Orthodontics & Craniofacial Research

ORIGINAL ARTICLE

N. S. Luu
M-A. Mandich
C. Flores-Mir
T. El-Bialy
G. Heo
J. P. Carey
P. W. Major

The validity, reliability, and time requirement of study model analysis using cone-beam computed tomography–generated virtual study models

Authors' affiliations:
N. S. Luu, M-A. Mandich, Private practice, Dynamic Orthodontics, Leduc, AB, Canada
C. Flores-Mir, T. El-Bialy, G. Heo, Orthodontic Graduate Program, Faculty of Medicine and Dentistry, University of Alberta, Edmonton, AB, Canada
J. P. Carey, Department of Mechanical Engineering, Faculty of Engineering, University of Alberta, Edmonton, AB, Canada
P. W. Major, Faculty of Medicine and Dentistry, University of Alberta, Edmonton, AB, Canada

Correspondence to:
Dr Paul W. Major
Faculty of Medicine and Dentistry
5-478, Edmonton Clinic Health Academy
University of Alberta
11405-87 Avenue NW
Edmonton,
AB Canada T6G 1C9
E-mail: major@ualberta.ca

Date:
Accepted 24 March 2013

DOI: 10.1111/ocr.12024

© 2013 John Wiley & Sons A/S.
Published by John Wiley & Sons Ltd

Luu N. S., Mandich M-A., Flores-Mir C., El-Bialy T., Heo G., Carey J. P., Major P. W. The validity, reliability, and time requirement of study model analysis using cone-beam computed tomography–generated virtual study models
Orthod Craniofac Res 2014; **17**: 14–26. © 2013 John Wiley & Sons A/S. Published by John Wiley & Sons Ltd

Structured Abstract

Objectives – To investigate the validity, reliability, and time spent to perform a full orthodontic study model analysis (SMA) on cone-beam computed tomography (CBCT)-generated dental models (Anatomodels) compared with conventional plaster models and a subset of extracted premolars.

Setting and Sample Population – A retrospective sample of 30 consecutive patient records with fully erupted permanent dentition, good-quality plaster study models, and CBCT scans. Twenty-two extracted premolars were available from eleven of these patients.

Materials and Methods – Five evaluators participated in the inter-rater reliability study and one evaluator for the intrarater reliability and validity studies. Agreement was assessed by ICC and cross-tabulations, while mean differences were investigated using paired-sample *t*-tests and repeated-measures ANOVA.

Results – For all three modalities studied, intrarater reliability was excellent, inter-rater reliability was moderate to excellent, validity was poor to moderate, and performing SMA on Anatomodels took twice as long as on plaster.

Conclusions – Study model analysis using CBCT-generated study models was reliable but not always valid and required more time to perform when compared with plaster models.

Key words: cone-beam computed tomography; dental models; imaging, three-dimensional; odontometry; reproducibility of results

Introduction

Study model analysis (SMA) is an important process for diagnosis and treatment planning in dentistry (1). Conventionally, this is performed using plaster models, but the current trend is moving toward using virtual models. Because SMA has customarily been performed on plaster models, such measurements can be considered the reference standard as alternative to measurements on live teeth.

A current diagnostic tool is cone-beam computed tomography (CBCT), which is a theoretically undistorted (2) radiographic approach to visualizing anatomy in 3D and allows manual digital tooth segmentation including the roots (3). Anatomage (San Jose, CA, USA) can create Anatomodels, virtual models generated from CBCT scans that offer an intriguing alternative to plaster models. A full SMA using CBCT-generated virtual study models has not yet been reported in the literature, but some parameters have been previously validated against laser-acquired (4) and plaster (5) study models.

The purpose of this study was to investigate the reliability, validity, and time requirements of quantitative and qualitative measurements in a full SMA using Anatomodels compared with plaster dental study models as well as a subset of extracted premolars.

Materials and methods

Ethical approval was obtained through the Health Research Ethics Board at the University of Alberta (Pro00010202). Based on the data from a previous study comparing virtual models to plaster (6), we took a standard deviation, σ, of 0.58 and set a statistical power, $1-\beta$, of 0.9 to detect a difference, δ, between Anatomodels and plaster of 0.5 mm at a significance level, α, of 0.05. The sample size was calculated applying 0.5 mm as the mean difference as specified by Rosner (7). A minimum of 27 patients was calculated. Ultimately, we investigated a retrospective sample of 30 consecutive patients chosen from the University of Alberta Graduate Orthodontic Clinic between February 2007 and November 2009. The inclusion criteria were patients with fully erupted permanent dentition whose diagnostic records included good-quality plaster study models and CBCT scans.

As this was a retrospective study, orthodontic treatment planning was independently completed and 11 patients were prescribed premolar extraction therapy. A total of twenty-two extracted premolars used in a separate study (8) were available for direct assessment of mesiodistal widths in this study.

Commonly used qualitative and quantitative parameters were included in the full SMA used for this study (Table 1).

A flowchart for the research plan is illustrated in Fig. 1. Three modalities were compared in this study: plaster dental study models, CBCT-generated study models, and matched samples of extracted premolars. Our experimental design had three arms of study across all three modalities: intrarater reliability, inter-rater reliability, and validity. In addition, the time spent to perform each SMA on plaster and Anatomodels was tracked.

Linear measurements on plaster study models and extracted teeth were performed using the same digital caliper (Model IP67, Mitutoyo Canada, Mississauga, ON, Canada) for all evaluators. The product specifications stated a resolution of 0.01 mm and an accuracy of $\pm$0.02 mm. Measurements of overbite and overjet on plaster models were taken with a periodontal probe to the nearest 0.5 mm.

Cone-beam computed tomography scans for the subjects with extracted premolars were taken with the 12-bit iCAT (Imaging Sciences International, Hatfield, PA, USA) set to a 40-s scan, 120 kVp, 47 mAs, to allow image reconstruction into DICOM format at $0.25 \times 0.25 \times 0.25$ mm voxels. The rest of the CBCT scans using the same iCAT machine were prescribed at 120 kVp, 24 mAs, and voxel sizes of $0.30 \times 0.30 \times 0.30$ mm. The DICOM datasets were uploaded to Anatomage and processed into Anatomodels, the company's product name for CBCT-generated study models.

Anatomodels were viewed using the software InVivo 5.0 build 229 (Anatomage), and linear

Table 1. **Commonly used parameters in the full study model analysis used for this study: (A) qualitative; (B) quantitative, 2-landmarks; (C) quantitative, >2-landmarks**

Count	Parameter	Levels
(A) Qualitative		
1	Right molar angle classification	I, II, or III
1	Right canine angle classification	I, II, or III
1	Left molar angle classification	I, II, or III
1	Left canine angle classification	I, II, or III
1	Maxillary arch symmetry	Symmetric or asymmetric
1	Maxillary arch size	Narrow, average, or expanded
1	Maxillary arch shape	U-shaped, V-shaped, tapered, or squared
1	Mandibular arch symmetry	Symmetric or asymmetric
1	Mandibular arch size	Narrow, average, or expanded
1	Mandibular arch shape	U-shaped, V-shaped, tapered, or squared
(B) Quantitative, 2-landmarks		
1	Maxillary intermolar width	Mesiopalatal cusp tips 1-6 and 2-6
1	Maxillary intercanine width	Cusp tips 1-3 and 2-3
1	Mandibular intermolar width	Central fossa 3-6 and 4-6
1	Mandibular intercanine width	Cusp tips 3-3 and 4-3
12	Maxillary mesiodistal widths	Teeth 1-6, 1-5, 1-4, 1-3, 1-2, 1-1, 2-1, 2-2, 2-3, 2-4, 2-5, 2-6
12	Mandibular mesiodistal widths	Teeth 4-6, 4-5, 4-4, 4-3, 4-2, 4-1, 3-1, 3-2, 3-3, 3-4, 3-5, 3-6
(C) Quantitative, >2-landmarks		
1	Maxillary arch perimeter	Four segments, mesial to 1-6 and 2-6
1	Maxillary arch crowding	Mesial to 1-6 and 2-6
1	Mandibular arch perimeter	Four segments, mesial to 3-6 and 4-6
1	Mandibular arch crowding	Mesial to 3-6 and 4-6
1	Bolton 6	Anterior ratio in millimeters
1	Bolton 12	Overall ratio in millimeters

measurements were shown onscreen to the nearest 0.01 mm. All evaluators were given a 5-min tutorial on how to perform the measurements using the software and had a chance to practice on a sample Anatomodel not included in this study. Similarly with the plaster models, the desired landmarks for each parameter of the SMA on Anatomodels were reviewed before evaluators performed their measurements.

The time required to perform all of the intended measurements in a SMA was calculated by taking the difference between the recorded start and finish times.

Intrarater reliability was assessed over five trials by the primary author. Ten subjects were randomly chosen from the subset of eleven subjects who had extracted premolars so that useful comparisons across the three modalities could be made. For both plaster and Anatomodels, timed SMA was repeated for 10 subjects five times at intervals of 10 days apart, with assessments limited to five unique cases per day in random order to minimize bias due to fatigue. Similarly, twenty-two extracted premolars were measured in random order from the subset of 11 subjects, repeated five times at 10-day intervals.

Inter-rater reliability was assessed across five evaluators: one senior orthodontic resident and orthodontists of 0.5, 1, 16, and 23 years of clinical experience. Time to complete the analysis was recorded.

For the validity studies, the principal investigator performed timed SMA on 30 subjects in

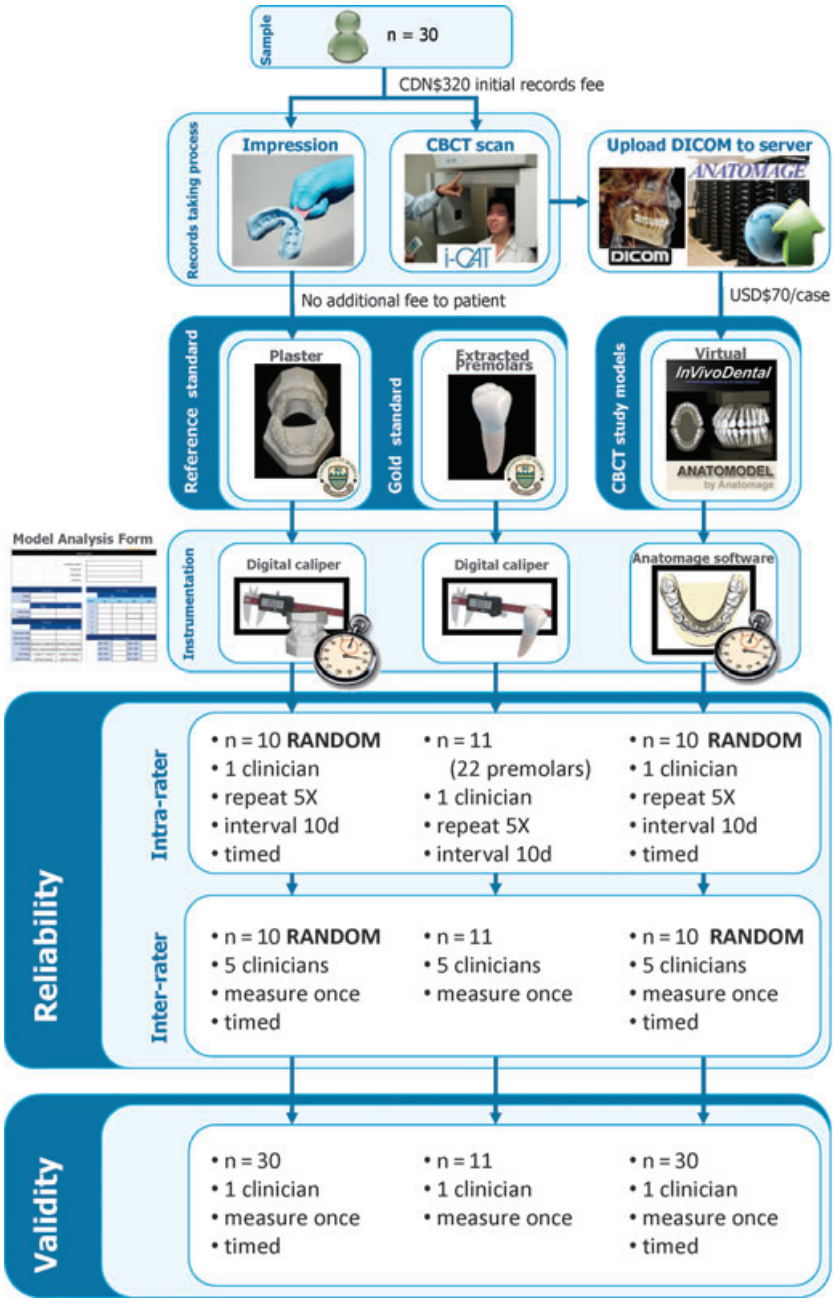

Fig. 1. Study flowchart.

random order on Anatomodels and then plaster, limited to only five cases per day. Twenty-two extracted premolars were measured in random order from the subset of 11 subjects.

The time to perform the full SMA in each trial of the studies was recorded by the principle investigator.

Agreement of measurements was assessed by way of intraclass correlation coefficient (ICC) and cross-tabulations for the reliability and validity studies (SPSS version 16, IBM, Armonk, NY, USA). Mean difference of measurements was investigated by way of paired *t*-tests in the validity studies and repeated-measures ANOVA in the reliability studies.

In this study, we will consider all ICC values below 0.6 to be poor, above 0.6 to be moderate, above 0.7 to be good, and above 0.8 as excellent. We assumed thresholds for clinically relevant mean differences for 2-landmark linear measurements of 0.5 mm and for >2-landmark linear measurements of 2.0 mm.

Results

Statistical model assumptions were checked and satisfied prior to performing the statistical tests.

A summary of the intrarater reliability results from ICC and repeated-measures ANOVA tests is presented for Anatomodels in Table 2, plaster in Table 3, and extracted premolars in Table 4. A summary of the inter-rater reliability results from ICC and repeated-measures ANOVA tests is presented for Anatomodels in Table 5, plaster in Table 6, and extracted premolars in Table 7.

The validity of measurements on 30 Anatomodels compared with plaster (Table 8) was mostly poor to moderate in terms of agreement but with low mean differences. A number of parameters had ICC values below 0.6 and wide 95% confidence intervals including teeth 1-1, 1-3, 2-3, 2-5, 3-4, 3-5, 3-6, 4-5, and 4-6, maxillary arch perimeter, and Bolton anterior and Bolton overall measurements. There was statistical evidence (p-value <0.05) to show that differences existed between Anatomodels and plaster for the mean measurements of teeth 1-1, 1-2, 1-3, 1-5, 2-1, 2-2, 2-3, 2-4, 2-5, mandibular intermolar width, maxillary and mandibular arch perimeter and crowding, and Bolton anterior and overall measurements; however, only maxillary arch perimeter had a magnitude of mean difference, 3.38 mm and 95% CI (2.48, 4.28), that exceeded the clinically significant threshold.

Compared to extracted premolars (Table 9), Anatomodels had ICC values well above 0.9 and measurements on average up to 0.08 mm larger, while plaster (Table 10) had ICC values only slightly above 0.7 with measurements on average up to 0.17 mm smaller. All of the p-values were above 0.05. Again, analysis was only attempted for teeth 1-4 and 2-4 because the sample sizes for these teeth were not too small.

Within one evaluator over 10 subjects during the intrarater reliability study, at worst, it took on average an additional 5.91 min longer than the best trial to perform a SMA on Anatomodels. The same comparison in plaster revealed on average only 2.34 additional minutes over the

Table 2. **Intrarater, Anatomodels: ICC and repeated-measures ANOVA mean differences shown for each parameter, grouped by linear measurements requiring two landmarks, and those requiring more than two landmarks**

		Intrarater reliability		Mean differences (mm)		
Parameter	N	ICC	95% CI	Worst	Best	p-value
Anatomodels, linear measurements, 2 landmarks						
Overjet	10	0.905	0.788, 0.971	0.43	0.02	0.113
Overbite	10	0.947	0.871, 0.985	0.53	0.01	0.016*
Tooth 1-1	10	0.871	0.723, 0.960	0.10	0.01	0.728
Tooth 1-2	10	0.975	0.940, 0.993	0.03	0.00	0.987
Tooth 1-3	10	0.916	0.813, 0.975	0.10	0.01	0.412
Tooth 1-4	10	0.919	0.818, 0.976	0.06	0.00	0.869
Tooth 1-5	10	0.927	0.835, 0.978	0.09	0.00	0.432
Tooth 1-6	10	0.913	0.799, 0.974	0.28	0.05	0.015*
Tooth 2-1	10	0.962	0.911, 0.989	0.12	0.00	0.337
Tooth 2-2	10	0.965	0.917, 0.990	0.12	0.00	0.334
Tooth 2-3	10	0.920	0.820, 0.976	0.09	0.00	0.288
Tooth 2-4	10	0.915	0.809, 0.975	0.06	0.00	0.614
Tooth 2-5	10	0.898	0.777, 0.969	0.10	0.00	0.344
Tooth 2-6	10	0.863	0.711, 0.958	0.19	0.01	0.310
Tooth 3-1	10	0.962	0.909, 0.989	0.04	0.00	0.957
Tooth 3-2	10	0.905	0.790, 0.972	0.06	0.01	0.647
Tooth 3-3	10	0.876	0.734, 0.962	0.12	0.01	0.568
Tooth 3-4	10	0.813	0.623, 0.940	0.13	0.01	0.213
Tooth 3-5	10	0.894	0.767, 0.968	0.09	0.01	0.563
Tooth 3-6	10	0.919	0.819, 0.976	0.08	0.00	0.726
Tooth 4-1	10	0.945	0.873, 0.984	0.07	0.01	0.569
Tooth 4-2	10	0.957	0.900, 0.987	0.06	0.00	0.618
Tooth 4-2a	1†	–	–	0.56	0.01	–
Tooth 4-3	10	0.866	0.716, 0.959	0.11	0.00	0.482
Tooth 4-4	10	0.867	0.717, 0.959	0.17	0.00	0.136
Tooth 4-5	10	0.977	0.944, 0.993	0.07	0.01	0.212
Tooth 4-6	10	0.902	0.784, 0.970	0.15	0.00	0.335
Mx_IMW	10	0.984	0.959, 0.995	0.52	0.09	0.010*
Mx_ICW	10	0.934	0.849, 0.980	0.31	0.03	0.482
Mn_IMW	10	0.965	0.918, 0.990	0.27	0.00	0.639
Mn_ICW	10	0.968	0.924, 0.991	0.32	0.02	0.198
Anatomodels, linear measurements, >2 landmarks						
Mx_Perim	10	0.929	0.802, 0.980	1.71	0.04	<0.001*
Mx_Crowd	10	0.889	0.746, 0.967	1.47	0.03	0.029*
Mn_Perim	10	0.934	0.850, 0.981	0.82	0.09	0.193
Mn_Crowd	10	0.920	0.820, 0.976	1.10	0.01	0.093

Table 2. **(continued)**

		Intrarater reliability		Mean differences (mm)		
Parameter	N	ICC	95% CI	Worst	Best	*p*-value
Bolton 6	10	0.936	0.853, 0.981	0.30	0.01	0.575
Bolton 12	10	0.887	0.775, 0.966	0.78	0.13	0.165

ICC, intraclass correlation coefficient; CI, confidence interval; Mx_, maxillary; Mn_, mandibular; IMW, intermolar width; ICW, intercanine width; Perim, arch perimeter; Crowd, crowding if negative; Bolton 6/Bolton 12, Bolton millimeter, positive when mandibular excess

†Cannot be computed because the sum of caseweights is less than or equal 1.

*p-value <0.05.

best trial. Between 5 evaluators over 10 subjects during the inter-rater reliability study, at worst, the slowest evaluator may take 8.15 and 2.33 min longer than the fastest evaluator for Anatomodels and plaster, respectively.

Across all 30 subjects during the validity study, the average time to perform all of the component measurements in the SMA was about 10 min using Anatomodels and 6 min using plaster. There was convincing evidence to show a statistical difference in the mean time to perform the same SMA in Anatomodels of 3.96 min and 95% CI (3.44, 4.48) longer than with plaster.

Discussion

This study sought to investigate the performance of SMA using Anatomodels compared with plaster study models. A comprehensive analysis was performed on the validity, intrarater reliability, and inter-rater reliability using ten nominal (categorical) parameters, thirty-scale (linear) 2-landmark parameters, and six-scale (linear) >2-landmark parameters over three modalities.

Differences of 0.5 mm for tooth widths and 5% for larger measurements were determined to be clinically significant by Asquith et al. (9). Furthermore, Goonewardene et al. (10) argued that extraction vs. non-extraction treatment plans could be influenced by variations of 1–2 mm in crowding measurements. But at less than 1.5 mm of tooth structure discrepancy in an arch, Mullen et al. (11) decided that this could

Table 3. **Intrarater, plaster: ICC and repeated-measures ANOVA mean differences shown for each parameter, grouped by linear measurements requiring two landmarks, and those requiring more than two landmarks**

		Intrarater reliability		Mean differences (mm)		
Parameter	N	ICC	95% CI	Worst	Best	*p*-value
Plaster models, linear measurements, 2 landmarks						
Overjet	10	0.926	0.832, 0.978	0.20	0.05	0.420
Overbite	10	0.935	0.852, 0.981	0.20	0.00	0.723
Tooth 1-1	10	0.980	0.946, 0.995	0.13	0.01	0.001*
Tooth 1-2	10	0.989	0.973, 0.997	0.07	0.00	0.358
Tooth 1-3	10	0.940	0.862, 0.982	0.08	0.00	0.203
Tooth 1-4	10	0.946	0.875, 0.984	0.05	0.01	0.688
Tooth 1-5	10	0.949	0.878, 0.985	0.12	0.02	0.028*
Tooth 1-6	10	0.931	0.841, 0.980	0.13	0.01	0.054
Tooth 2-1	10	0.990	0.976, 0.997	0.05	0.00	0.459
Tooth 2-2	10	0.980	0.949, 0.994	0.14	0.03	0.027*
Tooth 2-3	10	0.936	0.855, 0.981	0.08	0.00	0.224
Tooth 2-4	10	0.830	0.651, 0.946	0.15	0.01	0.274
Tooth 2-5	10	0.960	0.906, 0.988	0.09	0.00	0.252
Tooth 2-6	10	0.784	0.573, 0.930	0.08	0.00	0.860
Tooth 3-1	10	0.971	0.932, 0.992	0.05	0.00	0.503
Tooth 3-2	10	0.970	0.925, 0.991	0.09	0.01	0.012*
Tooth 3-3	10	0.942	0.867, 0.983	0.10	0.00	0.120
Tooth 3-4	10	0.901	0.781, 0.970	0.04	0.00	0.889
Tooth 3-5	10	0.847	0.677, 0.952	0.05	0.00	0.813
Tooth 3-6	10	0.937	0.852, 0.982	0.21	0.02	0.089
Tooth 4-1	10	0.902	0.782, 0.970	0.07	0.00	0.600
Tooth 4-2	10	0.941	0.864, 0.982	0.06	0.00	0.420
Tooth 4-2a	1†	–	–	0.09	0.01	–
Tooth 4-3	10	0.952	0.887, 0.986	0.12	0.01	0.062
Tooth 4-4	10	0.945	0.873, 0.984	0.05	0.00	0.689
Tooth 4-5	10	0.901	0.782, 0.970	0.14	0.00	0.118
Tooth 4-6	10	0.964	0.916, 0.990	0.10	0.01	0.274
Mx_IMW	10	0.992	0.981, 0.998	0.18	0.04	0.648
Mx_ICW	10	0.985	0.963, 0.996	0.11	0.01	0.843
Mn_IMW	10	0.977	0.945, 0.993	0.13	0.00	0.886
Mn_ICW	10	0.978	0.947, 0.994	0.15	0.01	0.750
Plaster models, linear measurements, >2 landmarks						
Mx_Perim	10	0.794	0.593, 0.934	1.33	0.04	0.358
Mx_Crowd	10	0.735	0.502, 0.934	1.31	0.16	0.394
Mn_Perim	10	0.927	0.833, 0.978	0.76	0.02	0.095
Mn_Crowd	10	0.915	0.799, 0.975	1.27	0.16	0.006*

Table 3. **(continued)**

		Intrarater reliability		Mean differences (mm)		
Parameter	N	ICC	95% CI	Worst	Best	*p*-value
Bolton 6	10	0.934	0.848, 0.980	0.18	0.01	0.700
Bolton 12	10	0.899	0.779, 0.970	0.36	0.04	0.623

ICC, intraclass correlation coefficient; CI, confidence interval; Mx_, maxillary; Mn_, mandibular; IMW, intermolar width; ICW, intercanine width; Perim, arch perimeter; Crowd, crowding if negative; Bolton 6/Bolton 12, Bolton millimeter, positive when mandibular excess.
†Cannot be computed because the sum of caseweights is less than or equal 1.
*p-value < 0.05.

Table 4. **Intrarater, extracted premolars: ICC and repeated-measures ANOVA mean differences shown for mesiodistal width measurements of each extracted premolar**

		Intrarater reliability		Mean differences (mm)		
Parameter	N	ICC	95% CI	Worst	Best	*p*-value
Extracted premolars						
Tooth 14	8	0.998	0.995, 1.000	0.02	0.00	0.532
Tooth 15	3†					
Tooth 24	9	0.999	0.997, 1.000	0.02	0.00	0.177
Tooth 25	2†					

ICC, intraclass correlation coefficient; CI, confidence interval.
†Test values were not reported due to low sample size.

be clinically insignificant. Although very few publications state a level of clinical significance, our proposed thresholds for clinical relevance of 0.5 mm for 2-landmark linear measurements and 2.0 mm for >2-landmark linear measurements, then, would be in line with these authors. If the clinically relevant thresholds were reduced, our study would reveal a greater number of linear measurements that should be interpreted with caution.

The act of performing measurements has an element of uncertainty and is subject to error (12). Uncertainty can be the result of random or systematic effects. Error can arise from problems with the measuring instrument, instability of the item being measured, difficulties in the measurement process, improper calibration, operator skill, sampling biases, and environmental factors (12).

Table 5. **Inter-rater, Anatomodels: ICC and repeated-measures ANOVA mean differences shown for each parameter, grouped by linear measurements requiring two landmarks, and those requiring more than two landmarks**

		Inter-rater reliability		Mean differences (mm)		
Parameter	N	ICC	95% CI	Worst	Best	*p*-value
Anatomodels, linear measurements, 2 landmarks						
Overjet	10	0.864	0.710, 0.958	0.21	0.02	0.808
Overbite	10	0.906	0.785, 0.972	0.69	0.04	0.021*
Tooth 1-1	10	0.909	0.797, 0.973	0.09	0.01	0.765
Tooth 1-2	10	0.939	0.857, 0.982	0.24	0.02	0.050
Tooth 1-3	10	0.771	0.556, 0.925	0.19	0.01	0.445
Tooth 1-4	10	0.795	0.592, 0.934	0.23	0.01	0.145
Tooth 1-5	10	0.818	0.631, 0.942	0.21	0.02	0.138
Tooth 1-6	10	0.867	0.718, 0.959	0.16	0.01	0.615
Tooth 2-1	10	0.904	0.787, 0.971	0.17	0.02	0.356
Tooth 2-2	10	0.619	0.351, 0.862	0.38	0.01	0.438
Tooth 2-3	10	0.771	0.505, 0.928	0.34	0.03	0.001*
Tooth 2-4	10	0.734	0.500, 0.910	0.17	0.01	0.564
Tooth 2-5	10	0.691	0.431, 0.894	0.32	0.01	0.005*
Tooth 2-6	10	0.735	0.497, 0.912	0.11	0.00	0.964
Tooth 3-1	10	0.942	0.866, 0.983	0.15	0.01	0.129
Tooth 3-2	10	0.684	0.412, 0.892	0.31	0.01	0.013*
Tooth 3-3	10	0.850	0.684, 0.954	0.07	0.01	0.927
Tooth 3-4	10	0.559	0.289, 0.830	0.32	0.00	0.080
Tooth 3-5	10	0.771	0.556, 0.925	0.21	0.00	0.313
Tooth 3-6	10	0.815	0.627, 0.941	0.23	0.01	0.186
Tooth 4-1	10	0.808	0.610, 0.939	0.23	0.00	0.035*
Tooth 4-2	10	0.827	0.647, 0.945	0.18	0.01	0.309
Tooth 4-2a	1†	–	–	0.41	0.00	–
Tooth 4-3	10	0.799	0.600, 0.935	0.13	0.03	0.610
Tooth 4-4	10	0.739	0.508, 0.913	0.17	0.01	0.511
Tooth 4-5	10	0.818	0.631, 0.942	0.21	0.03	0.362
Tooth 4-6	10	0.729	0.495, 0.908	0.28	0.04	0.215
Mx_IMW	10	0.837	0.663, 0.949	0.75	0.05	0.021*
Mx_ICW	10	0.749	0.524, 0.916	1.11	0.07	0.178
Mn_IMW	10	0.860	0.701, 0.957	1.17	0.02	0.120
Mn_ICW	10	0.934	0.832, 0.981	0.72	0.02	0.001*
Anatomodels, linear measurements, >2 landmarks						
Mx_Perim	10	0.679	0.268, 0.902	5.72	0.13	<0.001*
Mx_Crowd	10	0.566	0.188, 0.849	5.51	0.87	<0.001*
Mn_Perim	10	0.776	0.495, 0.931	3.03	0.08	<0.001*
Mn_Crowd	10	0.742	0.481, 0.916	3.00	0.11	<0.001*

Table 5. **(continued)**

		Inter-rater reliability		Mean differences (mm)		
Parameter	N	ICC	95% CI	Worst	Best	*p*-value
Bolton 6	10	0.884	0.749, 0.964	0.39	0.06	0.437
Bolton 12	10	0.833	0.654, 0.947	0.54	0.01	0.815

ICC, intraclass correlation coefficient; CI, confidence interval; Mx_, maxillary; Mn_, mandibular; IMW, intermolar width; ICW, intercanine width; Perim, arch perimeter; Crowd, crowding if negative; Bolton 6/Bolton 12, Bolton millimeter, positive when mandibular excess.
†Cannot be computed because the sum of caseweights is less than or equal 1.
*p-value < 0.05.

In this study, we found that parameters utilizing only two landmarks had much lower and often clinically insignificant mean differences compared with parameters requiring more than two landmarks.

It is interesting that within the >2-landmark parameters for both the reliability and validity studies, the mean differences in arch crowding, Bolton anterior, and Bolton overall, which use upwards to twenty-four component measurements, were better than arch perimeter that uses only four component measurements. When calculating multiple measurements, it is possible that component measurements provide not only for greater opportunities for variation but also for errors to cancel each other out.

For most linear parameters, inter-rater reliability using Anatomodels, plaster models, and extracted premolars had moderate to excellent agreement and clinically insignificant mean differences. This suggests that for most parameters, the mean measurements were consistent and acceptable.

Overall, intrarater reliability was better than inter-rater reliability: mean differences were smaller, and agreement and concordances were higher. The mean differences (0.02 mm) were consistent with the stated accuracy of the digital calipers.

Based on the ICC values, agreement in the validity studies was worse than in the reliability studies. If a bias due to fatigue was present, this might be addressed by increasing the interval

Table 6. **Inter-rater, plaster: ICC and repeated-measures ANOVA mean differences shown for each parameter, grouped by linear measurements requiring two landmarks, and those requiring more than two landmarks**

		Inter-rater reliability		Mean differences (mm)		
Parameter	N	ICC	95% CI	Worst	Best	*p*-value
Plaster models, linear measurements, 2 landmarks						
Overjet	10	0.550	0.249, 0.829	1.35	0.05	<0.001*
Overbite	10	0.771	0.460, 0.931	1.65	0.05	<0.001*
Tooth 1-1	10	0.923	0.826, 0.977	0.15	0.01	0.195
Tooth 1-2	10	0.639	0.377, 0.870	0.39	0.01	0.330
Tooth 1-3	10	0.712	0.470, 0.902	0.13	0.02	0.451
Tooth 1-4	10	0.830	0.649, 0.946	0.22	0.03	0.078
Tooth 1-5	10	0.856	0.697, 0.955	0.18	0.00	0.189
Tooth 1-6	10	0.823	0.624, 0.945	0.34	0.01	0.030*
Tooth 2-1	10	0.949	0.883, 0.985	0.11	0.01	0.401
Tooth 2-2	10	0.965	0.917, 0.990	0.07	0.00	0.837
Tooth 2-3	10	0.587	0.318, 0.845	0.25	0.00	0.168
Tooth 2-4	10	0.841	0.657, 0.951	0.22	0.01	0.032*
Tooth 2-5	10	0.866	0.715, 0.959	0.14	0.01	0.183
Tooth 2-6	10	0.702	0.458, 0.897	0.31	0.06	0.125
Tooth 3-1	10	0.842	0.669, 0.951	0.06	0.00	0.847
Tooth 3-2	10	0.900	0.780, 0.970	0.13	0.00	0.280
Tooth 3-3	10	0.890	0.760, 0.967	0.07	0.01	0.762
Tooth 3-4	10	0.874	0.704, 0.963	0.23	0.02	0.001*
Tooth 3-5	10	0.843	0.672, 0.951	0.20	0.00	0.077
Tooth 3-6	10	0.810	0.597, 0.941	0.45	0.06	0.002*
Tooth 4-1	10	0.935	0.851, 0.981	0.10	0.02	0.129
Tooth 4-2	10	0.855	0.693, 0.955	0.06	0.00	0.764
Tooth 4-2a	1†	–	–	0.54	0.00	–
Tooth 4-3	10	0.888	0.753, 0.966	0.19	0.03	0.055
Tooth 4-4	10	0.875	0.719, 0.962	0.22	0.01	0.027*
Tooth 4-5	10	0.834	0.659, 0.948	0.18	0.01	0.203
Tooth 4-6	10	0.826	0.613, 0.947	0.36	0.01	0.001*
Mx_IMW	10	0.854	0.683, 0.955	1.93	0.05	0.081
Mx_ICW	10	0.957	0.893, 0.988	0.52	0.09	0.013*
Mn_IMW	10	0.939	0.859, 0.982	0.45	0.04	0.131
Mn_ICW	10	0.905	0.775, 0.972	0.92	0.06	0.027*
Plaster models, linear measurements, >2 landmarks						
Mx_Perim	10	0.838	0.551, 0.955	3.07	0.01	<0.001*
Mx_Crowd	10	0.787	0.548, 0.933	1.94	0.01	<0.001*
Mn_Perim	10	0.522	0.195, 0.819	4.66	0.32	<0.001*
Mn_Crowd	10	0.655	0.345, 0.882	3.66	0.15	<0.001*

Table 6. **(continued)**

		Inter-rater reliability		Mean differences (mm)		
Parameter	N	ICC	95% CI	Worst	Best	*p*-value
Bolton 6	10	0.721	0.476, 0.906	0.10	0.01	0.994
Bolton 12	10	0.811	0.620, 0.939	0.75	0.02	0.274

ICC, intraclass correlation coefficient; CI, confidence interval; Mx_, maxillary; Mn_, mandibular; IMW, intermolar width; ICW, intercanine width; Perim, arch perimeter; Crowd, crowding if negative; Bolton 6/Bolton 12, Bolton millimeter, positive when mandibular excess.
[†]Cannot be computed because the sum of caseweights is less than or equal 1.
*p-value <0.05

Table 7. **Inter-rater, extracted premolars: ICC and repeated-measures ANOVA mean differences shown for mesiodistal width measurements of each extracted premolar**

		Inter-rater reliability		Mean differences (mm)		
Parameter	N	ICC	95% CI	Worst	Best	*p*-value
Extracted premolars						
Tooth 14	8	0.938	0.845, 0.985	0.17	0.03	0.275
Tooth 15	3[†]					
Tooth 24	9	0.913	0.799, 0.976	0.15	0.02	0.372
Tooth 25	2[†]					

ICC, intraclass correlation coefficient; CI, confidence interval.
[†]Test values not reported due to low sample size.

between measurements to greater than 10 days or reducing the number of models assessed per day.

This study showed that performing SMA on Anatomodels can take about 4 min longer than using plaster. This is similar to a study by Horton et al. (13) and Tarazona et al. (5) who reported that mesiodistal measurements on virtual models took about 3 min longer than using plaster. In contrast, earlier studies reported an opposite trend of about one to 3 min faster using virtual models (11, 14) compared with plaster. Our study evaluated more time-consuming measurements by dynamically manipulating the models, showing and hiding teeth to reveal the interproximal contact areas.

A discussion of the extra time for SMA on Anatomodels should be considered in the context of the total time and costs involved compared with plaster. A thorough analysis on the resources, time, and related costs involved is beyond the scope of this article but the resources that need to be considered for traditional in-house records include both time and costs for panoramic and cephalometric radiographs, clinic chair time, laboratory time, sterilization, materials and overhead, and finally the time to perform SMA on plaster models. The comparable resources for CBCT-generated digital models, assuming they are outsourced, involve practically no time from the practice but possibly only the related costs for the referral to the imaging center, which may or may not have the cost included for a radiologist report and for the Anatomodels, and then, there is the time spent to perform SMA on Anatomodels. Additionally, it was noted that the Anatomage turnaround times were unreliable, for our sample, with a case coming back after 5 months.

Because Anatomodels are produced via a proprietary process, there is an underlying assumption that when teeth are segmented from CBCT scans, it is done correctly along true anatomic contours. Any differences arising from this segmentation process, then, will contribute to systematic error (15). The error of this process as it relates to segmenting human teeth has not been fully studied.

In the absence of complicating factors (16) such as partial volume average, noise, artifacts, and threshold settings, it is theoretically possible to define a single point by selecting only one voxel. Additional voxels may help to identify the single voxel of interest but they are not necessary in the act of selecting a single voxel. When defining the true boundary of an object, at best, the line for this boundary will cross directly through the center of a voxel. But when attempting to select a boundary that truly goes between voxels, one is forced to select the center of one of the surrounding voxels. At worst, then, the accuracy for the selection of a single voxel of 0.3 mm sides will be unavoidably off by the equivalent of half the diagonal of the voxel or

Table 8. **Validity, Anatomodels vs. plaster: ICC and paired-sample mean differences shown for each parameter, grouped by linear measurements requiring two landmarks, and those requiring more than two landmarks**

		Agreement		Difference (mm)†		
Parameter	N	ICC	95% CI	Mean	95% CI	*p*-value
Anatomodels vs. plaster, linear measurements, 2 landmarks						
Overjet	18	0.927	0.815, 0.972	0.02	−0.31, 0.35	0.905
Overbite	18	0.925	0.808, 0.971	0.27	−0.17, 0.70	0.219
Tooth 1-1	30	0.558	0.159, 0.781	0.35	0.16, 0.54	0.001*
Tooth 1-2	29	0.772	0.552, 0.889	0.19	0.02, 0.37	0.031*
Tooth 1-3	30	0.532	0.196, 0.752	0.29	0.09, 0.49	0.007*
Tooth 1-4	30	0.749	0.540, 0.872	0.09	−0.04, 0.22	0.166
Tooth 1-5	30	0.611	0.319, 0.796	0.18	0.02, 0.33	0.026*
Tooth 1-6	30	0.724	0.499, 0.858	0.07	−0.12, 0.27	0.454
Tooth 2-1	30	0.630	0.108, 0.844	0.47	0.27, 0.67	<0.001*
Tooth 2-2	30	0.863	0.654, 0.941	0.21	0.08, 0.34	0.003*
Tooth 2-3	30	0.549	0.214, 0.763	0.23	0.07, 0.39	0.007*
Tooth 2-4	30	0.773	0.540, 0.890	0.16	0.03, 0.28	0.014*
Tooth 2-5	30	0.569	0.271, 0.768	0.17	0.00, 0.34	0.046*
Tooth 2-6	30	0.714	0.478, 0.853	−0.01	−0.21, 0.19	0.908
Tooth 3-1	30	0.648	0.385, 0.815	−0.08	−0.23, 0.07	0.269
Tooth 3-2	30	0.727	0.501, 0.860	−0.04	−0.16, 0.09	0.559
Tooth 3-3	30	0.718	0.485, 0.855	−0.02	−0.18, 0.14	0.796
Tooth 3-4	30	0.595	0.312, 0.783	−0.12	−0.28, 0.03	0.120
Tooth 3-5	30	0.429	0.099, 0.677	0.14	−0.05, 0.34	0.143
Tooth 3-6	30	0.560	0.251, 0.764	0.04	−0.22, 0.30	0.772
Tooth 4-1	30	0.682	0.437, 0.834	−0.08	−0.20, 0.03	0.159
Tooth 4-2	30	0.704	0.466, 0.847	−0.05	−0.18, 0.09	0.500
Tooth 4-2a	1‡	–	–			
Tooth 4-3	30	0.612	0.325, 0.795	0.03	−0.14, 0.20	0.733
Tooth 4-4	30	0.660	0.397, 0.823	0.03	−0.14, 0.21	0.694
Tooth 4-5	30	0.367	0.027, 0.635	0.14	−0.05, 0.33	0.147
Tooth 4-6	30	0.518	0.209, 0.736	0.23	−0.00, 0.47	0.051
Mx_IMW	30	0.953	0.900, 0.978	0.17	−0.19, 0.54	0.339
Mx_ICW	30	0.873	0.750, 0.937	0.14	−0.42, 0.71	0.607
Mn_IMW	30	0.949	0.885, 0.976	0.39	0.05, 0.73	0.026*
Mn_ICW	30	0.954	0.907, 0.978	0.29	−0.01, 0.59	0.061
Anatomodels vs. plaster, linear measurements, >2 landmarks						
Mx_Perim	30	0.536	−0.092, 0.824	3.38	2.48, 4.28	<0.001*
Mx_Crowd	30	0.864	0.710, 0.936	1.06	0.18, 1.94	0.020*
Mn_Perim	30	0.777	0.371, 0.909	1.71	0.88, 2.54	<0.001*
Mn_Crowd	30	0.718	0.205, 0.888	1.75	1.00, 2.49	<0.001*
Bolton 6	30	0.506	−0.097, 0.807	−1.57	−1.99, −1.16	<0.001*
Bolton 12	30	0.504	0.006, 0.770	−1.95	−2.73, −1.17	<0.001*

ICC, intraclass correlation coefficient; CI, confidence interval; Mx_, maxillary; Mn_, mandibular; IMW, intermolar width; ICW, intercanine width; Perim, arch perimeter; Crowd, crowding if negative; Bolton 6/Bolton 12, Bolton millimeter, positive when mandibular excess.
†Positive mean difference when measurements from Anatomodel are larger.
‡Cannot be computed because the sum of caseweights is less than or equal 1.
*p-value <0.05.

Table 9. **Validity, Anatomodels vs. extracted premolar: ICC and paired-sample mean differences for each premolar**

		Agreement		Difference (mm)†		
Parameter	N	ICC	95% CI	Mean	95% CI	*p*-value
Anatomodels vs. extracted premolars						
Tooth 14	8	0.963	0.842, 0.992	0.08	−0.07, 0.22	0.245
Tooth 15	3‡					
Tooth 24	9	0.957	0.835, 0.990	0.05	−0.09, 0.20	0.400
Tooth 25	2‡					

ICC, intraclass correlation coefficient; CI, confidence interval.
†Positive mean difference when measurements from Anatomodels are larger.
‡Test values not reported due to low sample size.

Table 10. **Validity, plaster vs. extracted premolar: ICC and paired-sample mean differences for each premolar**

		Agreement		Difference (mm) †		
Parameter	N	ICC	95% CI	Mean	95% CI	*p*-value
Plaster vs. extracted premolars						
Tooth 14	8	0.731	0.185, 0.938	−0.17	−0.51, 0.17	0.286
Tooth 15	3‡					
Tooth 24	9	0.755	0.243, 0.939	−0.08	−0.36, 0.20	0.531
Tooth 25	2‡					

ICC, intraclass correlation coefficient; CI, confidence interval.
†Positive mean difference when measurements from plaster are larger.
‡Test values not reported due to low sample size.

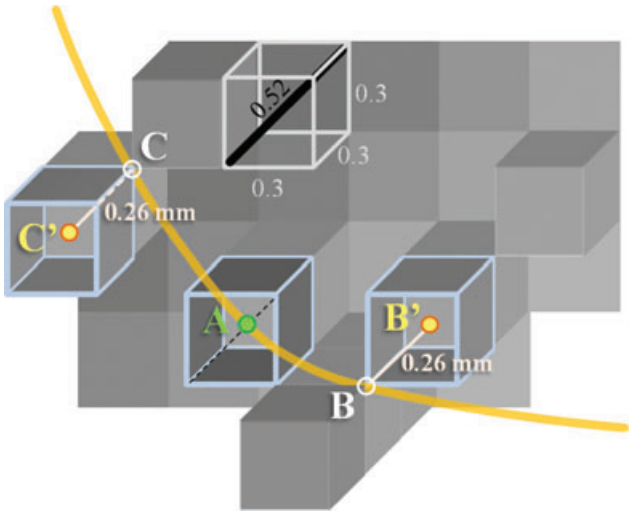

Fig. 2. The accuracy of selecting voxels, outlined in blue, for the boundary of an object that follows a path (orange line) through points A, B, and C. Selecting point A (green circle) is perfectly accurate because the orange line goes through the center of the voxel. But, in attempting to select points B and C, we are forced to select a neighboring voxel that centers at point B' and C' (yellow circles), respectively. Because the diagonal of a voxel with 0.3 mm sides is 0.52 mm, point B' has as much as 0.26 mm error from the true point B. Taking into account the error for point C', one can note that the accuracy of selecting two voxels can have a total error of much as about 0.5 mm.

0.26 mm. Given that a 2-landmark measurement will require the selection of two voxels, then, we would expect errors in accuracy to be as much as two half-diagonal distances or around 0.5 mm (Fig. 2).

The tendency toward positive mean differences of Anatomodels compared with extracted premolars suggests that conservative segmentation of voxel datasets in CBCT-generated virtual models occurred resulting in larger than expected measurements on Anatomodels. On the other hand, the negative differences in average mesiodistal measurements of plaster compared with extracted premolars suggest dimensional changes in plaster such that measurements were systematically smaller than in reality. This may be the result of imbibition of water (17) causing the alginate impression material to expand, thus resulting in a slightly smaller than expected stone cast. Again, these statements should be interpreted with caution due to the relatively small sample of extracted premolars.

A few Anatomodels had defects due to possible patient movement or streak artifacts. Radiographically, dental fillings are strongly attenuating objects that cause metal streak artifacts that are seen in reconstructed images as dark streaks in the direction of highest attenuation (18) resulting in an incomplete reconstruction of a segmented tooth and such missing surfaces will challenge the veracity of measurements. The presence of full fixed orthodontic appliances would likely make use of Anatomodels ineffective.

Conclusions

Intrarater reliability was excellent. Inter-rater reliability was moderate to excellent for most parameters. Validity was poor to moderate for many parameters. Time spent on Anatomodels can be almost twice as long as that on plaster.

Clinical relevance

Diagnosis through quantitative and qualitative measurements on virtual dental study models (Anatomodels, Anatomage) extracted from volumetric radiographic CBCT scans of the oral region should be tested against reference standards. Portions of what might be considered a full SMA have previously been investigated, and this study presents more comprehensive findings on the reliability, validity, and time requirements of measurements on Anatomodels compared with plaster models and a subset of extracted premolars. Study model analysis using Anatomodels can take twice as long as on plaster, and many linear and categorical measurements should be interpreted with caution.

Acknowledgements: The author would like to thank and acknowledge The Fund for Dentistry at the University of Alberta and the Alpha Omega Foundation for their grants in support of this research.

References

1. Kahl-Nieke B, Fischbach H, Schwarze CW. Treatment and postretention changes in dental arch width dimensions–a long-term evaluation of influencing cofactors. *Am J Orthod Dentofacial Orthop* 1996;109:368–78.
2. Scarfe WC, Farman AG, Sukovic P. Clinical applications of cone-beam computed tomography in dental practice. *J Can Dent Assoc* 2006;72:75–80.
3. Macchi A, Carrafiello G, Cacciafesta V, Norcini A. Three-dimensional digital modeling and setup. *Am J Orthod Dentofacial Orthop* 2006;129:605–10.
4. Kau CH, Littlefield J, Rainy N, Nguyen JT, Creed B. Evaluation of CBCT digital models and traditional models using the Little's Index. *Angle Orthod* 2010;80:435–9.
5. Tarazona B, Llamas JM, Cibrian R, Gandia JL, Paredes V. A comparison between dental measurements taken from CBCT models and those taken from a Digital Method. *Eur J Orthod* 2013;35:1–6.
6. Stevens DR, Flores-Mir C, Nebbe B, Raboud DW, Geo H, Major PW. Validity, reliability, and reproducibility of plaster vs digital study models: comparison of peer assessment rating and Bolton analysis and their constituent measurements. *Am J Orthod Dentofacial Orthop* 2006;129:794–804.
7. Rosner B. *Fundamentals of Biostatistics*. Boston: Brooks/Cole, Cengage Learning; 2010.
8. Rosenblatt MR. Tooth length measurement accuracy and reliability with cone-beam CT and panoramic radiography [Master's Thesis]. Edmonton, AB: University of Alberta; 2010.

9. Asquith J, Gillgrass T, Mossey P. Three-dimensional imaging of orthodontic models: a pilot study. *Eur J Orthod* 2007;29:517–22.
10. Goonewardene RW, Goonewardene MS, Razza JM, Murray K. Accuracy and validity of space analysis and irregularity index measurements using digital models. *Aust Orthod J* 2008;24:83–90.
11. Mullen SR, Martin CA, Ngan P, Gladwin M. Accuracy of space analysis with emodels and plaster models. *Am J Orthod Dentofacial Orthop* 2007;132:346–52.
12. Bell S. In: Laboratory NP, editor. *A Beginner's Guide to Uncertainty of Measurement*. Teddington, Middlesex: Crown; 1999. pp. 7–8.
13. Horton HM, Miller JR, Gaillard PR, Larson BE. Technique comparison for efficient orthodontic tooth measurements using digital models. *Angle Orthod* 2009;80:254–61.
14. 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.
15. Halazonetis DJ. From 2-dimensional cephalograms to 3-dimensional computed tomography scans. *Am J Orthod Dentofacial Orthop* 2005;127:627–37.
16. Molen AD. Considerations in the use of cone-beam computed tomography for buccal bone measurements. *Am J Orthod Dentofacial Orthop* 2010;137:S130–5.
17. Nicholls JI. The measurement of distortion: theoretical considerations. *J Prosthe Dent* 1977;37:578–86.
18. De Man B, Nuyts J, Dupont P, Marchal G, Suetens P. Metal streak artifacts in X-ray computed tomography: a simulation study. *IEEE Trans Nucl Sci* 1999;46:691–6.

Copyright of Orthodontics & Craniofacial Research is the property of Wiley-Blackwell and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.