Development and Assessment of Discrimination Exercises for Faculty Calibration in Preclinical Operative Dentistry

Sumitha N. Ahmed, BDS, MS; John Sturdevant, DDS; Rebecca Wilder, RDH, MS; Vicki Kowlowitz, PhD; Lee Boushell, DMD, MS

Abstract: The aims of this study were to identify the level of interexaminer agreement among preclinical operative dentistry faculty members when grading Class II preparations performed by first-year dental students; to develop discrimination exercises for specific preparation components where interexaminer agreement was poor; and to evaluate if the discrimination exercises were able to improve inter- and intraexaminer agreement. In the preliminary phase of this study, 13 components of 32 Class II cavity preparations were assessed by eight course faculty members at one U.S. dental school. Analysis of average interexaminer agreement on these components revealed that six were below 60%. These were proximal contact clearance, retention groove placement, retention groove depth, preparation walls, preparation margins, and preparation toilet/debris. A 30-minute calibration session was subsequently developed to provide discrimination exercises utilizing 3-D models and digital images of various levels of student performance for five of the six components. Immediately following calibration, the course faculty assessed the same 32 preparations (Phase I) followed by a delayed assessment without calibration (Phase II) approximately six months later. The results showed that overall interexaminer reliability improved after calibration. Although there was a decline in interexaminer reliability after an interval of six months (Phase II), the degree of variation among examiners was lower than in the preliminary assessment. These findings support the use of discrimination exercises for preclinical operative dentistry course faculty to increase interexaminer agreement and thereby improve the consistency of faculty-student communication.

Dr. Ahmed is Assistant Professor, Department of Operative Dentistry, University of North Carolina at Chapel Hill School of Dentistry; Dr. Sturdevant is Associate Professor, Department of Operative Dentistry, University of North Carolina at Chapel Hill School of Dentistry; Prof. Wilder is Professor, Department of Dental Ecology, University of North Carolina at Chapel Hill School of Dentistry; Dr. Kowlowitz is Evaluation, Teaching, and Learning Consult, University of North Carolina at Chapel Hill School of Dentistry and School of Nursing; and Dr. Boushell is Associate Professor, Department of Operative Dentistry, University of North Carolina at Chapel Hill School of Dentistry. Direct correspondence to Dr. Sumitha N. Ahmed, Department of Operative Dentistry, CB #7450, School of Dentistry, University of North Carolina, 441 Brauer Hall, Chapel Hill, NC 27599-7450; 919-537-3146; sumitha_ahmed@unc.edu.

Keywords: dental education, operative dentistry, calibration, intraexaminer reliability, interexaminer reliability, faculty calibration

Submitted for publication 10/14/15; accepted 1/5/16

Operative dentistry concepts and techniques are typically introduced to first-year dental students in a preclinical operative dentistry course. In most dental schools, students rely on multiple faculty members for application, reinforcement, and enhancement of theoretical principles during a laboratory portion of the course. Faculty members teaching the course are expected to provide consistent formative and summative feedback on student performance. Inconsistencies in grading among faculty members may lead to confusion and frustration for students. Haj-Ali and Feil's attempt to increase faculty agreement through improved communication of specific performance criteria, rating scales, and/or training met with inconsistent results.[1] Sharaf et al. carefully analyzed each of the components of the evaluation system and then took targeted steps, through faculty calibration training, to improve agreement.[2]

Faculty reliability, also referred to as faculty calibration, may be defined as the level of agreement among multiple faculty members that occurs while assessing student performance.[3] Faculty calibration can be divided into interexaminer reliability and intraexaminer reliability. Interexaminer reliability measures the level of agreement among multiple examiners when they are examining the performance of the same group of students on the same task.[4] Intraexaminer reliability describes the consistency of a single examiner in grading the same sample on multiple occasions.[3] Studies in the field of faculty calibration have shown that establishing agreement among faculty members is difficult; this may be due to

inconsistent grading methods, differing rating scales, and differences in individual teaching philosophy.[5-9]

In the preliminary phase of our study, we evaluated the level of faculty agreement in assessing 13 components of Class II cavity preparations for amalgam restorations by first-year dental students at the University of North Carolina at Chapel Hill School of Dentistry. Each component of the Class II preparation had a set of specific criteria that defined clinically acceptable and clinically unacceptable levels of procedure accomplishment. It is widely accepted that levels of agreement should minimally exceed that which would happen by chance (50%) alone.[10-14] Therefore, for this study, 60% was arbitrarily set as the minimum level of agreement, and average percentage agreement that fell below 60% was considered poor. A calibration exercise was designed to improve faculty agreement on five of the components that had an average agreement below 60%.

First-year dental students depend heavily on a consistent message from their faculty as they learn multiple procedures during their preclinical courses.[8,12] As the preliminary phase of our study suggested areas of faculty inconsistency, we determined that further steps should be taken to identify areas of poor faculty calibration and to target those areas with strategies designed to enhance inter- and intraexaminer reliability in assessing student performance. Therefore, the aims of this study were as follows: 1) to identify the level of interexaminer reliability among the preclinical operative dentistry faculty members when assessing Class II cavity preparations performed by first-year dental students; 2) to develop targeted exercises designed to enhance faculty members' ability to discriminate among levels of student performance (discrimination exercises), and to organize and present these discrimination exercises to individual faculty members as part of a calibration session; 3) to evaluate the effectiveness of discrimination exercises (as revealed by inter- and intraexaminer reliability) in increasing initial levels of faculty calibration (Phase I); and 4) to evaluate the effectiveness of discrimination exercises (as revealed by inter- and intraexaminer reliability) in sustaining an increase in levels of faculty calibration over a time interval of at least six months (Phase II).

Materials and Methods

The University of North Carolina Institutional Review Board determined that this study was exempt from review (#12-0262). The longitudinal, non-randomized cohort study was conducted from 2011 to 2013 at the University of North Carolina at Chapel Hill (UNC) School of Dentistry. The dentiform teeth used were a model of tooth #30 with simulated MOD caries, model # A27A-46U, Kilgore International Inc. (Coldwater, MI, USA). The participants were eight course faculty members in the UNC Department of Operative Dentistry. The principal investigator (SNA) conducted the individual calibration sessions with each examiner.

In the preliminary phase of the study, 32 Class II preparations representing ideal (n=8), acceptable (n=8), correctable (n=8), and unacceptable (n=8) student performance were randomly selected from a pool of 82 preparations. The same 32 Class II preparations were assessed by the examiners in the preliminary, Phase I, and Phase II parts of the study. The cavity preparations had been completed by first-year dental students as part of the preclinical operative dentistry course. In this course, students were instructed to prepare an ideal (according to specific criteria) MOD cavity preparation for restoration with dental amalgam, with complete removal of simulated caries. The preparations were done in a preclinical simulation laboratory designed to replicate the clinical setting. The student assessment form used in the study is shown in Table 1. Dentiform tooth #30 was placed in the dentiform with adjacent teeth (#29 and #31) forming proximal contacts during preparation. The criteria for cavity preparation were adopted from *Sturdevant's Art and Science of Operative Dentistry,* 5th edition.[13]

The duration of the calibration session was 20-40 minutes and utilized discrimination exercises to bring clarity to various levels of student performance on components identified in the preliminary assessment as having low faculty calibration. The discrimination exercises included 3D demonstration models of actual dentiform teeth, which represented various levels of student performance of preparation components and digital images of their ideal counterparts organized in the form of a PowerPoint presentation. At the end of the presentation, a detailed discussion was conducted with each faculty member regarding specific criteria outlined for each component of Class II cavity preparation as it appeared on the evaluation form.

Haj-Ali and Feil used photographs of ideal and variations of ideal to calibrate faculty in their study.[1] In our study, visual and tactile exercises were designed with 3D models in an effort to enhance identi-

fication of ideal performance of individual procedural components and discrimination of variations from the ideal. Discrimination exercises were designed for five of the components identified in the preliminary assessment as having poor interexaminer reliability: 1) proximal contact clearance, 2) retention groove placement, 3) retention groove depth, 4) preparation walls, and 5) preparation margins. A discussion

Table 1. Class II amalgam procedure performance rubric used for preliminary, Phase I, and Phase II assessments

	External Outline
Caries removal	Complete removal at the DEJ Incomplete removal at DEJ
Isthmus width	Less than 1 mm Between 1 mm and 1/3 of intercuspal distance Between 1/3 and 1/2 of intercuspal distance Greater than 1/2 of intercuspal distance
Proximal contact clearance	No clearance Open up to 0.5 mm in all directions Open between 0.5 and 0.75 mm in any direction Open more than 0.75 mm in any direction
Adjacent tooth damage	No damage Requires recontouring Requires restoration
	Internal Form
Enamel present	None Less than or equal to 50% of preparation Greater than 50% of preparation
Primary pulpal/axial wall	Less than or equal to 0.5 mm internal to DEJ 0.5 to 1.5 mm internal to DEJ 2.0 to 2.5 mm internal to DEJ Greater than 2.5 mm internal to DEJ
Caries removal	Incomplete Complete Complete with excessive dentin removal
	Retention Form
External walls	Occlusal convergence with ~90° cavosurface margins Excessive occlusal convergence with <90° cavosurface margins External walls parallel External walls diverge occlusally
Retention groove placement	Undermined enamel =0.2 mm internal to DEJ Between 0.2 and 1 mm internal to DEJ Greater than 1 mm internal to DEJ Not visible
Retention groove depth	Undetectable Between 0.1 and 0.5 mm Greater than 0.5 mm
	Finishing
Preparation walls	Smooth, gentle transitions Rough, abrupt transitions
Preparation margins	Unsupported enamel (<80°) Supported enamel (90°) Enamel margin >100°
Preparation toilet	Debris present Clean

of specific criteria for the component "preparation toilet" was completed, but no other discrimination exercise was developed for this component. The term "preparation toilet" was changed to "preparation debris" after the preliminary assessment.

Design of the five discrimination exercises, based on *Sturdevant's Art and Science of Operative Dentistry,*[13] was as follows:

1) Proximal contact clearance: In an attempt to standardize the measurement of proximal clearance, the UNC periodontal probe was used as an assessment tool. Using Image J software (National Institutes of Health, Bethesda, MD, USA), we assessed high-quality digital images of 82 probes to determine their mean diameter (mm). They were found to be 0.5 (±0.02) mm at the 4-5 mm mark and 0.75 (±0.02) mm at the 11-12 mm mark (Figure 1, panel a). The discrimination exercise included a tactile demonstration of how the proximal contact clearance can be assessed using the UNC periodontal probe (Figure 1, panels b and c).

2) Retention groove placement: The discrimination exercise included a series of sagittal sectioned 3D models of Class II preparations with different axial wall depths with variations of retention groove placement that were representative of types of errors found in student preparations (Figure 2, panels a-g).

3) Retention groove depth: The examiners were provided an explorer and three sagittal sectioned dentiform teeth with Class II preparations that contained retention grooves that were >0.5 mm in depth (unacceptable), 0.1-0.5 mm in depth (ideal), and <0.1

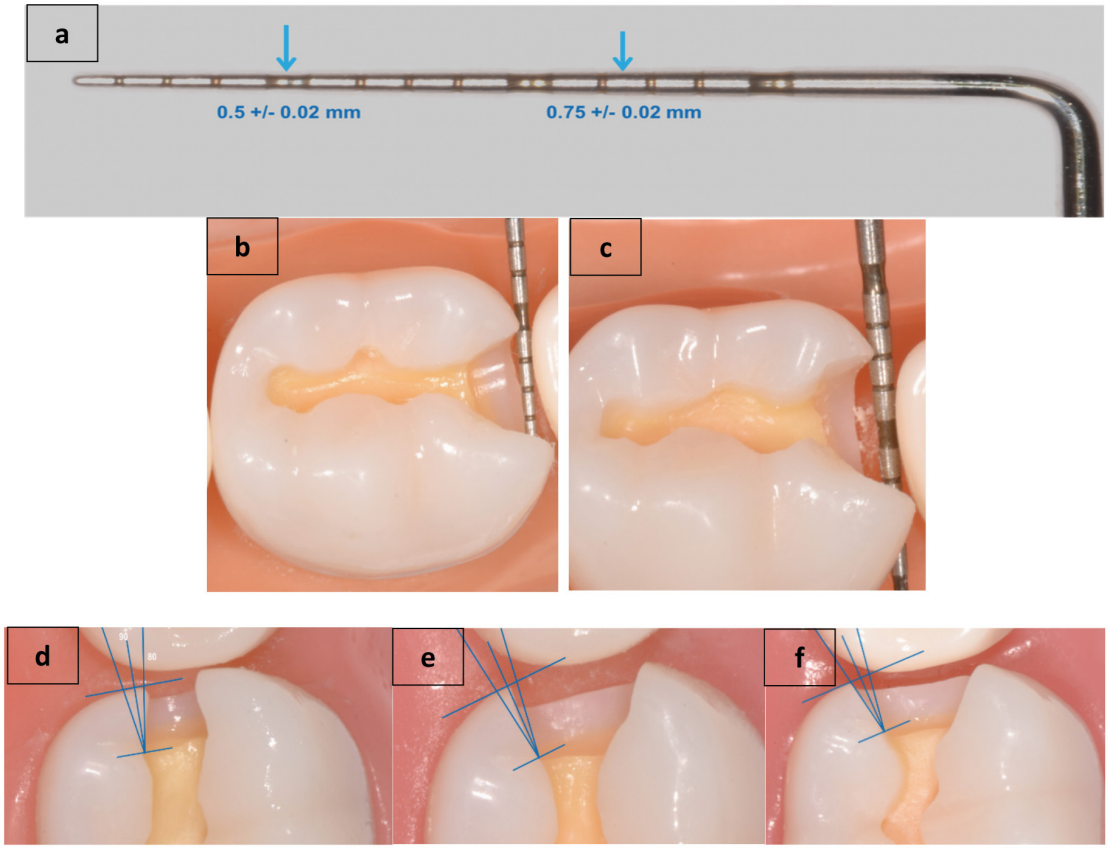

Figure 1. Probe and 3D models used in exercises on proximal contact clearance (panels a, b, c) and preparation margins (panels d, e, f)

Note: Mean diameter of UNC 15 probe at 4-5 mm mark was 0.5 ±0.02 and 0.75 ±0.02 at 11-12 mm mark (panel a). Measurements were made using Image J software. 3D models were used to demonstrate clinically acceptable (panel b) and unacceptable (panel c) examples of proximal contact clearance. Images of 3D model with a superimposed protractor demonstrate clinically unacceptable preparation wall orientation of <80° (panel d), clinically acceptable wall orientation of 90° (panel e), and clinically unacceptable wall orientation of >100° (panel f).

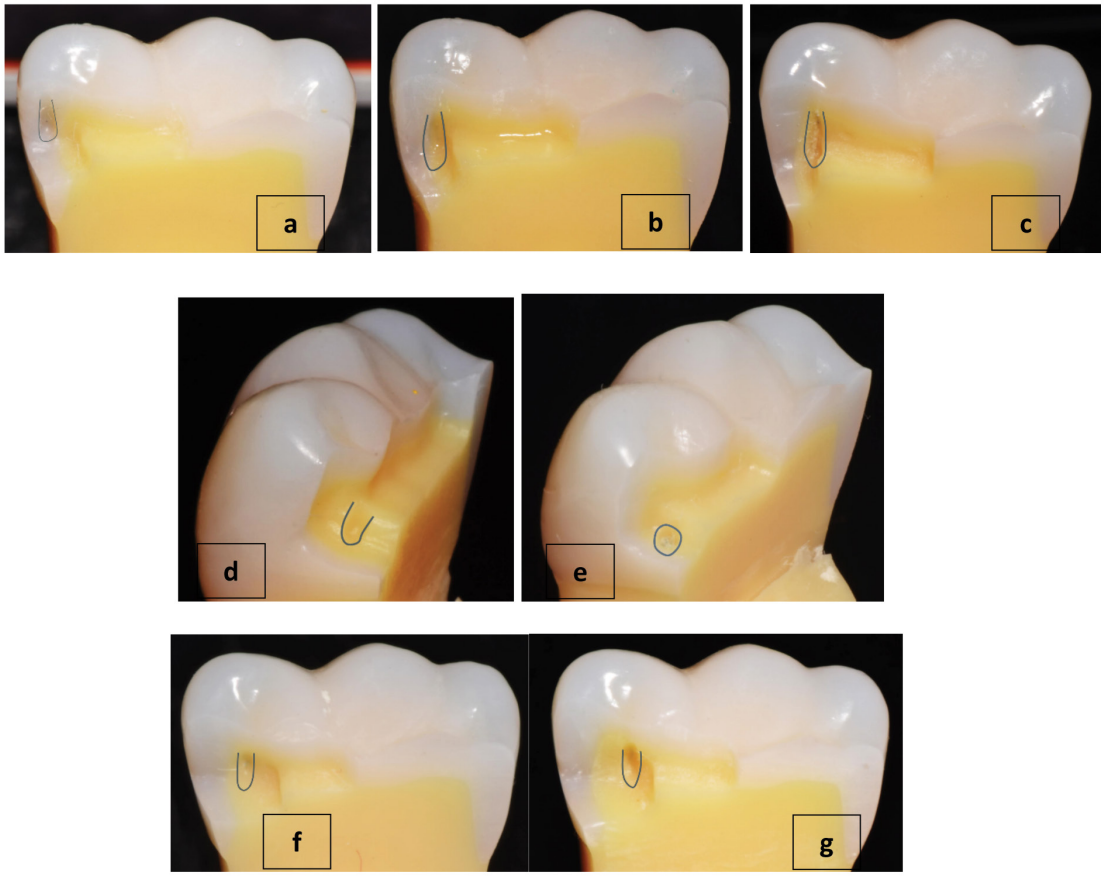

Figure 2. 3D models used in exercise on retention groove placement

Note: 3D models with ideal axial wall depth demonstrate unacceptable retention groove placement in enamel facial wall of a preparation (panel a); unacceptable retention groove placement at DEJ of facial wall of a preparation (panel b); acceptable retention groove placement ~0.2 mm internal to DEJ such that it is partially in dentin facial wall and partially in adjacent axial wall of a preparation (panel c); clinically unacceptable retention groove placement in axial wall of a preparation (panel d); unacceptable retention groove placement in gingival wall of a preparation (panel e); acceptable retention groove placement ~0.2 mm internal to DEJ in facial dentin wall of a preparation (panel f); and clinically unacceptable retention groove placement in line angle of facial and axial walls of a preparation (panel g).

mm in depth (unacceptable). No digital images were used to enhance discrimination of the various levels of performance in this component.

4) Preparation walls: 3D models of four Class II cavity preparations that contained the various combinations of levels of clinically acceptable or unacceptable preparation wall finishing, typical of student performance, were used (Figure 3, panels a-h).

5) Preparation margins: Discrimination exercises consisted of digital images and 3D models of clinically acceptable (90°) and unacceptable (<80° or >100°) cavosurface margin orientations (Figure 1, panels d-f).

Detailed discussion of the 13 components of a Class II cavity preparation and the specific criteria defining levels of student performance for each component listed on the assessment rubric form was conducted following the PowerPoint presentation. Examiners were given the opportunity to ask questions throughout the calibration session.

In the Phase I assessment, each examiner was asked to assess the 32 Class II cavity preparations immediately after completion of the calibration session. The assessment required two to three hours and was accomplished in one sitting. In the Phase II assessment, the examiners completed a second assessment of the same 32 Class II preparations after an average interval of six months. This assessment was conducted using the same controlled settings as the preliminary and Phase I assessments. However,

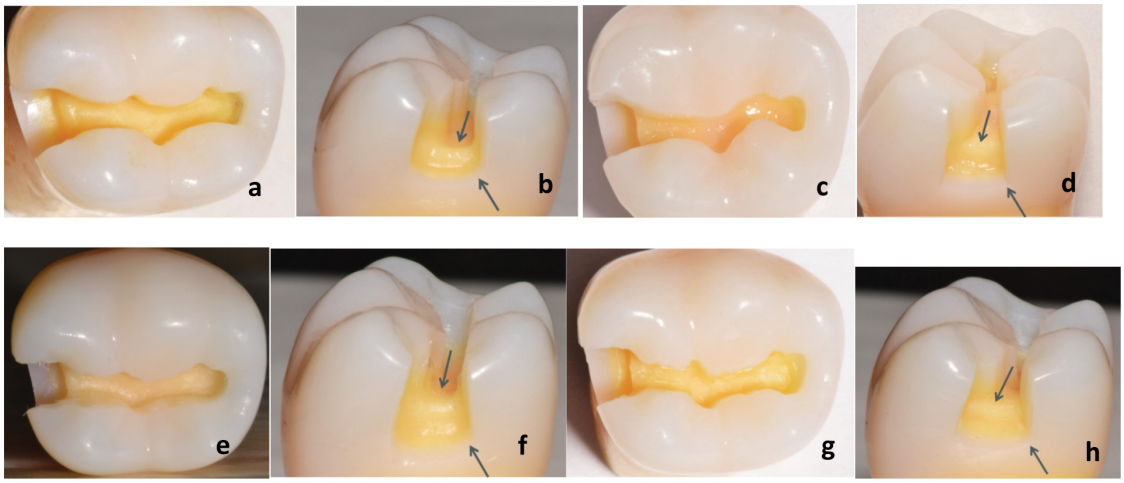

Figure 3. 3D models used in exercise on preparation walls

Note: 3D models demonstrating clinically acceptable preparation finish of smooth walls and gentle transitions (panels a and b); clinically unacceptable preparation finish of smooth walls and abrupt transitions (panels c and d); clinically unacceptable preparation finish of rough walls and gentle transitions (panels e and f); and clinically unacceptable preparation finish of rough walls and abrupt transitions (panels g and h).

no calibration session was provided. The purpose of the Phase II assessment was to evaluate the impact of the passage of time on levels of examiner calibration.

The Class II preparation assessment point values were transferred to a digital file, and the names of the examiners were coded with letters A to H by an independent investigator so that the principal investigator was blinded to examiner identity. The degree of agreement (i.e., the level of concordance and discordance) between each pair of examiners (interexaminer agreement) and for each examiner with himself or herself (intraexaminer agreement) were analyzed using Weighted Kappa and McNemar analysis. The interexaminer reliability was reported as the average percentage agreement among the eight examiners for the preliminary, Phase I, and Phase II assessments. The 95% confidence interval (CI) was calculated for each component of the cavity preparation for all three assessment sessions.

Results

Interexaminer Reliability: Phase I

The interexaminer reliability, reported as average percentage agreement, among the examiners increased for seven of the 13 components when compared to the results of the preliminary assessment (Figure 4). However, for three components (adjacent tooth damage, enamel present, and primary/axial wall depth), the average percentage agreement did not change. For another three components (external walls, retention groove placement, and retention groove depth), there was a decline in average percentage agreement.

Assessment of some components targeted with discrimination exercises showed increased levels of faculty agreement (proximal contact clearance, preparation walls, and preparation margins), whereas assessment of other components did not (retention groove placement and retention groove depth). The 95% CI for each of the 13 components in the preliminary, Phase I, and Phase II assessments is shown in Table 2.

Intraexaminer Reliability: Preliminary to Phase I

The average intraexaminer agreement (agreement of each examiner with himself or herself) among the course faculty was 74% (±5%) when comparing the preliminary assessment to Phase I. Assessment of the 13 preparation components varied greatly from the preliminary assessment to Phase I. The component "enamel present" had the lowest intraexaminer variation, and the component "proximal contact clearance" had the highest intraexaminer

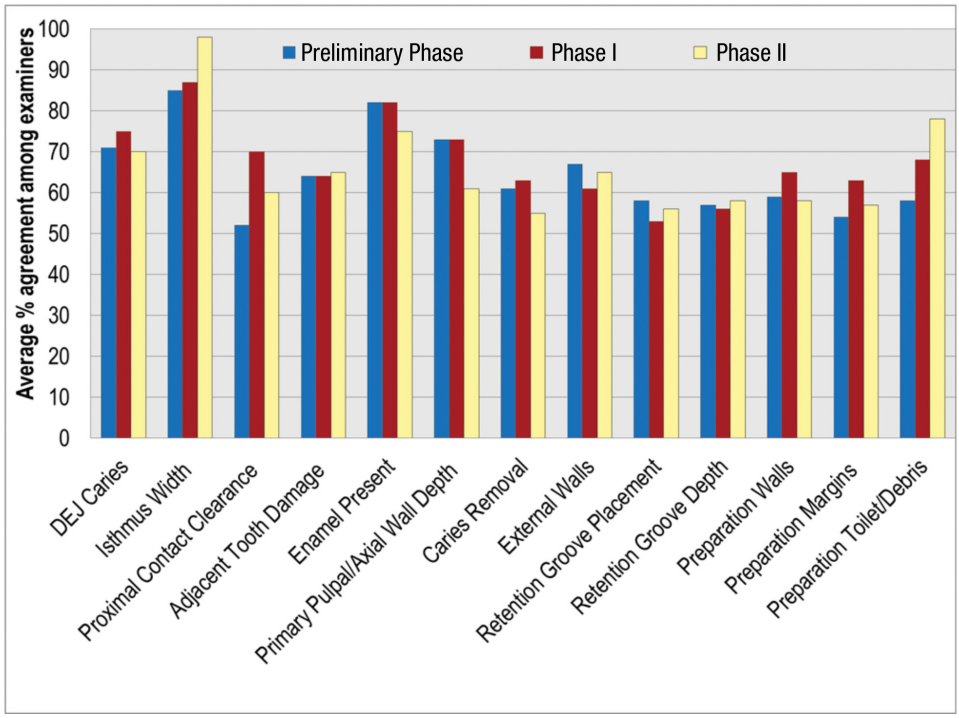

Figure 4. Average percentage agreement among examiners after assessment of 13 procedural components of Class II cavity preparations at three phases of study

Table 2. Assessment of 13 components of Class II cavity preparation during preliminary, Phase I, and Phase II assessments, by 95% confidence interval (CI)

	Preliminary	Phase I	Phase II
Component	95% CI (Lower, Upper)	95% CI (Lower, Upper)	95% CI (Lower, Upper)
DEJ caries	(67.29, 75.01)	(68.79, 80.54)	(61.11, 76.62)
Isthmus width	(80.56, 88.90)	(83.98, 90.35)	(96.53, 98.78)
Proximal contact clearance	(42.00, 62.03)	(67.84, 73.02)	(53.38, 67.17)
Adjacent tooth damage	(58.47, 69.44)	(59.78, 68.81)	(59.15, 70.99)
Enamel present	(72.74, 92.21)	(70.47, 93.15)	(66.15, 90.33)
Primary pulpal/axial wall	(68.56, 77.87)	(66.92, 78.18)	(54.53, 72.94)
Caries removal	(52.81, 64.38)	(54.79, 71.11)	(43.88, 66.17)
External walls	(62.45, 71.49)	(52.19, 70.15)	(56.35, 75.80)
Retention groove placement	(54.72, 62.03)	(46.15, 59.97)	(46.51, 59.97)
Retention groove depth	(53.07, 61.67)	(51.03, 61.03)	(50.51, 61.33)
Preparation walls	(57.04, 63.96)	(54.16, 68.62)	(53.93, 62.15)
Preparation margins	(46.21, 61.83)	(58.13, 67.32)	(50.36, 60.59)
Preparation toilet	(52.52, 64.23)	(63.80, 72.81)	(74.15, 86.35)

variation. These results suggest that the training sessions had improved the examiners' assessment skills in the short term, although to a varying degree across assessment components.

Interexaminer and Intraexaminer Reliability: Phase II

The interexaminer agreement began to decline for eight of the 13 components after an average inter-

val of six months (Figure 4). For three components (adjacent tooth damage, retention groove placement, and retention groove depth), there was a slight increase (1-3%) in average percentage agreement among the examiners. The average interexaminer agreement among the examiners stayed the same for one component (external walls). However, for two components (isthmus width and preparation toilet/debris), the interexaminer reliability continued to increase. The 95% CI for assessment of each component of the cavity preparation during Phase II is shown in Table 2.

The average intraexaminer reliability among the course faculty was 77% (±7%) when comparing the preliminary assessment to Phase II and 76% (±8%) when comparing Phase I to Phase II. While some examiners remained consistent in their assessment patterns from Phase I to Phase II, others reverted to assessing with a variation of 30-40% for a few components. This variation in intraexaminer reliability was lowest for the assessment "isthmus width" from the preliminary assessment to Phase I to Phase II and highest for the assessment "proximal contact clearance."

Discussion

It has been well documented that improving the level of agreement among faculty members when performing student assessments is not an easy task.[5,7,14] The overarching purposes of our study were to determine faculty interexaminer and intraexaminer reliability in assessing 13 components of a preclinical operative procedure completed by first-year dental students and to increase faculty agreement in areas where it was low. The results of the preliminary assessment confirmed that there were areas of low interexaminer agreement among the faculty. As part of the study, exercises were developed and presented in a calibration session for the purpose of increasing the faculty members' ability to discriminate among various levels of student performance. The effectiveness of these exercises was evaluated through the use of immediate (Phase I) and delayed (Phase II) inter- and intraexaminer reliability testing.

The concordance and discordance between each pair of examiners were analyzed using Weighted Kappa and McNemar analysis. A weakness of the standard Kappa statistic is that all disagreements are treated equally. Unlike the standard Kappa analysis, the Weighted Kappa statistic measures the degree of agreement. For example, as explained by Viera and Garrett, we may not care whether one radiologist categorizes a mammogram finding as normal and another categorizes it as benign, but we do care if one categorizes it as normal and the other as cancer.[15] Similarly for our study, if we look at the component on proximal contact clearance, a disagreement between "no clearance" and "open up to 0.5 mm in all directions" is not as severe as that between "no clearance" and "open more than 0.75 mm in any direction."

Proximal contact clearance is traditionally identified by the appearance of a visually open area between the proximal surfaces of adjacent teeth at the proximal height of contour. Assessment of the distance of proximal clearance is vague and subject to personal bias. Dimitrijevic et al. examined dentists' and dental students' abilities to estimate small depths and distances and found that individual perceptual abilities varied widely.[16] The results from Phase I in our study showed that the interexaminer reliability improved with the use of discrimination exercises for this component. Introduction of a specific instrument to assess this component may have contributed to the increase in interexaminer reliability and limited the influence of personal examiner bias.

There was also an increase in interexaminer reliability for components such as isthmus width and preparation debris, for which no discrimination exercises were designed. Detailed discussions regarding the specific criteria outlined for all 13 components of the Class II preparation may have produced increased understanding of preparation criteria and limited subjective interpretation.

The two components for which there was no increase in agreement among the examiners were retention groove placement and retention groove depth. This was in spite of faculty participation in carefully designed discrimination exercises. Although studies have been done on the significance, ideal position, and ideal depth of retention grooves, those researchers were not able to achieve consensus.[16,17] A survey by Moore published in 1992 investigated the teaching of proximal retention grooves in Class II cavity preparations for amalgam restoration.[17] Among the 59 U.S. and Canadian schools included in his study (64 total schools; response rate 92%), only 61% (36 schools) reported teaching retention grooves. These findings suggest there was at that time a lack of consensus with regard to the use of retention grooves, a result we found among the faculty members in our study. Our results showed a decrease in interexaminer

agreement (from preliminary assessment to Phase I to Phase II) for the components retention groove placement and retention groove depth. Variations in the examiners' professional attitudes about the use of retention grooves may have limited our ability to increase the level of agreement in assessing this component. It is important to note that both the placement and accurate assessment of retention grooves are technically challenging.

Although there was an increase in average percentage agreement among the examiners after the calibration, the results of the Phase II grading session revealed a definite decline in interexaminer reliability. The study did not test for the point in time at which the level of reliability started to decline, but only detected that there was a decline at an average interval of six months. This information is valuable and can be interpreted as indicating a need for frequent calibration sessions throughout the academic year.

Our study also did not evaluate intraexaminer variation based on the clinical and teaching experience of the examiners, which may be a limitation of the study. Jenkins et al. found that the level of pass/fail differences in their study seemed to be unrelated to the experience of the examiner, with even the senior examiner recording differences of 17%.[8] Mackenzie defined clinical competence in terms of quality, quantity, and need for performance criteria and explained the problem of intraexaminer variability as follows: "A good student finished a preparation in an ivorine tooth and took it to the instructor. The instructor said, 'Fine work,' and gave him an A. A little while later a poor student took the A-graded tooth to the same instructor. The instructor looked at it, said, 'Hm-mm, OK,' and gave him a C."[18] Efforts to improve consistent, unbiased application of defined criteria in grading remain essential.

Another limitation of this study was that the results may have been negatively influenced by examiner fatigue. Having to grade a large number of samples at one time may cause the examiner to lose focus. As suggested by Dhuru et al., future studies should limit the number of preparation samples or have the examiners take frequent breaks after grading 10-15 samples.[3]

The results of our study indicate that discrimination exercises can be beneficial for faculty calibration and also highlight a need for more frequent calibration sessions. Future efforts in the field of faculty calibration should focus on designing or identifying instruments available to all students and faculty that may aid in objective formative and summative evaluation of student performance. In general, our findings support the use of discrimination exercises for faculty calibration to improve the consistency of faculty-student communication. The study findings revealed an increase in inter- and intraexaminer reliability after calibration. Although there was a decline in interexaminer reliability after six months (Phase II), the degree of variation among examiners even then was lower than in the preliminary assessment, suggesting that some improvements were retained.

Conclusion

The results of this study suggest that overall interexaminer reliability may be improved through a focused process of faculty calibration and that improving faculty calibration in assessments of at least some tooth preparation components may benefit from targeted discrimination exercises. We found that the use of an objective means of measuring depth and distance (the periodontal probe) increased interexaminer reliability. However, our results suggest that frequent calibration sessions are necessary for maintaining at least minimum levels of faculty calibration as the positive effects of our calibration sessions were largely lost after six months.

Acknowledgments

The authors would like to thank Dr. Ceib Phillips and Ms. Debbie Price for their time and guidance with the statistical analysis.

REFERENCES

1. Haj-Ali R, Feil P. Rater reliability: short- and long-term effects of calibration training. J Dent Educ 2006;70(4):428-33.
2. Sharaf AA, AbdelAziz AM, El Meligy OA. Intra- and interexaminer variability in evaluating preclinical pediatric dentistry operative procedures. J Dent Educ 2007;71(4):540-4.
3. Dhuru VB, Rypel TS, Johnston WM. Criterion-oriented grading system for preclinical operative dentistry laboratory course. J Dent Educ 1978;42(9):528-31.
4. Brown G, Manogue M, Martin M. The validity and reliability of an OSCE in dentistry. Eur J Dent Educ 1999;3(3):117-25.
5. Houpt MI, Kress G. Accuracy of measurement of clinical performance in dentistry. J Dent Educ 1973;37(7):34-46.
6. Houpt MI, Kress G. Evaluation training: training raters to evaluate clinical performance. Dent Digest 1979;10(11):67-76.
7. Hinkelman KW, Long NK. Method for decreasing subjective evaluation in preclinical restorative dentistry. J Dent Educ 1973;37(9):13-8.

8. Jenkins SM, Dummer PM, Gilmour AS, et al. Evaluating undergraduate preclinical operative skill: use of a glance and grade marking system. J Dent 1998;26(8):679-84.
9. Lilley JD, ten Bruggen Cate HJ, Holloway PJ, et al. Reliability of practical tests in operative dentistry. Br Dent J 1968;125(5):194-7.
10. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20(1):37-46
11. Cohen J. Weighed kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968;70(4):213-20.
12. Salvendy G, Hinton WM, Ferguson GW, Cunningham PR. Pilot study on criteria in cavity preparation: facts or artifacts? J Dent Educ 1973;37(11):27-31.
13. Roberson TM, Swift EJ Jr. Sturdevant's art and science of operative dentistry, 5th ed. St. Louis, MO: Mosby, 2006: 353-409.
14. Scruggs RR, Daniel SJ, Larkin A, Stoltz RF. Effects of specific criteria and calibration on examiner reliability. J Dent Hyg 1989;63(3):125-9.
15. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med 2005;37(5):360-3.
16. Dimitrijevic T, Kahler B, Evans G, et al. Depth and distance perception of dentists and dental students. Oper Dent 2011;36(5):467-77.
17. Moore DL. Current teaching of proximal retention grooves for Class II amalgam preparations. J Dent Educ 1992;56(2):131-4.
18. Mackenzie RS. Defining clinical competence in terms of quality, quantity, and need for performance criteria. J Dent Educ 1973;37(9):37-44.