Dental Hygiene Faculty Calibration in the Evaluation of Calculus Detection


Abstract: The purpose of this pilot study was to explore the impact of faculty calibration training on intra- and interrater reliability regarding calculus detection. After IRB approval, twelve dental hygiene faculty members were recruited from a pool of twenty-two for voluntary participation and randomized into two groups. All subjects provided two pre- and two posttest scorings of calculus deposits on each of three typodonts by recording yes or no indicating if they detected calculus. Accuracy and consistency of calculus detection were evaluated using an answer key. The experimental group received three two-hour training sessions to practice a prescribed exploring sequence and technique for calculus detection. Participants immediately corrected their answers, received feedback from the trainer, and reconciled missed areas. Intra- and interrater reliability (pre- and posttest) was determined using Cohen’s Kappa and compared between groups using repeated measures (split-plot) ANOVA. The groups did not differ from pre- to posttraining (intrarater reliability \( p = 0.64 \); interrater reliability \( p = 0.20 \). Training had no effect on reliability levels for simulated calculus detection in this study. Recommendations for future studies of faculty calibration when evaluating students include using patients for assessing rater reliability, employing larger samples at multiple sites, and assessing the impact on students’ attitudes and learning outcomes.

Prof. Garland is currently Assistant Professor, Department of Dental Hygiene, Idaho State University, and formerly Adjunct Assistant Professor, Division of Dental Hygiene, Department of Primary Dental Care, University of Minnesota School of Dentistry; Dr. Newell is Associate Professor, Division of Dental Hygiene, Department of Primary Dental Care, University of Minnesota School of Dentistry. Direct correspondence and requests for reprints to Prof. Kandis Garland, 920 South 8th Avenue, Stop 8048, Pocatello, ID 83209; 208-282-6461 phone; 208-282-4071 fax; garlkand@isu.edu.

Key words: dental hygiene faculty, calibration, reliability, dental calculus, dental education

Submitted for publication 4/25/08; accepted 11/14/08

Dental and dental hygiene educators have been interested in the topic of faculty calibration for many years. Although the dental literature is extensive, little has been published specific to dental hygiene education about this issue.1 The literature indicates there is low agreement and much variability in clinical decision making among dental educators.1-8 Conflicting information from faculty members can easily confuse and frustrate students.1,9 Inconsistency among faculty members may detract from student learning, reduce students’ incentive to perform at higher levels (if they believe grading is capricious and inconsistent), decrease students’ satisfaction with their education, and could ultimately affect patient care.

Training or calibrating faculty is one way to help alleviate the problem of inconsistency. Calibration includes determining a standard based on criteria to be used when evaluating students (e.g., calculus present or not present; performance clearly acceptable, borderline, clearly unacceptable) and then reproducing the standard time and time again with repeated measures.10 In essence, to be considered reliable or consistent, faculty members must understand the designated criteria; each individual must apply the criteria the same way each time a student’s performance is evaluated; and multiple evaluators must repeatedly make similar qualitative assessments based on those criteria.

Student ratings are valuable tools for faculty development when used to improve teaching methodologies and behaviors. Understanding students’ opinions about their education as well as their learning styles helps administrators develop ways to improve student satisfaction.11,12 Few studies have examined the opinions of dental hygiene students about their educational experience. Anecdotal evidence, provided by students from the institution serving as the site of this study, indicates that dental hygiene students perceived a lack of consistency among their faculty members in exploring sequence and technique at the evaluation of calculus.13 Dental hygiene educators need to be cognizant of the fact that students might perceive us as inconsistent in grading their perfor-
formance. Calibration of faculty has been suggested as one way to enhance student satisfaction.\(^{14}\)

Several researchers agree that a calibration training program should include criteria development, a discussion of concepts, an explanation of the rating technique, practice with the rating technique, clearly defined criteria, concrete examples, a collection of pretraining scores, use of a gold standard, and a limited number of points on a rating scale.\(^{1,6,7,10,14}\)

Although it appears that faculty members can become more consistent through calibration training, the literature contains mixed results ranging from slightly effective to not at all effective.\(^{5,6,15,16}\) There is, however, consensus in the literature on the appropriate frequency of calibration. It should be ongoing and held at regular intervals.\(^{1,15,16}\) Calibration can be difficult and time-consuming, but it is achievable through hard work, repetition, and maintenance.\(^{1,7,15-17}\)

Few attempts have been made to improve intra- and interrater reliability levels with regard to calculus detection. Previous studies indicate a low level of agreement between faculty members, with Kappa averages of 0.34 on typodonts.\(^{17}\) This finding is consistent with other studies that demonstrated difficulty in detecting calculus.\(^{5,18-20}\) Currently, the explorer is the standard instrument used for calculus detection, but it can be subjective. Pippin and Feil suggested developing a less subjective or alternate method for calculus detection in lieu of an explorer.\(^{17}\) Emerging technologies such as endoscopy or red light emitting diode (LED) systems for calculus mapping could be used as an enhanced gold standard. Because few studies have examined the effects of calibration of faculty members with regard to calculus detection, further study is warranted. The purpose of this pilot study was to determine if a training program designed to enhance calibration of dental hygiene faculty members in scoring of calculus detection using an ODU 11/12 explorer affected intra- and interrater reliability levels for faculty receiving training in comparison with faculty members who did not receive training.

---

**Materials and Methods**

This pilot study utilized a two-group randomized experimental design to evaluate the effectiveness of a calibration training program on intra- and interrater reliability levels with regard to calculus detection. Following Institutional Review Board approval, a convenience sample of twelve study subjects was recruited from a pool of twenty-two clinical dental hygiene faculty members (five full-time and seventeen part-time) at the University of Minnesota School of Dentistry. Subjects volunteered and were randomly assigned to experimental and control groups. Six subjects participated in each group. All of the subjects had many years of clinical dental hygiene experience. The subjects had a wide range of teaching experience, ranging from a few months to more than twenty years. The study subjects included two full-time and ten part-time faculty members. Four of the participants were new faculty members, having begun clinical teaching within the academic year in which the study commenced. During the informed consent meeting with each participant, the investigator explained study protocols that emphasized refraining from discussing any aspects of the study. Study subjects agreed to adhere to the protocols and signed the informed consent statement.

Testing and training were performed on six typodonts (Kilgore model \#D85SDP-200[GSF]) with soft rubber gingiva, twenty-eight teeth without anatomically correct roots, and a flesh-colored oral cavity cover. Three typodonts were used exclusively for testing and three for training. Varying amounts of simulated calculus (sixteen to twenty-six tooth surfaces per typodont) were placed on the typodont teeth. Typodont 1 had the fewest surfaces of simulated calculus (sixteen out of 112 surfaces); typodont 2 had twenty surfaces out of 112; and typodont 3 had the most surfaces of simulated calculus with twenty-six out of 112 surfaces. All calculus detection was performed using ODU 11/12 explorers. Each tooth surface was explored and evaluated (four sites per tooth \([M, D, L, F]=112\) total surfaces). The size of the simulated calculus deposits varied intentionally on the three typodonts from small spicules to large ledges; however, subjects did not quantify the size of the deposit. They explored the typodonts and responded with a yes (calculus detected) or no (calculus not detected) answer for each tooth surface, and marked it accordingly on the answer sheet developed by the investigator.

All study subjects explored each of the three typodonts twice before the training (pretest/baseline during week one) and twice after the training (posttest during week four). Subjects in the control group did not receive training, but provided pre- and posttest scorings of calculus deposits. Each experimental and control subject’s attempt one and attempt two ratings were compared to determine intrarater (self) reliability during week one and week four. Subjects’
ratings for both attempts on all three typodonts were compared during week one and again during week four to assess interrater (between rater) reliability.

Training for experimental group subjects occurred during weeks two and three of the study. Training consisted of three two-hour sessions. Session one consisted of a discussion and operationally defining calibration. Specific exploring sequence and technique that subjects would be required to use during the training and posttest also were demonstrated. Exploring sequences are usually developed by personal preference. The investigators developed a sequence they deemed logical and efficient, as well as one that all faculty members could potentially use in their teaching and demonstration to students. The exploring technique was one that all the faculty at this institution already used; for example, the distal surfaces were explored first, then the mesial surfaces were explored. A research article was assigned for reading about the importance of calibration.10 The first session ended with subjects practicing the prescribed sequence and technique.

The second training session included a review of the exploring sequence and technique. Each subject explored each of the three typodonts once with the goal of achieving 80 percent accuracy when compared with the answer key. When 80 percent accuracy was not achieved, the subjects were required to repeatedly explore the calculus on the typodont until 80 percent accuracy was achieved. The investigators in this study paralleled a process used by the Central Regional Dental Testing Service (CRDTS), which uses 80 percent accuracy when calibrating its examiners.21 CRDTS is a regional testing service that administers a clinical dental hygiene licensing examination to candidates who are preparing for licensure. Two subjects did not initially achieve 80 percent accuracy and were required to repeat the process. Subjects were required to reconcile missed areas against the answer key first by re-exploring the area. They also could receive feedback from the trainer and unscrew the tooth to visually detect the calculus. Identical methods comprised the third training session one week later.

Accuracy was assessed against an answer key. Cohen’s Kappa coefficient was used to analyze all scores of calculus detection between the experimental and control groups. Kappa measures rater reliability and is a more robust measure than percent agreement alone because it takes into account the agreement occurring by chance.22 Kappa includes measurements between zero and one, with zero equaling no agreement and one equaling perfect agreement. Kappa averages closer to one are considered significant. Kappa averages ranging from 0.61 to 0.80 are considered in the full agreement range; anything higher than that indicates almost perfect agreement.22 Kappa is an appropriate test for nominal data. All faculty members’ Kappa averages before and after training were compared to evaluate the effectiveness of the training. Each subject explored and evaluated each of the three typodonts twice at both the pretest and posttest. The following comparisons of Kappa averages were made: attempts one and two, typodont 1 vs. typodont 2 vs. typodont 3, and experimental versus control group. There were seventy-two total Kappa values for intrarater reliability and 144 total Kappa values for interrater reliability. ANOVA (split-plot) was used to analyze if differences existed in each of these comparisons.

### Results

Table 1 shows the F-statistic, p-values, and associated Kappa averages measuring intrarater (self) reliability. The data suggest training sessions did not affect the difference between the intrarater reliability levels of the treatment and control groups’ Kappa averages measuring self-agreement. No differences were found between the treatment and control groups when comparing intrarater reliability levels of all calculus detection scores (three typodonts, two attempts, both groups). Analysis of pretest vs. posttest scores of all subjects also indicates no difference in subjects’ self-agreement (f-stat=0.05, p=0.82). There also was no significant interaction between pre/posttest, treatment/control groups’ Kappa averages (f-stat=0.24, p=0.64). A significant difference in intrarater reliability levels was found between scorings of typodonts 1, 2, and 3 (f-stat 8.29, p<0.002). Kappa averages declined as the number of surfaces of calculus increased.

Table 2 shows the F-statistic, p-values, and associated Kappa averages measuring agreement with the answer key (interrater reliability). Data suggest training did not result in a statistically significant difference between the treatment and control groups’ Kappa averages measuring agreement with the answer key. No differences were found between the treatment and control groups when comparing interrater reliability levels of all calculus detection scores (three typodonts, two attempts, both groups). There also was no significant interaction between
### Table 1. Intrarater (self) reliability of calculus detection scores

<table>
<thead>
<tr>
<th>Test</th>
<th>T/C Group</th>
<th>Pre/Post</th>
<th>Typodont</th>
<th>Mean Kappa Average</th>
<th>Standard Error</th>
<th>F-Statistic</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>C</td>
<td></td>
<td></td>
<td>0.812</td>
<td>0.030</td>
<td>2.90</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td></td>
<td></td>
<td>0.740</td>
<td>0.030</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>All</td>
<td>Pre</td>
<td></td>
<td>0.854†</td>
<td>0.029</td>
<td>8.29</td>
<td>&lt;0.002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Post</td>
<td></td>
<td>0.774</td>
<td>0.024</td>
<td>0.05</td>
<td>0.82</td>
</tr>
<tr>
<td>3</td>
<td>C</td>
<td>Pre</td>
<td></td>
<td>0.804</td>
<td>0.034</td>
<td>0.24</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>Post</td>
<td></td>
<td>0.820</td>
<td>0.034</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>Pre</td>
<td></td>
<td>0.743</td>
<td>0.034</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>Post</td>
<td></td>
<td>0.737</td>
<td>0.034</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>C</td>
<td>Pre</td>
<td></td>
<td>0.804</td>
<td>0.034</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>Post</td>
<td></td>
<td>0.820</td>
<td>0.034</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>Pre</td>
<td></td>
<td>0.743</td>
<td>0.034</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>Post</td>
<td></td>
<td>0.737</td>
<td>0.034</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

†Typodont 1 had the fewest surfaces of simulated calculus with sixteen out of 112 surfaces; typodont 2 had twenty out of 112 surfaces; and typodont 3 had the most surfaces of simulated calculus with twenty-six out of 112 surfaces. Kappa averages decreased as the number of surfaces of simulated calculus increased. The significant difference is in bold.

### Table 2. Interrater reliability of calculus detection scores

<table>
<thead>
<tr>
<th>Test</th>
<th>T/C Group</th>
<th>Pre/Post</th>
<th>Typodont</th>
<th>Mean Kappa Average</th>
<th>Standard Error</th>
<th>F-Statistic</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>C</td>
<td></td>
<td></td>
<td>0.804</td>
<td>0.021</td>
<td>1.20</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td></td>
<td></td>
<td>0.772</td>
<td>0.21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>All</td>
<td>Pre</td>
<td></td>
<td>0.868†</td>
<td>0.019</td>
<td>24.05</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Post</td>
<td></td>
<td>0.780†</td>
<td>0.019</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>All</td>
<td>Post</td>
<td></td>
<td>0.796</td>
<td>0.016</td>
<td>1.51</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pre</td>
<td></td>
<td>0.780</td>
<td>0.016</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>All</td>
<td>All</td>
<td>1</td>
<td>0.783</td>
<td>0.016</td>
<td>0.53</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>0.793</td>
<td>0.016</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>C</td>
<td>Post</td>
<td></td>
<td>0.083</td>
<td>0.023</td>
<td>1.90</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>Pre</td>
<td></td>
<td>0.8005</td>
<td>0.023</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>Post</td>
<td></td>
<td>0.789</td>
<td>0.023</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>Pre</td>
<td></td>
<td>0.755</td>
<td>0.023</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

†Typodont 1 had the fewest surfaces of simulated calculus with sixteen out of 112 surfaces; typodont 2 had twenty out of 112 surfaces; and typodont 3 had the most surfaces of simulated calculus with twenty-six out of 112 surfaces. Kappa averages decreased as the number of surfaces of simulated calculus increased. The significant difference is in bold.
pre/posttest, treatment/control groups’ Kappa averages (f-stat=1.90, p=0.20). The p-value (0.20) was not statistically significant at the predetermined 0.05 alpha level. A significant difference in interrater reliability levels was found between typodonts 1, 2, and 3 (f-stat=24.05, p=.0001). Kappa averages declined as the number of surfaces of calculus increased. All subjects had the highest Kappa average for typodont 1 at 0.868, the second highest for typodont 2 at 0.780, and the lowest for typodont 3 at 0.716. Subjects had better agreement with the answer key on the typodonts with fewer surfaces of simulated calculus, while having less agreement with the answer key on the typodonts with more surfaces of simulated calculus.

Discussion

Calibration might enhance the consistency of faculty members’ clinical teaching and evaluation. To determine the effectiveness of faculty calibration training programs, examiner scores should be compared before and after any training program as well as to reproducibility levels cited in the literature.

This pilot study evaluated the effectiveness of a training program on intra- and interrater reliability levels with regard to calculus detection using an 11/12 explorer on typodonts. Intra- and interrater reliability levels were assessed by comparing faculty calculus detection scores to an answer key using Cohen’s Kappa and ANOVA. The aim was to determine if a series of training sessions could enhance faculty members’ self-agreement and agreement with each other in an effort to enhance consistency when evaluating students. Calibration has been suggested as a way to enhance students’ learning and attitude.14

Few studies of this nature exist in the dental hygiene literature, so utilizing a small convenience sample served as a pilot study of the training method before a larger scale study would be undertaken. Intra- and interrater reliability levels did not improve significantly following training. However, the lack of improvement does not negate some potential benefits of the training program. After the study ended, faculty members in the experimental group anecdotally reported becoming more aware of their exploring skills and attuned to the students’ perceptions of lack of consistency among faculty.

The training sessions in this study did not have a significant effect on intrarater reliability levels, a finding consistent with several other studies in the dental literature.23-28 It appears that increased reliability levels with regard to self-agreement are difficult to achieve. All study subjects had moderate to high Kappa averages at the pretest ranging from 0.442 to 1.00. Only two faculty members scored below the 0.61–0.80 full agreement range: one new faculty member (0.442) and one faculty member with more than twenty years of teaching experience (0.462). These data indicate that, even before the study commenced, most subjects were already in the full agreement range. This initial level of skill leaves little room for improvement in intrarater reliability. This finding indicates that calibration training such as this might be more beneficial for faculty members with lower pretest Kappa averages or for new faculty members who need to calibrate with more experienced faculty members. Four of the study subjects were new faculty members, although they had many years of clinical dental hygiene experience.

The training sessions in this study did not have a significant effect on interrater reliability levels either. This finding is also consistent with similar studies in the dental literature. Pippin and Feil used a similar method (manikins with calculus) that showed poor interrater reliability levels, yielding a low Kappa average of 0.34.17 Other than this study by Pippin and Feil, no published literature reported results of calibration studies utilizing typodonts with simulated calculus. Research within restorative dentistry used other calibration methods such as use of human subjects, use of rubrics with visual criteria, and questionnaires that indicated interrater reliability levels can increase with training.2,3,16,17,24-27 Numerous studies show that increased interrater reliability levels are possible with strong training programs that are maintained regularly over time.2-4,7,16,17,24-28

One of the most interesting aspects of the study was the finding that all subjects (regardless of which group assignment) had better agreement with the answer key on the typodonts with fewer surfaces of simulated calculus, while having less agreement with the answer key on the typodonts with more surfaces of simulated calculus. In essence, the number of surfaces of calculus deposits per typodont apparently influenced the subjects’ judgment. All subjects’ Kappa averages declined as the number of surfaces of calculus increased. Mean Kappa averages for each typodont are indicated on both Tables 1 and 2. As stated earlier, Typodont 1 had the fewest surfaces of simulated calculus (sixteen out of 112 surfaces); typodont 2 had twenty surfaces out of 112; and typodont 3 had the most surfaces of simulated calculus with twenty-six out of 112 surfaces. Future
studies of this nature should include more surfaces of calculus.

Limitations of this study might have affected the results. The most important element of the methodology to change in future studies of calibration in calculus detection is the typodonts used for testing and training. Subjects indicated that the models used in the study were not adequate for realistic simulation of calculus and for exploring. Subjects had difficulty exploring due to the typodont’s anatomy. They reported the subgingival area did not feel realistic and they had a hard time distinguishing the calculus from the “bone” (the junction of the simulated alveolar ridge).

Timing of the testing and training in the late afternoon following the subjects’ normal work day also might have been a limitation of this pilot study. Ideally, training and testing would take place earlier in the day when subjects would be well rested, alert, and have better tactile sensation.

Although the results of this pilot study did not produce significant findings with regard to rater reliability levels, recommendations for future study of faculty calibration and detection of calculus were identified. Future research should include the use of larger and more varied samples (including faculty from multiple sites) as well as the development and use of alternate calculus detection methods. Studies might utilize human subjects to provide realistic tactile sensations and emerging technologies such as endoscopy or red LED systems for calculus mapping as an enhanced gold standard. More realistic typodonts for teaching and evaluating calculus detection are needed. Future research also should include follow-up training sessions scheduled at regular intervals due to the fact that the effects (if any are found) of calibration training drop off over time.

Calibration of faculty remains a topic of interest for research as well as for the education of dental hygiene and dental students. Accuracy and consistency in clinical teaching are necessary to enhance student learning, motivation, and satisfaction. Calibration is one way to enhance consistency. Although the results of this study did not produce increased intra- and interrater reliability levels, we believe that future calibration studies should be conducted in an attempt to increase accuracy and consistency when evaluating students and to determine the impact calibration has on students’ learning, attitude, and satisfaction.

Acknowledgments

The authors would like to thank Ms. Patricia Lenton for her assistance with the development of the training sessions and overall calibration knowledge; Dr. James Hodges and graduate student Qun Shi for their assistance with the study design and statistical analysis; Professor Denise Bowen (Idaho State University) for her assistance with manuscript preparation; and the dental hygiene faculty members at the University of Minnesota School of Dentistry for their participation in this project. Additionally, the authors would like to thank the Department of Dental Hygiene at the University of Minnesota School of Dentistry and the Minnesota Dental Hygienists’ Association for their support of this project.

REFERENCES