E4D Compare Software: An Alternative to Faculty Grading in Dental Education

Walter G. Renne, D.M.D.; S. Theodore McGill, D.M.D.; Anthony S. Mennito, D.D.S.; Bethany J. Wolf, Ph.D.; Nicole M. Marlow, M.S.P.H.; Stephanie Shaftman, M.Sc., M.S.; J. Robert Holmes, D.D.S., M.S., M.Ed.

Abstract: The traditional method of evaluating student tooth preparations in preclinical courses has relied on the judgment of experienced clinicians primarily utilizing visual inspection. At times, certain aids such as reduction matrices or reduction instruments of known dimension are used to assist the evaluator in determining the grade. Despite the skill and experience of the evaluator, there is still a significant element of uncertainty and inconsistency in these methods. Students may perceive this inconsistency as a form of subjective, arbitrary, and empirical evaluation, which often results in students' focusing more on the grade than the actual learning or developing skills necessary to accomplish the preparation properly. Perceptions of favoritism, discrimination, and unfairness (whether verbalized or not) may interfere with the learning process. This study reports the use of a new experimental scanning and evaluation software program (E4D Compare) that can consistently and reliably scan a student's tooth preparation and compare it to a known (faculty-determined) standardized preparation. An actual numerical evaluation is generated by the E4D Compare software, thereby making subjective judgments by the faculty unnecessary. In this study, the computer-generated result was found to be more precise than the hand-graded method.

Dr. Renne is Assistant Professor, Division of Restorative Dentistry, College of Dental Medicine, Medical University of South Carolina; Dr. McGill is Assistant Professor, Division of Restorative Dentistry, College of Dental Medicine, Medical University of South Carolina; Dr. Mennito is Instructor, Division of Restorative Dentistry, College of Dental Medicine, Medical University of South Carolina; Dr. Wolf is Assistant Professor, College of Medicine, Medical University of South Carolina; Ms. Marlow is Research Associate, College of Medicine, Medical University of South Carolina; Ms. Shaftman is Research Associate, College of Medicine, Medical University of South Carolina; and Dr. Holmes is Professor and Director, Division of Restorative Dentistry, College of Dental Medicine, Medical University of South Carolina. Direct correspondence and requests for reprints to Dr. Walter Renne, College of Dental Medicine, Medical University of South Carolina, 173 Ashley Avenue, BSB-545A, Charleston, SC 29425; 843-792-2503; renne@musc.edu.

Keywords: dental education, assessment, educational technology, tooth preparation

Submitted for publication 2/4/12; accepted 4/11/12

Accurate assessment of student work and ultimately translation of that assessment to the student are arguably the most critical components of dental education and paradoxically also its greatest weakness. In preclinical dental education, it is imperative that students receive consistent and accurate feedback from faculty so they can use this knowledge in order to achieve a higher level of performance before advancing to the clinics. Unfortunately, consistent feedback is very difficult to obtain, with many sources contributing to disagreement about student work including grading scale, rater calibration, training, and subjective influences.[1] In 1982, Mackenzie et al. went so far as to describe sixteen areas where inconsistencies can arise.[2]

It is widely agreed that faculty members should be calibrated in an attempt to overcome variability in assessment. However, there are significant problems that arise when calibrating faculty. Haj-Ali and Feil found that when trying to assess student work as simply acceptable or unacceptable after calibration, instructors often deemed the work as acceptable when it was actually unacceptable.[3] Furthermore, they concluded that for categorizing work as acceptable or unacceptable, seemingly the simplest grading scale, faculty members were not able to provide consistent feedback almost half the time. Not surprisingly, three separate studies conducted independently found significant disagreement between graders when evaluating dental work.[4-6] Furthermore, these studies found high levels of intra-examiner variability, in which the same examiner evaluating the same work on separate occasions each time gave a different grade. More recently, Sharaf et al. conducted a study to evaluate consistency in preclinical grading and found in almost all preparations there was significant disagreement between examiners.[7] Furthermore, attempts in that study to limit the grading scale—changing it from 1-10 to 1-5—did not help inter-examiner reliability. Interestingly, many dental schools still use grading scales from 0 to 100 to assess student preclinical work, counter to the consensus in

the literature indicating calibration is difficult for a large grading scale.[3-7] Students quickly learn which faculty members are "hawk" (hard) and "dove" (easy) graders. Students may perceive this inconsistency as subjective, arbitrary, or empirical grading. In our experience, this often results in students' focusing more on the grade than actual learning or developing skills necessary to accomplish the stated objective. Thoughts of favoritism, discrimination, or lack of fairness (whether verbalized or not) may interfere with the learning process. If highly trained and calibrated faculty members cannot provide consistent feedback, one would not expect dental students to have the ability to evaluate themselves accurately. Cho et al. found that "A" students are more likely to underestimate their work, while "D" and "F" students overestimate their work.[8] Therefore, the weaker students who need consistency in feedback are not getting it from faculty or through self-assessment; neither are "A" students getting consistent positive reinforcement.

Knight in a landmark article specified several rules with clearly defined grading criteria that, when followed, may help provide consistent and accurate feedback.[9] He concluded that valid and objective criteria along with rigorous faculty calibrations tied into promotion and tenure would help resolve the grading crisis in dental education. Recently, there has been great interest in the development of grading forms. The Commission on Dental Accreditation has suggested new standards for U.S. dental schools that relate to evaluation forms.[10] These standards mandate that evaluation forms be predetermined, standardized, reliable, and valid, and they suggest that faculty members be calibrated on how to be consistent when utilizing evaluation forms. Although this is certainly a step in the right direction, some investigators have concluded that if we are going to truly achieve accurate feedback, we need to remove the human element from evaluation and develop objective evaluation methods.[6,11]

The purpose of this study was to evaluate a new and revolutionary experimental software called E4D Compare developed by D4D Technologies (Richardson, TX, USA) in conjunction with dental educators around the United States. This software is in its unreleased Beta version and was still in the experimental phase at the time of this study. The hypothesis of our study was that the E4D Compare software is more consistent and therefore less variable when evaluating student preparations compared to three calibrated clinicians.

Methods

Fifty teeth were prepared by sophomore dental students at the Medical University of South Carolina (MUSC) as part of a preclinical fixed prosthodontics course. The preparations were done as a "practical examination" after didactic instruction on the proper parameters necessary to accomplish an ideal preparation. Laboratory practice (to include access to ideal examples of the preparation) was also a part of the training prior to the practical examination. The tooth preparation was an all-ceramic preparation on tooth #3 (maxillary right first molar) using a Kilgore Series 200 typodont (Nissan Dental Products, Kyoto, Japan) and Brasseler diamond burs 857KR 018 and 5368-023 (Brasseler USA, Savannah, GA, USA). The students were allowed one hour for the preparations.

Preparations were then graded (double blind) by three experienced and calibrated faculty members involved in teaching the course. The preparations were graded on a 0-100 scale in five-point increments. Faculty graders were calibrated to grade against the ideal "gold standard" preparation. Calibration was done with two separate hour-long lectures on what constitutes an ideal preparation and how to score deviations from ideal. Furthermore, we evaluated a sample of different preparations to ensure that all evaluators agreed independently as to what constituted grades of 60, 70, 80, 90, and 100. The gold standard preparation was agreed upon by the course faculty members and used as the example during student training prior to the practical examination. This agreed-upon gold standard preparation was based on visual inspection of preparation aspects such as taper, reduction, and quality of finish line.

The gold standard preparation was then scanned into the program with a laser scanner (D4D Technologies, Richardson, TX, USA) as the faculty ideal preparation. Next, the student's preparation was scanned, and a high-quality 3D model was generated (Figure 1). Using pinpoint precision, the two digital models were aligned based on common anatomical features of the adjacent teeth (Figure 2). E4D Compare software allows verification of proper alignment using a cross section of the aligned models to ensure proper "stitching" of the student model with the gold standard (Figure 3). Once proper alignment was verified, the faculty member marked the finish line of the student preparation and the gold standard, utilizing intuitive automatic margin finding tools and further refinement with manual tools (Figure 4).

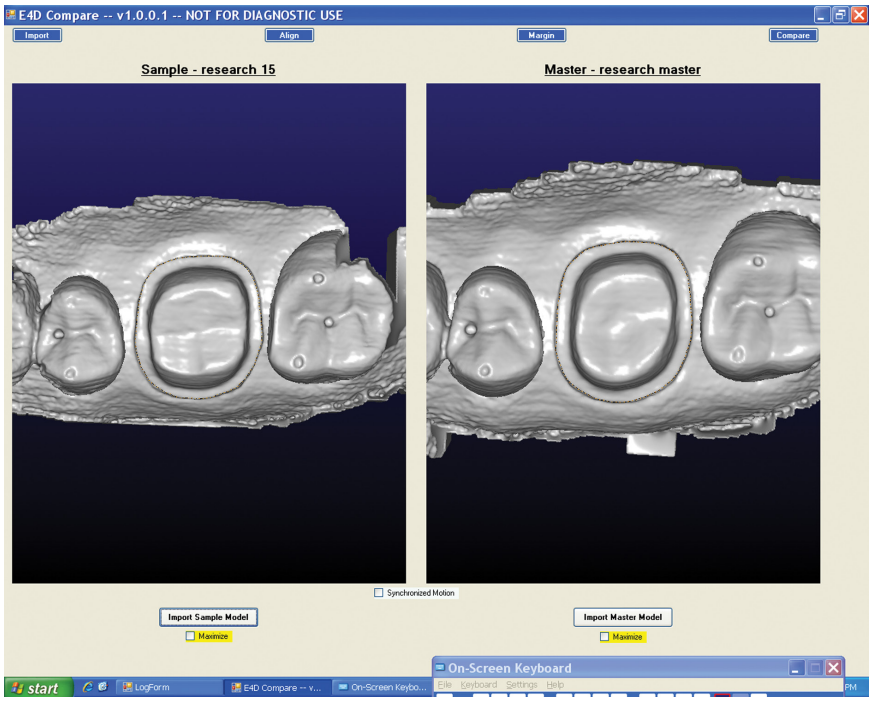

Figure 1. Sample student preparation next to faculty-determined gold standard

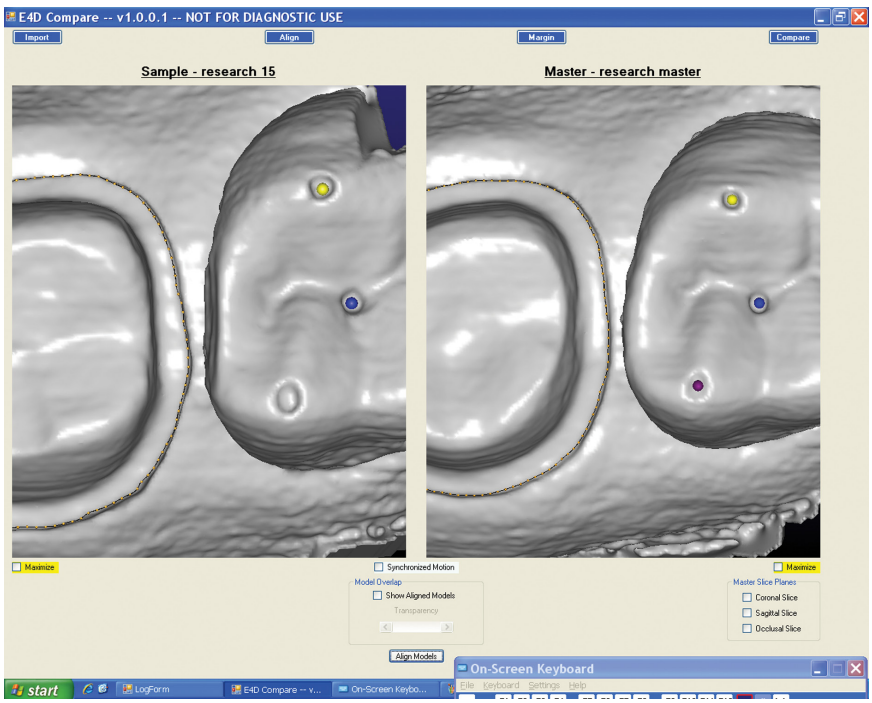

Figure 2. Alignment of sample student preparation and gold standard

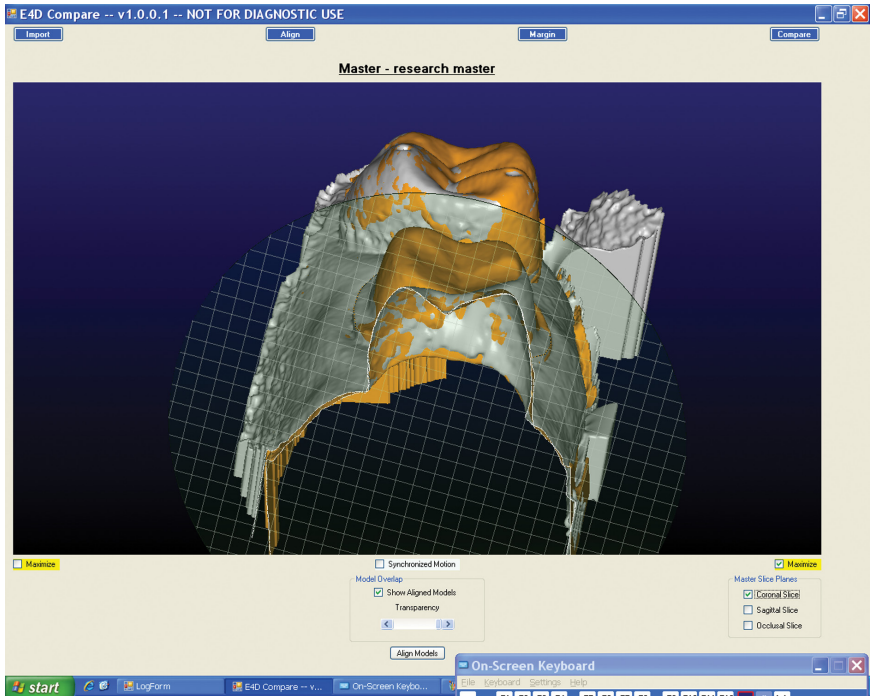

Figure 3. Cross section of student preparation aligned with gold standard

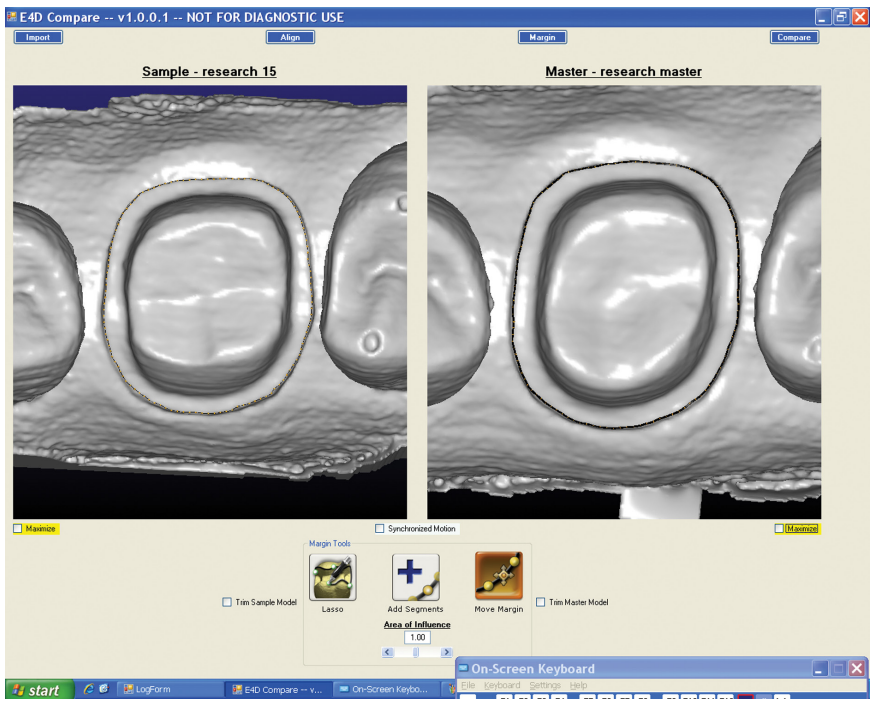

Figure 4. Finish line of student preparation and gold standard marked by faculty member

Next, the software measured any discrepancy in reduction (overreduction or underreduction) and displayed this discrepancy as a particular color. Areas within tolerances were displayed as green, areas underreduced appeared blue, and areas overreduced were shown in red. The software calculated the percentage of the surface area of the student preparation that was green and thus within the set tolerable range of discrepancy from ideal. The E4D Compare software automatically calculated the percent surface area of each color and displayed it as a numerical value (Figure 5). For this study, the numerical value for the percent surface area that was within the set range of 300µm from the ideal was recorded. The distance threshold (difference between the master and the student preparation) can be set at any level desired. In this study, 300µm was chosen as the acceptable range that student preparations can vary from the ideal (distance threshold) and still be scored green based on a pilot study. This pilot study found that when 300µm was used, the E4D Compare grade most closely correlated to faculty grades.

These two methods of evaluating a student's ability to prepare a tooth for an all-ceramic preparation were compared. The first method involved a rater comparing the student's tooth preparation to the gold standard preparation and provided a grade that ranged from 0 to 100 in increments of five units. The second method utilized the E4D Compare software to create a 3D image of the student's tooth preparation and compared it to the 3D gold standard preparation. It was hypothesized that the E4D Compare software would be more precise than the older hand-grading method described above.

The study included grades for fifty students. Three raters graded each randomized student preparation once by each of the two methods. The reduction evaluation provided by the software is a continuous measure, ranging from 0 to 100 and is the surface area of the student preparation that was within the set distance threshold of the ideal preparation. The mean difference in rater scores for each method was considered, as was the variability of scores within each method. In order to adequately compare the two methods, the dataset arising from the newer E4D evaluation method was rounded to the nearest unit of 5 (e.g., if the score was 62.3, it was rounded to 60; if the score was 64.5, it was rounded to 65). For comparison of the methods, both rounded and unrounded E4D measures were considered. Since

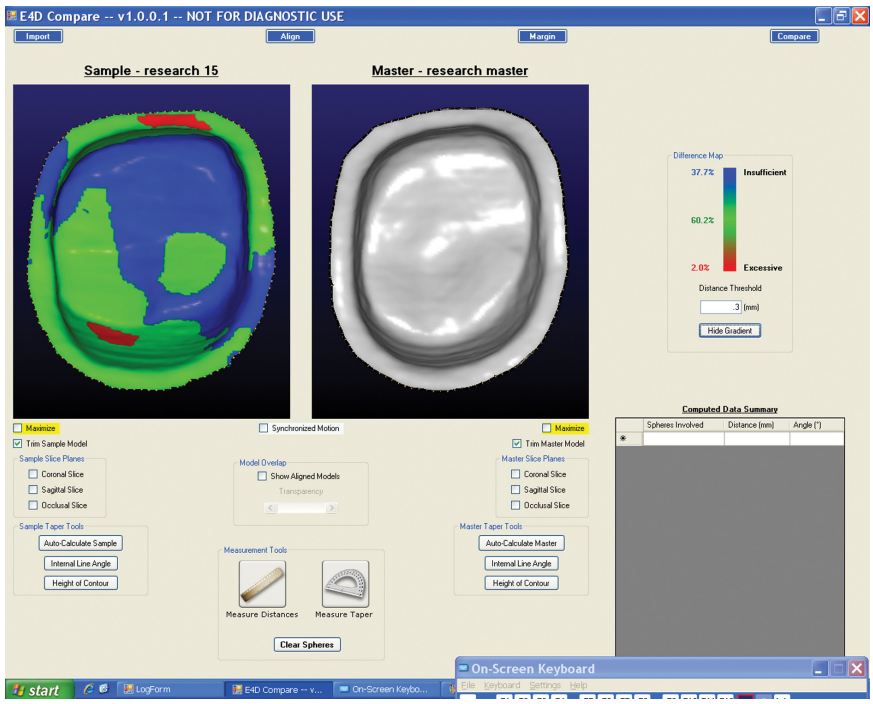

Figure 5. Software-calculated discrepancy between student preparation and gold standard

there was not a significant difference, the results for the unrounded E4D measures were reported.

For all calculations, the software SAS version 9.2 was used. Differences in rater scores for each evaluation method were calculated by taking the absolute value of the difference in each rater pair. For example, for tooth 1 within the hand-grading method, the differences between raters 1 and 2, raters 1 and 3, and raters 2 and 3 were calculated. Linear mixed models were applied to the data to examine differences between the methods in the mean rater differences. Linear mixed models were also used on the raw score data to obtain the variance-covariance estimates to examine the "within method" variability and overall variability of grades using intraclass correlations.

Results

The mean difference between raters' grades was significantly higher for the hand-grading method relative to the E4D Compare method ($p<0.001$). The mean difference was 8.07 (95 percent CI 6.98-9.15) for the hand-grading method and 2.23 (95 percent CI 1.14-3.31) for the E4D Compare method. Intraclass correlation coefficients (ICC) for each evaluation method were estimated to determine the relative precision of each method. The ICC for the hand-grading method was 0.620, while the ICC for the E4D evaluation method was 0.975. Thus, for the hand-grading method, 62 percent of the variability in students' grades resulted from which student's crown preparation was being graded, while 38 percent of the variability in grades was due to which rater graded the preparation. However, for the E4D method, >97 percent of the variability was due to variability in scores across student grades, while only 3 percent was due to which rater evaluated the preparation.

Thus, for a group of fifty students' crown preparations, using the E4D method, only 3 percent of the variability in the students' score was due to which rater evaluated the tooth, while 97 percent of the variability in score was due to which tooth was being graded. These results suggest that the E4D method is a significantly more precise method for assessing crown preparations than the hand-grading method.

Discussion

In many situations today, it seems students trust technology more than human judgment.[12] They have been raised with technology in every part of their lives and are frustrated when they are evaluated in what they perceive to be a subjective manner. It does not matter that faculty members may have tremendous experience and a finely honed ability to discriminate minute differences between various tooth preparations. They may even possess a high degree of consistency in evaluating things over time, and students with limited knowledge and experience may not be able to recognize or appreciate the significance of errors even when highlighted by a faculty evaluator. What is important is, in our experience, that students tend to distrust this evaluation and spend inordinate amounts of time questioning and/or challenging the grade or the grading criteria itself.

Often lost in this process is the focus on what the grade (however determined) actually represents. It should represent a deviation from the ideal and should encourage the student to try to discern any deficiencies and work to improve. Many times it does just the opposite. If students receive what they consider to be a good grade, they happily accept it and go on. If students receive what they consider to be a bad grade, they often attribute it to some form of bias, subjectivity, discrimination, or lack of evaluation ability on the part of the faculty member. It is not uncommon to hear such students say they do not plan to go back and try again as they will surely only get another unjust bad grade. The very students who would benefit from additional practice are thus often the first to give up or quit.

We acknowledge that this study has limitations and that further research including validation of accuracy and actual translation to student abilities in the clinic is needed. Furthermore, many institutions may need to invest considerable money in scanners to accommodate students with an appropriate ration of one scanner for every ten students.

Nevertheless, this study has shown that the scanning technique and comparison software used in our study takes the subjectivity out of the assessment process. Preparation evaluation can truly be generated in a nonthreatening, objective, and repeatable manner. This allows the student and faculty member to avoid wasting time dealing with questions of grade legitimacy and concentrate more on the student's weakness or lack of understanding about the procedure itself. This revolutionary software provides consistent and accurate assessment of students' preparations, allowing them to focus on improvement rather than arguing the validity of their grade.

The other tremendous benefit is that this system allows students to work independently. It is no longer

necessary for a faculty member to be present for the student to get valuable feedback. Students can work and practice independently outside of established laboratory times utilizing E4D Compare software as a self-assessment tool. Previously, students might be practicing the wrong things without accurate feedback, and rather than gain experience they would simply repeatedly reinforce errors.

Dr. Frank Medio, former director of graduate medical education at MUSC, once posed a question to the dental school faculty in a seminar. He asked, "What is the most important job of the faculty?" After faculty members uniformly answered that the most important job was to teach students, Medio disagreed. He stated that the most important job was "to provide accurate feedback, because if students received accurate feedback, they could teach themselves." The increased accuracy, precision, reliability, and consistency of the E4D Compare software in evaluating tooth preparations should allow students to learn and develop these skills more efficiently and in a shorter period of time. While our study did not address this issue, further research should be undertaken to answer this question.

The basic concept of this software is to compare the student's preparation in terms of overreduction and/or underreduction from a known standard. Obviously, other factors are important in terms of good tooth preparation. Smoothness of the surfaces, finish line configuration, and damage of the adjacent teeth are also very important. At this time, the software used in our study is still in development, and there are a number of other parameters that in the future may possibly be automatically calculated without subjective faculty evaluation. Currently, the software can calculate and display taper, total occlusal convergence (TOC), reduction lingual wall and axial wall height, and undercuts. The information provided by the software makes it easier for faculty members to provide accurate feedback to students. The finish line is obviously of great importance in any preparation and may have to be evaluated separately using cross-sections of the preparation at four or eight points around the tooth. Still, underdevelopment is a way to automatically evaluate the marginal configuration of the student preparation as compared to the ideal.

Comparison criteria, such as the distance threshold, can be determined by individual course instructors and changed as necessary. Faculty calibration can be easily done to help a larger group of clinicians to become more consistent in their perceptions and subsequent teaching of students. A true evaluation of the written parameters of an ideal preparation can be determined by using a preoperative scan of the tooth prior to preparation as compared to the final preparation. The cross-section tool is used to accurately assess in three dimensions the difference between the gold standard preparation and the unprepared tooth to ensure a perfect master is utilized for the comparison.

We have been utilizing the E4D Compare software as a tool for faculty evaluation for a semester and have evaluated over 500 preparations. Students seem to accept this feedback, trust it, and focus on improvement. We have not seen an improvement in student preparations of this magnitude in such a short period of time with conventional feedback mechanisms. On the first practical, the class average was 67, and three weeks later the average on the second practical was 98. E4D Compare has been a wonderful supplement to provide 3D assessment of student work beyond the conventional grade sheet. Potential implications for this technology extend well beyond the predoctoral dental classroom. Objective evaluations by state board examiners and testing agencies could ensure uniform results as well as permanent records of candidates' attempts during the examination.

Conclusions

This study demonstrated a reliable method of scanning and comparing student tooth preparations to a known ideal preparation. Using this method makes it feasible to accurately and consistently assess student work without dependence on subjective evaluation criteria. More research needs to be done to further improve assessment of student work and evaluate ways to reduce subjectivity. Future research evaluating E4D Compare software can include different methods to calibrate faculty, intra-examiner reliability, and accuracy.

Acknowledgments

The authors would like to thank Dr. Gary Severance and Mr. John Hinton (D4D Technologies, Richardson, TX) for their help with this study. The authors do not receive benefits from sales of the software discussed in this article and did not receive free products for this study. The university pays a yearly license fee for all E4D software. This work was

conducted with support from the National Institutes of Health National Center of Research Resources grant P20 RR017696–South Carolina COBRE in Oral Health.

REFERENCES

1. Feil PH, Gatti JJ. Validation of a motor skills performance theory with applications for dental education. J Dent Educ 1993;57(8):628-33.
2. Mackenzie RS, Antonson DE, Weldy PL, Welsch BB, Simpson WJ. Analysis of disagreement in the evaluation of clinical products. J Dent Educ 1982;46(5):284-9.
3. Haj-Ali R, Feil P. Rater reliability: short- and long-term effects of calibration training. J Dent Educ 2006;70(4):428-33.
4. Lilley JD, Bruggen Cate HJ, Holloway PJ, Holt JK, Start KB. Reliability of practical tests in operative dentistry. Br Dent J 1968;125(5):194-7.
5. Fuller JL. The effects of training and criterion models on inter-judge reliability. J Dent Educ 1972;36(4):19-22.
6. Salvendy G, Hinton WM, Ferguson GW, Cunningham PR. Pilot study on criteria in cavity preparation. J Dent Educ 1973;37(10):27-31.
7. Sharaf AA, AbdelAziz AM, El Meligy OAS. Intra- and inter-examiner variability in evaluating preclinical pediatric dentistry operative procedures. J Dent Educ 2007;71(4):540-4.
8. Cho GC, Chee WWL, Tan DT. Dental students' ability to evaluate themselves in fixed prosthodontics. J Dent Educ 2010;74(11):1237-42.
9. Knight GW. Toward faculty calibration. J Dent Educ 1997;61(11):941-6.
10. Commission on Dental Accreditation. Accreditation standards for dental education programs. Chicago: American Dental Association, 2006.
11. Schiff AJ, Salvendy G, Root CM, Ferguson GW, Cunningham PR. Objective evaluation of quality in cavity preparation. J Dent Educ 1975;39(2):92-6.
12. Junco R, Mastrodicasa J. Connecting to the net generation: what higher education professionals need to know about today's students. Washington, DC: NASPA, 2007.