Imprecision in orthodontic diagnosis: Reliability of clinical measures of malocclusion

Stephen D. Keeling, DDS, MS; Susan McGorray, PhD; Timothy T. Wheeler, DDS, PhD; Gregory J. King, DMD, DMSc

The current national debate regarding health care, especially the concept of rationing care to the most deserving in a society with limited resources, has rekindled interest among United States orthodontists in determining the need for orthodontic care. Numerous methods have been developed over the past 40 years to assess treatment priority and malocclusion severity. These include several frequently-referenced methods: the Treatment Priority Index (TPI),[1] the Occlusal Index (OI),[2] the Peer Assessment Rating (PAR) index,[3] Handicapping Labio-lingual Deviations (HLD) index,[4] the Handicapping Malocclusion index,[5] and the Index of Orthodontic Treatment Need (IOTN).[6] All are based on specifying the magnitude of deviation from normal occlusion. None has gained wide acceptance in the U.S. for determining treatment need.

These methods of assessing treatment need and malocclusion severity are based, in varying degrees, upon a description of individual morphologic malocclusion traits. Generalized use of an index by individual members of the specialty or by reviewers employed by health alliances would depend, in part, upon the reliability of the various descriptors of malocclusion. Havoc would ensue, if, for example, treatment need depended on a patient having an end-on molar relationship and there was poor consensus on an end-on classification. Such uncertainty would necessitate a rigid calibration of the decision-makers or an improved methodology to describe malocclusion.

Abstract

The study examined the reliability among seven orthodontists in judging dental and facial aspects of malocclusion in a screening of elementary schoolchildren. Data included measures typically recorded during a clinical orthodontic examination: facial assessment of skeletal/incisor relationships and individual measures of morphologic malocclusion. Interexaminer reliability data were collected on 52 children. Pairwise comparisons between orthodontists were made using exact percent agreement and agreement within one category. Kappa statistics and one-sided Z-tests were used to evaluate observed agreement compared with agreement that would be expected by chance. Median Kappa statistics indicated that the reliability of maxillary and mandibular anteroposterior positions, incisor exposure, interlabial gap, and maxillary crowding was poor (K<0.40). Acceptable reliability existed for mandibular anterior crowding, facial convexity, overbite, overjet, and molar classification (median Kappas ranged from 0.48 to 0.72). Excellent reliability existed only for evaluating the presence of a posterior crossbite (K=0.79). The results caution that the language of clinical orthodontic diagnosis is imprecise.

Key Words

Reliability • Morphologic malocclusion • Facial features • Kappa statistics

Submitted: December 1994 **Revised and accepted:** August 1995 Angle Orthod 1996;66(5):381-392.

Health care providers are often unaware of the imperfect reliability of the methods and data of clinical practice.[7] Reliability of clinical measures in dentistry (periodontal health assessment, caries detection) remains as a major challenge to diagnosis, treatment planning, and assessing the outcome of differing therapies.[8,9] In clinical orthodontics, the assessment of malocclusion remains problematic.[10-13] Although reliability and validity are separate and distinct qualities, most clinicians would agree that when clinical findings can be determined with high agreement, the resultant diagnoses are more likely to be valid.[13]

The TPI, OI and PAR index scores have been shown to have acceptable reliability.[3,14,15] However, the reliability of the individual measures that compute these index scores has not been reported. Even the report from the largest study of malocclusion in children conducted by the U.S. Public Health Service using multiple examiners to determine TPI scores did not provide data describing interjudge reliability of the individual measures assessed.[16] Improvements in developing an index of malocclusion should begin with an understanding of the language used by orthodontists to describe individual components of malocclusion.

Unlike index scores, terms describing individual malocclusion traits, such as *molar class* and *overbite*, remain the language of clinical orthodontics. How precise is this language? Studies assessing the reliability of scoring components of malocclusion have most frequently been performed on diagnostic records (e.g., dental casts, photographs, radiographs).[3,11,12,17,18] Thus, reports on the reliability of assessing individual malocclusion traits during the clinical examination of subjects over the past 25 years are scarce but noteworthy.

The interexaminer reliability of Angle's classification scheme (Class I; Class II, Division 1; Class II, Division 2; and Class III) during clinical examination was judged in an early study[12] so poor to be of doubtful value to clinicians. In contrast, the classification of posterior occlusion (Class I, II, or III) was performed with good reliability by hygienists, after retraining, but poorly by experienced, uncalibrated dentists; both groups demonstrated excellent reliability in assessing overjet and overbite in millimeters.[13] A high level of intraexaminer consistency has been shown for malocclusion traits when the traits are recorded as dichotomized variables from dental casts.[11]

The reliability of judging children's facial profiles (orthognathic, convex, concave) from cephalometric tracings and facial photographs has been shown to be acceptable.[19] Reports on the reliability of clinical assessments of facial characteristics do not exist.

The purposes of this study were to determine if differences exist between examiners in describing individual components of malocclusion in children, and to identify clinically-determined measures of malocclusion that have poor interexaminer reliability.

Materials and methods

This study assessing between-examiner reliability of determining malocclusion traits was conducted during the initial phase of a prospective randomized, controlled trial (RCT) of early Class II treatment. In order to identify subjects for the RCT, a county-wide screening of third and fourth grade school children in the 23 public elementary schools of Alachua County, Florida, was performed by seven orthodontists from January 1990 to February 1992. Reliability testing occurred during this time.

Training of examiners

Prior to the start of the screenings, six examining orthodontists met over a period of 3 months to determine data to be collected, to design the screening form, and to practice using it. These experiences served as an internal calibration. An additional examiner (orthodontist #7) was added to the team a year later and did not participate in the original planning, design, and training.

Examination procedures

Examinations were performed in a room separated from the classroom (typically, a media center or empty classroom) with the student standing erect in front of a seated examiner. Each examiner used a hand- or head-held light, a millimeter ruler, gloves, and a tongue depressor for cheek retraction. For each child, the examiner completed a standardized screening form, which included demographic and malocclusion variables.

Demographic information included age, gender, race, and history of prior orthodontic care. If the student had not received previous orthodontic treatment, information on the following variables was obtained, with the child in habitual (centric) occlusion:

- Facial convexity: classified as Class I, Class II, or Class III, based on visual lateral assessment of the facial profile
- Maxillary anteroposterior (AP) position: classified as retrognathic, orthognathic, or prognathic, based on visual assessment
- Mandibular AP position: classified as retrognathic, orthognathic, or prognathic,

based on visual assessment

- Posterior crossbite: coded as present if any posterior teeth were in crossbite, or absent
- Incisor exposure: measured to the nearest millimeter with hand-held ruler with the lips at rest
- Interlabial gap: measured to the nearest millimeter with hand-held ruler with the lips at rest
- Overjet: measured to the nearest millimeter with hand-held ruler
- Overbite: classified as percent of lower incisor coverage: none (or negative), >0% and ≤33%, >33% and ≤66%, >66% and ≤100%, >100%
- Molar classification, right: the Angle classification system was used as follows: 0 = greater than full cusp Class II
  1 = full cusp Class II
  2 = 3/4 cusp Class II
  3 = 1/2 cusp Class II
  4 = 1/4 cusp Class II
  5 = Class I
  6 = 1/4 cusp Class III
  7 = 1/2 cusp Class III
  8 = 3/4 cusp Class III
  9 = full cusp Class III
  10 = greater than full cusp Class III
- Molar classification, left: same groupings as above
- Anterior crowding, upper: coded as
  0 = >6 millimeters space
  1 = >3 and ≤6 mm. space
  2 = >0 and ≤3 mm. space
  3 = no space or crowding
  4 = >0 and ≤3 mm. crowding
  5 = >3 and ≤6 mm. crowding
  6 = >6 mm crowding
- Anterior crowding, lower: classified as above

Collection of reliability data

Screenings were performed during three time periods, coinciding with the school year semesters: spring 1990, fall 1990, and fall 1991. At the beginning of each time period, data were collected to assess interjudge reliability; 18 students were examined during period one, 21 during period two, and 13 during period three, for a total of 52. These "reliability" students were randomly selected from among the students screened at each school and were examined along with regularly screened students who were not selected for reliability testing. At each session, only the reliability students were examined by each of the orthodontists present at the school that day; these students were directed for repeated examinations by the staff assistant. Following each reliability session, results were tabulated and differences were discussed. For the first two time periods, the six orthodontists who had trained together performed the screenings. In the third time period, two of the original orthodontists were not available, and the new orthodontist (#7) joined the group.

Data handling

After the orthodontists examined the children, the screening forms were examined for completeness by a staff assistant and children were reexamined as necessary. The data were double entered in a data management software package, operating on a microcomputer. Discrepancies between entries were corrected after reviewing the original data entry form.

Statistical analysis

The goals of the statistical analysis were twofold: (1) to identify differences between the seven orthodontists, and (2) to identify variables with poor reliability. To summarize the data, percent agreement was calculated for each of the 19 possible pairs of orthodontists. Pairwise comparisons were used because different pairs of orthodontists reviewed different numbers of students; four orthodontists reviewed all 52 students, two orthodontists reviewed 39 students, and one orthodontist reviewed 13 students. For continuous variables or variables with many ordered categories, exact agreement was determined, along with agreement within one category. For example, for the continuous measure of overjet, agreement within one category meant that calls of 5 mm and 6 mm were considered in agreement; for the ordinal measure of molar class, agreement within one category meant that calls of 1/4 cusp Class II and 1/2 cusp Class II were considered in agreement.

When appropriate, Kappa statistics[20] were used to evaluate observed agreement compared with agreement that would be expected by chance, with weighted Kappa statistics[21,22] used for ordinal variables. Weighted Kappas allow differential weighting of disagreements, because differences of three units of a measure are worse than differences of one unit. One-sided Z-tests were performed to determine the orthodontist pairs for whom agreement greater than that due to chance was present. Kappa values equal to zero represent agreement equivalent to that expected by chance and 1.00 represents perfect agreement. It has been suggested that Kappa values of less than 0.40 be interpreted to represent poor agreement beyond chance, 0.40 to 0.74 represent fair to good (moderate) agreement, and 0.75 and above excellent agreement.[23-25] The re-

Table 1
Characteristics of the reliability sample students and of all screened students.

	Reliability sample	Screened students
Sample Size	52	6281
% Male	48.08	51.43
% White	75.00	62.25
Mean Age	9.38	9.30
S.D. Age	0.77	0.76

sult and discussion sections will use this interpretation.

To determine if individual orthodontists differed from the others, William's index[6] was calculated. This index is based on exact agreement and compares the average agreement of one orthodontist with all others to the average agreement of other pairs of orthodontists. A value of 1.00 indicates the agreement of the orthodontist under consideration was similar to that of the rest of the group. Using a jackknife estimate of variance,[26] 95% confidence intervals were constructed; an interval with an upper bound of less than 1.00 indicates the orthodontist differs from the group.

The target sample size for each session was 20 students. This was based on consideration of agreement between two orthodontists regarding a binary variable. A sample size of 20 yields adequate power to detect moderate agreement, with a probability of 0.72 of detecting an underlying Kappa value of 0.50. Using the combined sample size of 52, there is good power (0.88) to detect fair (or better) agreement (underlying Kappa value of 0.40).

All procedures were performed on a Unix workstation using SAS[27] and S-PLUS® statistical software.[28] For all analyses, a P value of 0.05 or less was considered statistically significant.

Results

Characteristics of the reliability sample students and all screened students are shown in Table 1. This report focuses on the interexaminer measurements obtained from the reliability sample. As seen in the table, the randomly selected reliability sample students were similar in age, gender, and race to the population of students from which they were selected.

Nineteen pairs of orthodontists reviewed students in common; six pairs reviewed all 52 students, nine pairs reviewed 39 students, and four pairs reviewed 13 students in common. Separate analyses were performed on the data from each of the three reliability sessions; results were similar in each session. The combined results from all three sessions are presented here.

A summary of exact agreement and agreement within one category for each variable is presented in Table 2; minimum, median, and maximum values are listed. As shown, based on consideration of median values, the highest exact agreement (97.44%) occurred for the dichotomous (yes/no) measure of posterior crossbite. The continuous measures of incisor exposure (median = 20.51%), overjet (46.15%), and interlabial gap (48.72%) had lower exact agreement levels than the ordinally scored measures of anterior crowding (upper, 50.00%; lower 57.89%), molar classification (right, 55.56%; left, 68.42%), mandibular AP position (64.85%), overbite (69.23%), and maxillary AP position and facial convexity (both 76.92%). Agreement improved considerably when agreement within one category was allowed. (This was only possible when there were more than 2 ordered categories). As shown, median agreements within one category were greater than 87% for overjet, overbite, molar classification, and anterior crowding. Incisor exposure and interlabial gap had the lowest median agreements within one category (<65%).

Since Kappa statistics are affected by unbalanced marginal distributions, these distributions were examined and reasonable balance was observed. This is displayed in Table 3, with mean and standard deviation estimates used to summarize the distributions for each variable and each orthodontist. In this table, categorical variables are coded sequentially (e.g., maxillary AP position is coded retrognathic = 1, orthognathic = 2, prognathic = 3).

Table 4 presents a summary of the Kappa statistics with a column indicating the number of pairs of orthodontists for which we could not detect agreement better than chance. Incisor exposure and interlabial gap were deleted from the form before the final set of screenings, hence a reduced number of pairs of orthodontists is available for these variables. Kappa statistics cannot be calculated when there is no variability in the data (for example, both orthodontists agree that no student reviewed in common had a posterior crossbite). For 10 of 19 orthodontist pairs, we could not detect agreement better than that expected by chance for determining maxillary AP position (retro, ortho, prog). This is in contrast to the findings for posterior crossbite, overjet, incisor exposure, molar classification, and lower anterior crowding where all orthodontist pairs had agreement significantly better than

Table 2
Pairwise percent exact agreement, and percent agreement within one category.

	No. pairs orthodontists	Exact agreement Median	Minimum	Maximum	Agreement within 1 category Median	Minimum	Maximum
Facial convexity	19	76.92	69.23	92.31			
Maxillary AP position	19	76.92	64.10	100.00			
Mandibular AP position	19	64.85	46.15	78.85			
Posterior crossbite	19	97.44	94.23	100.00			
Overbite	19	69.23	51.28	79.49	97.44	92.31	100.00
Overjet	19	46.15	23.07	64.10	92.31	82.69	97.44
Incisor exposure	15	20.51	7.69	76.92	58.97	38.46	89.74
Interlabial gap	15	48.72	5.13	64.10	64.10	25.64	76.92
Molar class, right	19	55.56	46.00	76.00	96.08	84.21	100.00
Molar class, left	19	68.41	46.15	81.08	94.74	86.84	100.00
Anterior crowding, upper	19	50.00	34.21	64.10	87.18	75.00	94.18
Anterior crowding, lower	19	57.89	30.77	66.67	92.31	76.92	100.00

chance. Of the 26 total pairs where we could not detect agreement greater than expected by chance, 14 cases involved the orthodontist (#7) who had not participated in the initial planning and had evaluated fewer students (individual orthodontist data not shown).

As shown in Table 4, median Kappa scores indicated that, while the facial convexity variable had moderate agreement (0.48), measures of maxillary and mandibular AP positions had poor agreement (0.22 and 0.25). The dichotomous variable posterior crossbite, with a median Kappa score of 0.79, was characterized as having excellent agreement. The median weighted Kappa statistics suggested poor agreement for incisor exposure (0.24), interlabial gap (0.26) and upper anterior crowding (0.36); moderate agreement for lower anterior crowding (0.45), overbite (0.59), overjet (0.67), and molar classification (right, 0.68; left, 0.72).

Molar classification discrepancies were further examined by plotting each orthodontist's decision for each student. Recall that molar class is coded in 1/4 cusp increments and that all seven orthodontists did not examine each child. The data for left molar class are displayed in Figure 1. Each line represents one student, numbered arbitrarily, and each orthodontist's decision is identified. All orthodontists agreed exactly for 23 of the 52 students (44%). A range of two classifications was observed for 17 students (33%), and a range of three classifications was observed for 11 students (21%). The most extreme case of a range of five classifications was observed for one student (2%). Also note that of the 29 students for which there were disagreements, 13 (45%) involved only one orthodontist disagreeing with all others.

Discrepancies in anterior crowding were also further examined by plotting each orthodontist's decision for each student. The decisions for mandibular anterior crowding, upper, are displayed in Figure 2. Note that exact agreement among all orthodontists was present for only five of the 52 students (10%). A range of two classifications was observed for 25 students (48%), and a range of three classifications was observed for 12 students (23%). One student's crowding was classified in 5 of the 7 possible categories, including the two most extreme ones.

Based on 95% confidence intervals for William's index (Table 5), agreement between each orthodontist and the others was consistent for facial convexity, posterior crossbite, molar classification, and anterior crowding classification. Differences were detected between one orthodontist and the rest of the group for maxillary and mandibular AP position and overbite; however, the estimates of William's index are all greater than 0.80 in these cases. For overjet, agreement percentages differed significantly for one orthodontist, with an estimated William's index of 0.70 (95% confidence interval [0.49, 0.90]). This orthodontist had the highest mean overjet value (Table 3). If we consider "exact" agreement as ±1 mm, the William's index for this

Table 3
Marginal distributions of orthodontic variables. Mean and standard deviation are listed.

Orthodontist (n)	1 (52)	2 (52)	3 (52)	4 (52)	5 (39)	6 (39)	7 (13)
Facial convexity							
mean	1.48	1.29	1.37	1.31	1.49	1.36	1.38
s.d.	0.57	0.50	0.49	0.47	0.51	0.54	0.51
Maxillary AP position							
mean	2.08	2.27	2.02	2.12	2.33	2.05	2.31
s.d.	0.39	0.45	0.14	0.32	0.48	0.32	0.48
Mandibular AP position							
mean	1.70	1.96	1.63	1.79	1.85	1.69	1.77
s.d.	0.58	0.34	0.49	0.50	0.67	0.52	0.73
Posterior crossbite							
mean	0.04	0.06	0.08	0.06	0.08	0.08	0.00
s.d.	0.19	0.24	0.27	0.24	0.27	0.27	0.00
Overbite							
mean	1.58	1.65	1.60	1.42	1.64	1.38	1.69
s.d.	1.04	0.97	0.80	0.85	0.87	0.91	0.75
Overjet							
mean	3.44	3.96	3.88	3.98	3.33	3.49	3.15
s.d.	2.30	2.43	2.45	2.11	2.64	2.51	1.68
Incisor exposure							
mean	0.77	0.62	2.51	1.05	2.28	0.97	—
s.d.	1.51	1.43	1.59	1.99	1.26	1.40	—
Interlabial gap							
mean	1.56	0.72	1.49	1.46	2.95	1.08	—
s.d.	2.35	1.61	1.97	2.62	1.19	1.75	—
Molar class, right							
mean	4.20	4.16	4.25	4.20	4.33	4.38	4.46
s.d.	1.37	1.25	1.36	1.18	1.26	1.27	1.45
Molar class, left							
mean	3.78	4.04	4.23	4.26	4.13	4.31	4.46
s.d.	1.51	1.25	1.41	1.24	1.65	1.30	1.13
Anterior crowding, upper							
mean	2.65	2.85	2.55	2.88	3.10	2.85	2.92
s.d.	1.37	1.50	0.99	0.94	0.94	0.87	1.19
Anterior crowding, lower							
mean	3.56	3.35	3.45	3.37	3.85	3.62	3.85
s.d.	1.04	1.27	1.21	1.09	0.87	0.75	0.80

orthodontist was 0.97. Discrepancies are apparent in judging incisor exposure, with two orthodontists differing from the rest (William's indices of 0.47 and 0.49). A serious discrepancy was detected between one orthodontist and the rest for assessing interlabial gap, with an estimated William's index of 0.15 (95% confidence interval [0.06, 0.25]). This is apparent in Table 3, with orthodontist 5 having a mean value 1.4 mm larger than any other orthodontist's mean value.

We also considered the possibility that discrepancies could have been related to mandibular positioning (centric relation/centric occlusion discrepancies). Two orthodontists might disagree on both molar classification and overjet because of differences in registering mandibular position (for example, one orthodontist using a more retruded mandibular position than an-

other). Table 6 presents cross-classified data for molar class discrepancy and overjet discrepancy. All decisions for all pairs of orthodontists are included in this table. Overjet discrepancy was defined as the absolute value of the difference between the two orthodontists' overjet calls. Molar class discrepancy was defined as the sum of the absolute value of the right and left differences. For example, if orthodontist A recorded molar class right and left as 3 and 2, while orthodontist B coded 4 and 1, the molar class discrepancy for this pair is 2 (abs[3-4] + abs[2-1]).

As shown in Table 6, 135 pairs of orthodontists agreed exactly, with no discrepancy in molar class or overjet; 96 pairs had a discrepancy of 1 unit in molar class decisions and 1 mm in overjet calls; 152 pairs had no discrepancy in molar class decisions but 1 mm discrepancy in overjet decisions; and 103 pairs had 1 unit of molar class discrepancy, but no discrepancy in overjet decisions. If a discrepancy in molar class was related to a discrepancy in overjet, we might expect no discrepancy in molar class to correspond with no discrepancy in overjet, and a large discrepancy in overjet to correspond with a large discrepancy in molar class. However, the Spearman's rank correlation coefficient estimate for this data is -0.06, suggesting that this is not the case. Since differences in positioning might occur between a specific pair of orthodontists, we tested each pair separately to assess whether a linear relationship was present (i.e., larger discrepancy in overjet associated with larger discrepancy in molar class); this was not detected in 18 of the 19 pairs of orthodontists. (The same procedure was used to evaluate molar class-overbite discrepancies; this did not appear to be a common occurrence.) These findings suggest that differences between orthodontists were not related to differences in mandibular positioning.

Table 4
Kappa statistics assessing pairwise agreement

	No. pairs ortho-dontists	No. pairs chance agreement*	Kappa Median	Minimum	Maximum
Facial convexity	19	4	0.48	0.30	0.85
Maxillary AP position	19	10	0.22	-0.06	1.00
Mandibular AP position	19	5	0.25	0.02	0.50
Posterior crossbite	15	0	0.79	0.37	1.00
Overbite**	19	1	0.59	0.29	0.78
Overjet**	19	0	0.67	0.45	0.84
Incisor exposure**	15	0	0.24	0.15	0.60
Interlabial gap**	15	4	0.26	0.05	0.45
Molar class, right**	19	0	0.68	0.47	0.81
Molar class, left**	19	0	0.72	0.39	0.82
Anterior crowding, upper**	19	2	0.36	0.22	0.51
Anterior crowding, lower**	19	0	0.45	0.23	0.58

* Number of pairs of orthodontists in which agreement was not statistically significantly better than that expected by chance
** Weighted Kappa statistic used

Discussion

Third and fourth grade school children in Alachua County, Florida, were screened to identify subjects for a prospective clinical trial evaluating early Class II orthodontic treatment. The variables of interest that were used as inclusion/exclusion criteria for the clinical trial (molar relationship, overjet, and overbite) were assessed during the screening procedure. Additional data were collected to characterize the population.

Since the orthodontists were not randomly selected from a larger pool, the results on reliability of individual measures cannot be generalized to all practicing orthodontists. However, the results are the only available estimates of what a large-scale study might find. Due to the number of variables under consideration, the number of orthodontists, and the varying number of subjects each orthodontist examined, a large number of statistical tests were performed. Because of this, type 1 errors (rejecting the null hypothesis when it is true) likely occurred. Also, because of the varying sample sizes, power differed over the comparisons, with less power to detect differences from the null hypothesis for orthodontist 7. When considering the reliability of the variables under review, results of all estimation (e.g., percent agreement, William's index) and testing were considered, and our conclusions are not based on the occurrence of a single significant test result.

Six of the seven examining orthodontists met to design the screening form and participated in training sessions to review the forms and examination procedures. Orthodontist 7, who generally had poorer agreement with the other orthodontists, did not train with them. This emphasizes the importance of joint training when collecting clinical data involving multiple examiners, and supports the observations of others.[13,29] We did not detect agreement differences comparing orthodontist 7 with the others (Table 5). However, the sample size for these comparisons was small, resulting in limited power to detect differences.

The imperfect reliability of several measures assessing malocclusion indicates that the assess-

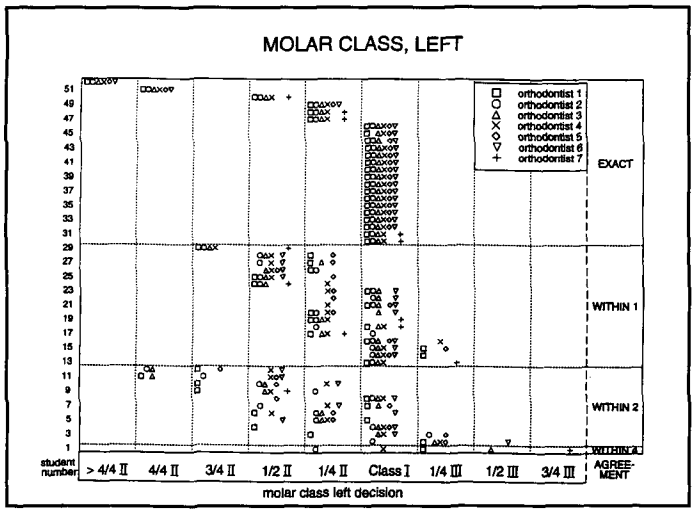

Figure 1

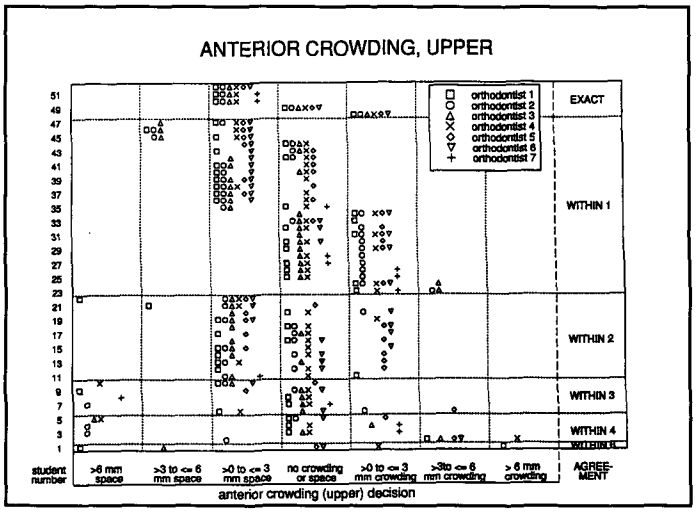

Figure 2

Figure 1
Left side molar classification score of each orthodontist for each of the children. Each horizontal line represents one child, N=52. For each child, the molar classification score for the left side as determined by each orthodontist is indicated. For example, orthodontist #4 scored the left molar relation as a Class I in child 1 (Line 1), while orthodontist #7 scored this relation as a 3/4 Class III for the same child. Not all orthodontists examined each child.

Figure 2
Maxillary anterior spacing/crowding scores of each orthodontist for each of the children. Each horizontal line represents one child, N=52. For each child, the maxillary spacing/crowding score as determined by each orthodontist is indicated. For example, orthodontist #1 scored the maxillary spacing/crowding as greater than 6 mm of space in child 1 (Line 1), while orthodontists #5 and #6 scored no maxillary spacing/crowding for the same child. Not all orthodontists examined each child.

ment of prevalence rates of malocclusion traits remains problematic. For measures having poor reliability, a population characterized by one examiner or group of examiners cannot be directly compared with a different population examined by others; differences could be due to differences between examiners and not populations. The precision and power of a study to detect differences will be attenuated if considerable error variance is due to examiner differences; such differences tend to obscure treatment effects when they exist because they result in less precise estimates of population parameters.[30]

Achieving reliability in clinical examinations is difficult because both the patient and clinician must be checked.[31] We speculated that several important signs of malocclusion (molar class, overjet, overbite) may change spontaneously due to differences in mandibular position as determined by the patient and/or examiner. However, our data, based upon the use of habitual (centric) occlusion, indicated that differences in registering mandibular position most likely did not contribute to interjudge errors, as disagreements in molar class were not associated with disagreements in overjet or overbite. Thus, differences in molar classification, overjet, and overbite among examiners were more likely due to differences in opinion than differences in examination technique. Whether the use of centric relation to register mandibular position would be more or less reliable was not examined.

Dworkin et al.[13] found the average reliability of judging molar class (Class I, II, or III) in adults by hygienists who have been retrained and who use calibrated methods, to be excellent (Kappa = 0.78), while uncalibrated dentists showed poor reliability (Kappa = 0.37). Our data on children suggest that orthodontists who have trained together have a very acceptable level of reliability (median Kappas of 0.68 and 0.72 on the two sides) in judging molar class, even when using a more rigorous scale divided into 1/4 cusp increments.

Dworkin's group also reported good agreement using interclass correlation (ICC) among retrained hygienists (ICC = 0.88) and among uncalibrated dentists (ICC = 0.79) in measuring horizontal overjet in millimeters,[13] where ICCs greater than 0.75 are interpreted as acceptable agreement.[32] Our data, when converted to ICCs, also support acceptable reliability of judging overjet (0.85) and molar class (0.86) in the clinical setting.

As a result of the earlier reliability sessions, data on incisor exposure and interlabial gap were dropped from the latter sessions because of poor agreement. Assessment of maxillary and mandibular horizontal positions also had poor reliability. Although this could arise due to

Table 5
William's indices for each orthodontist and variable. An * indicates orthodontist agreement differs from group, with a 95% confidence interval upper bound of less than 1.

Orthodontist (n)	1 (52)	2 (52)	3 (52)	4 (52)	5 (39)	6 (39)	7 (13)
Facial convexity	0.97	0.98	1.04	0.97	1.04	1.04	0.67
Maxillary AP position	1.03	0.93	1.03	1.06	0.85*	1.10	0.99
Mandibular AP position	0.94	1.03	1.09	1.09	0.82*	1.09	0.76
Posterior crossbite	0.99	0.97	0.99	1.02	1.02	1.02	1.00
Overbite	0.82*	0.96	1.12	1.10	1.00	1.03	1.01
Overjet	1.11	1.04	1.17	0.70*	0.85	1.28	0.81
Incisor exposure	1.39	1.53	0.47*	1.12	0.49*	1.39	—
Interlabial gap	1.21	1.50	1.23	1.12	0.15*	1.25	—
Molar class, right	0.88	1.06	1.10	1.08	0.95	0.94	0.89
Molar class, left	0.95	0.96	1.15	1.02	0.96	1.02	0.80
Anterior crowding, upper	1.04	1.01	0.97	1.02	0.80	1.18	0.98
Anterior crowding, lower	0.87	1.15	1.11	0.84	0.96	1.01	1.39

misinterpretation of instructions or form completion errors, we agree with others[18] that this very poor level of agreement suggests that clinicians have conceptual and evaluative differences in the interpretation of a patient's problems. This may reflect differences in treatment philosophy, treatment technique, or training among examiners. This, of course, raises an issue the specialty must someday resolve—the lack of a diagnostic "gold standard," which permits one's treatment philosophy or treatment technique preferences to drive one's diagnostic view. Whether these differences actually reflected different treatment philosophies or technique preferences was not directly examined.

The poor to fair reliability of the crowding measure may be related to the ordinal manner of scoring crowding. Examiners obviously applied different cut-offs. Displacement scores to estimate crowding as used in the Treatment Priority Index[1] and the Occlusal Index[2] should be examined in the clinical setting. Interexaminer errors also can be attributed to errors in completing the screening form. Large differences in judging spacing/crowding in Figure 2 are most likely due to form completion errors and not true disagreements.

These findings of poor to moderate reliability among orthodontists for many of the common malocclusion descriptors reveal that clinical orthodontic diagnostic language is imprecise. If these data are representative of the usual use of these diagnostic terms by the specialty, these findings suggest that methods to improve diagnostic terms should be considered. These may include altering measurement scales, more rigid definitions and/or a more systematic approach to training orthodontists. The lack of a very precise language may limit the specialty's ability to construct an index of malocclusion to assess treatment need or malocclusion severity.

Finally, it should be pointed out that these data reflect interexaminer reliability during mass

Table 6
Number of paired decisions by orthodontists with specified overjet-molar class discrepancy; percent of total given in parentheses

Molar class discrepancey	Overjet discrepancy 0	1	2	3	4
0	135 (19.54)	152 (22.00)	26 (3.76)	9 (1.20)	3 (0.43)
1	103 (14.91)	96 (13.89)	20 (2.89)	6 (0.87)	0 (0.00)
2	54 (7.81)	38 (5.50)	8 (1.16)	0 (0.00)	0 (0.00)
3	8 (1.16)	7 (1.01)	5 (0.72)	0 (0.00)	0 (0.00)
4	10 (1.45)	6 (0.87)	2 (0.29)	1 (0.14)	1 (0.14)
5	0 (0.00)	1 (0.14)	0 (0.00)	0 (0.00)	0 (0.00)

screenings of children in their schools with the child standing erect in front of a seated examiner. Improved reliability might result if the examinations had been conducted in a practice setting (dental chair, better lighting, more time per student).

Conclusions

The results obtained from studying the interexaminer reliability of clinical measures of malocclusion during the screening of children support the following conclusions:

1. The reliability of determining maxillary and mandibular anteroposterior positions, maxillary incisor exposure and interlabial gap during profile and frontal clinical examination is poor.
2. Acceptable, but moderate, reliability exists for judging facial convexity, overbite, overjet, and molar classification.
3. Excellent reliability exists for judging posterior crossbites.

Finally, these findings suggest that examiners should be trained prior to and during the collection of clinical malocclusion data to increase reliability and reduce bias. Further studies are indicated to improve the reliability of current clinical measures describing malocclusion.

Acknowledgments

We wish to thank the school children of Alachua County, Florida who participated, the county school administrators, especially Dr. Mel Lucas, and the classroom teachers who provided access and help during the screenings. In addition, we thank Ms. Sarah Garrigues-Jones, Dr. Sal Cabassa, Dr. Richard Hocevar, Dr. Michael Kania, Dr. Debra Sappington, Dr. Janet Pappas, Ms. Laurel Johnson, and Mr. Nirander Nangia for their assistance during the screenings and for data management tasks.

Supported by NIH-NIDR Grant DE 08715 to Stephen D. Keeling

Author Address

Stephen D. Keeling, DDS, MS
Department of Orthodontics
College of Dentistry
Box 100444, JHMHC, University of Florida
Gainesville, FL 32610-444

Stephen D. Keeling, Department of Orthodontics, College of Dentistry, University of Florida.

Susan McGorray, Division of Biostatistics, Department of Statistics, College of Liberal Arts and Sciences, University of Florida.

Timothy T. Wheeler, Department of Orthodontics, College of Dentistry, University of Florida.

Gregory J. King, Department of Orthodontics, College of Dentistry, University of Florida.

References

1. Grainger RM. Orthodontic treatment priority index.Washington, D.C.: U.S. Government Printing Office, 19671000 PHSPN, ed. 2 (#25).
2. Summers CJ. The Occlusal Index: A system for identifying and scoring occlusal disorders. Am J Orthod 1971;59:552-567.
3. Richmond S, Shaw W, O'Brien K, et al. The development of the PAR index (Peer Assessment Rating): Reliability and validity. Eur J Orthod 1992;14:125-139.
4. Draker DL. Handicapping labio-lingual deviations: A proposed index for public health purposes. Am J Orthod 1960;46:295-305.
5. Saltzmann JA. Handicapping malocclusion assessment to establish treatment priority. Am J Orthod 1968;54:749-765.
6. Brook PH, Shaw WC. The development of an index of orthodontic treatment priority. Eur J Orthod 1989;11:309-320.
7. Koran LM. The reliability of clinical methods, data and judgments. N Engl J Med 1975;293(13):642-646.
8. Fleiss JL, Mann J, Paik M, Goultchin J, Chilton NW. A study of inter- and intra-examiner reliability of pocket depth and attachment level. J Period Res 1991;26:122-128.
9. Mauriello SM, Bader JD, Disney JA, Graves RC. Examiner agreement between hygienists and dentists for caries prevalence examinations. J Public Health Dent 1990;50(1):32-37.
10. Solow B, Helm S. A method for tabulation and statistical evaluation of epidemiological malocclusion data. Acta Odont Scand 1968;26:63-88.
11. Helm S. Intra-examiner reliability of epidemiologic registrations of malocclusion. Acta Odont Scand 1976;35:161-165.
12. Gravely JF, Johnson DB. Angle's classification of malocclusion: An assessment of reliability. Br J Orthod 1974;1(3):79-86.
13. Dworkin SF, LeResche L, DeRoucen T, Von Korff M. Assessing clinical signs of temporomandibular disorders: Reliability of clinical examiners. J Prosthet Dent 1990;63(5):574-579.
14. Grewe JM, Hagan DV. Malocclusion indices: A comparative evaluation. Am J Orthod 1972;61(3):286-294.
15. Lewis EA, Albino JE, Cunat JJ, Tedesco LA. Reliability and validity of clinical assessments of malocclusion. Am J Orthod 1982;81(6):473-477.
16. Kelly JE, Sanchez M, Van Kirk LE. An Assessment of the Occlusion of the Teeth of Children 6-11 Years, United States. In: Vital and Health Statistics. Rockville: U. S. Department of Health, Education, and Welfare, 1973: 1-48. vol 11, #130.
17. Rudge SJ, Jones PT, Hepenstal S, Bowden DEJ. The reliability of study model measurement in the evaluation of crowding. Eur J Orthod 1983;5:225-231.
18. Phillips C, Bailey L, Sieger R. Level of agreement in clinicians' perception of Class II malocclusions. J Oral Maxillofac Surg 1994;52:565-571.
19. Fields HW, Vann JWF, Vig KW. Reliability of soft tissue profile analysis in children. Angle Orthod 1982;52(2):159-165.
20. Cohen J. A coefficient of agreement for nominal scales. Educ Psych Meas 1960;20:37-46.
21. Cohen J. Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psych Bull 1968;70:213-220.
22. Fleiss JL, Cohen J, Everitt BS. Large sample standard errors of Kappa and weighted Kappa. Psych Bull 1969;72:323.
23. Hunt RJ. Percent agreement, Pearson's correlation, and Kappa as measures of inter-examiner reliability. J Dent Res 1986;65(2):128-130.
24. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-174.
25. Fleiss JL. Statistical methods for rates and proportions. (2nd ed.) New York: John Wiley and Sons, 1981:216-218.
26. Posner KL, Sampson PD, Caplan RA, Ward RJ, Cheney FW. Measuring interrater reliability among multiple raters: An example of methods for nominal data. Stat Med 1990;8:1103-1115.
27. SAS Institute. SAS/STAT user's guide, Version 6. (4th ed.) Cary, NC: SAS Institute, 1990; vol 1-2.
28. SAS Institute. S-Plus user's manual. Seattle: Statistical Sciences, 1991
29. Dworkin SF, LeResche L, Von Korff MR. Diagnostic studies of temporomandibular disorders: challenges from an epidemiologic perspective. Anesth Prog 1990;37:147-154.
30. Fleiss JL, Shrout PE. The effects of measurement errors on some multivariate procedures. Am J Public Health 1977;67:1188-1191.
31. 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.
32. Shrout PE, Fleiss JL. Interclass correlations: Uses in assessing rater reliability. Psych Bull 1979;86:420-428.