Interobserver and intraobserver agreement of clinical orthodontic judgments based on intraoral and extraoral photographs

Isabelle Lauweryns, LDS; Nathalie Van Cauwenberghe, LDS; Carine Carels, DDS, PhD

Careful patient examination is important for diagnosis and orthodontic treatment planning, along with the quantitative analysis of patient records performed on study casts, cephalograms and other radiographs.[1] Extraoral and intraoral examinations have also grown in importance in the assessment of progress during treatment.[2] A number of diagnostic analyses are based upon facial characteristics[3,4] and an orthodontic treatment evaluation is considered incomplete if facial photographs are omitted. Furthermore, producing photographs is relatively inexpensive and noninvasive. Facial esthetics can also be evaluated from facial photographs.

Recently, however, Han and colleagues[5] claimed that study casts alone provided adequate information for orthodontic diagnosis in 55% of cases. The addition of other records, such as roentgenograms and facial photographs, enhanced interobserver agreement on treatment decisions by only 5%. Information obtained from facial photographs does not always help in making a diagnosis or determining a treatment strategy as normative values are not available. Recently, however, attempts have been made to reduce disagreements derived from facial photographs by standardization of the patient head position.[3] Other investigators have been concerned about the standardization of judgments evaluated from facial photographs to enhance interobserver agreement.[6] Judgments based on rating scales can be useful to reduce observer variation.[7] If orthodontists really start to use information from photographs in orthodontic diagnosis and treatment

Abstract

The purpose of this study was to calculate the agreement between and within observers for orthodontic judgments based on intraoral and extraoral photographs at two separate occasions in ten twin pairs. Eighteen variables were scored by two orthodontic students according to well-defined rating scales. Interobserver and intraobserver proportion of agreement as well as the agreement which could be expected only by chance and the remaining agreement beyond chance were calculated. The calculated agreement beyond chance was not significant (α=0.05%) for middle and upper facial height, anterior apical area in lower jaw and posterior apical area in both jaws within the first observer and for upper facial height within the second observer. Interobserver reliability was not acceptable at the 5% level for judging asymmetry, facial animation, posterior apical area in the upper and lower jaws, sagittal lip position and the upper facial height. Lower facial height, sagittal lip position and middle apical area in the lower jaw agreed significantly at this level for only one interobserver comparison.

Key Words

Orthodontics • Clinical judgments • Photographs • Interobserver agreement • Intraobserver agreement.

Submitted: January 1993 **Revised and accepted for publication:** May 1993 Angle Orthod 1994;64(1):23-30

Figure 1A

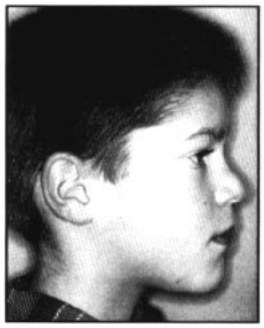

Figure 1B

Figure 1C

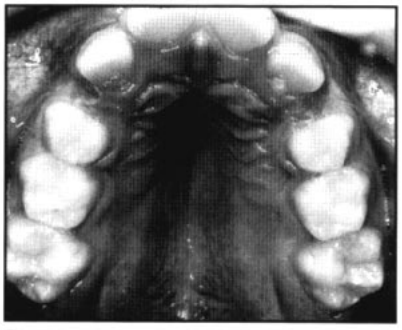

Figure 1D

Figure 1E

Figure 1F

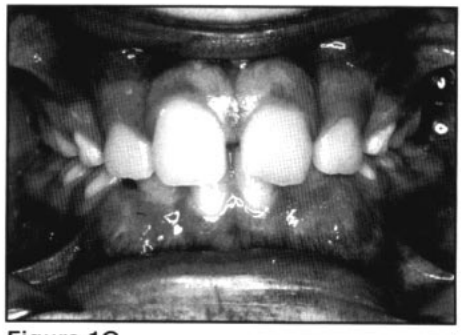

Figure 1G

Figure 1H

Figure 1A-C
Facial photographs, with the Frankfort Horizontal plane parallel to the floor, from frontal non-smiling (1A), lateral (1B) and oblique (1C) angles.

Figure 1D-E
Photographs of the upper (D) and lower (E) jaws.

Figure 1F-H
Photographs in maximal occlusion seen from right (F), left (H), and front (1G).

planning, the reliability of these judgments needs to be evaluated.

The aim of the present investigation was to determine the agreement within and between two clinicians in assessing orthodontic judgments from facial and intraoral photographs.

Material and methods

Ten twin pairs, between 10 and 14 years of age, were randomly selected from an ongoing twin study at the Department of Orthodontics, Katholieke Universiteit Leuven, Belgium. Extraoral photos were taken with the face positioned with Frankfort plane parallel to the floor in frontal (non-smiling) (Figure 1A), lateral (Figure 1B) and oblique (Figure 1C) views. Intraoral photographs were taken of the upper and the lower jaws separately (Figure 1D-E) as well as in occlusion (Figures 1F-H). All photographs were taken by the same clinician who had no further involvement in the study.

From a standard orthodontic clinical examination list, proposed by van der Linden and Boersma,[8] the two clinicians involved in the judgments selected those variables which could be judged on the photographs (18 judgments, listed in Table 1). The etiology of clinical disagreement can be ascribed to three categories: the examiner, the examined and the examination.[9] In the present study the variation of the examiner was kept to a minimum as both clinicians studied dentistry at the same university and were both active in their third year of postgraduate orthodontic education. In consensus, the clinicians defined rating scales for each judgment (Table 1). Based on this protocol they evaluated the variables independently on the projected photographs. This was repeated one month later to exclude memory biases upon the judgments. As all judgments were based on the same set of photographs, the second source of variability, the examined, was minimized. The examinations, finally, were held in constant environmental conditions.

Judgments made at the same scoring time (time zero and one month later) were compared between both observers; this comparison is called the interobserver agreement. Each observer's initial and second set of judgments were compared; this comparison is called the intraobserver variation of the first and second clinician (Table 2).

For statistical analysis the outcome of first and second judgments were displayed in a contingency table, as illustrated in Table 3, for a judgment with two possible answers, A and B. The agreement between observations (p_1) was obtained by dividing the sum of the answers in the diagonal axis of the table by the number of observations. The proportion of clinical agreement which can be expected to arise by chance alone (p_2) is obtained according to the formula listed in Table 3. In case this proportion exceeded the observed agreement ($p_2 > p_1$) no further calculations were performed. In judgments where p_1 exceeded p_2 it was important to calculate the remaining agreement obtained beyond chance (p_3) (formula is listed in

Table 3).[10,11] Furthermore the standard error (SE p_3), being an indication of the uncertainty of the measure, was calculated as $\sqrt{[(p_3 X (1-p_3))/N]}$. The agreement beyond chance was subsequently divided by the standard error (p_3/SE). This value was then compared to the standard normal deviate (the Z value) at different levels of confidence in order to determine the level of significance. The 5% level, the criteria most often applied in statistics, was chosen to determine significant outcomes.[12,13]

Results

Table 4 lists the judgments of observer 1 (horizontally) and observer 2 (vertically) for the variable "profile" for the 20 subjects involved, at scoring time zero. According to the formulas listed in Table 3 the proportion of agreements (p_1, p_2, p_3) was calculated. Percentages (P_1, P_2, P_3) were obtained after multiplying these agreements by 100. Because the clinicians agreed that six subjects showed a convex profile and four subjects showed a straight profile, the observed percentage of agreement between both observers was 50% (interobserver agreement = p_1 = [6+4]/20 = 0.50). Because the agreement which could be expected by chance alone (p_2 = [[(6X11)/20 + (5X8)/20 + (1 X 1)/20] / 20] = 0.2675) was lower than the observed agreement, further analysis was meaningful. The percentage of interobserver agreement beyond chance (p_3 = [(50-26.75)/(100-26.75)] = 0.3174) was 0.32%. The standard error ([$\sqrt{}$ [0.32 X (1-0.32]] = 0.10) was very small. Inversing the standard error and multiplying by p_3 (SEp_3/p_3 = 3.05) procured the value which needed to be compared with the Standard normal deviate Z. As the Z value was exceeded at a confidence level of 99% (α=0.01, Z = 2.576), it can be concluded that the agreement between both observers beyond chance differed significantly from zero at a level α=0.01.

Table 5 lists the percentages of inter- and intraobserver agreements at scoring time zero and one month later for the 18 judgments, in alphabetical order. Interobserver agreements exceeding agreements expected by chance varied between 31.74% and 90.78%; intraobserver values between 39.17% and 65.99% for the first and between 36.17% and 92.37% for the second clinician.

In Table 6 the standard error SE, P_3/SE as well as the level of significance α (α=0.05%; α=0.01%; α=0.001%) are displayed for those judgments where the observed proportion of agreeing pairs exceeded the agreement which could be expected by chance. For the remaining judgments it was concluded that the observed agreement was not significant (NS). The standard error, a measure of

Table 1
List of variables* to be judged on extra- and intraoral photographs

Variables	Categories (N)	
Based on extra- oral photographs		
Asymmetry	2	Yes; No.
Facial Animation	3	High; Normal; Small.
Menton - Subnasale (lower facial height)	7	Numerical values from 0.6 to 1.2 mm (with steps of 0.1 mm)
Nasiolabial Angle	3	Acute; Right; Obtuse.
Nose	4	Large; Small; Normal; Straight.
Profile	7	(Light) Convex; (Light) Straight; Concave; Dished-in; Double Proposition.
Sagittal Chin Position	3	Normal; Proposition; Retroposition.
Sagittal Lip position	3	Normal; Inversed; Enlarged.
Subnasale - Trichion (middle facial height)	7	Numerical values from 0.6 to 1.2 mm (with steps of 0.1 mm)
Trichion - Ophrion (Upper facial height)	7	Numerical values from 0.6 to 1.2 mm (with steps of 0.1 mm)
Upper and Lower Lip	3	Normal; Abundant; Thin.
Based on intra- oral photographs		
Anterior Apical Area (upper and lower jaw)	3	Large; Medium; Small.
Middle Apical Area (lower and upper jaw)	3	Large; Medium; Small
Oral Hygiene	3	Optimal; Satisfying; Insufficient.
Posterior Apical Area (lower and upper jaw)	3	Large; Medium; Small.

*selected from van der Linden and Boersma, 1986.
N = Number of categories

Table 2
Design of the present study

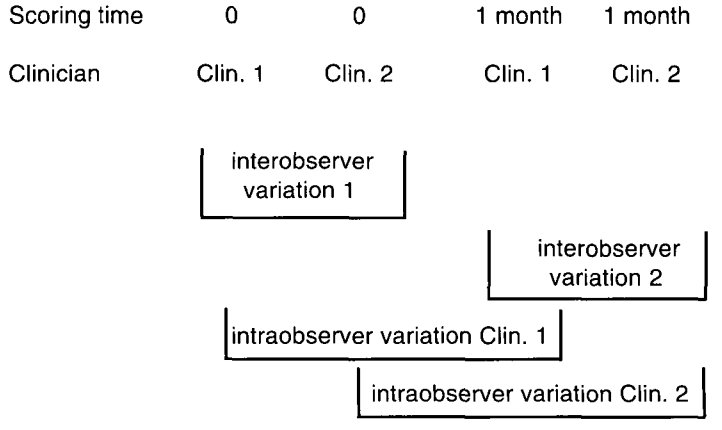

Table 3
Contingency table between judgments with two rating scales

	A	B	
A	c1r1	c2r1	r1
B	c1r2	c2r2	r2
	c1	c2	N

1. Calculation of agreements
   - observed proportion of agreeing pairs
     $= (c1r1 + c2r2) / N = p_1$
   - agreement expected by chance alone
     $= ((c1 \times r1) / N + (c2 \times r2) / N) / N = p_2$
2. Further statistical analysis in case $p_1 > p_2$
   -agreement beyond chance
     $= (p_1 - p_2) / (100 - p_2) = p_3$
   - Standard error p_3
     $= SE\ p_3 = \sqrt{[p_3 \times (1-p_3)/N]}$
   - $p_3 / SE\ p_3$

(The proportion p is multiplied by 100 to convert to a percentage P)

Table 4
Agreement between both observers scoring the variable "profile" at time zero

	X	S	V	DI	DP	MX	MV	N
X	6	1	0	0	0	4	0	11
S	0	4	0	1	0	3	0	8
V	0	0	0	0	0	0	0	0
DI	0	0	1	0	0	0	0	1
DP	0	0	0	0	0	0	0	0
MX	0	0	0	0	0	0	0	0
MV	0	0	0	0	0	0	0	0
N	6	5	1	1	0	7	0	20

observed proportion of agreeing pairs (p_1)
$= (6 + 4) / 20 = 0.50$

agreement expected by chance alone (p_2)
$= ((6 \times 11) / 20) + ((5 \times 8) / 20) + ((1 \times 1) / 20) / 20 = 0.2675$

actual agreement above chance (p_3)
$= (50 - 26.75) / (100 - 26.75) = 0.3174$

$SE\ p_3 = \sqrt{[0.3174 \times (1-0.3174)/N]} = 0.10$

$p_3 / SE\ p_3 = 3.05$ (significant at 99%)

Abbreviations of the categories:

X	Convex	DP	Double proposition
S	Straight	MX	Moderate Convex
V	Concave	MV	Moderate Concave
DI	Dished - in		

(N = number of subjects)

the uncertainty in a simple statistic, varied between similar values (0.07-0.11) for all agreements beyond chance, except the anterior apical area in the lower jaw where 0.26 was found for the interobserver agreement, scored after one month. While both interobserver agreements were not significant for asymmetry, facial animation and sagittal lip position, significant levels were obtained for these judgments within the same observers. Profile, lower facial height and sagittal chin position were significant in one interobserver agreement, but at a lower confidence level when compared to both intraobserver agreements.

The agreement between the judgments of the second clinician was significantly different from zero at a higher level than the agreement between the judgments of the first clinician for the anterior and the middle apical area in the lower jaw as well as for the posterior apical area in both jaws and for middle facial height. For oral hygiene however the intraobserver agreement of the first observer exceeded the intraobserver agreement of the second observer and was similar to the interobserver value. The first clinician's intraobserver agreement was not significant for the anterior apical area in lower jaw and the middle facial height, while significance was attained for these judgments between both observers and for the second observer. Only the secondclinician's intraobserver agreement for the posterior apical area in the lower and upper jaws was significant.

Within as well as between observers, agreement was significant at a 0.001% level for the middle apical area in upper jaw, the nasiolabial angle and upper and lower lips.

Judging the morphology of the nose and the anterior apical area in the upper jaw seemed significant at a 0.001% level for all agreements except for one interobserver agreement, where a level of 0.05% was attained. No significance was found for agreements concerning the upper facial height.

Table 5
Interobserver and intraobserver agreement. Judgments are listed alphabetically.

	Interobserver (Time = 0)			Interobserver (Time = 1 Month)			Intraobserver Clin 1			Intraobserver Clin 2		
Judgments	P_1	P_2	P_3	P_1	P_2	P_3	P_1	P_2	P_3	P_1	P_2	P_3
Anterior Apical Area (lower jaw)	60.00	40.00	33.33	65.00	52.5	26.32	60.00	52.50	15.79	90.00	43.00	82.46
Anterior Apical Area (upper jaw)	55.00	44.00	19.64	95.00	62.5	86.67	70.00	46.25	44.19	80.00	59.00	51.22
Asymmetry	40.00	32.50	11.11	35.00	32.00	4.41	75.00	60.00	37.50	90.00	81.50	45.95
Facial Animation	35.00	40.25	/	55.00	62.50	/	80.00	66.00	41.18	60.00	43.75	28.89
Lower Facial Height	55.00	33.00	32.84	40.00	31.00	13.00	65.00	33.00	47.76	65.00	28.75	50.88
Middle Apical Area (lower jaw)	55.00	47.25	14.69	80.00	72.00	28.57	70.00	62.25	20.53	90.00	53.50	78.49
Middle Apical Area (upper jaw)	65.00	46.00	35.19	75.00	55.50	42.86	75.00	54.00	45.65	75.00	47.50	52.38
Middle Facial Height	60.00	44.50	27.93	75.00	57.25	41.52	55.00	47.75	13.88	70.00	52.50	36.84
Nasiolabial Angle	95.00	45.75	90.78	85.00	51.00	69.39	80.00	49.00	60.78	90.00	47.75	80.86
Nose	65.00	46.00	35.19	45.00	33.25	17.60	75.00	50.25	49.75	55.00	29.50	36.17
Oral Hygiene	70.00	38.50	51.22	60.00	35.25	38.52	60.00	34.25	39.17	60.00	39.25	34.16
Posterior Apical Area (lower jaw)	60.00	63.50	/	70.00	70.00	/	70.00	70.00	/	80.00	65.00	42.86
Posterior Apical Area (upper jaw)	55.00	59.00	/	70.00	70.00	/	65.00	65.00	/	80.00	65.00	42.86
Profile	50.00	26.75	31.74	30.00	27.50	3.50	75.00	26.50	65.99	65.00	45.00	36.36
Sagittal Chin Position	40.00	33.50	9.70	30.00	12.50	20.00	70.00	43.50	46.90	75.00	50.75	49.24
Sagittal Lip position	25.00	35.00	/	40.00	39.00	1.64	85.00	64.00	58.33	75.00	47.50	52.38
Upper Facial Height	95.00	95.00	/	90.00	90.00	/	95.00	95.00	/	90.00	90.00	/
Upper and Lower Lip	75.00	32.00	63.24	60.00	29.75	43.06	70.00	38.50	51.22	95.00	34.50	92.37

Legend
P_1 = observed agreement
P_2 = agreement expected by chance
P_3 = agreement beyond chance
Values are listed as percentages (P = p X 100)

Discussion

If orthodontists claim to adapt their diagnosis and treatment planning strategies to intraoral and extraoral photographs, more investigation is needed to evaluate the reliability of judgments on the basis of this material. It is not acceptable to make definite conclusions from judgments which appear to change thoroughly in time or from judgments which appear to vary from individual to individual.

In the present investigation, two orthodontic students with the same educational level (third year of postgraduate orthodontic school) were used to minimize interobserver differences. Special attention was also given to patient positioning and photographing and to the examination procedure.

Neither interobserver agreement was satisfactory for the judgments of asymmetry, facial animation, posterior apical area in both jaws, sagittal lip and chin position and upper facial height. Moreover, the anterior apical area in the lower jaw showed an excessive standard error for the second interobserver agreement. For lower facial height and middle apical area in the lower jaw a significance level of 5% was obtained only at one scoring time. Judging these variables seemed related to the observer. The first clinician was inconsistent in scoring the middle and upper facial height, the anterior apical area in the lower jaw and the posterior apical areas in both jaws. The second clinician on the other hand was not consistent at a 5% level for assessing the upper facial height. The middle apical area in upper jaw, the nasiolabial angle and upper and lower lip seemed to be evaluated very accurately (0.001% of significance) within as well as between both observers. Of course, as the number of rating scales was not identical for all 18 variables involved in this study, no information relating to the sensitivity of the judgments could be calculated.

It should be mentioned that software has been

Table 6
Level of significance for the observer's agreement beyond chance. Judgments are listed alphabetically.

	Interobserver (Time = 0)			Interobserver (Time = 1 Month)			Intraobserver Clin 1			Intraobserver Clin 2		
Judgments	SE P_3	P_3/SE	α	SE P_3	P_3/SE	α	SE P_3	P_3/SE	α	SE P_3	P_3/SE	α
Anterior Apical Area (lower jaw)	0.11	3.16	0.01	0.26	2.67	0.01	0.08	1.94	NS	0.09	9.70	0.001
Anterior Apical Area (upper jaw)	0.09	2.21	0.05	0.08	11.40	0.001	0.11	3.98	0.001	0.11	4.58	0.001
Asymmetry	0.07	1.58	NS			NS	0.11	3.46	0.001	0.11	4.12	0.001
Facial Animation			NS			NS	0.11	3.74	0.001	0.10	2.85	0.01
Lower Facial Height	0.11	3.13	0.01	0.08	1.73	NS	0.11	4.28	0.001	0.11	4.55	0.001
Middle Apical Area (lower jaw)	0.08	1.86	NS	0.10	2.83	0.01	0.09	2.27	0.05	0.09	8.54	0.001
Middle Apical Area (upper jaw)	0.11	3.30	0.001	0.11	3.95	0.001	0.11	4.10	0.001	0.11	4.69	0.001
Middle Facial Height	0.10	2.78	0.01	0.11	3.77	0.001	0.08	1.80	NS	0.11	3.42	0.001
Nasiolabial Angle	0.06	14.04	0.001	0.10	6.73	0.001	0.11	5.57	0.001	0.09	9.19	0.001
Nose	0.11	3.30	0.001	0.09	2.07	0.05	0.11	4.45	0.001	0.11	3.37	0.001
Oral Hygiene	0.11	4.58	0.001	0.11	3.52	0.001	0.11	3.59	0.001	0.11	3.22	0.01
Posterior Apical Area (lower jaw)			NS			NS			NS	0.11	3.78	0.001
Posterior Apical Area (upper jaw)			NS			NS			NS	0.11	3.87	0.001
Profile	0.10	3.05	0.01			NS	0.11	6.21	0.001	0.11	3.38	0.001
Sagittal Chin Position			NS	0.08	2.02	0.05	0.11	4.20	0.001	0.11	4.40	0.001
Sagittal Lip Position			NS			NS	0.11	5.29	0.001	0.11	4.69	0.001
Upper Facial Height			NS			NS			NS			NS
Upper And Lower Lip	0.11	5.87	0.001	0.11	3.89	0.001	0.11	4.58	0.001	0.06	15.56	0.001

Legend
SE = standard error; p_3 = agreement beyond chance; NS = Not Significant at 5% level.
According α and standard score (Z) for the specific confidence levels:
95% (α=0.05; Z=1.960); 99%(α=0.01; Z=2.576); 99,9% (α=0.001; Z=3.29).[12,13]

developed to perform facial analysis on photographs visualized by means of a camera, connected to the computer. This computerized video imaging allows a comparison and superimposition of lateral and profile extra-oral photographs with RX images.[14]

Involving quantitative measurements based on radiographic and dental cast analysis seems essential to check clinical findings concluded from photographic material.

Keeping the present findings in mind, a degree of caution is urged when decisions are made upon first time observations of facial and intra-oral photographs. It is advisable to compare clinical judgments within and between clinicians as well as to compare them to results obtained from other records.

Conclusions

Intraobserver agreement for the first observer was found to be significant at a confidence level of at least 95% for 13 judgments; the second observer agreed significantly for 17 judgments. Interobserver agreement was satisfactory for only eight judgments. Clinical orthodontic patient analysis based on extraoral and intraoral photographs, taken in a standardized way, should be used with caution. It is advisable to compare observations made on extraoral and intraoral photographs with other records providing quantitative and objective values, as clinicians are not always consistent from one day or week to the next, and are even less consistent with colleagues for these judgments.

Acknowledgements

I. Lauweryns holds a fellowship at the National Fund for Scientific Research (N.F.W.O.) - Belgium. This study was supported by the Fund for Medical Scientific Research (F.G.W.O.). The authors deeply appreciate the participation of the twins. They are also grateful to Mrs. D. Ural for taking the photographs.

Author Address

Isabelle Lauweryns, LDS
Katholieke Universiteit Leuven
School voor Tandheelkunde, Mondziekten en Kaakchirurgie
Afdeling Orthodontie
Kapucijnenvoer 7
B-3000 Leuven
Belgium

I. Lauweryns is a Research Fellow of the National Fund for Scientific Research (Belgium), Department of Orthodontics.

N. Van Cauwenberghe is an Assistant in the Department of Orthodontics.

C. Carels is Professor and Head of the Department of Orthodontics.

References

1. Sackett DL. The science of the art of clinical management. In: Science and Clinical Judgment in Orthodontics. P.S. Vig and K.A. Ribbens (eds.), Monograph 19, Craniofacial Growth Series, Center for Human Growth and Development, The University of Michigan, Ann Arbor, 1986; 237-251.
2. Turpin DL. The orthodontic examination. Angle Orthod 1990; 60(1): 3-4.
3. Bass NM. The aesthetic analysis of the face. Eur J Orthod1991; 13: 343-350.
4. Ricketts RM, Bench RW, Roth RH, Chaconas SJ, Schulhof RJ and Engel GA. Orthodontic diagnosis and planning, Rocky Mountain/Orthodontics, Vol 1 and Vol 2, 1982.
5. Unae Kim Han, Vig KWL, Weintraub JA, Vig PS and Kowalski CJ. Consistency of orthodontic treatment decisions relative to diagnostic records. Am J Orthod Dentofac Orthop 1991; 100(3): 212-219.
6. Peerlings RHJ. Orthodontie en dento-faciale esthetiek. Ph. D., Thesis, 1992, Catholic University Nijmegen, The Netherlands.
7. Feinstein AR. The clinician as scientist. In: Science and Clinical Judgment in Orthodontics. P.S. Vig and K.A. Ribbens (eds.), Monograph 19, Craniofacial Growth Series, Center for Human Growth and Development, The University of Michigan, Ann Arbor, 1986; 1-14.
8. van der Linden FGPM and Boersma H. Onderzoek van de patient. In: van der Linden FGPM and Boersma H, editors. Diagnostiek en behandelingsplanning in de orthodontie. Samson staphleu Alphen aan den Rijn, 1986; 73-92.
9. Department of Clinical Epidemiology and Biostatistics. Clinical disagreement: I. How often it occurs and why. Can Med Assoc J. 1980; 123: 499-504.
10. Light RJ. Measures of response agreement for qualitative data: some generalizations and alternatives. Psychological Bulletin 1971; 76(5): 365-377.
11. Aoki N, Horibe H, Ohno Y, Hayakawa N, Kondo R and Okada H. Epidemiological evaluation of funduscopic findings in cerebrovascular diseases. III. Observer variability and reproducibility for funduscopic findings. Japanese Circulation Journal 1977; 41: 11-17.
12. Gardner MJ and Altman DG. (eds.) Statistics with confidence - confidence intervals and statistical guidelines, The University Press (Belfast) Ltd., 1989.
13. Snedecor GW and Cochran WG. (eds.) Statistical Methods, The Iowa State University Press, Ames, Iowa, U.S.A., 1987.
14. Blaseio G. Quick Ceph Image * Reference Guide, version 4.0, 1986-1993, Orthodontic processing, 386 East H Street, Ste. 209-404, Chula Vista, CA 91910.