular joint. *J Craniomandib Disord Facial Oral Pain.* 1992;6:202–207.
20. Wahlund K, List T, Dworkin SF. Temporomandibular disorders in children and adolescents: reliability of a questionnaire, clinical examination, and diagnosis. *J Orofacial Pain.* 1998;12:42–51.
21. Grant JM. The fetal heart rate trace is normal, isn't it? Observer agreement of categorical assessments. *Lancet.* 1991;337:215–218.
22. Svedström-Oristo A-L, Pietilä T, Pietilä I, Alanen P, Varrela J. Morphological, functional and aesthetic criteria of acceptable mature occlusion. *Eur J Orthod.* 2001;23:373–381.
23. Dawson PE. *Evaluation, Diagnosis, and Treatment of Occlusal Problems,* 2nd edition. St Louis, Mo: Mosby; 1989:29, 41–44.
24. Lehmann EL. *Nonparametrics: Statistical Methods Based on Ranks.* San Francisco: Holden Day Inc; 1975:120–201, 260–285.
25. Haas M, Nyiendo J, Peterson C, Thiel H, Sellers T, Cassidy D, Young-Hing K. Interrater reliability of roentgenological evaluation of the lumbar spine in lateral bending. *J Manipulative Physiol Ther.* 1990;13:179–189.
26. Pett MA. *Nonparametric Statistics for Health Care Research, Statistics for Small Samples and Unusual Distributions.* London: Sage Publications; 1997:237–248.
27. Pietilä T, Pietilä I, Widström E, Varrela J, Alanen P. Extent and provision of orthodontic services for children and adolescents in Finland. *Commun Dent Oral Epidemiol.* 1997;25:150–155.
28. Smith JP. Observer variation in the clinical diagnosis of mandibular pain dysfunction syndrome. *Commun Dent Oral Epidemiol.* 1977;5:91–93.
29. Scivier GA, Menezes DM, Parker CD. A pilot study to assess the validity of the orthodontic treatment priority index in English schoolchildren. *Community Dent Oral Epidemiol.* 1974;2:246–252.
30. Richmond S, Shaw WC, O'Brien K, Buchanan IB, Jones R, Stephens CD, Roberts CT, Andrews M. The development of the PAR Index (Peer Assessment Rating): reliability and validity. *Eur J Orthod.* 1992;14:125–139.
31. Kay E, Nuttal N. Clinical decision making—an art or a science? Part II: making sense of treatment decisions. *Br Dent J.* 1995; 178:113–116.