ORIGINAL ARTICLE

Evaluation of the accuracy of digital model analysis for the American Board of Orthodontics objective grading system for dental casts

Peter A. Costalos,[a] Keivan Sarraf,[b] Thomas J. Cangialosi,[c] and Stella Efstratiadis[d]
New York, NY

Introduction: In 1999, after 3 years of field testing, the American Board of Orthodontics (ABO) implemented a grading system for posttreatment orthodontic models and panoramic radiographs, to make the phase III examination both fair and objective. In the ABO's objective grading system, 7 occlusal criteria (tooth alignment, vertical positioning of marginal ridges, buccolingual inclination of posterior teeth, occlusal relationship, occlusal contacts, overjet, and interproximal contacts) are measured on plaster models to assess a patient's final occlusion. To date, no study has evaluated the ABO grading system for use on digital models. The purpose of this study was to determine whether digital models can be used with reasonable accuracy and reliability for assessing patients' final occlusions. **Methods:** Plaster and digital (OrthoCAD, Cadent Inc, Carlstadt, NJ) posttreatment models of 24 patients were gathered from the postgraduate orthodontic clinic at Columbia University School of Dental and Oral Surgery. The plaster models were scored by using the ABO measuring gauge and the 7 criteria of the ABO grading system. A second analysis was done on the digital models. To determine interexaminer error, a fourth-year dental student at Columbia University served as a second examiner and repeated all the analyses. **Results:** The means of the total score and those for marginal ridges, occlusal contacts, occlusal relationships, overjet, and interproximal contacts were not significantly different between plaster and digital models. However, the means for alignment and buccolingual inclination were significantly different. In addition, the scorings of 2 examiners differed for the 2 methods. **Conclusions:** This finding suggests that alignment and buccolingual inclination should be reevaluated with both methods, and adequate calibration of the examiners is essential to achieve repeatability in both methods. Digital models might be acceptable for use in the ABO model examination. (Am J Orthod Dentofacial Orthop 2005;128:624-9)

In addition to diagnosis, treatment planning, and overall case management, a criterion for determining the acceptability of a case for the American Board of Orthodontics (ABO) phase III examination is final occlusion. A major component of grading these case reports depends on model analysis as a means of evaluating final occlusion.[1] In 1999, in an attempt to be fair and objective, the ABO implemented a grading system for orthodontic models, the objective grading system. This new system assures reliability and consistency. A portion of the overall score a candidate receives on a case report is the sum of scores given for 8 established occlusal criteria. The first 7 are made on orthodontic models: tooth alignment, vertical positioning of marginal ridges, buccolingual inclination of posterior teeth, occlusal relationship, occlusal contacts, overjet, and interproximal contacts. A special instrument, the ABO measuring gauge, was developed for taking these measurements. The eighth criterion, root angulation, is measured on a panoramic radiograph.[1,2] In general, a case report will fail if more than 30 points are deducted. If fewer than 20 points are deducted, the case report will pass.[1] In a study of orthodontic treatment outcomes, Yang-Powers et al[3] concluded that, since the ABO objective criteria were introduced in 1999, an average case that once passed might not pass under the new criteria. However, the new criteria allow candidates to grade their own cases and more easily identify those that are acceptable.

From Columbia University, School of Dental and Oral Surgery.
[a]Former postgraduate orthodontic student; private practice, New York, NY.
[b]Former dental student; postgraduate orthodontic student, New York University College of Dentistry.
[c]Professor, Division of Orthodontics, and chair, Section of Growth and Development.
[d]Professor, Division of Orthodontics.
Reprint requests to: Dr Stella Efstratiadis, Division of Orthodontics, School of Dental and Oral Surgery, Columbia University, 630 W 168th St, New York, NY 10032; e-mail, sse1@columbia.edu.
Submitted, March 2004; revised and accepted, August 2004.
0889-5406/$30.00

doi:10.1016/j.ajodo.2004.08.017

Today, many orthodontists are incorporating digital orthodontic records into their clinical practices and using computer software to assist with diagnosis and treatment planning as well as to alleviate storage constraints for records. Proffit[4] stated that an advantage of using analytical software is that the computer can provide a result quickly. Ho and Freer[5] used a digital version of their graphical analysis of tooth-width discrepancy and concluded that using digital calipers can eliminate errors due to measurement transfer and manual calculation.

Zilberman et al[6] tested the accuracy of measuring tooth size and arch width on conventional plaster models and OrthoCAD (Cadent Inc, Carlstadt, NJ) digital models and stated that both methods had clinically acceptable levels of accuracy and reproducibility. However, they stated that plaster models had a higher accuracy and reproducibility, and concluded that digital models are acceptable in clinical applications but might not be acceptable for research.[6]

Santoro et al[7] evaluated the accuracy of measuring tooth size, overbite, and overjet using OrthoCAD models compared with plaster models. They reported a statistically significant difference for tooth size and overbite, but the differences were considered clinically insignificant (<0.5 mm). There was no difference in overjet between plaster models and digital models.

Tomassetti et al[8] studied the accuracy and efficiency of doing Bolton's tooth-size analysis using manual measurements with a vernier caliper and 3 computerized methods, including OrthoCAD. Although they found no statistically significant differences among the tested methods, there were clinically significant differences (>1.5 mm) for all methods. They concluded that OrthoCAD was among the methods with the greatest differences.

More recently, Quimby et al[9] used a plastic model occlusion (dentoform) as a gold standard to evaluate the errors associated with plaster and computer-based models. They demonstrated that only measurements of maxillary and mandibular space available made on computer-based models differed from measurements made on the dentoform. They concluded that computer-based models appear to be a clinically acceptable alternative to conventional plaster models.

To date, no study has evaluated the ABO objective grading system for orthodontic models by using digital models. The purpose of this study was to determine whether digital models can be used with reasonable accuracy and reliability for assessing the final occlusion of patients at the end of orthodontic treatment.

MATERIAL AND METHODS

Alginate maxillary and mandibular impressions and a wax bite registration were taken on patients at the completion of their orthodontic treatment at the postgraduate orthodontic clinic of the Columbia University School of Dental and Oral Surgery and sent to OrthoCAD to construct plaster and digital models. Forty-eight sets of models from 24 patients were selected to be studied. These included 24 plaster models and 24 digital OrthoCAD models. These models were selected by using only 3 criteria: (1) no deciduous teeth were present, (2) no edentulous spaces were present, and (3) the plaster and digital models of each patient had acceptable molar and canine relationships, overjet, and overbite on visual inspection.

Examiner 1, a postgraduate orthodontic student (P.A.C.), was trained with the voice-over CD-ROM provided by the ABO and the information for the ABO phase III grading tool available in the OrthoCAD v. 2.17 software package.[2,10] Based on the ABO measuring gauge, the 7 criteria of the ABO grading system were scored on the 24 plaster models. At least 4 weeks after the plaster models were scored, a second analysis was done for each patient by using the digital models and grading software, and the same criteria and grading tool.[10] The scores for each of the 7 criteria were summed to derive a total score per model. Each plaster and digital model analysis was also timed from first measurement to final computation.

A senior predoctoral dental student (K.S.) at Columbia, trained as for examiner 1, served as examiner 2 for this study.[2,10] An attempt was made to calibrate the 2 examiners by having each score a selection of 2 plaster models and 2 digital OrthoCAD models independently and jointly review their scorings after each model analysis. Examiner 2 repeated the analyses on the 24 plaster models and the 24 digital models using the same methods as examiner 1.

Statistical analysis

Statistical analyses were applied to investigate the accuracy and repeatability of the methods studied with the aid of the statistical program SAS for Windows v. 8.0 (SAS, Cary, NC). Means and standard deviations of the 7 variables and the total score for plaster and digital models were calculated separately. Means and standard deviations of the difference of the total scores between plaster and digital models were also calculated. Analysis of variance (ANOVA) was applied to the data to see whether the 2 methods have equivalent means for the 7 variables and the total score. An intraclass correlation coefficient of reliability was calculated for

Table I. Data collected by examiner 1 from analyses of 24 plaster models

Patient	*A*	*MR*	*BLI*	*OC*	*OR*	*OJ*	*IPC*	*Score*
1	8	1	8	8	1	8	1	35
2	5	8	5	13	0	12	0	43
3	6	3	4	14	6	11	0	44
4	7	1	6	4	2	3	0	23
5	13	4	6	6	4	5	0	38
6	2	9	4	11	4	5	2	37
7	4	3	7	2	2	10	2	30
8	5	4	4	2	2	3	0	20
9	11	10	9	6	2	7	0	45
10	8	3	5	6	3	6	1	32
11	3	1	6	1	0	6	0	17
12	9	4	7	5	2	8	0	35
13	6	1	7	10	0	10	0	34
14	2	3	1	0	0	5	0	11
15	6	5	6	8	7	5	0	37
16	6	2	9	2	1	1	0	21
17	6	1	6	12	0	0	0	25
18	2	3	5	2	4	3	1	20
19	7	3	6	7	1	12	0	36
20	3	3	7	9	0	4	0	26
21	4	2	4	3	0	6	0	19
22	0	5	4	3	2	10	0	24
23	6	3	6	13	0	8	0	36
24	1	6	4	10	1	2	0	24

A, Alignment; *MR*, marginal ridges; *BLI*, buccolingual inclination; *OC*, occlusal contacts; *OR*, occlusal relationship; *OJ*, overjet; *IPC*, interproximal contacts.
Score, Sum of the 7 criteria (the total score).

Table II. Data collected by examiner 1 from analyses of 24 OrthoCAD models

Patient	*A*	*MR*	*BLI*	*OC*	*OR*	*OJ*	*IPC*	*Score*
1	11	1	11	7	3	4	2	39
2	7	7	6	2	0	3	4	29
3	7	6	5	17	7	5	3	45
4	8	2	8	3	3	7	0	31
5	13	4	8	1	2	6	0	34
6	8	9	1	10	9	5	2	44
7	6	1	2	1	0	6	3	19
8	10	5	6	0	1	5	0	27
9	14	7	11	10	2	2	2	48
10	7	4	7	8	5	1	0	32
11	9	0	2	1	0	1	0	13
12	16	7	6	9	6	1	0	45
13	8	4	10	5	2	2	0	31
14	1	4	4	0	0	5	0	14
15	8	5	12	1	6	8	0	40
16	9	3	12	7	2	5	0	38
17	9	4	7	19	0	4	0	43
18	0	0	8	1	3	12	0	24
19	9	9	6	11	0	4	1	40
20	2	3	7	7	0	4	0	23
21	5	3	5	2	0	3	1	19
22	5	4	3	0	0	9	0	21
23	11	0	7	2	0	8	0	28
24	3	4	7	4	1	2	0	21

A, Alignment; *MR*, marginal ridges; *BLI*, buccolingual inclination; *OC*, occlusal contacts; *OR*, occlusal relationship; *OJ*, overjet; *IPC*, interproximal contacts.
Score, Sum of the 7 criteria (the total score).

each examiner to assess the reliability of OrthoCAD models compared with plaster models.

Interexaminer error was evaluated, and the mean values of the measurements by each examiner were compared. Fixed-examiner and random-examiner models were used.

RESULTS

Tables I-IV contain the data collected by both examiners. The plaster and digital models analyzed by examiner 1 had a total score mean difference of 1.5, which is very small. Therefore, there was a high correlation between the total scores for both models. This finding is supported by a high intraclass correlation coefficient of reliability for the total score (Table V) for both examiners (examiner 1: R = 0.69; examiner 2: R = 0.86).

Table VI shows the *P* values from the ANOVA, which assesses whether plaster and digital models have equivalent means for the 7 criteria scored and for the total score. Based on the *P* values, it was concluded that means of the total score and those for marginal ridges, occlusal contacts, occlusal relationship, overjet, and interproximal contacts were not significantly different between plaster and digital models. A significant *P* value of .05 was used for each test. However, the means for alignment were significantly different between plaster models and digital models. The *P* value for buccolingual inclination was almost .05. Therefore, plaster models and digital models had good correlation except for alignment and buccolingual inclination.

ANOVA was used to examine whether there were differences between the 2 examiners. The *P* value was less than .0001, which demonstrates that there were statistically significant differences between the 2 examiners. The interclass reliability was generally moderate (Table VII). The plaster models had a moderate reliability (fixed-examiner model: R = 0.53; random-examiner model: R = 0.46); however, the reliability was slightly higher for the digital models (fixed-examiner model: R = 0.69; random-examiner model: R = 0.65).

DISCUSSION

The total score, the sum of the 7 measurements of the ABO grading system, was similar for plaster mod-

Table III. Data collected by examiner 2 from analyses of 24 plaster models

Patient	*A*	*MR*	*BLI*	*OC*	*OR*	*OJ*	*IPC*	*Score*
1	11	4	4	9	2	8	1	39
2	8	3	2	5	1	8	2	29
3	11	4	0	17	18	5	1	56
4	11	2	5	8	8	11	0	45
5	13	7	2	5	4	4	0	35
6	6	8	0	11	9	1	2	37
7	5	5	2	3	3	11	1	30
8	10	7	4	5	4	5	0	35
9	14	10	5	12	3	7	2	53
10	8	5	3	12	5	2	0	39
11	6	2	1	3	2	5	0	19
12	16	5	4	15	6	5	0	51
13	10	4	1	17	5	4	0	47
14	9	6	2	0	1	5	0	22
15	17	8	7	3	12	5	0	52
16	6	2	6	11	1	2	1	29
17	10	4	5	18	1	6	0	44
18	6	4	5	8	5	8	1	37
19	15	10	5	6	7	10	0	53
20	9	1	2	10	2	6	0	30
21	14	2	1	6	2	5	0	30
22	18	7	4	4	2	2	0	37
23	12	3	8	5	2	7	0	37
24	12	6	4	7	2	4	0	35

A, Alignment; *MR*, marginal ridges; *BLI*, buccolingual inclination; *OC*, occlusal contacts; *OR*, occlusal relationship; *OJ*, overjet; *IPC*, interproximal contacts.
Score, Sum of the 7 criteria (the total score).

Table IV. Data collected by examiner 2 from analyses of 24 OrthoCAD models

Patient	*A*	*MR*	*BLI*	*OC*	*OR*	*OJ*	*IPC*	*Score*
1	13	0	9	5	5	7	1	40
2	10	3	2	2	1	0	2	20
3	17	3	3	28	8	9	4	72
4	12	3	9	4	2	19	0	49
5	19	5	4	0	2	3	0	33
6	6	7	2	14	8	0	1	38
7	6	2	1	3	4	11	1	28
8	13	5	9	2	2	2	0	33
9	19	8	8	11	0	6	2	54
10	10	3	6	11	3	0	0	33
11	9	1	1	3	4	2	0	20
12	18	4	5	11	1	2	0	41
13	14	4	6	5	6	5	0	40
14	8	3	5	0	1	0	0	17
15	14	4	8	3	7	3	0	39
16	8	4	11	5	4	2	0	34
17	15	4	7	18	2	3	0	49
18	5	2	9	2	7	10	0	35
19	17	5	8	14	3	7	0	54
20	7	2	4	8	5	2	0	28
21	17	0	5	1	5	6	0	34
22	12	6	2	4	5	4	0	33
23	15	1	5	5	2	8	0	36
24	3	2	10	7	4	5	0	31

A, Alignment; *MR*, marginal ridges; *BLI*, buccolingual inclination; *OC*, occlusal contacts; *OR*, occlusal relationship; *OJ*, overjet; *IPC*, interproximal contacts.
Score, Sum of the 7 criteria (the total score).

els and digital models. This is supported by the high level of intraclass correlation coefficient of reliability and the small difference of the 2 means for total score. However, alignment and buccolingual inclination had statistically significant differences between plaster and digital models. Therefore, the differences for alignment and buccolingual inclination demonstrate measurement differences between the 2 techniques. This might be due to difficulty in identifying the same landmarks on the plaster and OrthoCAD models. The limitations in enlarging digital models significantly might have contributed to this problem. Houston[11] reported that 1 of the greatest sources of random error is the difficulty in identifying landmarks. Zilberman[6] stated that this is particularly a concern for digital models because a 3-dimensional structure is viewed as a 2-dimensional image, and identifying landmarks becomes more difficult.

During the analyses, 2 concerns were noted. In measuring alignment on the plaster models, generally a macroscopic assessment of alignment is done, and the amount of misalignment of adjacent teeth is measured with the ABO measuring gauge. However, the same measurement on the digital models gains in microscopic detail. Two points are placed on each tooth to assess alignment. Slightly moving each point does not significantly change the visual macroscopic alignment of the adjacent teeth; however, points might be deducted or spared. There appears to be a range wherein points can be spared or deducted depending on the ultimate position of the 2 selected points, even though the resulting line for alignment is in an acceptable position. Therefore, the examiners might find it difficult to pinpoint the exact mesial and distal points to be used to evaluate alignment, making it difficult to consistently

Table V. Intraclass correlation coefficient of reliability

	Intraclass correlation (R)
Examiner 1	0.69
Examiner 2	0.86

R has range of 0 (no reliability) to 1 (perfect reliability).
High intraclass correlation coefficient of reliability for both examiners (examiner 1 = 0.69, examiner 2 = 0.86) demonstrates that there is no significant difference between plaster and digital model evaluations when they are done by same examiner.

Table VI. Analysis of variance (ANOVA)

	SD_1	SD_2	$Mean_1$	$Mean_2$	F (1,23)	P
Alignment	3.89	3.11	7.75	5.42	22.21	<.0001
Marginal ridges	2.60	2.48	4.00	3.67	0.54	.4694
Buccolingual inclination	3.06	1.81	6.71	5.67	4.25	.0507
Occlusal contacts	5.31	4.24	5.33	6.54	1.61	.2169
Occlusal relationship	2.63	1.97	2.17	1.83	0.88	.3567
Overjet	2.75	3.42	4.67	6.25	2.80	.1077
Interproximal contacts	1.22	0.62	0.75	0.29	3.87	.0613
Total score	10.47	9.29	31.17	29.67	0.92	.3467

SD_1, Standard deviation of mean for plaster models; SD_2, standard deviation of mean for digital models; $Mean_1$, mean of 7 criteria and total score of plaster models; $Mean_2$, mean of 7 criteria and total score of digital models.
P from the ANOVA, which assesses whether plaster and digital models have equivalent means for 7 criteria scored and for total score.

Table VII. Interclass reliability

	Interclass reliability	
	Fixed-examiner (R)	*Random-examiner (R)*
Plaster models	0.53	0.46
Digital models	0.69	0.65

R has range of 0 (no reliability) to 1 (perfect reliability). Interclass reliability, which describes differences between examiners, was generally moderate. Plaster models had moderate reliability (fixed-examiner model R = 0.53; random-examiner model R = 0.46); however, reliability was slightly higher for digital models (fixed-examiner model R = 0.69; random-examiner model R = 0.65).

and accurately make measurements. Tomassetti et al[8] reached the same conclusion in using the OrthoCAD program for the Bolton tooth-size analysis. Quimby et al[9] suggested that the manufacturer's process of producing plaster casts and the process of scanning and recording data points from the plaster model might contribute to the differences in measurements between the plaster and computer-based models. In measuring buccolingual inclination, we noted a variation with OrthoCAD. The buccolingual inclination of a tooth is assessed from a plane created from a cusp tip on that tooth extended to the cusp tip of the contralateral tooth. Ordinarily, contralateral human teeth are not positioned parallel to each other. This is particularly true of premolars, which converge toward the anterior midline. Therefore, it is difficult to assess buccolingual inclination on digital models with only a single line because the line cannot symmetrically bisect the occlusal surface of each contralateral tooth at the cusp tip. This measurement on plaster models is different because the plaster models are truly 3-dimensional; therefore, buccolingual inclination is visualized from the plane created by the ABO measuring gauge. This variation can be avoided if this plane is formed by 2 lines joined to form an angle at the sagittal plane. If these lines are maintained in the same plane, they can be manipulated by the examiner to bisect each contralateral tooth at the cusp tip and therefore to more accurately measure buccolingual inclination.

The scorings did not correlate very well between examiner 1 (the postgraduate orthodontic student) and examiner 2 (the predoctoral dental student). The predoctoral dental student was not familiar with either the ABO model analysis or the OrthoCAD program before the study. Also, the calibration method, necessary to assure coordination among examiners, might have been ineffective or insufficient. Calibration is an important part of the process of preparing ABO examiners. Two months before and immediately after examinations, the examiners go through measurement calibration exercises to establish an effective confidence interval in scoring.[2]

The use of digital models is advantageous for clinical practice and has great potential if the models are reliable and easy to use, and further advances in the OrthoCAD system—and probably other digital systems—might be anticipated. At present, digital models have certain variations when compared with conventional plaster models; this agrees with previous studies.[6,8] However, these variations can be resolved to the benefit of the clinician. Moreover, the slightly higher reliability of digital models (fixed-examiner model: R = 0.69; random-examiner model: R = 0.65) versus the plaster models (fixed-examiner model: R = 0.53; random-examiner model: R = 0.46) makes them very promising.

CONCLUSIONS

1. In general, plaster and digital models showed similar values for total score and for 5 of the 7 criteria measured.
2. Some improvement in software accuracy or usage is necessary to attain greater reliability in measuring alignment and buccolingual inclination.

3. Interexaminer reliability was somewhat better with digital models.
4. Calibration of examiners is essential for both manual and digital methods of measurement.
5. The use of digital models produced by the OrthoCAD system seems to be a viable alternative to plaster models.

REFERENCES

1. Casko J, Vaden J, Kokich V, Damone J, James R, Cangialosi T, et al. Objective grading system for dental casts and panoramic radiographs. Am J Orthod Dentofacial Orthop 1998;114:589-99.
2. James RD. Objective cast and panoramic radiograph grading system. Am J Orthod Dentofacial Orthop 2002;122:450.
3. Yang-Powers LC, Sadowsky C, Rosenstein S, BeGole EA. Treatment outcome in a graduate orthodontic clinic using the American Board of Orthodontics grading system. Am J Orthod Dentofacial Orthop 2002;122:451-5.
4. Proffit WR. Contemporary orthodontics. 3rd ed. Saint Louis: Mosby; 2000. p. 169-70.
5. Ho CT, Freer TJ. A computerized tooth-width analysis. J Clin Orthod 1999;33:498-503.
6. Zilberman O, Huggare JAV, Parikakis KA. Evaluation of the validity of tooth size and arch width measurements using conventional and three-dimensional virtual orthodontic models. Angle Orthod 2003;73:301-6.
7. Santoro M, Galkin S, Teredesai M, Nicolay O, Cangialosi TJ. Comparison of measurements made on digital and plaster models. Am J Orthod Dentofacial Orthop 2003;124:101-5.
8. Tomassetti JJ, Taloumis LJ, Denny JM, Fischer JR Jr. A comparison of 3 computerized Bolton tooth-size analyses with a commonly used method. Angle Orthod 2001;71:351-7.
9. Quimby ML, Vig KWL, Rashid RG, Firestone AR. The accuracy and reliability of measurements made on computer-based digital models. Angle Orthod 2004;74:298-303.
10. OrthoCAD, USA. Available at: http://www.orthocad.com. Accessed December 5, 2002.
11. Houston WJ. The analysis of errors in orthodontic measurements. Am J Orthod 1983;83:382-90.