Utilizing Self-Assessment Software to Evaluate Student Wax-Ups in Dental Morphology

Karen R. McPherson, DMD; Anthony S. Mennito, DDS; Jompobe Vuthiganon, DMD; Yianne G. Kritzas, DMD; Richard A. McKinney, DMD; Bethany J. Wolf, PhD; Walter G. Renne, DMD

Abstract: Traditionally, evaluating student work in preclinical courses has relied on the judgment of experienced clinicians utilizing visual inspection. However, research has shown significant disagreement between different evaluators (interrater reliability) and between results from the same evaluator at different times (intrarater reliability). This study evaluated a new experimental software (E4D Compare) to compare 66 student-produced tooth wax-ups at one U.S. dental school to an ideal standard after both had been digitally scanned. Using 3D surface-mapping technology, a numerical evaluation was generated by calculating the surface area of the student’s work that was within a set range of the ideal. The aims of the study were to compare the reliability of faculty and software grades and to determine the ideal tolerance value for the software. The investigators hypothesized that the software would provide more consistent feedback than visual grading and that a tolerance value could be determined that closely correlated with the faculty grade. The results showed that a tolerance level of 450µm provided 96% agreement of grades compared with only 53% agreement for faculty. The results suggest that this software could be used by faculty members as a mechanism to evaluate student work and for students to use as a self-assessment tool.

Dr. McPherson is Assistant Professor, Department of Oral Rehabilitation, James B. Edwards College of Dental Medicine, Medical University of South Carolina; Dr. Mennito is Assistant Professor, James B. Edwards College of Dental Medicine, Medical University of South Carolina; Dr. Vuthiganon is Assistant Professor, James B. Edwards College of Dental Medicine, Medical University of South Carolina; Dr. Kritzas is an officer in the United States Naval Dental Corps; Dr. McKinney is an officer in the United States Naval Dental Corps; Dr. Wolf is Assistant Professor, James B. Edwards College of Dental Medicine, Medical University of South Carolina; and Dr. Renne is Associate Professor, James B. Edwards College of Dental Medicine, Medical University of South Carolina. Direct correspondence to Dr. Karen Runey McPherson, James B. Edwards College of Dental Medicine, Medical University of South Carolina, 173 Ashley Ave., MSC 507, Charleston, SC 29425; 843-792-5156; drkaren71@aol.com.

Keywords: dental education, grading, assessment, preclinical education, E4D Compare

Submitted for publication 4/17/14; accepted 10/7/14

Self-assessment, a learned skill that is essential for dental professionals to acquire and perfect for the purpose of self-regulation and lifelong learning, is also arguably one of the most difficult skills to obtain. At some point in their dental education, students are expected to acquire the ability to objectively and accurately evaluate their work against a prescribed norm. Once acquired and perfected, the skill of self-assessment allows for constant improvement of work and maintenance of a certain level of competence. Practicing clinicians who have not mastered self-assessment have little chance at improvement as they advance in their careers due to the isolation of working in a private practice setting. Without any feedback from outside sources, clinicians may become accustomed to repeating clinical errors.

Despite the importance of this skill, the general consensus in the medical and dental literature is that students are unable to accurately evaluate themselves. In an attempt to emphasize the importance of self-assessment, recent Commission on Dental Accreditation (CODA) standards require that “graduates recognize the role of lifelong learning and self-assessment in maintaining competence.” One alarming finding regarding self-assessment in dental education was reported by Cho et al., whose study found that A students were more likely to underestimate their work, while D and F students tended to overestimate their work. Part of the confusion students have regarding the quality of their work may stem from a lack of consistent feedback from faculty members as it is well known that there is little consistency between faculty grading, even after intense calibration. This lack of consistency has been found in both numerical grading and simplified scoring systems. In one study, when faculty members had only two categories of grading (acceptable and unacceptable), they failed to demonstrate consistency a staggering 50% of the time.
A more objective grading system is needed in dental education. Recently, new objective assessment software, Planmeca Compare version 1.0 (E4D Technologies, Richardson, TX, USA), has been implemented at the James B. Edwards College of Dental Medicine at the Medical University of South Carolina (MUSC). For a dental morphology course, a digital model of a student’s tooth wax-up is generated by scanning it with the Planmeca Planscan scanners (also from E4D Technologies). The software then automatically overlays a 3D model of an ideal wax-up on top of the student’s work. The faculty member or student enters one parameter into the Compare software indicating what variation from ideal is considered acceptable. This is called the tolerance value. For example, if 300 µm is entered, the 3D comparison map will display green anywhere the student was within 300 µm of the ideal faculty wax-up. Once that value is entered, the software instantly measures any differences between the two and highlights deviations from the ideal utilizing a color-coded display (Figure 1). The color-coded map of differences provides instant visual and objective feedback, showing where the student has made errors (Figure 2). The software codes green (i.e., within the tolerance range of ideal) as acceptable, red as overreduced or undercontoured, and blue as underreduced or overcontoured. The more surface area that is green, the closer the student is to the ideal. The percent surface area that is green is used as the grade. This visual feedback can not only aid students in evaluating their own work, but also help the faculty provide objective feedback on practical examinations.

When compared with calibrated faculty hand-grading, the use of Planmeca Compare software for evaluating crown preparations has been shown to be a more precise mechanism for both grading and feedback. While the benefits of utilizing the Planmeca Compare software have been demonstrated, the ideal tolerance value remains unknown. It is essential that the tolerance value used provides a fair assessment of student work, since a difference in the tolerance value changes the color percentages dramatically and results in a great variation of the student’s grade (Figure 3). Students utilizing the software for self-assessment need to have confidence that the feedback provided by the software is calibrated to a similar

---

**Figure 1.** Process by which E4D Technologies software calculates feedback on student wax-up

*Note:* The software scans the ideal and student models and uses millions of data points from the adjacent teeth and typodont base to combine and overlay the two. The differences are measured, and both numerical and visual feedback are instantly provided.
Figure 2. Color-coded map of differences showing where student has made errors

Note: The software codes green as acceptable, red to yellow as overreduced or undercontoured, and blue as underreduced or overcontoured. Numerical values of each color are presented as surface area. For this study, the percentage that was green was used as the grade. The instructor has control over the strictness of the tolerance value.

Figure 3. Differences in software tolerance values that result in varying amount of acceptable color (green)

Note: Alterations in the strictness or tolerance value generate different results. With a smaller number such as 0.2 mm entered as the acceptable tolerance, little green is displayed, and more errors are highlighted. If a higher number such as 0.5 mm is used, almost the entire tooth is green indicating the student was within 0.5 mm of the faculty wax-up.
strictness as the feedback they would receive from the faculty in that course. Previously, tolerance values were assigned arbitrarily.

The aims of this study were 1) to compare the repeatability, and therefore the reliability, of faculty and E4D Compare grades for 66 student-produced tooth wax-ups and 2) to determine the tolerance value for the software that provides scores most closely correlating with faculty grades (smallest average difference from faculty grades). The null hypothesis was that there is no difference in rater reliability when comparing one grading machine to another machine and one faculty member to another faculty member, the same faculty member or same machine at different times, or between faculty and machines. Additionally, no tolerance value for the grading software could be found that coincides with the average of the faculty grades.

Materials and Methods

Approval for this study was obtained from the Medical University of South Carolina Institutional Review Board. The study sought to evaluate current 3D machine-based grading software and compare it to calibrated faculty grading by evaluating interrater reliability, intrarater reliability, and intraclass evaluation. Due to use of the machines as additional evaluators in this study, however, this terminology was applied in a different manner than usual. We use the term “interrater evaluation” to compare the faculty grades to those given by the computer, the term “intrarater evaluation” to compare grades by the same evaluator (whether computer or faculty) on the same wax-up at two different times, and the term “intraclass evaluation” to compare grades given by one faculty evaluator to the grades given by the other faculty evaluators. The term “intraclass evaluation” was also used to compare the grades given by one of the machines to the grades given by the second machine. By evaluating these three degrees of agreement, we could determine how consistent faculty grading was with computer grading, how consistent each computer or faculty member was in grading at two different times, how consistent grading was among the four faculty members, and how consistent grading was between the two machines.

As part of normal classwork in a dental morphology course, students completed a wax-up of tooth #4 on typodont pegs inserted into a Kilgore series 200 typodont (Nissan Dental Products, Kyoto, Japan). Students were instructed to mimic the anatomy of the typodont tooth that they were waxing. A total of 66 deidentified student wax-ups were collected as part of a normal practical examination in dental morphology. The wax-ups were assigned a random number and given to four calibrated faculty members who were told to grade each tooth as they would for a normal practical exam utilizing a scale of 0-100. Each of the faculty members graded all 66 wax-ups initially and then again at eight weeks. Prior to the study, faculty members were calibrated with a one-hour training session and taught to grade against the ideal morphology as found on the identical typodont tooth.

A single team member who had received training from an E4D Planmeca trainer prior to the study scanned the 66 student wax-ups into the Compare software and graded each sample using tolerance values ranging from 200-800 µm, progressing in 25 µm increments, to determine what tolerance value yielded grading that most correlated with faculty grading. In the same manner as the faculty hand-grading, the software compared the morphology of the student wax-up to the ideal morphology of the typodont tooth. All 66 wax-ups were then scanned on a separate Planmeca Compare machine, and the study was repeated in order to examine the reliability between different machines. After eight weeks, the samples were regraded by the same four faculty members who graded the teeth initially in order to evaluate intrarater reliability. The data collected from the four faculty members consisted of the grades given to each student’s wax-up at two different time points. These data, along with the grades given by the two E4D machines, were then statistically analyzed.

We hypothesized that the E4D software would be more precise than the hand-grading method. Several comparisons were evaluated, including the mean difference between the average faculty grade and the average E4D grade, the mean difference between raters in each method (hand-grading or E4D), and the variability of scores in each method. All model assumptions were checked graphically and were found to be met. The mean difference between the average faculty grade and the E4D grades was determined from the linear mixed model using contrast statements, whereas the average faculty grade for a student was the average of the four faculty grades. The difference between the average faculty grade and the E4D machine grade was then determined, and the mean was taken across all wax-ups.
Linear mixed models were constructed including a random rater effect to account for correlation between repeated measures on the same wax-up. Contrast statements were used to estimate the mean difference between the average faculty grade and each tolerance value in order to determine which tolerance value provided a grade that was closest, on average, to the faculty grade. Contrast statements were also used to determine the mean difference in grades on each student wax-up by grading method (i.e., the difference among the four faculty grades on the same tooth and the difference between the two machines for the same tooth at each specific tolerance value). Bonferroni correction was used to adjust for multiple comparisons for all contrast statements. The linear mixed model also provided an estimate of the variance-covariance structure, which was used to estimate the intraclass correlations (ICC) between faculty members and between E4D machines at each different setting (200-800 µm). All analyses were conducted in SAS v. 9.3 (SAS Institute, Cary, NC, USA).

Results

The mean differences between faculty grades and the grades provided by the two E4D machines at each of the 25 tolerance values were compared to determine which tolerance value was closest to the faculty grades. The mean difference between the machine grades and faculty grades ranged from 35.9 grade points to -10.3 grade points. The largest differences were seen at the lowest tolerance value (200 µm), with the lower machine settings providing a lower grade on average relative to the faculty and the higher machine settings providing a higher grade on average than the faculty. The mean differences between grades for each machine setting compared with the faculty grades were significantly different from 0 for all tolerance values except values between 425 and 475 µm. The smallest mean difference from faculty scores occurred at the tolerance value of 450 µm. The mean difference between the faculty grades and the grades obtained from the machine set at a tolerance value of 450 µm was 0.20 points (95% CI -2.39, 2.98). The mean difference between the faculty grades and the machine grades for each setting is shown in Figure 4.

The mean difference between raters within each method was also evaluated (hand-grading and software grading). Every tooth was assigned a grade by each of the four faculty members and by both of the machines. The mean difference among the four faculty grades for each tooth was significantly higher when compared with the mean difference in grades found between the two machines’ grades (p<0.001). The mean difference among faculty grades across all teeth was 6.04 total grade points (95% CI 4.58-7.50), while the mean difference between machine grades across all teeth and all settings was 2.55 total grade points (95% CI 1.81-3.29) for the Compare software method.

To determine how well raters within each method agreed with one another, the intraclass correlation (ICC) was utilized. An ICC of 0.4 to 0.7 indicates fair to good agreement, while an ICC greater than 0.7 indicates excellent agreement. The ICC for the hand-grading method was 0.470 (95% CI: 0.239-0.935), while the ICC for the E4D method at the 450 setting was 0.958 (95% CI: 0.877-0.997). Thus, for the hand-grading method, 47% of the variability in grades resulted from which wax-up was being graded, while 53% of the variability was due to which faculty member graded the wax-up. For the E4D method, only 4% of the variability in the grades was due to which machine graded the wax-up, while 96% of the variability was due to which wax-up was being graded.

These results show that the E4D method was a significantly more precise method for grading than the hand-grading method. The tolerance value of 450 µm most closely correlated to the average of four calibrated faculty members. Faculty grades were found to vary significantly among the four raters. Compare software tolerance values that most closely matched Grader 1 (mean difference was not statistically different from 0) ranged from 400 to 475 µm; values that most closely matched Grader 2 ranged from 375 to 400 µm; values that most closely matched Grader 3 ranged from 475 to 575 µm; and values that most closely matched Grader 4 ranged from 500 to 550 µm. Averaging the ranges among the faculty tolerance values was considered to be the best way to arrive at a fair and accurate tolerance value; thus, a tolerance of 450 µm was found as the ideal in this study.

The E4D Compare software demonstrated greater consistency in student wax-up grades between multiple machines relative to hand-grading by different faculty members. The ICC values for all 25 machine settings were evaluated, and all were found to be between 0.94 and 0.97. An ICC of 0.96 indicated that there was excellent agreement between the
significant difference among faculty graders at different times (intrarater), and a statistically significant difference among different faculty grades (intraclass). However, no statistically significant difference was found between the grades given to a particular wax-up by the two machines or between the grades given by each machine between the first and second grading.

This study also attempted to correlate the average faculty grades to a tolerance level for the Compare software. The faculty average would serve to provide a baseline for this tolerance level rather than assigning it as an arbitrary number. Depending on the exercise or operator’s level of expertise, this number could be adjusted from the baseline to be more lenient or strict. A tolerance level of 450 µm provided grading that closely approximated that of faculty members in this study. The wide range of grades given by the faculty members shows that, even with proper calibration, variability in grading

two E4D machines, while an ICC of 0.47 indicated only fair agreement between the faculty grades for the same student wax-up (Table 1).

**Discussion**

The null hypothesis was that there was no difference in rater reliability when comparing one grading machine to another machine and one faculty member to another faculty member (intraclass), the same faculty and same machine at different times (intrarater), or between faculty and machines (interrater). This null hypothesis was partially rejected. Additionally, the null hypothesis that no tolerance value for the grading software could be found that coincided with the average of the faculty grades was rejected.

This study found a statistically significant difference in rater reliability when comparing grading machines and faculty (interrater), a statistically significant difference among faculty graders at different times (intrarater), and a statistically significant difference among different faculty grades (intraclass). However, no statistically significant difference was found between the grades given to a particular wax-up by the two machines or between the grades given by each machine between the first and second grading.

**Figure 4. Mean difference between faculty and machine grades for each tolerance value under consideration**

*Note:* Error bars represent the 95% confidence interval around the mean; 450 tolerance most closely correlates with faculty scores.
will still exist. Subsequently, this tolerance number may vary from school to school based on the relative strictness of individual graders. Repetition of this study at other dental schools would be helpful to establish a universal standard tolerance level for dental morphology. Establishing this tolerance level can provide additional consistency in the grading process and give students even more confidence that they are receiving an accurate level of feedback. This study was somewhat limited by the number of faculty graders. Utilizing a greater number of faculty members for grading would verify the ideal tolerance level.

Based on the results of this study, there may be several areas for further research. It would be interesting to determine if use of the software and its visual feedback when faculty members were not available would be of significant benefit to students. Additionally, the use of the software as an aid for faculty calibration should be explored. By establishing an ideal tolerance, faculty members can compare their grading techniques and adjust their methods of deductions accordingly.

### Conclusion

This study found that utilizing computer software to evaluate student wax-ups provided feedback that was both reliable and repeatable. The software was shown to provide evaluations that were significantly more precise than the hand-grading by faculty members. While 53% of the variability in grades was due to which faculty member graded the wax-up, only 4% of the variability in grades was due to which of the two machines graded the wax-up. The data showed that the machine tolerance value of 450 \( \mu m \) resulted in grades that most closely correlated to the faculty average. Further research should be conducted to evaluate the efficacy of this software as a self-assessment tool and as a method of faculty calibration in dental education.

### REFERENCES


### Table 1. Consistency in student wax-up grades between the software with multiple settings relative to average of faculty hand-grading

<table>
<thead>
<tr>
<th>Rater</th>
<th>ICC</th>
<th>Mean Difference from Faculty Average</th>
<th>Range of Differences from Faculty Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faculty</td>
<td>0.427</td>
<td>33.2 (14.2)</td>
<td>2.73, 70.3</td>
</tr>
<tr>
<td>M200</td>
<td>0.969</td>
<td>28.3 (14.5)</td>
<td>-3.12, 68.2</td>
</tr>
<tr>
<td>M225</td>
<td>0.969</td>
<td>23.8 (14.6)</td>
<td>-7.05, 66.4</td>
</tr>
<tr>
<td>M250</td>
<td>0.968</td>
<td>19.8 (14.6)</td>
<td>-8.65, 64.3</td>
</tr>
<tr>
<td>M275</td>
<td>0.967</td>
<td>16.1 (14.4)</td>
<td>-8.65, 64.3</td>
</tr>
<tr>
<td>M300</td>
<td>0.967</td>
<td>12.8 (14.1)</td>
<td>-9.55, 62.3</td>
</tr>
<tr>
<td>M325</td>
<td>0.966</td>
<td>9.74 (13.8)</td>
<td>-11.8, 60.2</td>
</tr>
<tr>
<td>M350</td>
<td>0.966</td>
<td>6.99 (13.4)</td>
<td>-13.7, 57.4</td>
</tr>
<tr>
<td>M375</td>
<td>0.964</td>
<td>4.52 (12.96)</td>
<td>-10.0, 53.5</td>
</tr>
<tr>
<td>M400</td>
<td>0.958</td>
<td>2.22 (12.5)</td>
<td>-16.0, 49.1</td>
</tr>
<tr>
<td>M425</td>
<td>0.958</td>
<td>0.259 (12.0)</td>
<td>-17.7, 44.9</td>
</tr>
<tr>
<td>M450</td>
<td>0.957</td>
<td>-1.54 (11.15)</td>
<td>-19.9, 40.8</td>
</tr>
<tr>
<td>M475</td>
<td>0.952</td>
<td>-3.17 (11.1)</td>
<td>-21.6, 36.6</td>
</tr>
<tr>
<td>M500</td>
<td>0.952</td>
<td>-4.66 (10.6)</td>
<td>-23.1, 32.8</td>
</tr>
<tr>
<td>M525</td>
<td>0.950</td>
<td>-5.98 (10.22)</td>
<td>-23.6, 29.9</td>
</tr>
<tr>
<td>M550</td>
<td>0.941</td>
<td>-7.05 (9.85)</td>
<td>-23.8, 26.9</td>
</tr>
<tr>
<td>M575</td>
<td>0.947</td>
<td>-8.12 (9.42)</td>
<td>-24.1, 22.0</td>
</tr>
<tr>
<td>M600</td>
<td>0.948</td>
<td>-9.02 (9.05)</td>
<td>-25.6, 20.1</td>
</tr>
<tr>
<td>M625</td>
<td>0.949</td>
<td>-9.79 (8.68)</td>
<td>-26.4, 18.1</td>
</tr>
<tr>
<td>M650</td>
<td>0.953</td>
<td>-10.5 (8.31)</td>
<td>-27.4, 16.1</td>
</tr>
<tr>
<td>M675</td>
<td>0.957</td>
<td>-11.1 (7.96)</td>
<td>-27.7, 13.9</td>
</tr>
<tr>
<td>M700</td>
<td>0.957</td>
<td>-11.6 (7.57)</td>
<td>-28.8, 11.2</td>
</tr>
</tbody>
</table>

Note: Column 1 shows the machine setting used. Column 2 shows the ICC for each grading method or machine setting. Column 3 shows the mean difference (and the standard deviation) between faculty average grades and the machine grade at each E4D machine setting. Column 4 shows the range of differences between the faculty average grades (using the lowest and the highest grades) and the machine grade at each E4D machine setting.