Comparison of automated grading of digital orthodontic models and hand grading of 3-dimensionally printed models

Jonathan D. Scott, Jeryl D. English, Benjamin E. Cozad, Catharine L. Borders, Lacey M. Harris, Audrey L. Moon, and F. Kurtis Kasper
Houston, Tex

Introduction: Emerging workflows in orthodontics enable automated analysis of digital models and production of physical study models from digital files for the evaluation of treatment outcomes. The objective of this study was to compare the automated assessment of digital orthodontic models and the hand grading of 3D-printed models with the use of the American Board of Orthodontics cast-radiograph evaluation (ABO CRE) system. **Methods:** Plaster models from 15 cases were scanned with the use of a desktop model scanner to create digital models from which physical models were produced with the use of a stereolithography-based 3D printer. All digital models from each case were graded with the use of an automated software tool (SureSmile), and 3D-printed models were scored by hand with the use of the ABO CRE grading system. All hand-graded models were scored a second time at least 2 weeks later. **Results:** SureSmile gave statistically significantly higher scores to alignment and rotations ($P < 0.001$), overjet ($P < 0.001$), occlusal contacts ($P < 0.001$), and total score ($P < 0.001$). Hand grading scored higher in buccolingual inclination ($P < 0.001$). No significant differences were found in marginal ridges, occlusal relationships, and interproximal contacts. **Conclusions:** Scores assessed in an automated manner by SureSmile are generally significantly greater than those assessed by hand grading. (Am J Orthod Dentofacial Orthop 2019;155:886-90)

Assessment of the clinical outcomes of orthodontic treatment provides a mechanism through which the thoughtful clinician can review and refine their treatment approach. Orthodontic residency programs, in particular, apply extensive evaluation of treatment outcomes in clinical training, and research studies commonly use case assessment metrics in the discovery of evidence to guide clinical practice. Although a variety of orthodontically relevant metrics present utility in assessing outcomes, the cast-radiograph evaluation (CRE) system of the American Board of Orthodontics (ABO) presently serves as a benchmark approach for measuring final treatment outcomes.[1]

Department of Orthodontics, School of Dentistry, University of Texas Health Science Center, Houston, Tex.
All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest, and none were reported.
Address correspondence to: F. Kurtis Kasper, Assistant Professor, Department of Orthodontics, School of Dentistry, The University of Texas Health Science Center, 7500 Cambridge Street, Suite 5130, Houston, TX 77054; e-mail, fred.k.kasper@uth.tmc.edu.
Submitted, June 2018; revised and accepted, November 2018.
0889-5406/$36.00

https://doi.org/10.1016/j.ajodo.2018.11.011

The CRE involves application of 8 criteria in case evaluation: alignment and rotations, marginal ridges, buccolingual inclination, occlusal relationships, occlusal contacts, overjet, interproximal contacts, and root angulation. Conventionally, case evaluation with the use of the CRE involves measurements taken by hand on plaster cast models with the use of a measuring gauge.[2] However, recent developments in digital technologies and associated workflows in orthodontics present new options for case evaluation. For example, intraoral scanning technologies enable the generation of digital models, which have been shown to be as accurate as plaster cast models.[3] Accordingly, clinicians can apply software-based tools to duplicate, manipulate, and analyze models digitally without requiring a physical cast. In addition, 3-dimensional (3D) printing technologies provide a mechanism for generation of a physical model from the digital file in cases where a tangible object is desired.

A number of software tools support analysis of digital models with user-defined landmarks, and these software tools have been shown to be limited only by landmark identification.[4] Comparison of CRE scores obtained

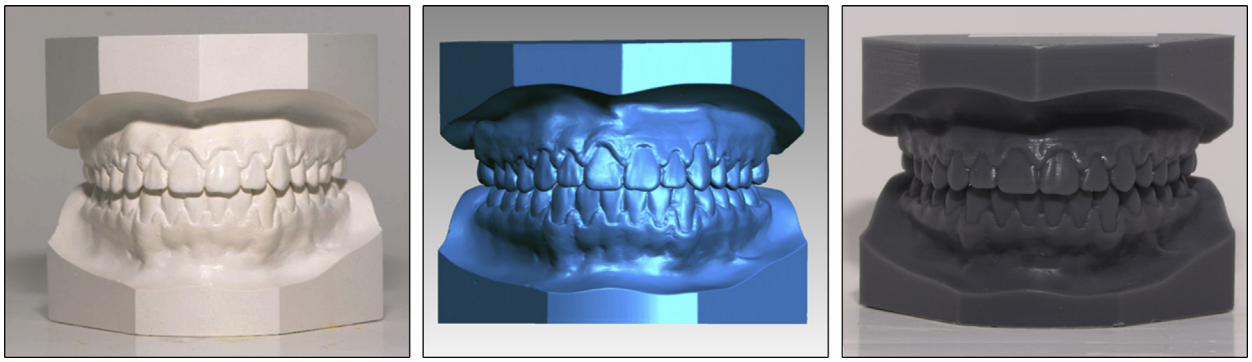

Fig. Photograph of a representative plaster model (left), screen capture of the corresponding digital model after scanning (middle), and photograph of the 3D-printed resin duplicate of the model (right).

with user-selected landmarks on digital models versus hand grading of cast models indicates a lack of correlation. The scores showed a difference in 3 criteria, namely, alignment and rotations, occlusal contacts, and overjet.[5] The current software allowing user-selected landmarks is not adequate for scoring all parameters as required by the ABO CRE.[6] Some software platforms now present the ability to conduct automated evaluation of digital models in accordance with the ABO CRE. For example, SureSmile (OraMetrix, Richardson, Texas) accepts models in multiple digital file formats and performs automated grading by means of computer algorithms to identify landmarks and assess digital casts.

The application of software-driven tools for automated grading of digital models presents clear potential advantages for clinicians and researchers in assessing treatment outcomes objectively, reproducibly, and efficiently. Nevertheless, in the dawning era of digital orthodontics, it remains unknown how the automated grading of digital models compares with traditional hand grading of physical models produced from the corresponding digital files via 3D printing. Elucidation of the potential differences in grading arising from these workflows will inform the clinical and research communities of the strengths and potential limitations of each approach. The specific objective of the present study was to compare the automated assessment of digital orthodontic models and the hand grading of 3D-printed models using the ABO CRE grading system.

MATERIAL AND METHODS

Fifteen sets of stone models that were presented previously to the ABO for CRE and received passing scores were scanned with the use of an R700 desktop scanner (3Shape, Copenhagen, Denmark). The study protocol was determined to qualify for exempt status according to 45 CFR 46.101(b) by the Institutional Review Board of The University of Texas Health Science Center at Houston (HSC-DB-17-0168). Each arch was scanned separately and then in occlusion, resulting in 30 files matched with information that would allow digital articulation of models in occlusion. The resulting digital files were converted to STL format. The 30 STL files, 1 maxillary and 1 mandibular file for each model, were used to 3D-print 30 arches in a Form 2 (Formlabs, Somerville, Mass) 3D printer and a gray photopolymer resin (GPGR03; Formlabs) at 50-μm print layer height (Fig). Formlabs Preform software version 2.11.1 was used to arrange the models on the build platform. All models were printed in their entirety as scanned, including the ABO base, and positioned so that the occlusal plane was parallel to the build platform and the occlusal surfaces were oriented away from the build platform.

The models were removed from the build platform after print completion and washed in 2 separate immersion baths of 2-propanol for 10 minutes in each, per the manufacturer's instructions.[7] The models were removed from the second bath and allowed to air dry at room temperature. The bottom of each model was labeled with a coded identifier.

The 3D-printed models for each of the 15 cases were graded by 8 ABO CRE–calibrated graders: 4 board-certified faculty members and 4 second-year orthodontic residents. Graders were calibrated in the use of the ABO measuring gauge with the ABO CRE instructions and YouTube videos published by ABO.[8] Each grader evaluated each model twice with at least 2 weeks between grading sessions. Scores for alignment and rotations, marginal ridges, buccolingual inclination, overjet, occlusal contacts, occlusal relationships, and interproximal contacts were recorded on ABO CRE worksheets.[2] Root angulation was not considered in this study, because presently there is no known automated means by which it can be evaluated. All scores were transferred to a spreadsheet for analysis.

Table I. Wilcoxon paired signed rank test: faculty versus residents

Metric	*Faculty*	*Residents*	*P*
Alignment and rotations	3.13	5.04	0.0007211*
Marginal ridges	3.00	3.55	0.02107†
Buccolingual inclination	3.97	3.93	0.5316
Overjet	1.61	1.30	0.1069
Occlusal contacts	2.58	1.66	0.004647†
Occlusal relationships	3.35	3.74	0.1392
Interproximal contacts	0.030	0.200	0.05447
Total	17.8	19.3	0.003971†

*$P < 0.001$; †$P < 0.05$.

Table II. Wilcoxon paired signed rank test: hand grading versus SureSmile

Metric	*Hand grading*	*Sure Smile*	*P*
Alignment and rotations	3.75	13.2	6.104e−05*
Marginal ridges	3.28	3.27	0.9321
Buccolingual inclination	3.95	2.27	0.0008898*
Overjet	1.45	6.47	6.104e−05*
Occlusal contacts	2.12	7.40	6.104e−05*
Occlusal relationships	3.55	4.06	0.3338
Interproximal contacts	0.110	0.000	0.05791
Total	18.5	36.7	0.00007247*

*$P < 0.001$.

Each of the 15 cases was submitted digitally to SureSmile and modeled as diagnostic models in preparation for automated grading. The SureSmile program is able to automatically select landmarks and render a score based on the ABO CRE. The landmarks are user adjustable, but were not manipulated for the purposes of this study. The ABO CRE scores were downloaded for each model and transferred to a spreadsheet for analysis.

Statistical analysis

Due to lack of normality, Wilcoxon paired signed rank tests were used to compare faculty and resident hand-grading results. Wilcoxon paired signed rank tests were also used to compare the score for each criteria assessed by SureSmile versus hand grading across both time points. All P values were 2 sided, and P values <0.05 were considered to be significant.

RESULTS

As Table I indicates, faculty and residents differed in some evaluation criteria but were similar in others. Residents assessed higher scores for alignment and rotations ($P < 0.001$), marginal ridges ($P < 0.05$), and total score ($P < 0.004$). Faculty assigned higher scores to occlusal contacts ($P < 0.005$). Criteria such as buccolingual inclination, overjet, occlusal relationships, and interproximal contacts were not significantly different between residents and faculty.

Comparing hand grading versus SureSmile grading, Table II indicates that SureSmile assessed statistically significantly higher scores to alignment and rotations ($P < 0.001$), overjet ($P < 0.001$), occlusal contacts ($P < 0.001$), and total score ($P < 0.001$). Hand grading scored higher in buccolingual inclination ($P < 0.001$). No significant differences were found in marginal ridges, occlusal relationships, and interproximal contacts.

DISCUSSION

The differences in scores for faculty and residents for the criteria of alignment and rotations, marginal ridges, and total score are possibly due to inexperience of residents in evaluation. This was not a surprise, despite their receiving the same calibration training. Faculty were more experienced graders who had all participated in the ABO certification process as applicants or graders.

The scores for marginal ridges, occlusal relationships, and interproximal contacts showed no significant difference between hand grading and SureSmile grading. This is perhaps due to the design of the measuring gauge used to evaluate items such as marginal ridges. If the marginal ridge blocks the passing of the gauge and correlates with the 1 mm mark, 1 point is scored. If it blocks the passing of the 2 mm mark, 2 points are scored. Occlusal relationships are graded by comparing cusp tips to landmarks by direct measure with the gauge. Interproximal contacts are evaluated by the visual presence of space between teeth and the ability to pass the narrow edge of the gauge through the contacts. The direct visualization may make it easier for hand graders to identify points effectively.

However, multiple metrics were significantly different between SureSmile and hand grading. Alignment and rotations scores were on average 9.45 points higher in SureSmile grading than hand grading ($P < 0.001$). This is likely due to the inability to use the measuring gauge to assess angles directly. Previous research on the subject of angle perception indicates that without reference, judging angles is challenging.[9] Accordingly, grading the discrepancy in alignment and rotations along an arc without a direct reference would present challenges for hand graders. On the other hand, SureSmile software likely is able to mark discreet points on that arc to calculate deviation accurately.

Buccolingual inclination was scored, on average, 1.68 points higher by hand grading than by SureSmile ($P < 0.001$). The ABO measuring gauge presents

millimetric steps that can be used to judge buccolingual inclination discrepancies on either the maxillary or mandibular arch. The grader is required to align the gauge with the cusp tips and then judge whether the steps on the gauge are able to clear the cusp to be graded. If the first step clears, 1 point is scored. If the second step clears, 2 points are scored. Grading with this measuring gauge may be challenging because it requires the grader to maintain contact with 2 separate points across the arch while assessing the scoring cusps. The gauge easily could be dislodged from the opposite cusp, finding rest on a lower part of the cusp, causing the gauge to angle away from the cusp to be graded and thus increasing the measurement. SureSmile likely is able to digitally mark the landmarks and assess them for these metrics, thereby mitigating error encountered in the identification and maintenance of landmarks when grading by hand.

Overjet was graded 5.02 points higher on average by SureSmile compared with hand grading ($P < 0.001$). The score for overjet is determined with the arches in occlusion. Anterior teeth are evaluated by visualization of the presence of contact between teeth; if the teeth do not contact, points are assessed based on the lack of contact (in millimeters) using the edge of the measuring gauge. In hand-articulated models, it is possible to shift the occlusion based on unintentional pressure placement. Such a shift could give the appearance that teeth are in contact when they normally would not be. Another evaluation metric is posterior overjet. The posterior overjet is measured from the mandibular buccal cusps to the central fossae of the maxillary teeth. This metric is difficult for hand graders to assess owing to the visual interference presented by the maxillary buccal cusps, which impede direct straight-line measurement from mandibular tooth to fossae.

Occlusal contacts were graded 5.28 points higher on average by SureSmile compared with hand grading ($P < 0.001$). Hand grading of occlusal contacts requires visualization of space between cusps and fossae while in occlusion. The same issue of pressure during hand-articulation may present an explanation for the discrepancy observed with this metric. Also, grading of the occlusal contacts of the lingual cusps requires viewing the contacts in occlusion as if one were standing on the tongue. Manipulating the casts and visualizing without shadows is challenging with the dark gray resin used in fabricating models in the present study. SureSmile presents the potential to mark the cusps and fossae and then calculate the distance between the points while digitally occluding the models.

The total score was calculated to be 18.2 points higher on average by SureSmile than by hand grading ($P < 0.001$). Given that many of the individual criteria were scored higher by SureSmile, it is not surprising the summation of the metrics is greater. In general, the automated grading by SureSmile appears to be far more rigorous and exacting than grading by hand. It is understandable that this is the case, given the ability of the software to precisely measure distances between selected points and variations from an archform. Many of the scores calculated by SureSmile were double the scores calculated by hand. However, it is important to note that root angulation was not included in the total score calculation in the present study, because no known automated method exists for scoring root angulation.

Metrics for which scores were significantly different were alignment and rotations, buccolingual inclination, overjet, and occlusal contacts. A previous study used OrthoCAD (Cadent, Fairview, NJ) to compare digital grading and hand grading of stone models and found results similar to ours. The use of OrthoCAD software requires the user to select landmarks on a digital model. The results indicated statistically significant differences in alignment, occlusal contacts, overjet, and total score.[5] That study indicated that the digital method is reproducible but does not give the same score for each criterion as determined by hand with the use of the ABO measuring gauge. The researchers concluded that there is no replacement for the manual scoring of the ABO objective grading system, which is the previous name for the CRE.[5,6]

Despite the evolution of the ABO certification process, at the present time there does not appear to be a suitable replacement for the CRE in evaluating case quality. As the field of orthodontics continues to embrace digital technologies, the value of case assessment can not be overstated in the digital workflow. The present study shows a clear difference in hand grading versus automated computer grading. The results provide insight as to what areas of assessment are different and the degree to which they differ. These findings open the door to future investigation as to how to standardize the automated score so it correlates with a traditional hand-graded score. This could take place with a modification of the scoring algorithm, an adjustment of the scores to reflect hand grading, or a new scale for passing in digitally graded cases.

Limitations of this study include recognition that the details of the proprietary methods by which diminutive cusp anatomy was considered and landmarks automatically identified in the software were not known to the investigators and that the landmarks were not adjusted by the user. However, digital grading by hand selection of landmarks has been evaluated in multiple previous investigations.[4,10] In the digital workflow, the automated selection of landmarks would save time for the clinician

or researcher, and the user could adjust specific landmarks if desired.

The use of SureSmile or similar digital grading tools in assessing outcomes would streamline the research process, eliminating what amounts to hours of work when grading multiple models by hand. Furthermore, it would allow a clinician to immediately identify deficient areas after a quick intraoral scan, make adjustments to treatment, and deliver results that would be considered excellent in the CRE.

CONCLUSIONS

1. Scores assessed by SureSmile are generally significantly greater than those assessed by hand grading, as reflected in the total score.
2. If SureSmile is to be used for evaluation of case quality, the digital scores should be comparable to values obtained by hand grading.
3. SureSmile scores need not match hand grading scores, if scaled properly to reflect the quality of the finish.
4. Future research could investigate methods of score adjustment to obtain consistency between methods as well as compare the landmarks selected by SureSmile and hand grading.

ACKNOWLEDGMENTS

The authors thank SureSmile for the automated grading of cases used in this study and J. Nathaniel Holland III and Ruby Benjamin-Gardner for assistance with statistical analysis.

REFERENCES

1. Casko JS, Vaden JL, Kokich VG, Damone J, James RD, Cangialosi TJ, et al. American Board of Orthodontics: objective grading system for dental casts and panoramic radiographs. Am J Orthod Dentofacial Orthop 1998;114:589-99.
2. American Board of Orthodontics. ABO cast-radiograph evaluation: 2010. Available at: https://www.americanboardortho.com/media/1184/cast-radiograph-worksheet-for-print.pdf. Accessed January 25 2018.
3. Santoro M, Galkin S, Teredesai M, Nicolay OF, Cangialosi TJ. Comparison of measurements made on digital and plaster models. Am J Orthod Dentofacial Orthop 2003;124:101-5.
4. Rossini G, Parrini S, Castroflorio T, Deregibus A, Debernardi CL. Diagnostic accuracy and measurement sensitivity of digital models for orthodontic purposes: a systematic review. Am J Orthod Dentofacial Orthop 2016;149:161-70.
5. Hildebrand JC, Palomo JM, Palomo L, Sivik M, Hans M. Evaluation of a software program for applying the American Board of Orthodontics objective grading system to digital casts. Am J Orthod Dentofacial Orthop 2008;133:283-9.
6. Okunami TR, Kusnoto B, BeGole E, Evans CA, Sadowsky C, Fadavi S. Assessing the American Board of Orthodontics objective grading system: digital vs plaster dental casts. Am J Orthod Dentofacial Orthop 2007;131:51-6.
7. Formlabs. Form 2 basic finishing steps: 2018. Available at: https://support.formlabs.com/hc/en-us/articles/115000024524-Form-2-Basic-Finishing-Steps. Accessed January 25 2018
8. American Board of Orthodontics. The American Board of Orthodontics grading system for dental casts and panoramic radiographs: 2012. Available at: https://www.americanboardortho.com/media/1191/grading-system-casts-radiographs.pdf. Accessed January 25 2018.
9. Erkelens CJ. Perception of perspective angles. Iperception 2015;6:2041669515593022.
10. Akyalcin S, Dyer DJ, English JD, Sar C. Comparison of 3-dimensional dental models from different sources: diagnostic accuracy and surface registration analysis. Am J Orthod Dentofacial Orthop 2013;144:831-7.