Inter- and Intrarater Reliability Using Different Software Versions of E4D Compare in Dental Education

Richard S. Callan, DMD, EdS; Jeril R. Cooper, DMD; Nancy B. Young, DMD; Anthony G. Mollica, DMD; Alan R. Furness, DMD; Stephen W. Looney, PhD

Abstract: The problems associated with intra- and interexaminer reliability when assessing preclinical performance continue to hinder dental educators’ ability to provide accurate and meaningful feedback to students. Many studies have been conducted to evaluate the validity of utilizing various technologies to assist educators in achieving that goal. The purpose of this study was to compare two different versions of E4D Compare software to determine if either could be expected to deliver consistent and reliable comparative results, independent of the individual utilizing the technology. Five faculty members obtained E4D digital images of students’ attempts (sample model) at ideal gold crown preparations for tooth #30 performed on typodont teeth. These images were compared to an ideal (master model) preparation utilizing two versions of E4D Compare software. The percent correlations between and within these faculty members were recorded and averaged. The intraclass correlation coefficient was used to measure both inter- and intrarater agreement among the examiners. The study found that using the older version of E4D Compare did not result in acceptable intra- or interrater agreement among the examiners. However, the newer version of E4D Compare, when combined with the Nevo scanner, resulted in a remarkable degree of agreement both between and within the examiners. These results suggest that consistent and reliable results can be expected when utilizing this technology under the protocol described in this study.

Dr. Callan is Associate Professor and Chair, Department of General Dentistry, College of Dental Medicine, Georgia Regents University; Dr. Cooper is Assistant Professor, Department of General Dentistry, College of Dental Medicine, Georgia Regents University; Dr. Young is Instructor, Department of General Dentistry, College of Dental Medicine, Georgia Regents University; Dr. Mollica is Assistant Professor, Department of General Dentistry, College of Dental Medicine, Georgia Regents University; Dr. Furness is Assistant Professor, Department of Oral Rehabilitation, College of Dental Medicine, Georgia Regents University; and Dr. Looney is Professor, Department of Biostatistics and Epidemiology, Medical College of Georgia, Georgia Regents University. Direct correspondence to Dr. Richard S. Callan, College of Dental Medicine, Georgia Regents University, Room GC 3080, 1430 John Wesley Gilbert Drive, Augusta, GA 30912-1290; 706-721-3881; rcallan@gru.edu.

Keywords: dental education, assessment, grading, interrater reliability, intrarater reliability, CAD/CAM

Submitted for publication 9/17/14; accepted 12/2/14

With many researchers lamenting the shortcomings of subjective assessments of a student’s performance, the desire for improvement has resulted in numerous strategies and protocols aimed at mitigating the subjectivity through increased accuracy and accountability. Another approach to decreasing the subjectivity of assessments is to develop more objective means of evaluating a student’s knowledge and ability. Although not unique to dental education, this problem is of particular concern to dental students and faculty alike when it comes to the assessment of a student’s performance in preclinical practical examinations involving preparation of typodont teeth for various types of restorations. Over the last few years, a number of articles have described dental schools’ attempts to implement technologies capable of providing objective feedback to students. The Virtual Reality Dental Simulator (Dent Sim, DenX Ltd., New York, NY, USA), Haptic Virtual Reality Simulator (IDEA Dental, Las Vegas, NV, USA), E4D Compare (E4D Technologies, Richardson, TX, USA), prepCheck (Sirona, Salzburg, Austria), Kavo PrepAssist (Charlotte, NC, USA), and most recently the iPad (Apple Inc., Cupertino, CA, USA) are all examples of technologies developed to assist practitioners and students in improving their clinical skills. While each has its own advantages, few dental schools have the resources to acquire them all. Even if they could, there would be no time to learn or implement them all in an already overcrowded curriculum. For a multitude of reasons, each institution is thus forced to decide which few technologies can effectively and efficiently be integrated into its educational process. Technologies that can serve multiple purposes throughout the entire spectrum of a student’s education, in both preclinical
and clinical settings, possess a clear advantage over those that do not. The skills learned through use of these technologies in the preclinical environment are directly applicable to the skills necessary for success in the clinic and, ultimately, private practice. In addition, the earlier students learn these skills and the more frequently they can apply them, the more proficient they will become over time.15,20-23

The E4D Design Center, combined with E4D Compare software, when implemented throughout the dental curriculum, provides students with the ability to assess their own progress while simultaneously exposing them to a technology that continues to expand its influence in dentistry. Since E4D Compare was originally reported as an “alternative to faculty grading in dental education,”13 it has undergone considerable reprogramming. This reprogramming is in response to feedback from the many initial users of the original version. Previous studies have described some of the parameters that must be considered when implementing a technology for the purpose of objective assessment.12,14,16,24 The results obtained must be reliable, reproducible, and valid. In addition, the use of the technology must be effective and efficient.3,9,12-15

To decisively determine the appropriate application of this evolving technology, we have, over the past four years, designed and conducted studies to test some of these parameters. In an earlier study, we demonstrated the reliability of CAD/CAM technology in assessing crown preparations in a preclinical environment.12 At the conclusion of that study, we stated that “the next logical step would be to test the alignment of various digital impressions using numerous faculties.” The first aim of this study was to conduct such a test. Its null hypothesis was that the comparison results obtained using E4D Compare version 2.0 would be the same (no more reliable and no more reproducible) than the results obtained using E4D Compare version 1.0. A secondary aim of this study was to determine the inter- and intrarater reliability of five faculty members using E4D Compare when assessing students’ ability to duplicate the ideal crown preparation on tooth #30 in a preclinical environment.

Methods

This study received approval from the Institutional Review Board of Georgia Regents University. Using a Kilgore dental typodont (Kilgore International, Inc., Coldwater, MI, USA) made specifically for the Georgia Regents University College of Dental Medicine, tooth #30 was prepared by a faculty member to represent what our students are taught to be the ideal preparation for a full gold crown. The same types of typodonts and teeth were utilized by the dental students to prepare tooth #30 to receive a full gold crown.

This study was first accomplished utilizing the E4D Design Center (version 4.6.0.40) with its associated wand. These results were determined utilizing E4D Compare (version 1.0). The study was then repeated with the NEVO scanner (version 5.0.1.6) and E4D Compare (version 2.0).

E4D Design Center and E4D Compare (version 1.0)

Using the E4D Design Center and its associated wand, multiple individual images were captured and stitched together, creating a model of a randomly chosen senior dental student’s preparation and the adjacent teeth. This model was then imported into E4D Compare as the “sample model.” The student’s preparation was removed from the dentoform and replaced by the ideal preparation completed by the faculty member. Individual images were captured of the ideal preparation and its adjacent teeth and gingiva. The resulting model was imported into E4D Compare and labeled the “master model.” Small dots were placed on the gingiva of the dentoform adjacent to the tooth being examined. Common landmarks on the adjacent teeth were also selected by the faculty examiner.

The models (sample and master) were then aligned using the pinpointing of the dots method described previously12 or the pinpointing of the common landmarks. The latter method is the one described in the E4D Compare user manual.25 The margins of the student’s preparation and the professor’s preparation were indicated on both models. To compare the sample and master models of the entire tooth, the margin of the preparation was drawn at the juncture of the tooth and the tissue, not the actual margin of the preparation. After the models were aligned and the margins indicated, the preparations were compared at tolerances of 0.1 mm, 0.2 mm, 0.3 mm, 0.4 mm, and 0.5 mm. In E4D terminology, tolerance is “the number of millimeters that you want to consider as being too much of a difference between the two models, NOT the amount of reduction.”25

Five examiners (the first author and four other faculty members), representing varying degrees of experience with the E4D Design Center and E4D
Figure 1 shows a comparison of the screen displays of the two software versions of E4D Compare. The image on the left side is from version 1.0. The green color represents areas of the preparation that are within the tolerance level chosen for comparison. The blue represents areas of underreduction, whereas the orange to red represents areas of overreduction in regards to the selected tolerance level. The difference map indicates the percentage of the preparation that falls within each area or is overreduced or underreduced within the tolerance level. The image on the right side of Figure 1 is from version 2.0. As with the image on the left, the green represents areas of the preparation within the tolerance level chosen for comparison, and the blue designates areas of underreduction. The percent comparison (chart on right side of image) shows the percentage of the surface area of the preparation that matches the ideal preparation within the chosen tolerance level. It also shows the maximum underreduction, maximum overreduction, and average error of the tooth preparation compared with the ideal at the chosen tolerance level.

**Statistical Analysis**

The intraclass correlation coefficient (ICC) was used to measure both interrater and intrarater agreement among the faculty examiners. The method of Gilder et al. was used to estimate each ICC. In general, an ICC of 0.75 or greater indicated adequate agreement. Interrater agreement was determined by asking this question: did the five examiners generally agree among themselves when using each method (dots prescan, dots scan, landmarks prescan, landmarks scan) with each of the five tolerances? Intrarater agreement was determined by asking this question: separately for each method and each tolerance, did the five examiners generally agree within themselves when using the E4D system on more than one occasion?

**Results**

**E4D Design Center and E4D Compare (version 1.0)**

Interrater agreement coefficients for each method (dots prescan, dots scan, landmarks prescan, landmarks scan) and each of the five tolerances are shown in Figure 2. Regardless of the method or tolerance, the five examiners did not agree with each
0.972, which was the interrater agreement at 0.1 mm tolerance using the scan method.

Discussion

E4D Design Center and E4D Compare (version 1.0)

The overall poor interrater reliability achieved in this part of our study indicated a great deal of inconsistency between and within the faculty examiners when utilizing this technology for assessment of students’ performance. A previous study (using E4D Design Center and E4D Compare version 1.0) reported an interrater agreement of 0.97. In that study, a tolerance level of 0.3 mm was used for the comparison. We were not able to duplicate those results.

The largest degree of interrater agreement achieved in our study was 0.66, which was achieved with a 0.1 mm tolerance using the landmarks prescan method. The highest agreement coefficient at a 0.3 mm tolerance level achieved in our study (0.64) was accomplished using the dots scan method. The large discrepancy between the two studies can be somewhat mitigated by examining both the methods utilized and the experience level of the faculty. Both studies described the use of dots: either dots indicating common anatomical features of the adjacent teeth or dots placed on the gingiva of the typodonts.

Nevo Scanner and E4D Compare (version 2.0)

The E4D Compare has the auto-align feature, which eliminates the need for various methods of alignment. The results obtained utilizing the Nevo scanner and the E4D Compare (version 2.0) showed a remarkable degree of agreement both between and within the examiners (interrater, Figure 4; intrarater, Figure 5). The lowest coefficient of agreement was 0.972, which was the interrater agreement at 0.1 mm tolerance using the scan method.
Table 1. Intrarater reliability average (standard deviation, SD) for raters 1-5 for various alignment methods using E4D Design Center and E4D Compare (version 1.0)

<table>
<thead>
<tr>
<th>Alignment Method</th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
<th>Rater 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dots prescan</td>
<td>0.93 (0.02)</td>
<td>0.89 (0.06)</td>
<td>0.54 (0.10)</td>
<td>0.59 (0.22)</td>
<td>0.91 (0.07)</td>
</tr>
<tr>
<td>Dots scan</td>
<td>0.96 (0.02)</td>
<td>0.90 (0.02)</td>
<td>0.53 (0.17)</td>
<td>0.64 (0.39)</td>
<td>0.92 (0.05)</td>
</tr>
<tr>
<td>Landmarks prescan</td>
<td>0.94 (0.02)</td>
<td>0.87 (0.07)</td>
<td>0.78 (0.09)</td>
<td>0.37 (0.24)</td>
<td>0.29 (0.20)</td>
</tr>
<tr>
<td>Landmarks scan</td>
<td>0.54 (0.20)</td>
<td>0.97 (0.03)</td>
<td>0.74 (0.06)</td>
<td>0.42 (0.39)</td>
<td>0.62 (0.16)</td>
</tr>
</tbody>
</table>

Note: A value of 0.75 was considered minimally acceptable in terms of adequate reliability.

Figure 2. Interrater reliability using E4D design center (version 4.6.0.40) and E4D Compare (version 1.0): 1=dots prescan, 2=dots scan, 3=landmarks prescan, 4=landmarks scan

Note: A value of 0.75 was considered minimally acceptable in terms of adequate reliability.

Figure 3. Intrarater reliability using E4D design center (version 4.6.0.40) and E4D Compare (version 1.0): 1=dots prescan, 2=dots scan, 3=landmarks prescan, 4=landmarks scan

Note: A value of 0.75 (shown in horizontal line) was considered minimally acceptable in terms of adequate reliability.
were very consistent within themselves. The interrater agreement across all five tolerance levels for the prescan method was 0.999; for the scan method, it ranged from 0.972 to 0.992. The criteria proposed by Morton and McCarter can be used to interpret the magnitude of the ICC: values between 0.00 and 0.20 in absolute value can be classified as negligible, values between 0.20 and 0.50 as weak, values between 0.50 and 0.80 as moderate, and values between 0.80 and 1.00 as strong. From this, we can conclude a strong intrarater agreement for the five raters and a strong interrater agreement across all five tolerance levels for both the prescan and scan methods. This means that regardless of who was performing the scans and/or the comparisons, the results statistically were the same, and they were consistent.

study was carried out by faculty members with varying levels of experience with the E4D technology. Raters #1, #2, and #5 had each used the technology for a minimum of two years. They achieved a high level of interrater agreement (0.87) using the dots prescan method at the 0.3 mm tolerance level. Raters #3 and #4 were novices and had approximately four hours of hands-on training prior to the study. Their interrater agreements were, on the average, much less than the other raters.

**Nevo Scanner and E4D Compare (version 2.0)**

The intrarater agreement for all raters combined was 0.990, which means that overall the five raters were very consistent within themselves. The intrarater agreement across all five tolerance levels for the prescan method was 0.999; for the scan method, it ranged from 0.972 to 0.992. The criteria proposed by Morton and McCarter can be used to interpret the magnitude of the ICC: values between 0.00 and 0.20 in absolute value can be classified as negligible, values between 0.20 and 0.50 as weak, values between 0.50 and 0.80 as moderate, and values between 0.80 and 1.00 as strong. From this, we can conclude a strong intrarater agreement for the five raters and a strong interrater agreement across all five tolerance levels for both the prescan and scan methods. This means that regardless of who was performing the scans and/or the comparisons, the results statistically were the same, and they were consistent.
The purpose of comparing the previous scanner and software to the new is to illustrate the importance of determining the reliability and reproducibility of a process prior to implementation in a real-life setting, such as using it for assigning a grade in a practical examination. E4D Compare version 2.0 has the ability to determine the shoulder width, total occlusal convergence, and axial wall height of a crown preparation. These three aspects are derived using measurements from the actual crown preparation and can be used as independent criteria for assessment.

The difference component of E4D Compare version 2.0 measures the difference between an ideal preparation and a student’s attempt to reproduce the ideal by calculating the percent comparison of the surface area between the two, within a selected tolerance level. The accuracy of this comparison has been confirmed, and this study confirmed the accuracy across different examiners. The validity of using percent comparison as a measure of competence is the topic of another study currently under way.

One must also take into consideration the selection of the tolerance level when applying this technology for quantitative evaluation of a student’s performance. A very low tolerance level may introduce a higher degree of “noise” (defined as false positives and/or false negatives), whereas a tolerance set at too high a level may yield meaningless results. The tolerance level should be low enough to give adequate feedback for proper assessment, but high enough to ensure confidence that the feedback is accurate and not misleading. Additional studies are currently under way to help define and determine the appropriate tolerance levels to be applied using the technology as it exists today.

The ability to offer accurate feedback to students who are practicing these much-needed skills is perhaps the greatest potential of this technology. Providing constructive feedback in a timely manner, opportunities for self-assessment, and repeated practice are all aspects of enhanced learning and continued improvement.\textsuperscript{4,21,22,28,32} Students who are permitted unlimited access to the Nevo scanners and E4D Compare software will have the opportunity to immediately evaluate their progress and make appropriate adjustments to their preparations upon subsequent attempts. Future studies have been designed to evaluate the effectiveness of the use of this technology on students’ performance on practical examinations. The level of trust a student has in the accuracy of the assessment process can also have a direct effect on his or her confidence and performance.\textsuperscript{5} The results of this study indicate a high degree of accuracy of the system, which can translate into a high level of trust from the students.

This study has demonstrated the benefits of the auto-align feature of E4D Compare (version 2.0) when used with digital images obtained via the Nevo scanner. The raters were the same for both sets of comparisons, and the comparisons were done in the same order. The interrater differences between the experienced raters and the novice raters noted in the first comparison were not evident in the second comparison. These results disprove the null hypothesis of this study. One can reasonably expect more consistent and more reliable results when utilizing the Nevo scanner and E4D Compare (version 2.0) than those resulting from the use of the E4D Design Center and E4D Compare (version 1.0). The resulting accuracy is not dependent on the operator, nor is it dependent on when the test is performed. However, the importance of adequate training, experience, and calibration for faculty examiners at any institution interested in implementing this technology for assessing students’ performance should not be disregarded.

This study did not attempt to identify the most appropriate tolerance level with which to assess the student models, nor did it address the technology’s ability to detect critical errors that may be evident in non-ideal preparations, so these may be considered limitations of the study. Although the percent comparison numbers are accurate and reproducible, further study is required to determine the validity of using these numbers as a qualitative assessment of a student’s preclinical performance.

**Conclusion**

The results of this study suggest that the auto-align feature available in E4D Compare (version 2.0) compared to the previous version (version 1.0) greatly increased interrater and intrarater agreement when comparing students’ preparation of a full gold crown to an ideal preparation. Further studies are necessary to determine the appropriate tolerance level to be applied when comparing student and ideal preparations. The relevance of any such comparison to improved student performance and learning must also be investigated more thoroughly.
Acknowledgments

We would like to thank Henry Schein, Inc., and E4D Technologies for use of the Nevo scanner (version 5.0.1.6) and E4D Compare (version 2.0).

REFERENCES
