DEVELOPMENT AND ASSESSMENT OF DISCRIMINATION EXCERCISES FOR FACULTY CALIBRATION IN PRECLINICAL OPERATIVE DENTISTRY

SUMITHA NAZAR AHMED

A thesis submitted to the Faculty of The University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Science in the Department of Operative Dentistry at the School of Dentistry

Chapel Hill
2014

Approved by:

Lee W. Boushell

John Sturdevant

Rebecca S. Wilder

Vicki Kowlowitz

ABSTRACT

Sumitha Nazar Ahmed: Development and Assessment of Discrimination Exercises
for Faculty calibration in Preclinical Operative Dentistry
(Under the direction of Lee Boushell)

The purpose of this research was to develop and assess methods for faculty
calibration in a preclinical operative dentistry course, to evaluate intra- and
interexaminer agreement based on the information provided during calibration
sessions, and to evaluate intra- and interexaminer reliability 6 months after initial
calibration. The average intraexaminer agreement among the course faculty was
74(+/-5) %. The overall interexaminer reliability improved after calibration. The
use of an instrument (UNC 15 periodontal probe) was introduced, for objective
measurement of a component. Visual discrimination exercises were developed and
tested for their ability to1) initially aid in increased faculty calibration while
assessing the Class II preparation and 2) provide sustained calibration among the
same faculty over a period of approximately 3 months. The results showed that, the
objective use of an instrument (UNC 15 periodontal probe) and the use of discrimination
exercises increased interexaminer reliability.

To my husband Yashin Ahmed, my children Adi and Ian,
your love and support made this dream come true.

To my parents Naziruddin and Jameela Nazar who taught me that success come
from following one's heart

To Rameez Nazar, Sunitha, Imthiaz Kadeer and my nephew Omi, thank you for
being there for me

ACKNOWLEDGEMENTS

To my Thesis Committee members, Dr. Lee Boushell, Dr. John Sturdevant, Ms. Rebecca Wilder and Dr. Vicki Kowlowitz, I am grateful for your time, encouragement and guidance

Dr. Terry Donovan, Dr. Harald Heymann, Dr. Andre Ritter and Dr. Edward Swift, it has been an honor to be your student

To the staff of the Department of Operative Dentistry, Shannon Tate, Dayna McNaught, Jamie Desoto, Rosanna Arrington, Barbara Walton and Cynthia Lambert. Thank you for always being there, ready to help.

I would like to thank Dr. Rick Walter and Dr. Leinfelder for taking the time to participate in my thesis project. Dr. May, Dr. Eidson and Dr. Zandona, thank you for making me feel like a part of the Operative family.

Dr. Ceib Phillips and Ms. Debbie Price, I thank you for your time and guidance with the statistical analysis. Dr. Wilder thank you for being so patient with me and all the help you have given me over the years.

Alex and Silvia you have been the best classmates. I will never forget these three years of sharing accomplishments, success and failures. You will be in my heart forever

To all the Operative residents who made this journey memorable, Ayesha, Roopsie, Upoma, Villi, Clay, Tan, Mohammad, Anmar and Taiseer. Thank you!

A special thank you to Dr. James Parker for all your help and advice

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

3D   3 Dimensional

CI   Confidence Interval

DEJ   Dentino-enamel Junction

EM   Electromagnetic

ICDAS   International Caries Detection and Assessment System

IRB   Institutional Review Board

MOD   Mesial Occlusal Distal

SD   Standard Deviation

UNC   University of North Carolina at Chapel Hill

US   The United States of America

GG   Glance and Grade

CL   Checklist

CLSC   Checklist with Specific Criteria

Introduction

Dental students undergo rigorous pre-clinical training in preparation for entrance into the clinical setting. Dental faculty utilizes simulated clinical settings in pre-clinical courses for the purpose of developing and assessing dental students' knowledge and skills pre-requisite for patient care. Conservative Operative Dentistry (DENT112) is an example of a pre-clinical operative dentistry taught at the UNC School of Dentistry.

In the didactic session of the pre-clinical operative course the students learn the theory behind different designs of cavity preparations. During the hands-on (or laboratory) session of the course, they use this theoretical information to develop the hand skill necessary for preparing and restoring cavities on dentiform teeth that are mounted in manikins. As part of the pre-clinical course, they come in contact with multiple faculty who grade their work. Learning Theory suggests that it is very important that the information given to the students remain consistent from one instructor to the other in order to avoid any confusion among the students.

As future dentists, dental students are expected to make clinical judgments based on their education. In any educational setting, students rely on the faculty to provide consistent formative and summative feedback. They also rely on the faculty to provide reinforcement and enhancement of concepts that they learn in their dental curriculum. Learning is optimal when multiple faculty

are consistent with each other as they communicate concepts to their students. Active steps must be taken to increase the levels of consistency, or agreement or calibration among faculty. Poor consistency among faculty may lead to confusion and frustration among the students.

Review of Literature

Evaluation methods can be broadly classified into subjective evaluation and objective evaluation. While the word subjective can change from person to person, objective means a fixed reality. Subjectivity associated with grading student performance can lead to confusion and stress among the students. Fuller states that, in an evaluation system guided by subjectivity, it is often difficult for the faculty to "defend" his/her score to the student's satisfaction or his/her own.[1]

According to Jenkins et al, the problems with faculty consistency may lead students to perceive that evaluation methods are somewhat arbitrary. They suggested that this concept can undermine the learning process and produce a negative effect on undergraduate's confidence and performance.[2] Their findings support the idea that consistency in the feedback provided to the students helps to improve their performance and also avoids ambiguity associated with grading procedures. Mackenzie recommends 'maximizing diagnostic feedback' (using objective evaluation methods), for the purpose of student learning.[3] This can only be accomplished by establishing a system that makes use of objective evaluation methods.

The consistency or degree of agreement between faculty members, performing the task of grading samples of student's work, is known as examiner reliability.[4] Examiner reliability can be divided into intraexaminer reliability

and interexaminer reliability. Intraexaminer reliability describes the consistency of a single examiner in grading the same sample on multiple occasions.[4] Interexaminer reliability measures the degree of agreement among the examiners when they evaluate the performance of the same group of students on the same task.[5] Studies in the field of faculty calibration have shown that establishing agreement among faculty members is not an easy task. This can be due to the inconsistent grading methods, differing rating scales, individual teaching philosophy and so on.

Clinical performance can be judged in terms of its 3 possible outcomes. Performance that: 1) is clinically acceptable, 2) needs improvement and 3) is clinically unacceptable. American dental schools have historically assessed performance in pre-clinical simulations of the clinical environment by using 3 methods: 1) Glance and Grade (GG), 2) Checklist (CL) and 3) Checklist with Specific Criteria (CLSC). The Glance and Grade method utilizes a subjective global assessment of the student performance without specific evaluation of each component of the skill. of a skill.[6] The Checklist and Checklist with Specific Criteria methods seek to utilize a more analytical form of evaluation where each component of the entire performance is evaluated separately on a written checklist. The Checklist with Specific Criteria method also defines specific levels of performance of each component.

Houpt and Kress suggests that global assessment can be used for the purpose of certification, i.e., in a situation where the student performance needs to be evaluated as 'satisfactory' or 'unsatisfactory'.[7] Mackenzie refers to global

assessment as a 'product approach' and points out that it should be supplemented with other forms of evaluation. He argues that looking only at the measure of the end- product, without analyzing the factors that contribute to the variations in quality of the product, results in wasteful repetition of the evaluation process.[3]

Both global and analytical approaches have been tested to determine the level of intra- and interexaminer reliability. Goepferd and Kerber[8] developed and tested the efficiency of an analytical system for evaluating class II cavity preparation on primary teeth. The intent of the analytical system was to decrease the subjectivity of clinical evaluation and to introduce objective measures in order to increase examiner reliability. This system was then compared with the traditional glance and grade method to determine degree of intra- and inter-examiner reliability with both methods. The study found that the overall interexaminer reliability improved with the analytical system; however the findings were not tested for statistical significance.

Additional research comparing GG, CL and CLSC methods was accomplished at the University of North Carolina in 1983. The intra- and interexaminer reliability in preclinical pedodontic grading was evaluated. It was concluded that no method yielded superior intra- or interexaminer reliability.[9]

While some tried to increase intra and interexaminer reliability through different evaluation methods (GG vs CL vs CLSC), others have tried to gain superior intra- and interexaminer reliability through altering the nature of the rating scale within the various methods. Houpt and Kress compared 3 different

rating scales: a two-point scale with two specified points, a five-point scale with end points specified and five-point scale with all end points defined. The findings of the study demonstrated that when criteria were defined specifically, examiners tended to be more accurate in their judgments. However, the use of a two-point rating scale was found to have more interexaminer agreement than the use of a five-point rating scale even though a five-point scale may be more beneficial for instructional purposes.[7] Similar results by Hinkelman and Long showed slightly more interexaminer agreement in a two-point evaluation system (pass- fail) than a three-point scaling system ('no improvement necessary', 'clinically acceptable' and 'clinically unacceptable and uncorrectable'). However, the authors suggest that the three-point evaluation system was more useful for ranking according to ability.[10]

Although the results of these studies indicate that a two-point scale leads to greater examiner reliability, they highlight that it gives little to no instructional advantage in teaching and identifying components of a student's performance. In response to concerns about reliability and instructional quality, Deranleau et al[11] conducted a study that investigated if the increase in intra- and interexaminer reliability gained from two-point scoring justified its use over the more instructionally advantageous three-point scoring. Porcelain jacket crown wax-ups and Class II MO wax-ups were assessed by 5 faculty members. All five examiners evaluated each project four times with a time interval of one week between grading sessions. The examiners were also asked about their preference regarding the grading system used and why. The results were similar to the

Hinkelman and Long study who found increased interexaminer reliability in their collapsed two-point scoring system over three-point scale. The examiners who liked the two-option scoring were those with less experience. They disliked the range of competency assessment options and the levels assigned to some criteria. Those who preferred the three-option scoring method reported it allowed them to recognize superior work and, at the same time, avoid unrealistic expectations of perfection.

Research in the field of faculty calibration has also sought to analyze specific criteria, used in assessment methods, in terms of clarity of the verbiage. Studies that have carefully defined checkpoints, designed to decrease ambiguity, have obtained fairly high agreement among the examiners.[12] As defined by Mackenzie et al, eliminating problems with the evaluation process includes, but are not limited to, items such as unstandardized aids to evaluate, incomplete operational definitions, discrepancies in visual acuity and inadequacy in verbal definitions. They concluded that the impact of these problems can be reduced by using checklists with definitions that describe how to make observations and how to categorize the observations with minimal inference or subjective bias.[13]

Sharaf et al, has suggested that the goal of educators should be to design a method of assessment that is both objective and reliable so as to reduce friction between students and faculty over the issue of grading.[14] In an attempt to make the evaluation process more objective, Schiff et al designed a device called the "pulpal floor measuring instrument" to measure the profile of preparations, including depth, smoothness, and flatness of the pulpal floor. The

authors reported significant improvement in operator consistency using this instrument.[15] Cardoso et al used the Kavo PrepAssistant® to objectively evaluate student tooth preparations. The machine was used to compare an ideal tooth preparation by the faculty member against a preparation completed by a student. Both the preparations were digitally scanned and a data sheet containing preparation analysis and feedback was then processed which was given to the student. The authors found the use of the machine in its full potential by the preclinical department to be very labor intensive. Even with a greater guarantee of objective evaluation, the machine was not able to assess all the components of the tooth preparation, leaving 30 % of the evaluation to be completed by the faculty. [16]

In the last two decades there have been an increasing number of studies on computer assisted learning and computer assisted simulation systems.[17-19] DentSim® is one such system that has been gaining popularity among the dental schools in the nation. In addition to teaching operative dentistry preparation techniques, DentSim is also used to teach crown preparations and endodontic access cavities. The impact of DentSim on dental education has been studied extensively.[20,21] Welk et al found that training with DentSim helped in early identification of students who needed more time to acquire the required skills. These students could be helped so as to avoid falling behind in the curriculum. They concluded that the time needed for training students to prepare specific cavities correctly in a conventional lab could be reduced so that the preparation tests of the operative dentistry course could be scheduled earlier than usual.[22]

Such devices help to ensure that the student receives an objective evaluation of their work in addition to consistent feedback. It should be noted that these systems currently cannot be used to provide formative and summative feedback for student performance of dental restorations.

In the absence of such devices the course director relies on the preclinical course faculty to provide consistent summative and formative feedback. Attempts to increase faculty consistency through improved communication of specific performance criteria, rating scales, and/or training have met with inconsistent results.[23] One study found that faculty meetings designed to resolve inconsistencies did not result in increased calibration.[24]

Natkin and Guild reported a statistically significant increase in interexaminer reliability after the training sessions. Each session consisted of the examination of ten randomly selected student projects followed by discussion of the criteria established for those projects.[25] A literature review by Patridge and Mast found that studies on faculty training yielded inconsistent results.[26] Sometimes, the inconsistency and lack of fair grading is due to the absence of clear guidelines on how the examiners should evaluate clinical performances.[27]

Jenkins et al found that the level of pass–fail differences seemed to be unrelated to the experience of the examiner, with even the senior examiner recording differences of 17%.[2] Lilley et al found similar results in their study with interexaminer variability fairly constant at 30 -50% between 3 grading sessions, despite consultation among each other. However, they found high

intraexaminer reliability particularly for the most experienced examiner.[28]
Philip Feil recommends the use of two examiners grading independently but
working in association with one another to develop a bipartisan approach, which
will theoretically facilitate a more uniform method of assessment by reducing
the effect of individual bias.[29]

Scruggs et al conducted a pilot study to investigate the use of specific
criteria and faculty calibration on the reliability of inexperienced examiners on
dental sealant evaluations. The objectives of the study were to identify
differences in calibrated and non-calibrated examiners. The examiners were
calibrated by an expert and a non-expert to assess reliability among the study
participants. They used an analytical, criteria based evaluation system. The
results showed that the calibrated group had a higher reliability than the non-
calibrated group. One important finding of the study was that the group
calibrated by the expert actually decreased in interexaminer reliability after
training. Based on these results, they concluded that the method used for
calibrating examiners may be of more importance than the experience of the
calibrator when planning a training session.[30]

Haj-Ali and Feil conducted a study of interexaminer reliability associated
with a three-point rating scale assessment of Class II amalgam preparations over
short and long term periods. They evaluated the immediate effects of calibration
on interexaminer reliability as compared with a gold standard and determined if
the effects could be sustained over a ten-week period. The examiners graded the
samples at three different occasions: prior to calibration training, immediately

following training, and ten weeks later. They concluded that, with training, interexaminer reliability with a gold standard can be improved and such improvement is reasonably resistant to deterioration after ten weeks.[23]

Salvendy et al, in their article from 1973, commented that "A dental student can acquire his skills most effectively when he is provided with and evaluated on objective and qualitative criteria measures".[31] The criterion oriented method, by definition, reduces the subjectivity associated with grading. If the description of the criteria leaves no room for subjective interpretation, then it only enhances the communication between the student and the examiner.

Review of salient literature to date lends support to the notion that assessment methods that utilize CLSC afford the best possibility of achieving reasonable levels of faculty calibration. However, use of CLSC alone does not ensure calibration among course faculty. Methods of communicating the meaning and application of each criterion must be identified. Incorporation of standardized measurement instruments where ever possible becomes essential. The Department of Operative Dentistry at the University of North Carolina uses CLSC to support the teaching of principles inherent to the disciplines of operative dentistry. The current research studied the level of calibration among faculty in a pre-clinical operative dentistry course by identifying the level of inter- and intraexaminer reliability while assessing each of 13 components of a Class II cavity preparation for amalgam. Preparation components that revealed low interexaminer reliability were identified. An instrument for objective component measurement as well as visual discrimination exercises were

developed and tested for their ability to1) initially aid in increased faculty

calibration while assessing the Class II preparation and 2) provide sustained

calibration among the same faculty over a period of approximately 3 months.

REFERENCES

1. Fuller JL. The effects of training and criterion models on interjudge reliability. *Journal of dental education.* Apr 1972;36(4):19-22.

2. Jenkins SM, Dummer PM, Gilmour AS, Edmunds DH, Hicks R, Ash P. Evaluating undergraduate preclinical operative skill; use of a glance and grade marking system. *Journal of dentistry.* Nov 1998;26(8):679-684.

3. Mackenzie RS. Defining clinical competence in terms of quality, quantity, and need for performance criteria. *Journal of dental education.* Sep 1973;37(9):37-44.

4. Dhuru VB, Rypel TS, Johnston WM. Criterion-oriented grading system for preclinical operative dentistry laboratory course. *Journal of dental education.* Sep 1978;42(9):528-531.

5. Brown G, Manogue M, Martin M. The validity and reliability of an OSCE in dentistry. *European journal of dental education : official journal of the Association for Dental Education in Europe.* Aug 1999;3(3):117-125.

6. Vanek HG. Objective evaluation of dental student technic products. *Journal of dental education.* Mar 1969;33(1):140-144.

7. Houpt MI, Kress G. Accuracy of measurement of clinical performance in dentistry. *Journal of dental education.* Jul 1973;37(7):34-46.

8. Goepferd SJ, Kerber PE. A comparison of two methods for evaluating primary class II cavity preparations. *Journal of dental education.* Sep 1980;44(9):537-542.

9. Vann WF, Machen JB, Hounshell PB. Effects of criteria and checklists on reliability in preclinical evaluation. *Journal of dental education.* Oct 1983;47(10):671-675.

10. Hinkelman KW, Long NK. Method for decreasing subjective evaluation in preclinical restorative dentistry. *Journal of dental education.* Sep 1973;37(9):13-18.

11. Deranleau NJ, Feiker JH, Beck M. Effect of percentage cut-off scores and scale point variation on preclinical project evaluation. *Journal of dental education.* Oct 1983;47(10):650-655.

12. Ryge G, Snyder M. Evaluating the clinical quality of restorations. *Journal of the American Dental Association.* Aug 1973;87(2):369-377.

13. Mackenzie RS, Antonson DE, Weldy PL, Welsch BB, Simpson WJ. Analysis of disagreement in the evaluation of clinical products. *Journal of dental education.* May 1982;46(5):284-289.

14.   Sharaf AA, AbdelAziz AM, El Meligy OA. Intra- and inter-examiner variability in evaluating preclinical pediatric dentistry operative procedures. *Journal of dental education.* Apr 2007;71(4):540-544.

15.   Schiff AJ, Salvendy G, Root CM, Ferguson GW, Cunningham PR. Objective evaluation of quality in cavity preparations. *Journal of dental education.* Feb 1975;39(2):92-96.

16.   Cardoso JA, Barbosa C, Fernandes S, Silva CL, Pinho A. Reducing subjectivity in the evaluation of pre-clinical dental preparations for fixed prosthodontics using the Kavo PrepAssistant. *European journal of dental education : official journal of the Association for Dental Education in Europe.* Aug 2006;10(3):149-156.

17.   Long AF, Mercer PE, Stephens CD, Grigg P. The evaluation of three computer-assisted learning packages for general dental practitioners. *British dental journal.* Dec 10-24 1994;177(11-12):410-415.

18.   Pollard DJ, Davenport JC. An evaluation of training general dental practitioners in partial denture design using a computer-assisted learning program. *British dental journal.* Dec 10-24 1994;177(11-12):405-409.

19.   Grigg P, Stephens CD. Computer-assisted learning in dentistry. A view from the UK. *Journal of dentistry.* Jul-Aug 1998;26(5-6):387-395.

20.   Buchanan JA. Experience with virtual reality-based technology in teaching restorative dental procedures. *Journal of dental education.* Dec 2004;68(12):1258-1265.

21.   LeBlanc VR, Urbankova A, Hadavi F, Lichtenthal RM. A preliminary study in using virtual reality to train dental students. *Journal of dental education.* Mar 2004;68(3):378-383.

22.   Welk A, Maggio MP, Simon JF, et al. Computer-assisted learning and simulation lab with 40 DentSim units. *International journal of computerized dentistry.* 2008;11(1):17-40.

23.   Haj-Ali R, Feil P. Rater reliability: short- and long-term effects of calibration training. *Journal of dental education.* Apr 2006;70(4):428-433.

24.   Robertello FJ, Pink FE. The effect of a training program on the reliability of examiners evaluating amalgam restorations. *Operative dentistry.* Mar-Apr 1997;22(2):57-65.

25.   Natkin E, Guild RE. Evaluation of preclinical laboratory performance: a systematic study. *Journal of dental education.* Jun 1967;31(2):152-161.

26.   Patridge MI, Mast TA. Dental clinical evaluation: a review of the research. *Journal of dental education.* Jun 1978;42(6):300-305.

27.   Polyzois I, McLouglin J, Kelly A, Claffey N. Clinical teaching in restorative dentistry and the variation between students' and supervisors' perceptions of its effectiveness.

European journal of dental education : official journal of the Association for Dental Education in Europe. May 2010;14(2):92-98.

28.   Lilley JD, ten Bruggen Cate HJ, Holloway PJ, Holt JK, Start KB. Reliability of practical tests in operative dentistry. *British dental journal.* Sep 3 1968;125(5):194-197.

29.   Feil PH. An analysis of the reliability of a laboratory evaluation system. *Journal of dental education.* Aug 1982;46(8):489-494.

30.   Scruggs RR, Daniel SJ, Larkin A, Stoltz RF. Effects of specific criteria and calibration on examiner reliability. *Journal of dental hygiene : JDH / American Dental Hygienists' Association.* Mar-Apr 1989;63(3):125-129.

31.   Salvendy G, Hinton WM, Ferguson GW, Cunningham PR. Pilot study on criteria in cavity preparation--facts or artifacts? *Journal of dental education.* Nov 1973;37(11):27-31.

MANUSCRIPT

1. Introduction

Operative dentistry concepts and techniques are initially introduced to first year dental students through participation in a preclinical operative dentistry course. The course contains a didactic portion during which students learn the theoretical aspects of operative dentistry. The students rely on multiple faculty for application, reinforcement and enhancement of theoretical principles during a simultaneous laboratory portion of the course. Faculty members teaching the course are expected to provide consistent formative and summative feedback of the student performance. Low agreement, with regard to assessment of student performance, among faculty may lead to student confusion and frustration. Attempts to increase faculty agreement through improved communication of specific performance criteria, rating scales, and/or training have met with inconsistent results.[1] Few studies have carefully analyzed each of the components of the evaluation system being used so as to identify specific areas of low agreement and then taken 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 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 the 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 the inconsistent grading methods, differing rating scales and individual teaching philosophy[5-9].

Operative Dentistry procedures accomplished in the clinical setting are generally assessed as either clinically acceptable or clinically unacceptable. However, the pre-clinical training of these procedures requires that they be further subdivided into individual components/steps so as to aid the student during the learning process. This allows assessment of conceptual understanding as well as the ability to implement each component of the procedure. Overall procedural competence may then be assessed as a net sum of all individual components.

A pilot study evaluating the level of faculty calibration, which occurred while assessing 13 components of first year dental student Class II amalgam preparations, was conducted at the University of North Carolina (UNC) in 2011 Each component of the Class II preparation had a set of specific criteria which defined clinically acceptable and clinically unacceptable levels of procedure accomplishment.. The student assessment form used for the study is shown in Table 1.

It is generally accepted that levels of agreement should minimally exceed that which would happen by chance (50%) alone. Therefore, for the purpose of this study, 60% was arbitrarily set as the minimum level of agreement and average percentage agreement that fell below 60% was considered poor. The average percentage agreement (mean and standard deviation with confidence intervals) for each of the 13 components assessed during the Pilot study are listed in Table 2.

Analysis of the assessments from the pilot study revealed six out of thirteen cavity preparation components had interexaminer reliability below 60% (Figure 1). The six components of Class II cavity preparation that needed further faculty calibration were as follows: 1) Proximal and Gingival Contact Clearance, 2) Retention Groove Placement, 3) Retention Groove Depth, 4) Preparation Margin Finish, 5) Preparation Margin Orientation and 6) Preparation Toilet. These various preparation features are described in further detail in materials and methods.

Evaluation of the pilot study suggested levels of inherent faculty calibration which were unacceptable for at least 6 of the 13 components of one Class II procedure. First year dental students learn multiple procedures during their pre-clinical courses. Therefore, it was deemed appropriate that further steps should be undertaken so as to identify areas of poor faculty calibration and target these areas with strategies designed to enhance agreement relative to the assessment of student performance.

2. **Specific Aims:**

    1. To identify the level of interexaminer reliability among pre-clinical operative dentistry faculty when assessing 32 (thirty two) Class II preparations performed by 1st year dental students.

    2. To develop targeted exercises designed to enhance the ability of faculty to discriminate among levels of student performance (discrimination exercises) of various procedural components where average agreement among faculty was found to be below 60% and to organize and present these discrimination exercises to individual faculty members as part of a calibration session.

    3. To evaluate the ability of discrimination exercises (as revealed by inter- and intraexaminer reliability) to increase initial levels of faculty calibration in assessment of the components of Class II preparation performed by first year dental students accomplished during their pre-clinical operative dentistry course (referred to as Phase I).

    4. To evaluate the ability of discrimination exercises (as revealed by inter- and intraexaminer reliability) to sustain an increase in levels of faculty calibration, in assessment of the components of Class II preparation performed by first year dental students accomplished during their pre-clinical operative dentistry course, over a time interval of at least 6 months (referred to as Phase II).

3. Materials and methods

This was a longitudinal, non-randomized cohort study conducted during the period of 2011 – 2013 at the University of North Carolina School of Dentistry. This study was IRB exempt (#12-0262) by the University of North Carolina's Institutional Review Board.

3.1 Dentiform teeth

According to the Webster's dictionary, the word 'Dentiform' is defined as having the shape of tooth or teeth. The dentiform teeth used for the current study were a model of tooth #30 with MOD caries, model # A27A-46U, Kilgore® International Inc. The dentiform teeth had been fabricated with a composite material crown, epoxy resin dentin contained simulated dentin caries. The enamel and dentin anatomy was morphologically similar to a natural mandibular first molar and the placement of the simulated caries was similar to that found in the Class II caries disease state.

3.2 Panel of Examiners

The panel of examiners consisted of 8 examiners of which, five were full time faculty members, one was part time faculty and two were graduate students in the Department of Operative Dentistry at the UNC School of Dentistry who were currently teaching principles of operative dentistry procedures, including Class II preparation design for amalgam restoration to first year dental students. The faculty had varied levels of teaching and clinical experience. The author conducted an individual calibration session with each examiner.

3.3 Selection of dentiform teeth

Thirty two 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. The same 32 Class II preparations were assessed by the examiners as part of the Pilot, Phase I and Phase II parts of the study. The cavity preparations were completed by first year dental students as part of the pre-clinical operative dentistry course. The students were instructed to prepare an ideal (according to specific criteria) MOD cavity preparation for amalgam restoration, with complete removal of the simulated caries lesion. The preparations were accomplished in the simulation laboratory designed to replicate the clinical setting. The dentiform tooth #30 was placed in the dentiform with adjacent teeth (tooth # 29 and # 31) forming proximal contact points, during the preparation. The criteria for cavity preparation were adopted from Sturdevant's Art and Science of Operative Dentistry, 5th Edition.[10]

3.4   Calibration Session

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 as having low faculty calibration. The discrimination exercises included 3D demonstration models of actual dentiform teeth with various levels of student performance of preparation components as well as and digital images of component performance organized in the form of a Microsoft PowerPoint presentation. At the end of the PowerPoint presentation, a detailed discussion was conducted with the faculty member regarding the

specific criteria outlined for each component of Class II cavity preparation as it appeared in the evaluation form.

3.5   Discrimination Exercise Rationale

Visual and tactile exercises were designed in an effort to enhance identification of ideal performance of individual procedural component and discrimination of variations from the ideal.

3.6   Discrimination Exercise Design

Discrimination exercises were designed for 5 out of 6 components that the pilot study had identified as having poor interexaminer reliability The o5 components were 1) Proximal and Gingival Contact Clearance, 2) Retention Groove Placement, 3) Retention Groove Depth, 4) Preparation Walls and 5) Preparation Margin Orientation. A discussion of the specific criteria for the component 'Preparation Toilet' was completed but no other discrimination exercise was developed for this component. The description and design of the five discrimination exercises are as follows:

3.6.1. Proximal Contact Clearance

A brief review of the desired final location of the facial and lingual walls of the proximal box was provided to each faculty member. A demonstration of how proximal contact clearance is measured at the proximal height of contour was provided. Review of the specific criteria defining levels of component performance was accomplished. Discussion included how the ideal proximal clearance is defined as visibly open and/or is open up to but not exceeding 0.75

mm on either one or both the proximal walls as well as the specifics of student performance that varied from the ideal and how this was to be assessed.[10]

In an attempt to find an objective means of standardizing the assessment of proximal clearance, digital images (Nikon D3100 camera, Nikkor 105mm Lens, Sigma EM-140 DG flash) of 82 Premier UNC 15 periodontal probes were taken. The mean diameter of the probes (mm) was assessed using Image J® (National Institutes of Health, Bethesda, Md) software.

The mean diameter of the UNC 15 probe was found to be 0.5 ($\pm$0.02)mm at the 4-5 mm mark and 0.75 ($\pm$0.02)mm at the 11-12 mm mark (Figure 2.1). The UNC-15 periodontal probe was therefore adopted as a standardized measurement device for the purpose of assessing proximal contact clearance. The discrimination exercise included a tactile demonstration of how the proximal contact clearance is be assessed as ideal , if the diameter of the periodontal probe at the 4/5mm mark (up to 0.5 mm) cannot pass between the proximal surfaces, acceptable if the clearance between the proximal surfaces was between the 5 mm mark and the 11/12 mm mark (between 0.5 mm and 0.75 mm) and/or clinically unacceptable, if the clearance between the proximal surfaces is greater than the diameter of the periodontal probe at the 11/12mm mark (> 0.75 mm) between the proximal contact clearances (Figures 2.2 and 2.3).

3.6.2. Retention Groove Placement

Retention grooves are considered secondary retentive features that may be included in preparation design for retention of anticipated restoration in the proximal areas of the preparation. The ideal placement of retention grooves

should theoretically be placed approximately 0.2mm internal to the dentinoenamel junction (DEJ) on the facial and lingual proximal preparation walls as this would ensure sufficient amount of supported enamel and avoid injury to the pulp.[10]

Students may incorrectly place retention grooves. Therefore the discrimination exercise included a series of Class II preparations with different axial wall depths that represented the various types of errors found in student preparations. The dentiform teeth were sectioned sagitally so as to allow visualization of the gingival and facial external walls, the axial and pulpal internal walls and various positions of retention grooves (Figures 3.1 - 3.7). The preparations were as follows:

Preparation#1: Ideal axial wall depth (0.5mm inside the DEJ) with a retention groove in the enamel facial wall indicating a clinically unacceptable position (Figure 3.1).

Preparation#2: Ideal axial wall depth (0.5mm inside the DEJ) with a retention groove placed at the DEJ indicating a clinically unacceptable position (Figure 3.2).

Preparation#3: Ideal axial wall depth (0.5mm inside the DEJ) with retention groove placed ~0.2mm internal to the DEJ indicating a clinically acceptable position (Figure 3.3).

Preparation#4: Ideal axial wall depth (0.5mm inside the DEJ) with retention groove placed in the axial wall indicating clinically an unacceptable position (Figure 3.4).

Preparation#5: Ideal axial wall depth (0.5mm inside the DEJ) with retention groove placed in the gingival floor indicating a clinically unacceptable position (Figure 3.5).

Preparation#6: Deep axial wall (~2mm inside the DEJ) with retention groove placed 0.2mm internal to the DEJ indicating a clinically acceptable position (Figure 3.6).

Preparation#7: Deep axial wall (2mm inside the DEJ) with retention groove placed at the axial and facial wall line angle indicating a clinically unacceptable position (Figure 3.7).

3.6.3. Retention Groove Depth:

Students may create a retention groove that has incorrect depth. The groove should ideally be 0.1-0.5 mm deep to ensure adequate retention.[10]

A discrimination exercise provided examiners with an explorer and three sagitally sectioned dentiform teeth, with Class II preparations, that contained retention grooves that were deep (> 0.5 mm in depth), ideal (0.1 – 0.5 mm in depth) and shallow (< 0.1 mm in depth, undetectable). Deep or shallow retention grooves were deemed unacceptable. No digital images were used to enhance discrimination of various levels of performance of this component.

3.6.4. Preparation Finish

Students may fail to pay close attention to subtle detail as they create cavity preparations. Unsupported enamel associated with cavosurface margins, rough preparation walls and sharp internal line angles may compromise the placement and longevity of the restoration. Preparation Finish   involves

smoothing of the external walls to eliminate any unsupported enamel and rounding the internal line angles to avoid stress concentration.[10]

A discrimination exercise provided the examiners with four Class II cavity preparations that contained the various combinations of student performance of preparation wall finish. Each combination was identified as clinically acceptable or unacceptable. The preparation finish combinations were as follows:

1) Preparation finish resulting in smooth walls and gentle transitions - clinically acceptable (Figures 4.1 & 4.2).

2) Preparation finish resulting in smooth walls and abrupt transitions - clinically unacceptable (Figures 4.3 & 4.4).

3) Preparation finish resulting in rough walls and abrupt transitions - clinically unacceptable (Figures 4.5 & 4.6).

4) Preparation finish resulting in rough walls and gentle transitions - clinically unacceptable (Figures 4.7 & 4.8).

3.6.5. Preparation Margins:

Students may create Class II cavity preparations with cavosurface margin angles that are <80 degrees (°), are 90° or are >100°. The cavosurface angle is the angle of the tooth structure formed by the junction of a prepared (cut) wall and the external surface of the tooth. The actual junction is referred to as cavosurface margin. Enamel cavosurface margins of Class II cavity preparations require dentin support. External preparation walls that are aligned parallel to

average enamel rod orientation will result in cavosurface margin angles that are approximately 90°.[10]

Proximal external walls that are oriented at 90° to a tangent aligned with the external curvature of the tooth are considered to have enamel that is supported by enamel. A 90° proximal wall orientation is considered clinically acceptable. Proximal external walls that are oriented < 80° to a tangent aligned with the external curvature of the tooth are considered to have enamel that is not supported by dentin. A proximal wall orientation that is < 80° is considered clinically unacceptable. Proximal external walls that are oriented > 100° to a tangent aligned with the external curvature of the tooth will result in an amalgam restoration margin that is fragile. A proximal wall orientation that is > 100° is considered clinically unacceptable.

Discrimination exercises consisted of digital images of clinically acceptable and unacceptable proximal wall orientations. A transparent protractor was superimposed over the images to enable more objective identification of wall orientations that were < 80°, were approximately 90° or were >100°. (Figures 5.1, 5.2 and 5.3). Examples of clinically acceptable and unacceptable cavosurface margin orientations, including ideal cavosurface margins (90°) were presented and discussed during this discrimination exercise.

3.7 Review of the Assessment Rubric

Following the Microsoft PowerPoint presentation, 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 completed. Examiners were provided opportunity to ask questions throughout the calibration session.

3.8 Timing of the Phase I and Phase II Assessments

3.8.1 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 approximately 2-3 hours and was accomplished in one sitting.

3.8.2 Phase II Assessment

The examiners assessed the same 32 Class II preparations after an average time interval of 6 months. The assessment was conducted using the same controlled settings as the Pilot and Phase I studies. However, 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.

3.9 Organization of the Data

The Class II preparation assessment point values were transferred by the primary investigator from the assessment rubrics to a digital file (Table 3). An independent investigator coded the names of the examiners using alphabets letters from 'A' to 'H' so that the primary investigator was blinded to examiner identity. The data was then subjected to statistical analysis.

4. Statistical Analysis

The concordance and discordance between each pair of examiners were analyzed using Weighted Kappa and McNemar analysis. The interexaminer reliability was as average percentage agreement among the 8 examiners for the

Pilot study, Phase I and Phase II. The confidence interval (CI) was calculated for each component of the cavity preparation for all 3 assessment sessions.

5. Results

5.1 Interexaminer Reliability- Phase I

The interexaminer reliability, reported as average percentage agreement, among the examiners increased for 7 out of the 13 components when compared to the results of the pilot study. However, for 3 components (Adjacent Tooth Damage, Unsupported Enamel and Primary/Axial Wall Depth) the average percentage agreement did not change and for another 3 components (Occlusal Convergence, Retention Groove Placement and Retention Groove Depth) there was a decline in average percentage agreement (Figure 6).

Assessment of some components targeted with discrimination exercises showed increase levels of faculty calibration whereas assessment of other components did not. The average percentage agreement increased for Proximal Contact Clearance from 52% before the calibration session to 70% after calibration session with 95% CI at [(68.79, 80.54)]. The average interexaminer agreement for Retention Groove Placement decreased 58 to 53% [95% CI (46.15, 59.97)] and for Retention Groove Depth decreased from 57 to 56% [95% CI (51.03, 61.03)]. The average interexaminer agreement for Preparation Wall Finish increased from 59 to 65% [95% CI (54.16, 68.62)],, for Preparation Margin Orientation increased from 54 to 63% [95% CI (58.13, 67.32)] and for Preparation Debris increased from 58 to 68% [95% CI (63.80, 72.81)]. The

confidence interval for each of the 13 components after the calibration session is reported in Table 4.

5.2 Intraexaminer Reliability– Pilot Study to Phase I

The average intraexaminer agreement among the course faculty was 74 (+/-5) %. The intraexaminer reliability (agreement of each examiner with themselves) was calculated as average intraexaminer agreement (%) when comparing Pilot study to Phase I. The results are listed in Table 5. Assessment of the 13 preparation components varied greatly from Pilot to Phase I. The component 'Enamel Present' had the lowest intraexaminer variation and the component 'Proximal Contact Clearance' had the highest intraexaminer variation (Figures 7 and 8). Table 6 shows the level of intra-examiner variation for each examiner for all 13 components between the Pilot study and Phase I.

5.3 Interexaminer Reliability - Phase II

The interexaminer agreement began to decline for 8 out of the 13 components after an average time interval of 6 months. For 3 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 percentage agreement among the examiners stayed the same for one component (Occlusal Convergence). However, for 2 components (Isthmus Width and Preparation Toilet) the inter examiner reliability continued to increase compared to Pilot study and Phase I (Figure 9).

The results of the components for which discrimination exercises were designed are as follows: average interexaminer agreement decreased for

41

'Proximal Contact Clearance' from 70% after calibration exercises to 60% [95% CI (53.38, 67.17)] after an average time interval of 6-7 months. The average interexaminer agreement decreased from 58 to 53% [95% CI (46.15, 59.97)] for 'Retention Groove Placement' and 57 to 56% [95% CI (51.03, 61.03)] for 'Retention Groove Depth'. The average interexaminer agreement increased from 59 to 64% [95% CI (54.16, 68.62)] for 'Preparation Wall Finish', 54 to 63% [95% CI (58.13, 67.32)] for 'Preparation Margin Orientation' and 58 to 68% [95% CI (63.80, 72.81)] for 'Preparation Debris'. The average percentage agreement and confidence interval for assessment of each component of the cavity preparation during Phase II is reported in Table 7.

5.4 Intraexaminer Reliability - Phase II

The average intraexaminer agreement among the course faculty was 77 (+/-7) % when comparing Pilot study to Phase II and 76(+/-8) % when comparing Phase I to Phase II. The intraexaminer reliability was calculated for Pilot Study: Phase I and Phase I: Phase II comparisons (Tables 8 & 9). While some examiners remained consistent in their assessment patterns from Phase I to Phase II, the others reverted back to assessing with a variation of 30-40% for a few components. Assessment of 'Isthmus Width' had the lowest intraexaminer variation and assessment of 'Proximal Contact Clearance' had the highest intraexaminer variation (Figures 10 and 11). Comparison of the intraexaminer variation that occurred while assessing all 13 components when comparing Phase I: Phase II and the Pilot Study: Phase II and Phase I: Phase II respectively.

The level of intraexaminer variation (%) for all 13 components of Class II cavity preparation for each examiner when comparing Pilot study to Phase II is shown in Table 10 and the level of intraexaminer variation (%) for all 13 components of Class II cavity preparation for each examiner when comparing Phase I to Phase II is shown in Table 11.

6. Discussion:

It has been well documented that improving the level of agreement among faculty members is not an easy task. [5,7,11]. The over-arching goal of the current study was to determine the interexaminer and intraexaminer reliability while assessing 13 components of a preclinical operative procedure completed by first year dental students and to seek to increase faculty agreement in areas where it was low. The results of the pilot study confirmed that there were areas of low interexaminer agreement among the faculty. As part of the study design, specific exercises were developed and presented in a calibration session so as to increase the ability of faculty to discriminate among various levels of student performance. The efficiency 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 stated by Viera and Garett, 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[12]. Similarly for the current study, if we look at the component Proximal contact clearance a disagreement between 'No Clearance' and 'Open up to 0.5mm in all directions' is not as severe as one 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 clearance 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 established that individual perceptual abilities vary widely. This study highlighted that some dentists and many dental students, particularly early in their course, have great difficulty in accurately gauging depths and distances.[13] Course faculty must ensure students receive objective, unbiased assessment of their performance.

The results from Phase I showed that the interexaminer reliability improved with the use of discrimination exercises. As part of the discrimination exercises, the use of UNC 15 periodontal probe was introduced in a novel way so as to more objectively assess proximal contact clearance. The average percentage agreement among the examiners increased from 52% before calibration to 70% after calibration for this component. After an average time interval of 6 months, the percentage of inter examiner agreement decreased to 60% for this component but still remained higher than the Pilot study. It may be

that introduction of a specific instrument to assess this component contributed to the increase in interexaminer reliability, and limited the influence of personal examiner bias associated with the assessment of this component.

There was also an increase in interexaminer reliability for components such as 'Isthmus Width' and 'Preparation Debris' for which no discrimination exercises were designed. The authors speculate that detailed discussions regarding the specific criteria outlined for all 13 components of the Class II preparation may have helped to limit mis-understanding of subjective interpretation of these criteria.

The two components for which there was no increase in reliability among the examiners were retention groove placement and retention groove depth. This was in spite of examiner participation in carefully designed discrimination exercises. Although studies have been done on the significance, ideal position and ideal depth of retention grooves, researchers have not been able to achieve consensus.[11-13]

A survey by David Moore in 1992 investigated the teaching in dental schools regarding proximal retention grooves in Class II cavity preparations for amalgam restoration. He included 64 schools across the US and Canada. The response rate was 92% (59 schools) out of which 61% (36 schools) of the schools responded 'YES' to teaching retention grooves and 39% (23 schools) of the schools responded 'NO' to teaching this technique[14].

The results of the current study also showed increased intraexaminer variations for the components 'Retention groove placement' and 'Retention

groove depth'. These increased variations prompted us to conduct an informal2013 survey of the nation's dental schools to inquire if they regularly taught the placement of retention grooves in conservative Class II cavity preparations for amalgam restoration. The questionnaire was emailed to 65 dental schools across the nation. The response rate was 53 (82%). Of the 53 schools that responded 24 of them replied 'YES' and 29 of them replied 'NO' to the teaching of retention grooves. It may be that variations in the professional philosophy (on the use of retention grooves) of the examiners included in this study limited the ability to increase the level of agreement in assessing these components. It is important to note that both the placement and accurate assessment of retention grooves is difficult.

Although there was a rise 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 when the level of reliability started to decline but only detected that there was a decline at an average time interval of 6 months. This piece of information is valuable and can be translated as a need for frequent calibration sessions throughout the academic year.

A limitation of this study is that it does not evaluate intraexaminer variation based on the clinical and teaching experience of the examiners. Jenkins et al found that the level of pass–fail differences (intraexaminer variability) in their study seemed to be unrelated to the experience of the examiner, with even the senior examiner recording differences of 17%[8]. Another

study reported that it is theoretically possible that examiner variability alone may result in some unfortunate undergraduate student failing on more than one occasion particularly if acknowledged that on a different day, the same cavity, assessed by the same examiner, may well have passed[8]. An excerpt from Mackenzie's article on defining clinical competence in terms of quality, quantity and need for performance criteria, clearly states the problem of intra-examiner variability; "*A dental student reported the following event. A good student and a poor student were sitting next to each other in a basic technics laboratory. The good student finished the 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 the poor student tooth the 'A graded' tooth to the same instructor. The instructor looked at it, said, "Hm-mm, OK," and gave him a 'C'*[15].

Another limitation of this study is that the results may be negatively influenced from examiner fatigue. Examiner fatigue may play a major role in the decline of interexaminer reliability. Having to grade large number of samples at one time may cause the examiner to lose focus. As suggested by Dhuru, Rypel and Johnston[3] future studies should limit the number of preparation samples or have the examiners take frequent breaks after grading 10-15 samples.

The results of the current study indicate that discrimination exercises are beneficial for faculty calibration. Future efforts in the field of faculty calibration should focus on designing instruments to aid in objective evaluation of student performance. The results also highlighted a need for more frequent

calibration sessions. There is a potential benefit in designing calibration sessions as online CE courses, because, they can be used by students and faculty as often as needed for self-analysis.

In summary, the overall interexaminer reliability improved after calibration. The results showed that, the objective use of an instrument (UNC 15 periodontal probe) increased interexaminer reliability. The study supports the use of discrimination exercises for faculty calibration in order to improve the consistency of faculty-student communication. The study findings reveal a decrease in inter and intra-examiner reliability at 6 months. Wide intraexaminer variation was noted from Pilot Study to Phase I. Even though variations were noted in intraexaminer reliability after a time interval of 6 months, the degree of variation was lower than Pilot Study: Phase I.

7. Conclusions:

-   Overall interexaminer reliability improved after calibration.

-   The objective use of an instrument (UNC 15 periodontal probe) increased interexaminer reliability.

-   The study findings reveal a decrease in inter and intra-rater reliability at 6 months. This suggests that having more frequent calibration sessions may be beneficial for maintaining an optimum level of calibration among the course faculty. However this notion was not tested.

TABLE 1 - Class II Amalgam Procedure Performance Assessment Rubric Used For The Pilot Study, Phase I And Phase II

| 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 0.5mm 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 re-contouring
	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 - 1.5 mm internal to DEJ
	2.0 - 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 mm and 1 mm internal to DEJ
	Greater than 1mm internal to DEJ
	Not Visible
Retention Groove Depth	Undetectable
	Between 0.1 mm 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
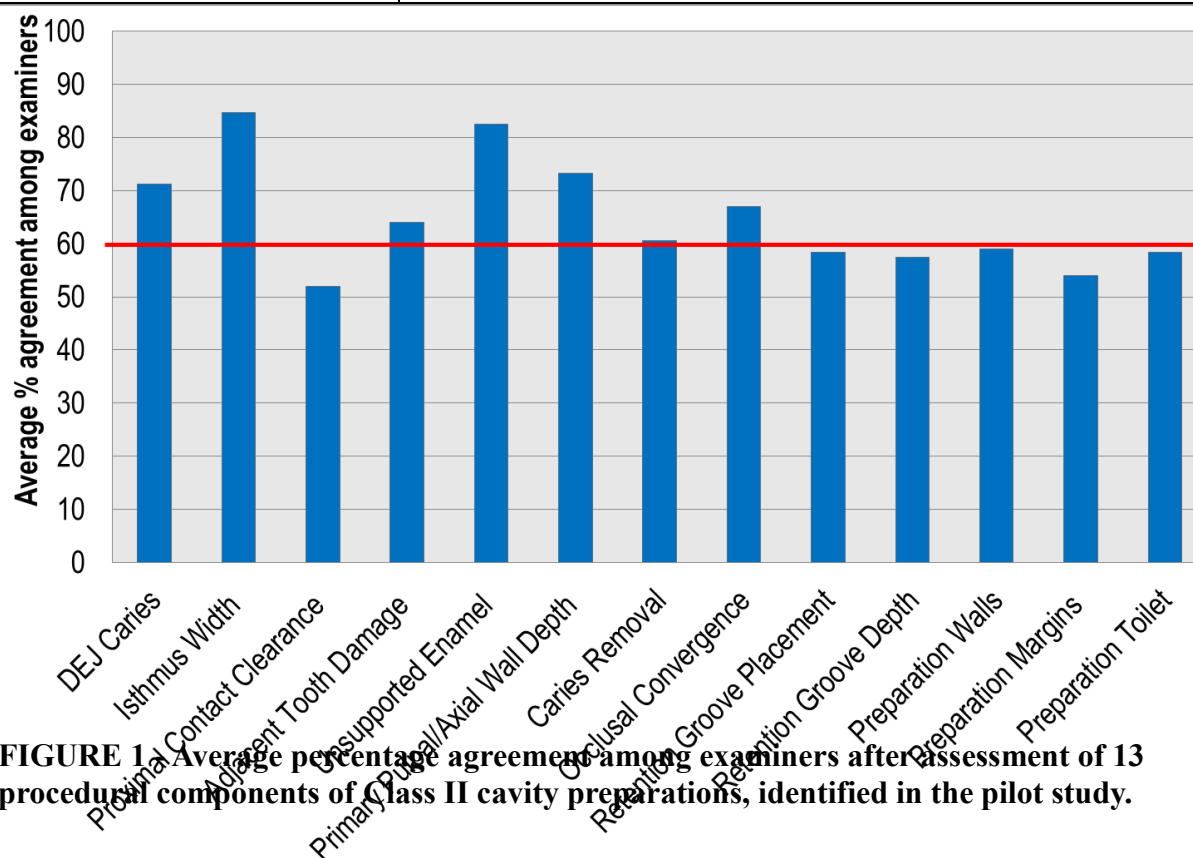
FIGURE 1. Average percentage agreement among examiners after assessment of 13 procedural components of Class II cavity preparations, identified in the pilot study.

TABLE 2 – Average percent agreement (mean +/- sd) and 95% CI for assessment of components of Class II cavity preparation during the Pilot study

COMPONENTS OF CLASS II CAVITY PREPARATION	MEAN	STD DEV	95% CI (LOWER, UPPER)
DEJ Caries	71.15	9.96	(67.29, 75.01)
Isthmus Width	84.73	10.76	(80.56, 88.90)
Proximal Contact Clearance	52.01	25.83	(42.00, 62.03)
Adjacent Tooth Damage	63.95	14.14	(58.47, 69.44)
Enamel Present	82.48	25.10	(72.74, 92.21)
Primary Pulpal Axial Wall	73.22	12.01	(68.56, 77.87)
Caries Removal	60.60	14.92	(52.81, 64.38)
External Walls	66.97	11.66	(62.45,71.49)
Retentive Groove Placement	58.37	9.43	(54.72, 62.03)
Retention Groove Depth	57.37	11.10	(53.07, 61.67)
Preparation Walls	59.01	8.93	(57.04, 63.96)
Preparation Margins	54.02	20.14	(46.21, 61.83)
Preparation Toilet	58.38	15.10	(52.52, 64.23)
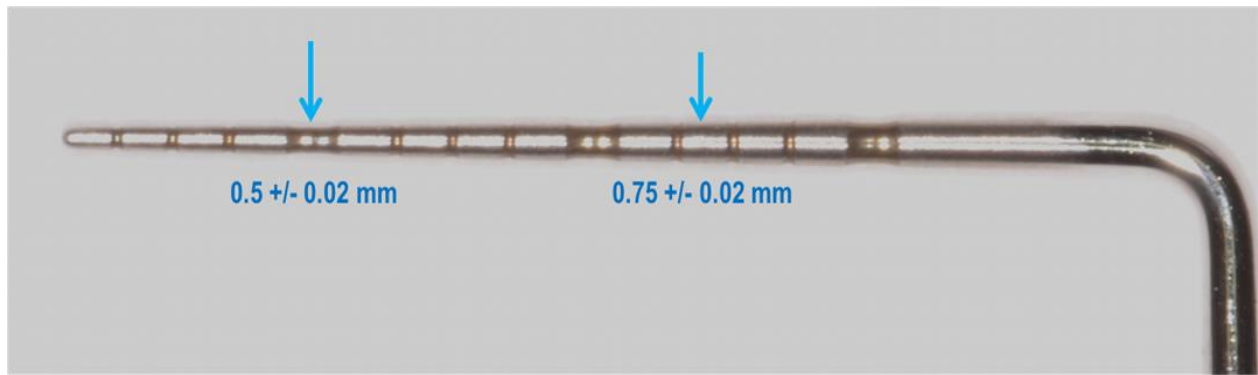
FIGURE 2.1- The mean diameter of the UNC 15 probe at the 4-5 mm mark was 0.5 +/-0.02 and 0.75 +/-0.02 at the 11-12 mm mark. The measurements were made using the Image J ® software.

FIGURE 2.2- 3D model used to demonstrate clinically acceptable example of proximal contact clearance

FIGURE 2.3- 3D model used to demonstrate clinically unacceptable example of proximal contact clearance
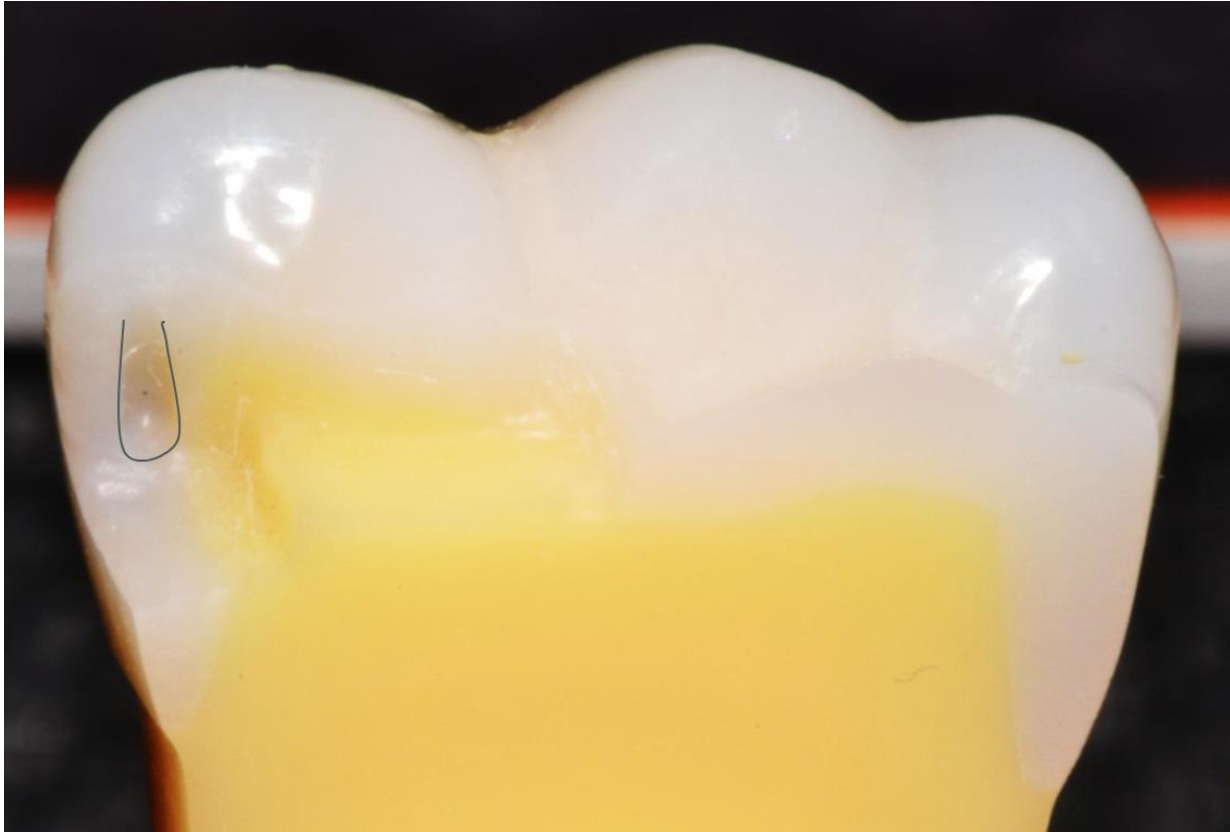
FIGURE 3.1- 3D model demonstrating clinically unacceptable retention groove placemen in the enamel facial wall of a preparation with ideal axial wall depth.

FIGURE 3.2- 3D model demonstrating clinically unacceptable retention groove placement at the DEJ of the facial wall of a preparation with ideal axial wall depth.

FIGURE 3.3- 3D model demonstrating clinically acceptable retention groove placement ~0.2 mm internal to the DEJ such that it is partially in the dentin facial wall and partially in the adjacent axial wall of a preparation with ideal axial wall depth.

FIGURES 3.4- 3D model demonstrating clinically unacceptable retention groove placement in the axial wall of a preparation with ideal axial wall depth.

FIGURES 3.5- 3D model demonstrating clinically unacceptable retention groove placement in the gingival wall of a preparation with ideal axial wall depth .
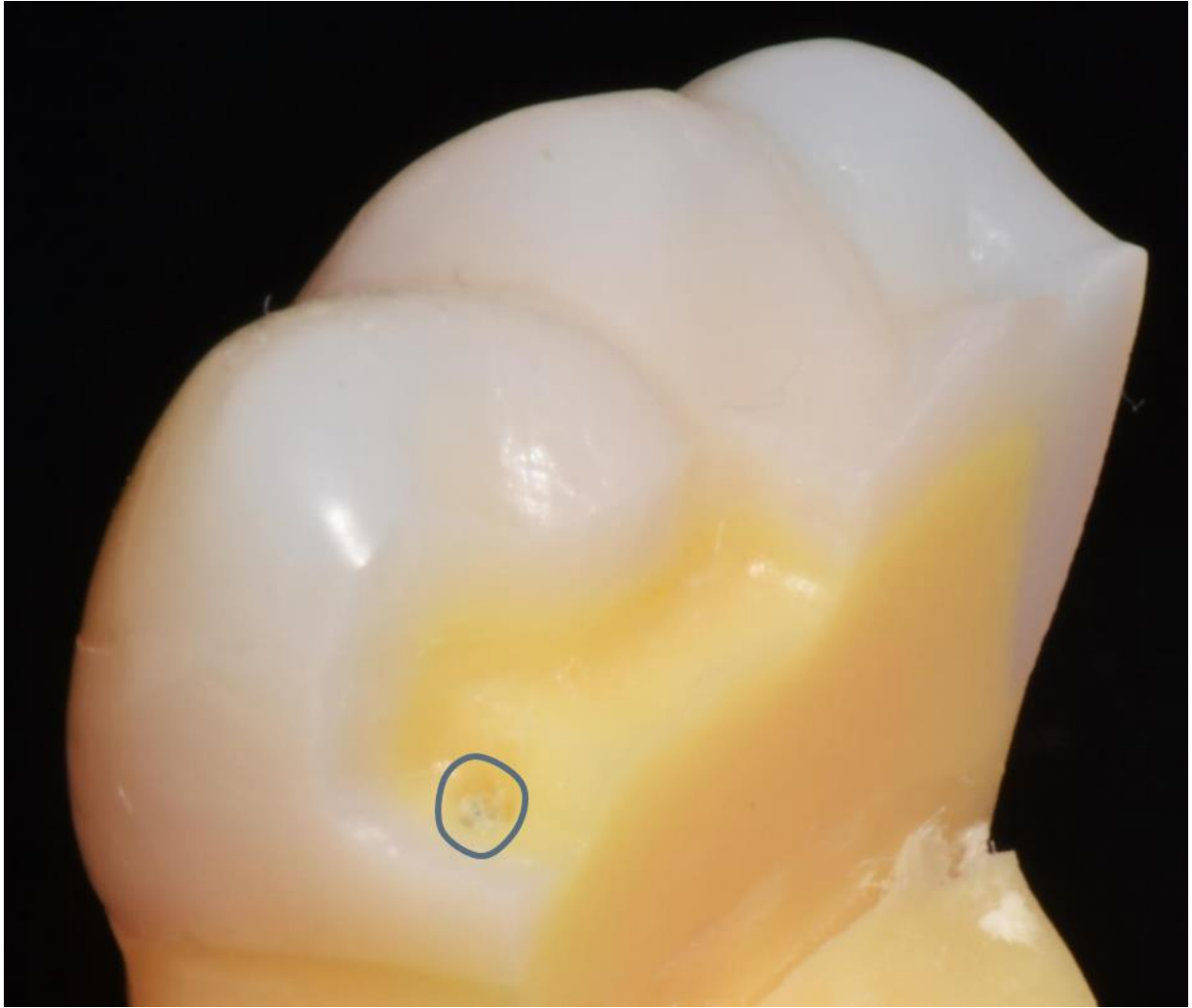
FIGURES 3.6- 3D model demonstrating clinically acceptable retention groove placement ~0.2 mm internal to the DEJ in the facial dentin wall of a preparation with deep axial wall depth.

FIGURES 3.7- 3D model demonstrating clinically unacceptable retention groove placement in the line angle of the facial and axial walls of a preparation with deep axial wall depth.

Figure 4.1
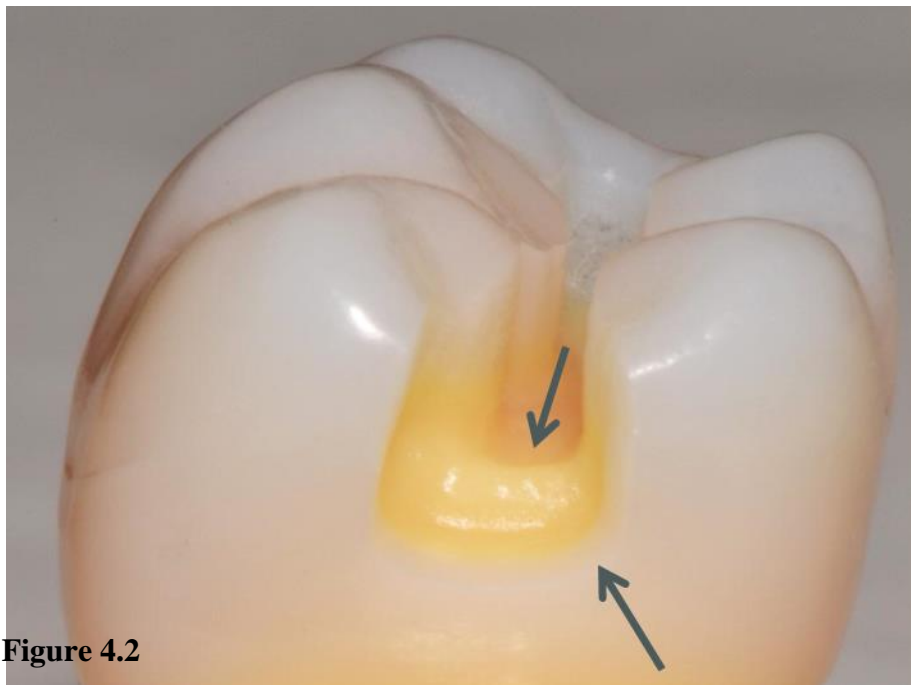
Figure 4.2

FIGURES 4.1 & 4.2 – 3D model demonstrating clinically acceptable preparation finish of smooth walls and gentle transitions.
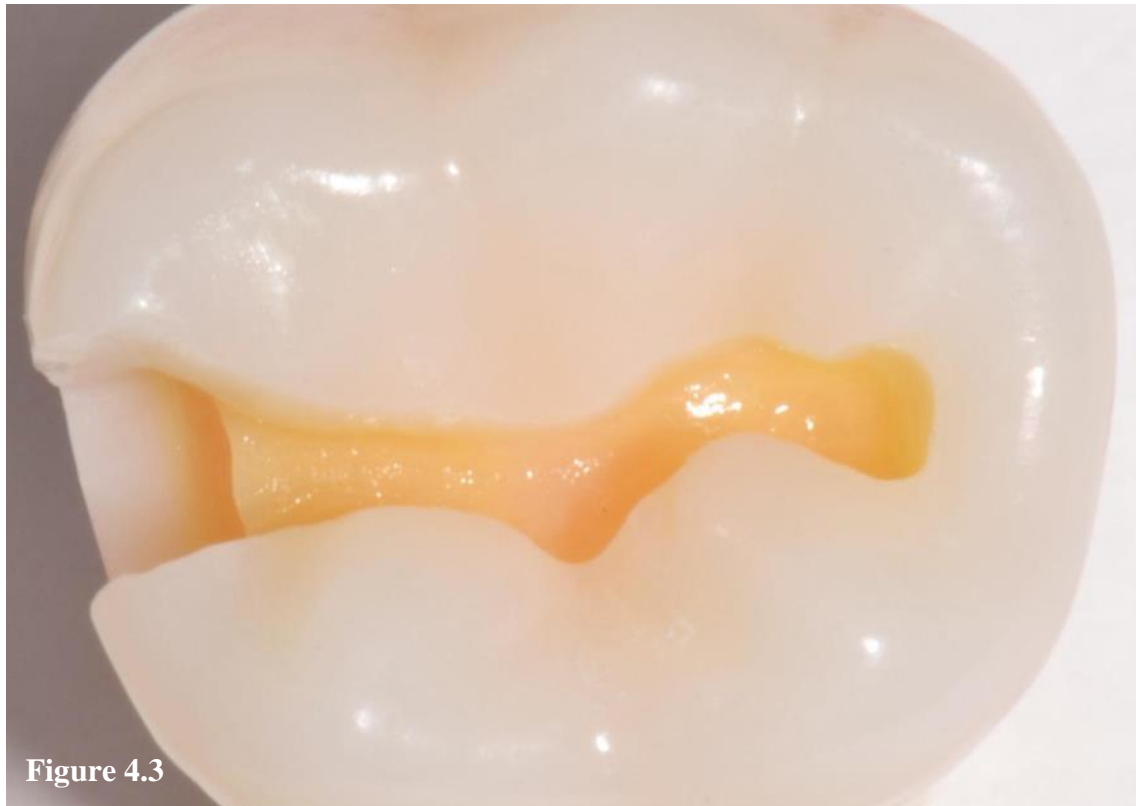
Figure 4.3
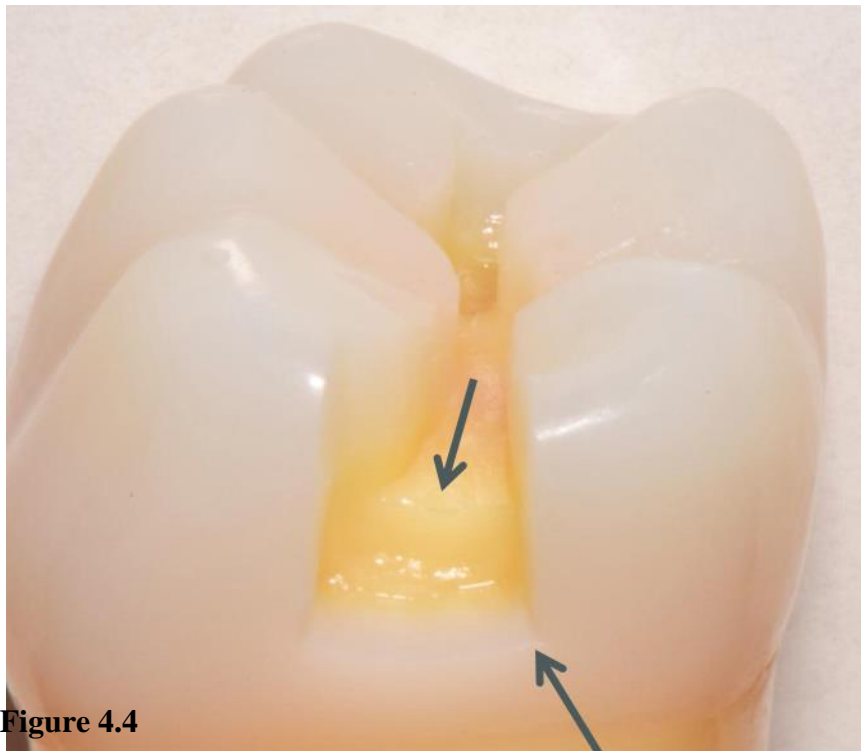
Figure 4.4

FIGURES 4.3 & 4.4 – 3D model demonstrating clinically unacceptable preparation finish of smooth walls and abrupt transitions.
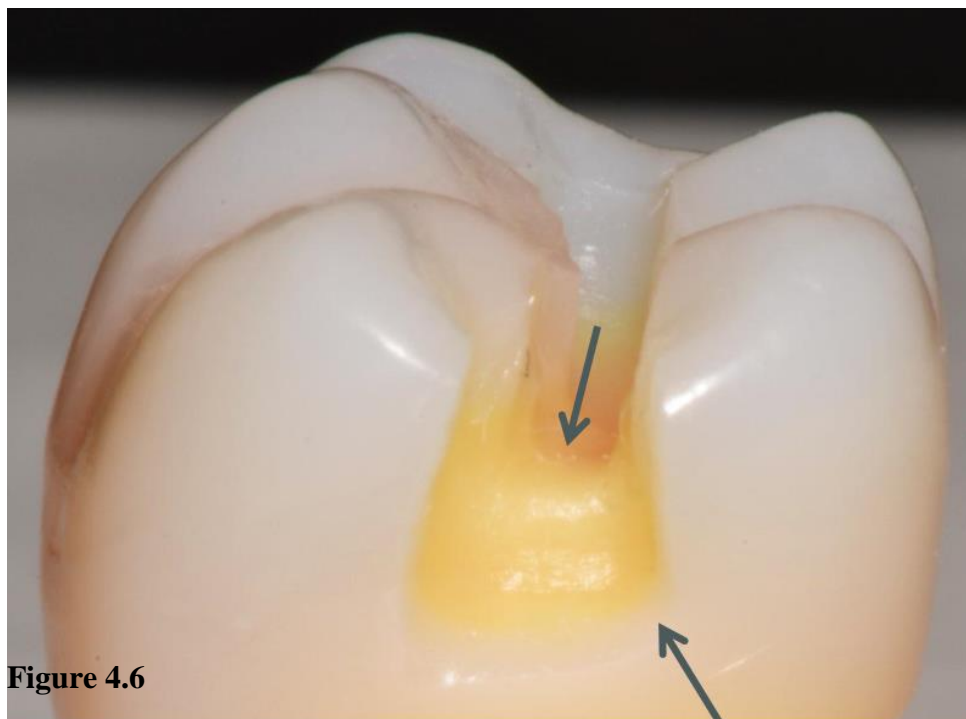
Figure 4.5

Figure 4.6

FIGURES 4.5 & 4.6 – 3D model demonstrating clinically unacceptable preparation finish of rough walls and gentle transitions.

Figure 4.7
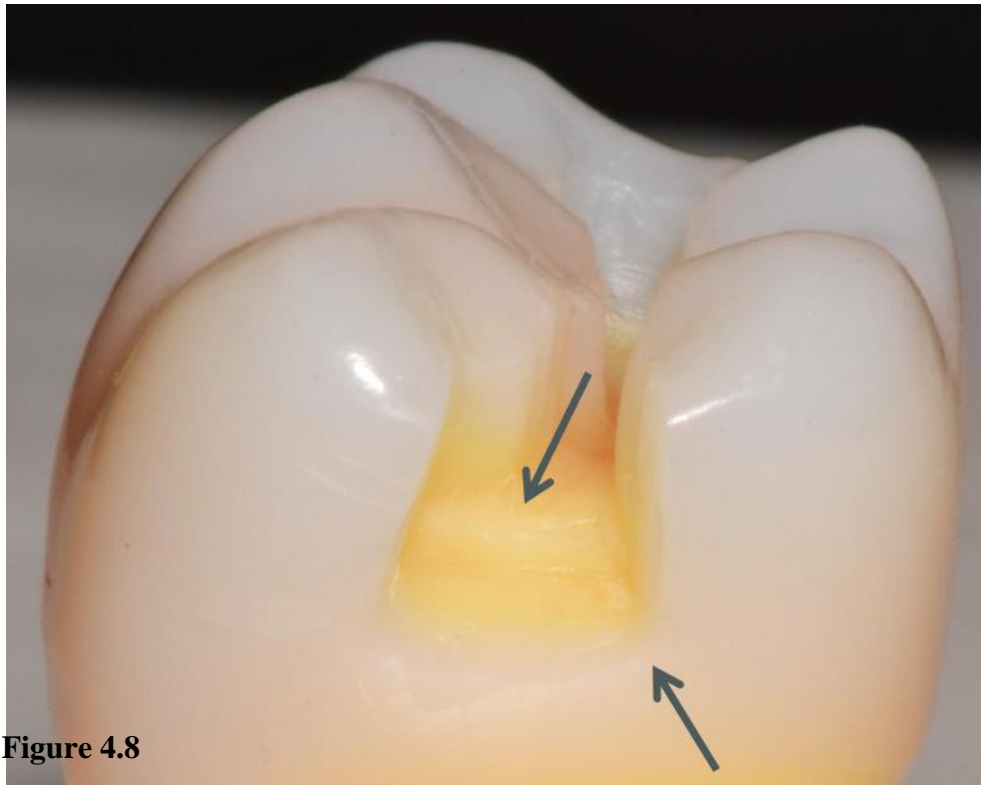
Figure 4.8

FIGURES 4.7 & 4.8 – 3D model demonstrating clinically unacceptable preparation finish of rough walls and abrupt transitions.
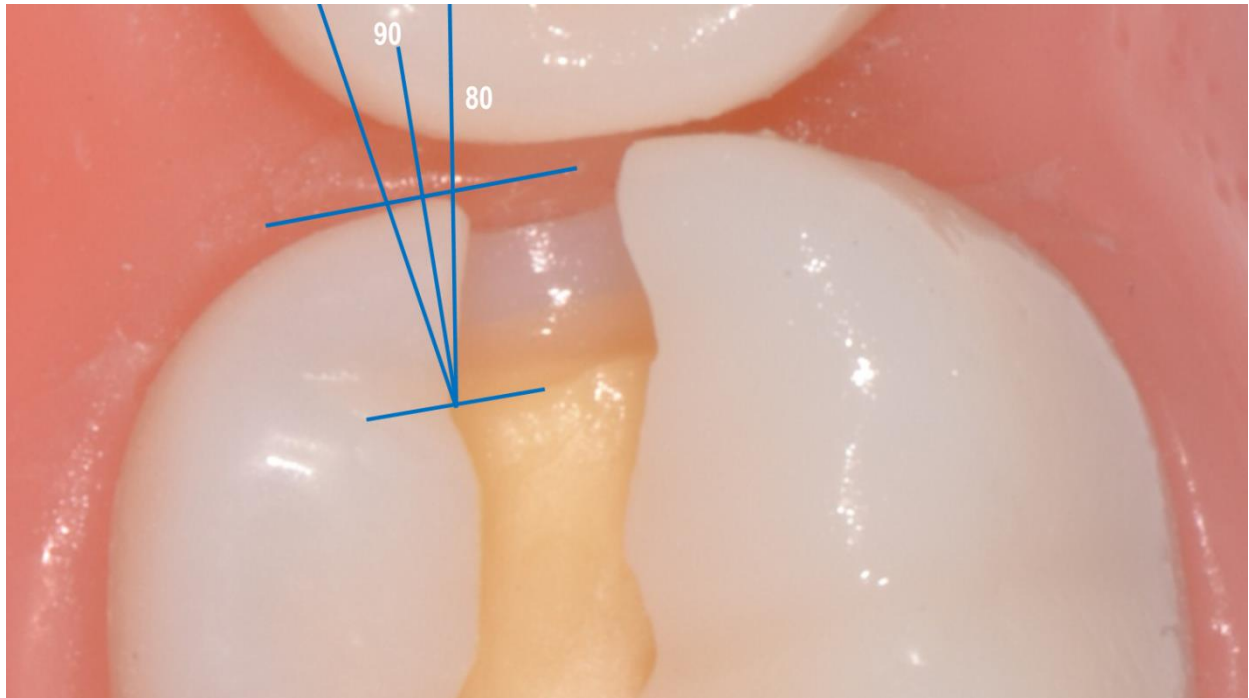
FIGURE 5.1- Image of the 3D model with a superimposed protractor that demonstrates clinically unacceptable preparation wall orientation of $< 80^0$.
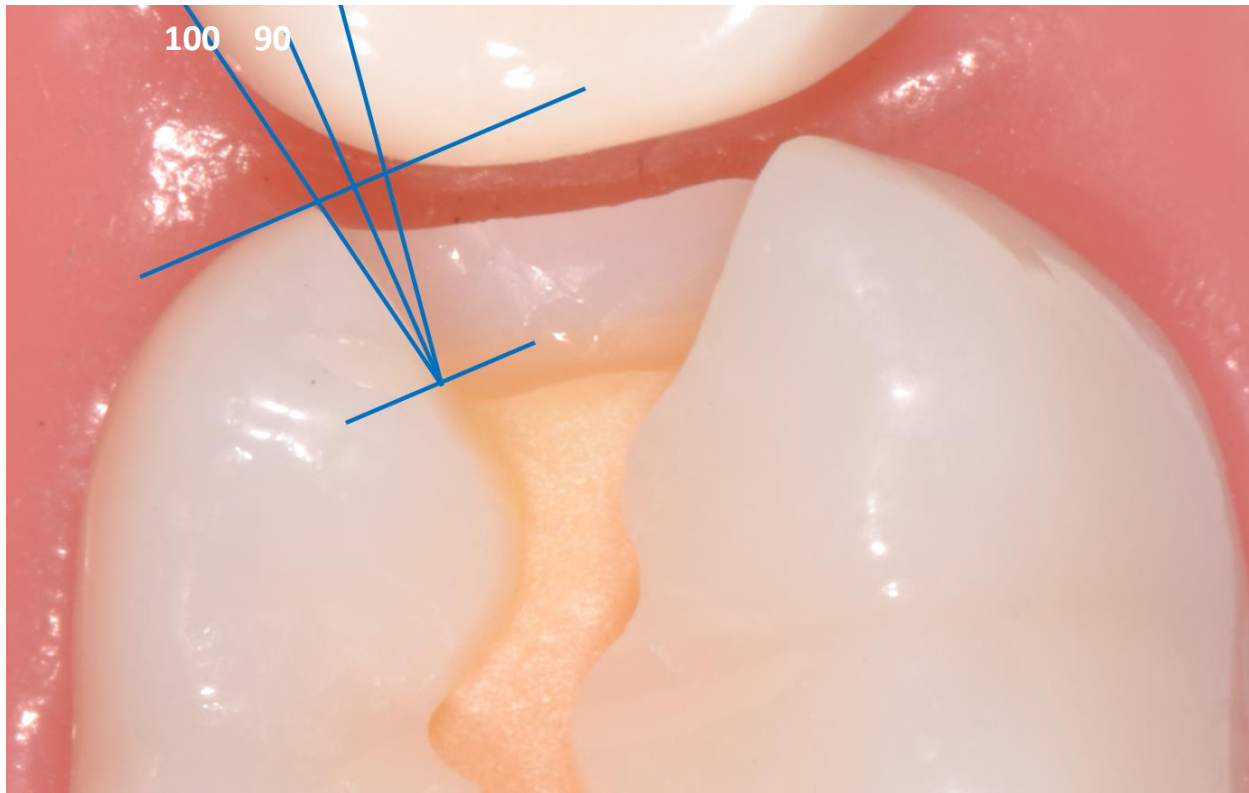
FIGURE 5.2- Image of the 3D model with a superimposed protractor that demonstrates clinically unacceptable preparation wall orientation of $>100^{0}$.
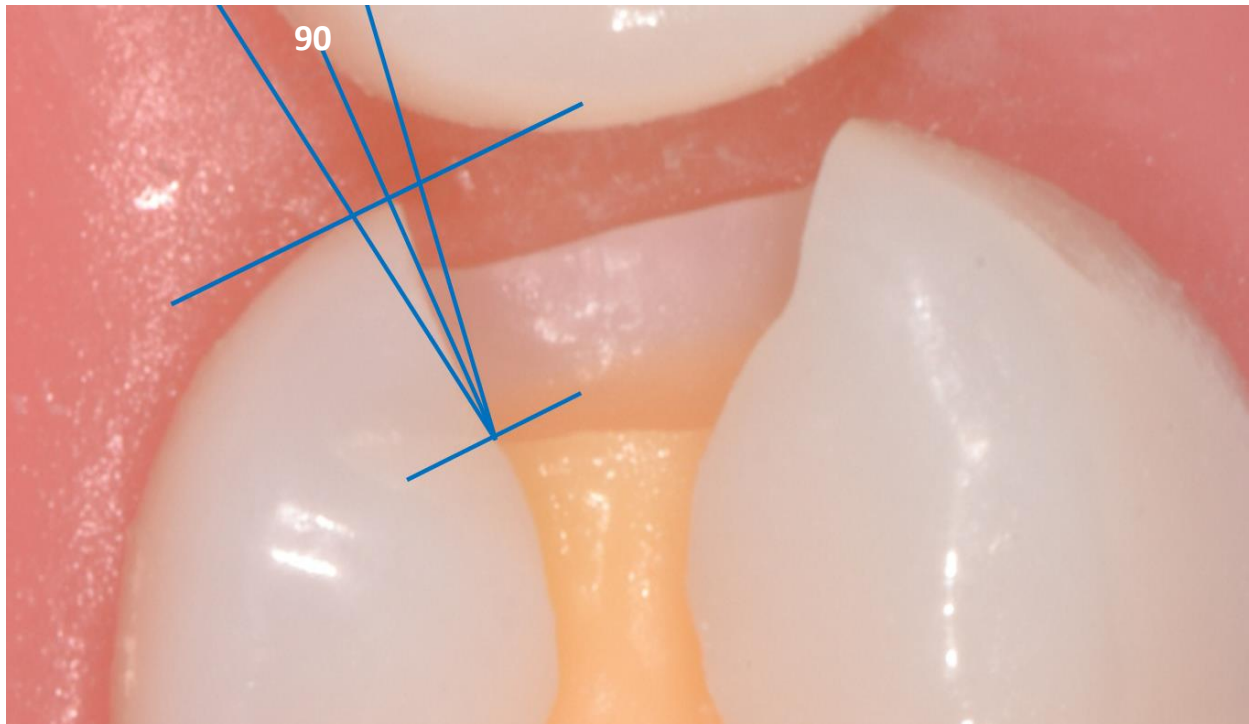
FIGURE 5.3- Image of the 3D model with a superimposed protractor that demonstrates clinically acceptable preparation wall orientation of approximately 90^{0}.

TABLE 3 - Class II amalgam procedure assessment rubric that was used to define the level of student performance and associated point values for each component

PREPARATION COMPONENTS	SPECIFIC CRITERIA FOR EACH COMPONENT	POINT VALUES FOR EACH CRITERIA (Clinically acceptable-CA) (Clinically unacceptable-CU)
EXTERNAL OUTLINE		
DEJ Caries	Complete removal at the DEJ	1 (CA)
	Incomplete removal at the DEJ	0 (CU)
Isthmus Width	Less than 1 mm	1 (CU)
	Between 1 mm and 1/3 of intercuspal distance	3 CA)
	Between 1/3 and 1/2 of intercuspal distance	2 (CA)
	Greater than 1/2 of intercuspal distance	0 (CU)
Proximal Contact Clearance	No Clearance	1 (CU)
	Open up to 0.5mm in all directions	3 (CA)
	Open between 0.5 and 0.75 mm in any direction	2 (CA)
	Open more than 0.75 mm in any direction	0 (CU)
Adjacent Tooth Damage	No damage	2 (CA)
	Requires re-contouring	1 (CU)
	Requires restoration	0 (CU)
INTERNAL FORM		
Enamel Present	None	2 (CA)
	Less than or equal to 50% of preparation	1 (CU)
	Greater than 50% of preparation	0 (CU)
Primary Pulpal/Axial Wall	Less than or equal to 0.5 mm internal to DEJ	3 (CA)
	0.5 - 1.5 mm internal to DEJ	2 (CA)
	2.0 - 2.5 mm internal to DEJ	1 (CU)
	Greater than 2.5 mm internal	0 (CU)

	to DEJ	
Caries Removal	Incomplete	0 (CU)
	Complete	2 (CA)
	Complete with excessive dentin removal	1 (CU)
RETENTION FORM		
External Walls	Occlusal convergence with ~90° cavosurface margin	3 (CA)
	Excessive occlusal convergence with <90° cavosurface margin	2 (CA)
	External walls parallel	1 (CU)
	External walls diverge occlusally	0 (CU)
Retention Groove Placement	Undermined enamel	0 (CU)
	≈ 0.2 mm internal to DEJ	3 (CA)
	Between 0.2 mm and 1 mm internal to DEJ	2 (CA)
	Greater than 1mm internal to DEJ	1 (CU)
	Not Visible	-1 (CU)
Retention Groove Depth	Undetectable	0 (CU)
	Between 0.1 mm and 0.5 mm	2 (CA)
	Greater than 0.5 mm	1 (CU)
PREPARATION FINISH		
Preparation Walls	Smooth, gentle transitions	1 (CA)
	Rough, abrupt transitions	0 (CU)
Preparation Margins	Unsupported enamel (< 80°)	0 (CU)
	Supported enamel (90°)	2 (CA)
	Enamel margin (> 100°)	1 (CU)
Preparation Toilet	Debris present	0 (CU)
	Clean	1 (CA)

FIGURE 6- Average percentage agreement after assessment of 13 procedural components of Class II cavity preparations among examiners at Pilot and Phase I.

TABLE 4 – Average percent agreement (mean +/- sd) and 95% CI for assessment of components of Class II cavity preparation during Phase I

COMPONENTS OF CLASS II CAVITY PREPARATION	MEAN	STD DEV	95% CI – PHASE I (LOWER, UPPER)
DEJ Caries	74.67	15.15	(68.79, 80.54)
Isthmus Width	87.17	8.22	(83.98, 90.35)
Proximal Contact Clearance	70.43	6.67	(67.84, 73.02)
Adjacent Tooth Damage	64.29	11.64	(59.78, 68.81)
Enamel Present	81.81	29.25	(70.47, 93.15)
Primary Pulpal/Axial Wall	72.55	14.52	(66.92, 78.18)
Caries Removal	62.95	21.04	(54.79, 71.11)
External Walls	61.17	23.16	(52.19, 70.15)
Retentive Groove Placement	53.06	17.82	(46.15, 59.97)
Retention Groove Depth	56.03	12.90	(51.03, 61.03)
Preparation Walls	64.89	18.66	(54.16, 68.62)
Preparation Margins	62.73	11.84	(58.13, 67.32)
Preparation Toilet	68.31	11.62	(63.80, 72.81)

TABLE 5 - Average percentage intra examiner agreement for each examiner when comparing the results from Pilot Study to Phase I

EXAMINERS: PILOT STUDY – PHASE I	AVERAGE INTRA-EXAMINER AGREEMENT (%)	AVERAGE KAPPA
Examiner A Pilot - Phase I	68	0.32
Examiner B Pilot -Phase I	68	0.33
Examiner C Pilot -Phase I	74	0.50
Examiner D Pilot -Phase I	77	0.45
Examiner E Pilot -Phase I	72	0.49
Examiner F Pilot -Phase I	83	0.58
Examiner G Pilot -Phase I	72	0.41
Examiner H Pilot -Phase I	80	0.56
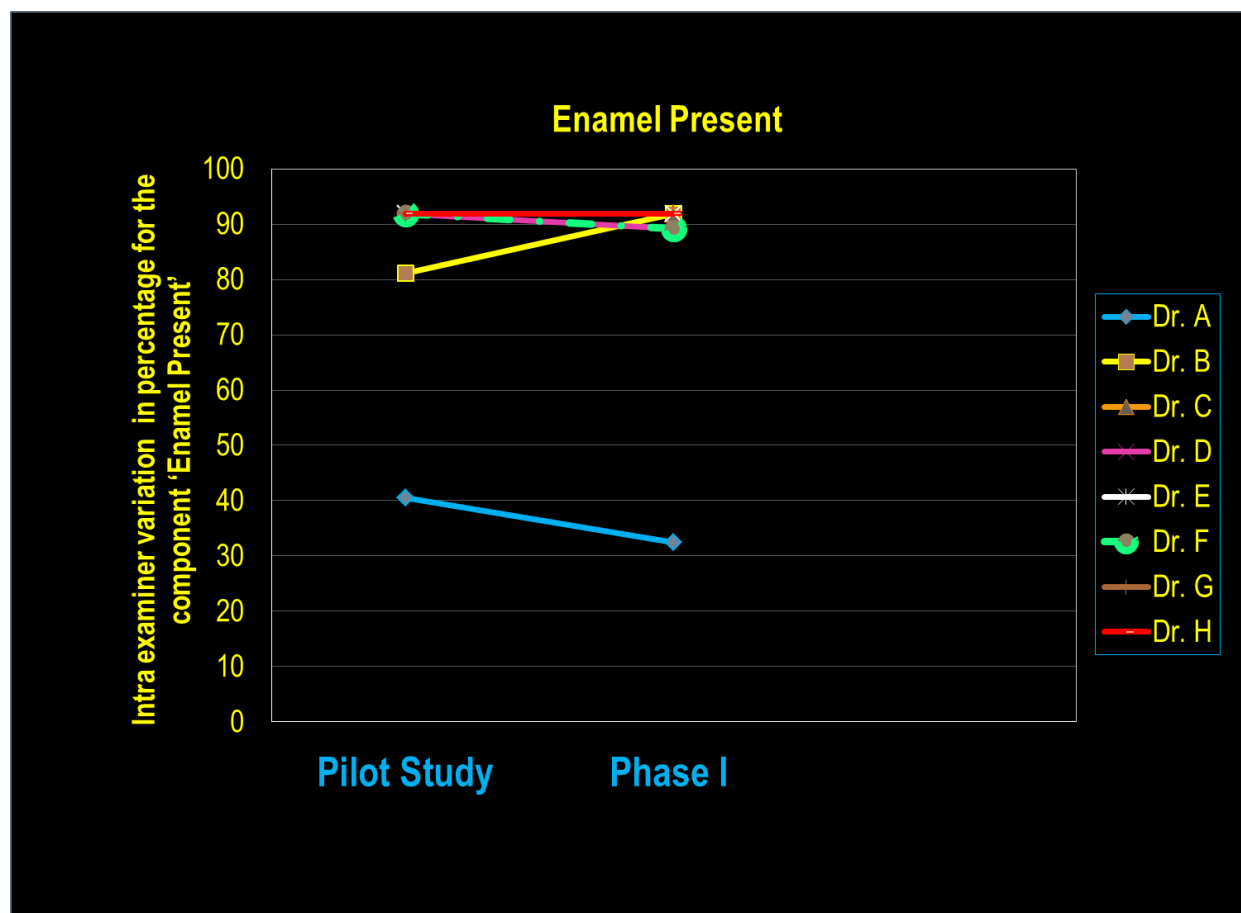
FIGURE 7 - The component 'Enamel Present' had the lowest intraexaminer variation from   Pilot study to Phase I
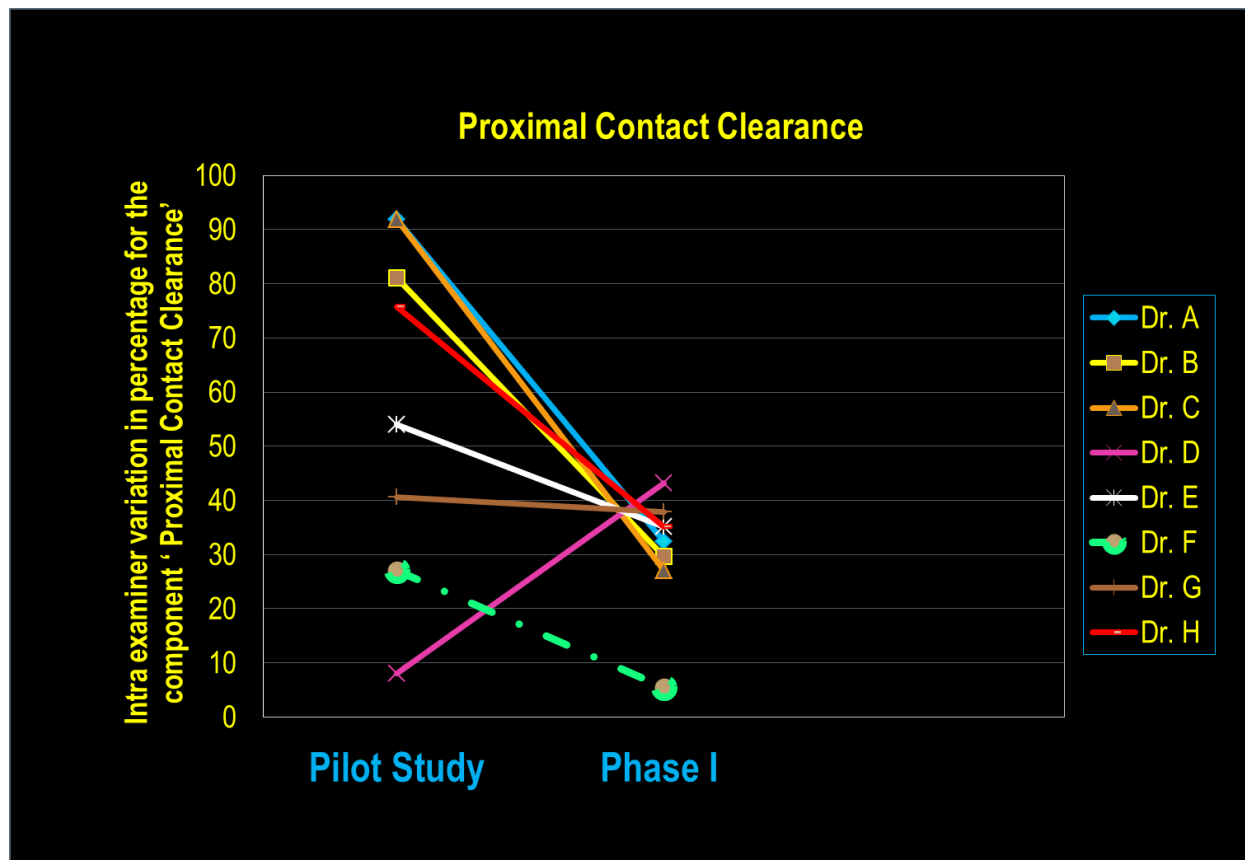
FIGURE 8 - The component 'Proximal Contact Clearance' had the highest intraexaminer variation from   Pilot study to Phase I

TABLE 6- The level of intraexaminer variation (%) for all 13 components of Class II cavity preparation for each examiner when comparing Pilot study to Phase I

Components of Class II cavity preparation	Degree of intra examiner variation (%) between Pilot study and Phase I for each examiner								Mean (%)	Std Dev
	Examiners A through H									
	A	B	C	D	E	F	G	H		
DEJ Caries	0.0	21.6	2.7	10.8	62.2	5.4	29.7	8.1	17.6	19
Isthmus Width	10.8	5.4	10.8	13.5	2.7	8.1	18.9	0.0	8.8	6
Proximal Contact Clearance	59.5	51.4	64.9	35.1	18.9	21.6	2.7	40.5	36.8	20
Adjacent Tooth Damage	2.7	29.7	0.0	5.4	8.1	10.8	5.4	0.0	7.8	9
Enamel Present	8.1	10.8	0.0	2.7	0.0	2.7	0.0	0.0	3.0	4
Primary Pulpal/Axial Wall	5.4	21.6	43.3	16.2	18.9	0.0	13.5	5.4	15.5	13
Caries Removal	18.9	5.4	0.0	5.4	13.5	18.9	8.1	2.7	9.1	7
External Walls	16.2	5.4	13.5	2.7	18.9	8.1	62.2	13.5	17.6	18
Retentive Groove Placement	10.8	18.9	2.7	27.0	18.9	13.5	13.5	35.1	17.6	9
Retention Groove Depth	5.4	21.6	29.7	29.7	2.7	2.7	13.5	13.5	14.9	10
Preparation Walls	18.9	16.2	21.6	27.0	16.2	8.1	45.9	48.7	25.3	14
Preparation Margins	13.5	13.5	35.1	8.1	27.0	5.4	43.2	16.2	20.3	13
Preparation Toilet	29.7	16.2	37.8	16.2	8.1	5.4	2.7	0.0	14.5	13

Row label (vertical): Degree of intra examiner variation (%) between Pilot study and Phase I for each component
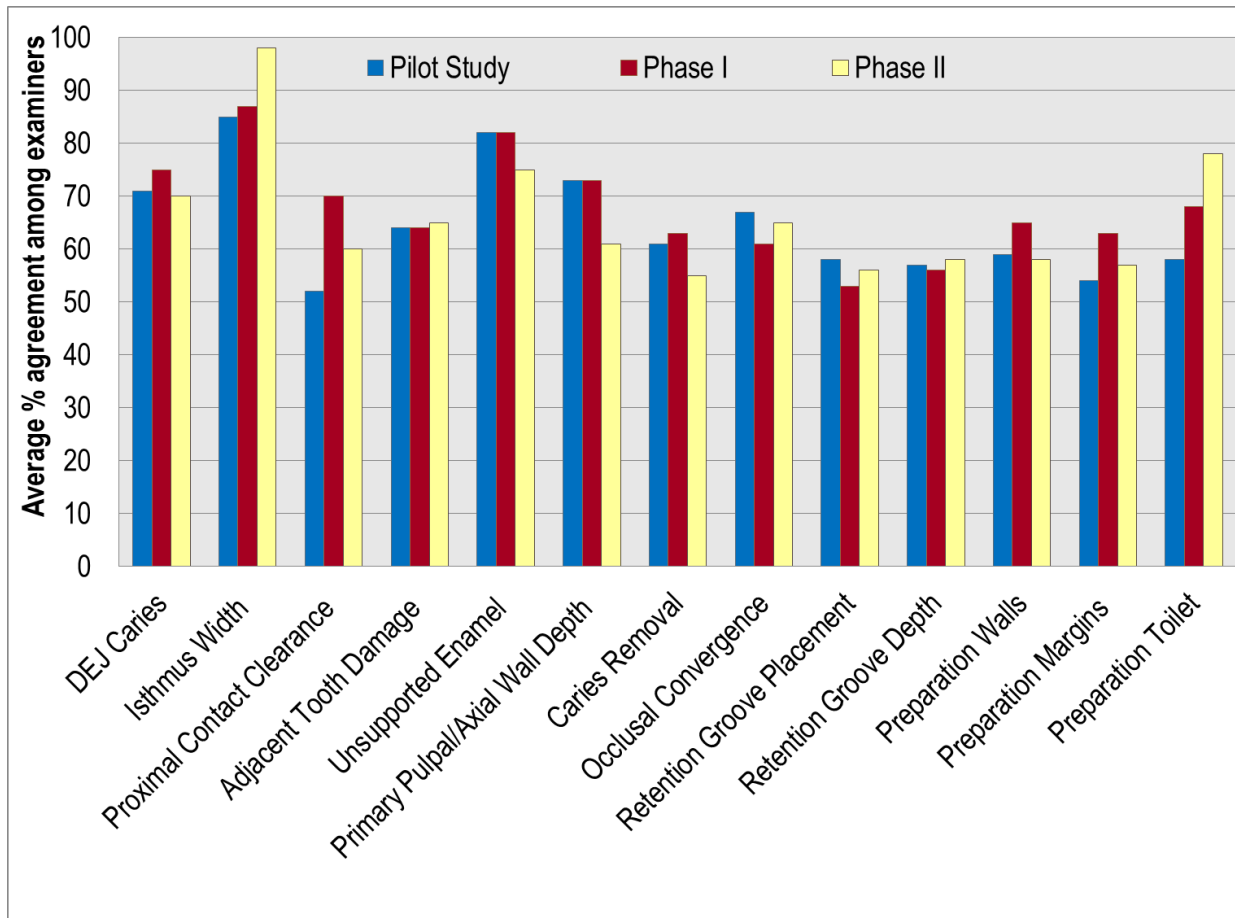
FIGURE 9- Average percentage agreement among examiners after assessment of 13 procedural components of Class II cavity preparations, identified in the post

TABLE 7 – Average percent agreement (mean +/- sd) and 95% CI for assessment of components of Class II cavity preparation during Phase II

COMPONENTS OF CLASS II CAVITY PREPARATION	MEAN	STD DEV	95% CI – PHASE II (LOWER , UPPER)
DEJ Caries	68.87	20.01	(61.11, 76.62)
Isthmus Width	97.66	2.90	(96.53, 98.78)
Proximal Contact Clearance	60.27	17.78	(53.38, 67.17)
Adjacent Tooth Damage	65.07	15.26	(59.15, 70.99)
Enamel Present	78.24	31.19	(66.15, 90.33)
Primary Pulpal/Axial Wall	63.73	23.74	(54.53, 72.94)
Caries Removal	55.03	28.73	(43.88, 66.17)
External Walls	66.08	25.08	(56.35, 75.80)
Retentive Groove Placement	53.24	17.36	(46.51, 59.97)
Retention Groove Depth	55.92	13.96	(50.51, 61.33)
Preparation Walls	58.04	10.60	(53.93, 62.15)
Preparation Margins	55.47	13.19	(50.36, 60.59)
Preparation Toilet	80.25	15.73	(74.15, 86.35)

TABLE 8 - Average percentage intraexaminer agreement for each examiner when comparing the results from Pilot Study to Phase II

EXAMINERS: PILOT STUDY - PHASE II	AVERAGE INTRA EXAMINER AGREEMENT (%)	AVERAGE KAPPA
Examiner A Pilot - Phase II	67	0.12
Examiner B Pilot - Phase II	65	0.23
Examiner C Pilot - Phase II	86	0.26
Examiner D Pilot - Phase II	81	0.37
Examiner E Pilot - Phase II	77	0.33
Examiner F Pilot - Phase II	82	0.45
Examiner G Pilot - Phase II	78	0.42
Examiner H Pilot - Phase II	81	0.31

TABLE 9 - Average percentage intraexaminer agreement for each examiner when comparing the results from Phase I to Phase II

EXAMINERS- PHASE I : PHASE II	AVERAGE INTRA EXAMINER AGREEMENT (%)	AVERAGE KAPPA
Examiner A Phase I - Phase II	71	0.28
Examiner B Phase I - Phase II	62	0.18
Examiner C Phase I - Phase II	81	0.17
Examiner D Phase I - Phase II	78	0.27
Examiner E Phase I - Phase II	83	0.42
Examiner F Phase I - Phase II	86	0.47
Examiner G Phase I - Phase II	66	0.19
Examiner H Phase I - Phase II	79	0.24
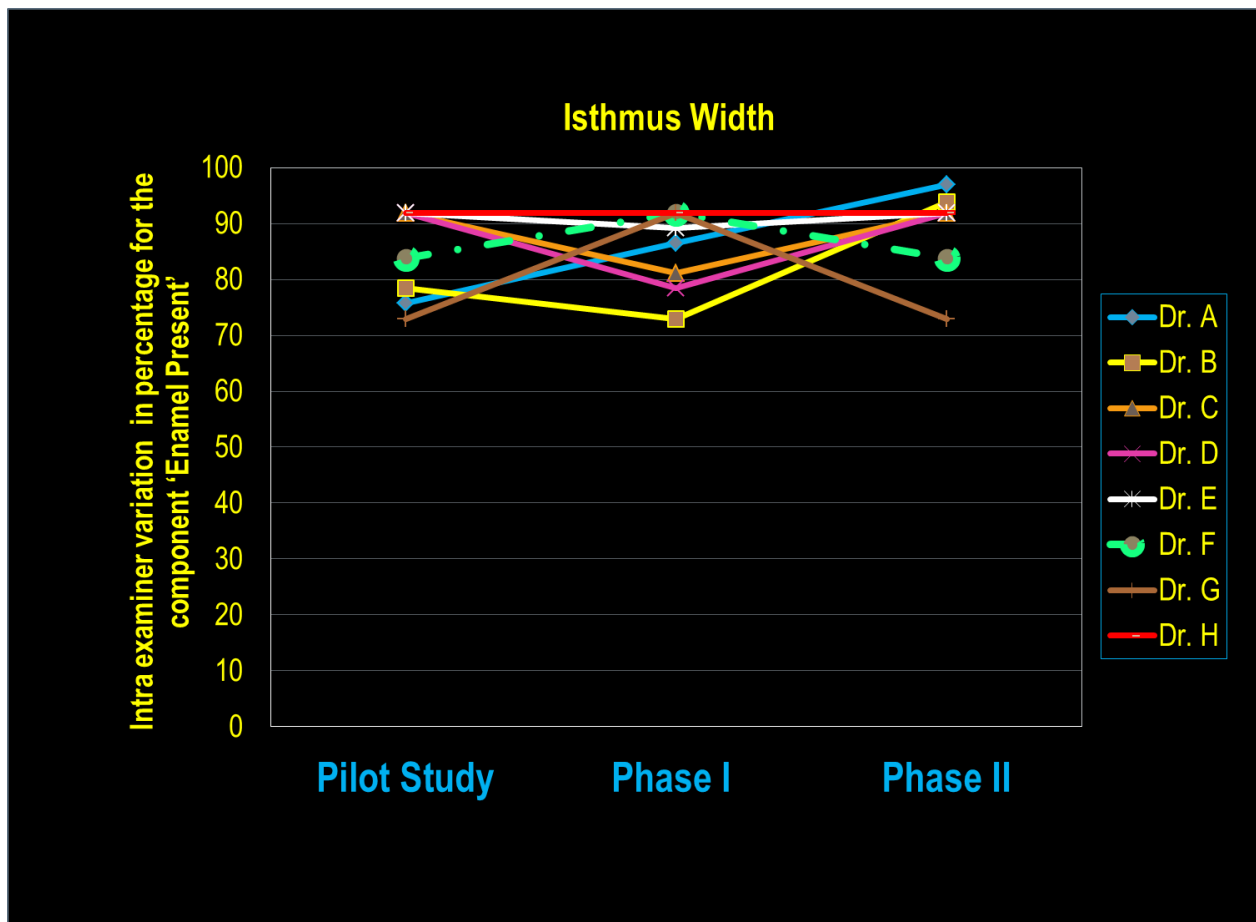
FIGURE 10 - The component 'Isthmus Width' had the lowest intraexaminer variation from   Pilot Study: Phase I: Phase II.

FIGURE 11 - The component 'Proximal Contact Clearance' had the highest intraexaminer variation from   Pilot Study: Phase I: Phase II

TABLE 10 - The level of intraexaminer variation (%) for all 13 components of Class II cavity preparation for each examiner when comparing the Pilot Study to Phase II

| Components of Class II cavity preparation | Degree of intra examiner variation (%) between Pilot study and Phase II for each examiner | | | | | | | | Mean (%) | Std Dev |
| | Examiners A through H | | | | | | | | | |
	A	B	C	D	E	F	G	H		
DEJ Caries	20.53	29.05	0	10.1	68.58	0.25	21.37	12.67	**20.3**	**19**
Isthmus Width	21.2	15.37	0	0	0	0	0	0	**4.6**	**8**
Proximal Contact Clearance	29.39	31.08	32.51	4.39	41.55	23.9	18.84	22.55	**25.5**	**10**
Adjacent Tooth Damage	28.04	5.07	4.56	3.21	4.64	3.21	14.11	7.69	**8.8**	**8**
Enamel Present	43.84	56.08	0	0	0	0	0	0	**12.5**	**21**
Primary Pulpal/Axial Wall	0	33.02	2.11	15.4	10.56	1.86	5.91	14.53	**10.4**	**10**
Caries Removal	30.07	38.85	0	23.7	19.93	0	8.7	7.26	**16.1**	**13**
External Walls	17.23	32.6	0	14.7	23.91	14.11	2.45	3.3	**13.5**	**10**
Retentive Groove Placement	16.81	16.97	5.57	7.01	7.43	14.27	40.29	46.11	**19.3**	**14**
Retention Groove Depth	6.17	13.42	38.01	8.02	13.85	1.35	46.96	59.63	**23.4**	**19**
Preparation Walls	15.54	9.04	17.4	19.1	1.18	14.7	13.42	11.32	**12.7**	**5**
Preparation Margins	8.69	4.3	57.94	6.93	13.77	10.98	15.54	5.07	**15.4**	**16**
Preparation Toilet	64.02	17.57	0	29.3	21.37	16.39	26.61	0	**21.9**	**18**

Row group label (rotated, left margin): Degree of intra examiner variation (%) between Pilot study and Phase II for each component

TABLE 11- The level of intraexaminer variation (%) for all 13 components of Class II cavity preparation for each examiner when comparing Phase I to Phase II

	Components of Class II cavity preparation	Degree of intra examiner variation (%) between Phase I and Phase II for each examiner								Mean (%)	Std. Dev
		Examiners A through H									
		A	B	C	D	E	F	G	H		
Degree of intra examiner variation (%) between Phase I and Phase II for each component	DEJ Caries	20.53	50.68	0	20.95	6.42	5.16	8.36	4.56	**14.6**	**14**
	Isthmus Width	10.39	50.68	0.01	0	0	0	0	0	**7.6**	**16**
	Proximal Contact Clearance	30.07	50.68	32.35	30.74	22.64	2.28	21.54	17.99	**26.0**	**12**
	Adjacent Tooth Damage	25.34	50.68	4.56	8.61	3.47	7.6	19.52	7.69	**15.9**	**14**
	Enamel Present	51.95	50.68	0	0	0	0	0	0	**12.8**	**21**
	Primary Pulpal/Axial Wall	0	50.68	41.14	31.59	8.36	1.86	19.43	9.12	**20.3**	**17**
	Caries Removal	11.15	50.68	0	18.24	6.42	0	0.6	4.56	**11.5**	**15**
	External Walls	1.01	50.68	0	11.99	4.99	22.22	64.62	16.81	**21.5**	**21**
	Retentive Groove	6	50.68	8.28	20.02	26.35	0.76	53.81	10.98	**22.1**	**18**
	Retention Groove Depth	0.77	50.68	8.28	21.71	11.15	1.35	60.47	46.11	**25.1**	**21**
	Preparation Walls	34.46	50.68	4.22	7.94	15.03	22.8	32.52	37.33	**25.6**	**14**
	Preparation Margins	22.21	50.68	22.8	1.18	13.26	5.57	58.78	11.15	**23.2**	**18**
	Preparation Toilet	34.29	50.68	0	13.1	29.48	10.98	29.31	0	**21.0**	**16**

REFERENCES

1.   Haj-Ali R, Feil P. Rater reliability: short- and long-term effects of calibration training. *Journal of dental education.* Apr 2006;70(4):428-433.

2.   Sharaf AA, AbdelAziz AM, El Meligy OA. Intra- and inter-examiner variability in evaluating preclinical pediatric dentistry operative procedures. *Journal of dental education.* Apr 2007;71(4):540-544.

3.   Dhuru VB, Rypel TS, Johnston WM. Criterion-oriented grading system for preclinical operative dentistry laboratory course. *Journal of dental education.* Sep 1978;42(9):528-531.

4.   Brown G, Manogue M, Martin M. The validity and reliability of an OSCE in dentistry. *European journal of dental education : official journal of the Association for Dental Education in Europe.* Aug 1999;3(3):117-125.

5.   Houpt MI, Kress G. Accuracy of measurement of clinical performance in dentistry. *Journal of dental education.* Jul 1973;37(7):34-46.

6.   Houpt MI, Kress G. Evaluation training: training raters to evaluate clinical performance. *Quintessence international, dental digest.* Nov 1979;10(11):67-76.

7.   Hinkelman KW, Long NK. Method for decreasing subjective evaluation in preclinical restorative dentistry. *Journal of dental education.* Sep 1973;37(9):13-18.

8.   Jenkins SM, Dummer PM, Gilmour AS, Edmunds DH, Hicks R, Ash P. Evaluating undergraduate preclinical operative skill; use of a glance and grade marking system. *Journal of dentistry.* Nov 1998;26(8):679-684.

9.   Lilley JD, ten Bruggen Cate HJ, Holloway PJ, Holt JK, Start KB. Reliability of practical tests in operative dentistry. *British dental journal.* Sep 3 1968;125(5):194-197.

10.   Theodore M. Roberson HOH, Edward J. Swift, Jr. Sturdevant's Art and Science of Operative Dentistry 2006.

11.   Scruggs RR, Daniel SJ, Larkin A, Stoltz RF. Effects of specific criteria and calibration on examiner reliability. *Journal of dental hygiene : JDH / American Dental Hygienists' Association.* Mar-Apr 1989;63(3):125-129.

12.   Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. *Family medicine.* May 2005;37(5):360-363.

13.   Dimitrijevic T, Kahler B, Evans G, Collins M, Moule A. Depth and distance perception of dentists and dental students. *Operative dentistry.* Sep-Oct 2011;36(5):467-477.

14.   Moore DL. Current teaching of proximal retention grooves for Class II amalgam preparations. *Journal of dental education.* Feb 1992;56(2):131-134.

15.   Mackenzie RS. Defining clinical competence in terms of quality, quantity, and need for performance criteria. *Journal of dental education.* Sep 1973;37(9):37-44.