University of Dundee

Evaluating and Improving the Assessment and Consistency of Feedback within the Clinical Skills Laboratory at Dundee Dental School

Ahmed Amru Mhanni

DDS, MDSc

A thesis submitted for the degree of Doctor of Philosophy

University of Dundee

January 2018

بسم الله الرحمن الرحيم

وقل رب زدني علما

“O my lord! Advance me in Knowledge”

Taha (20:114) سورة طه

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

First and foremost, I would like to express my special appreciation and thanks to my supervisors Dr. Andrew Hall (AFH), Professor David Ricketts (DNJR) and Professor Graham Chadwick (RGC) you have been tremendous mentors for me. I would like to thank you for motivating my research and for allowing me to grow as a research scientist. Your advices on my research have been priceless. In addition, I owe my deepest gratitude to Dr. Andrew Hall (AFH), who has been constant, unwavering source from time I commenced my PhD. Dr Hall took me under his wing over four years. I cannot thank his enough. Undoubtedly, my doctoral studies would not have been possible without my supervisors' assistance. I could not have wished for a better supervisory team. They have allowed me the freedom and space to decide on the board structure of my project, while always being at hand to refine and shape it.

I would like to thank the staff in the clinical skills laboratory at Dundee dental school for their support throughout the PhD course especially Sandra Ogilvie. I would like to thank my colleagues and friends who gave their valuable time to help me; also to Dr Gordon Gray, Bristol University for permitting through my supervisors the use of his class II feedback sheet (Gray feedback sheet).

I would like to thank my family. Words cannot express how grateful I am to my father (Amru), my brothers (Salaheddin and Mohammed) my sisters (Asma and Hajer) and for all of the sacrifices that you have made on my behalf. Your prayer was what sustained me thus far. I would like to send my regards to sole of my beloved mother who passed away last year. Thank you my mother.

At the end words are not enough to express my appreciation to my beloved wife (Tahani) who was always my support in the moments when there was no one to answer my queries and for taking care of our two lovely children (Taym and Rancy).

DECLARATION

I hereby declare that, all the work described in this thesis is my own original work and that I have consulted all the references cited. This work has been carried out in the Clinical Skills Laboratory at Dundee Dental School, under the supervision of Dr. Andrew Hall, Professor David Ricketts and Professor Graham Chadwick. This work has not previously been submitted for a higher degree in this or any other university.

Ahmed Amru Mhanni (AM)

CERTIFICATE

I hereby certify that **Ahmed Amru Mhanni** has fulfilled the condition of Ordinance 39 of the University of Dundee and is qualified to submit this thesis for degree of doctor of philosophy.

Dr A. F. Hall (AFH)

Professor D. N. Ricketts (DNJR)

Professor R. G. Chadwick (RGC)

DEDICATATION

This thesis is dedicated in loving memory to my mother, who suddenly passed away on 13 February 2016.

I also dedicate this thesis to my family and my wife.

Thank you very much for your support without you this would not have been possible

MAY GOD BLESS YOU ALL

ABSTRACT

Due to the increasing demand from teaching institutions and the General Dental Council to provide dental students with accurate assessment and feedback, the focus of this thesis is to evaluate and improve assessment and feedback at Dundee Dental School. The aim is to determine and appraise assessment tools used in evaluating the clinical skills of dental students in laboratory setting. In addition, the purpose of this thesis is to further develop the assessment tools to provide valid and reliable assessment and feedback on students' performance.

Dental students practise clinical procedures in clinical skills laboratories which are evaluated by qualified staff. Effective evaluation should be valid (accurate and reliable) and produce consistently useful feedback. In this thesis, assessment of experienced (senior) examiners demonstrated, unrepeatable (intra-examiner) and reproducible (inter-examiner) evaluation of class II amalgam cavity and full veneer gold shell crown preparations in a clinical skills laboratory (Chapters 3 and 4).

Further assessment of a wider group of additional teaching staff also demonstrated poor levels of intra-examiner reliability (repeatability) and inter-examiner reliability (reproducibility) (Chapter 4).

Concentrating on the senior examiners, consistent methods, including the use of novel tools, were then devised and tested to improve intra-examiner repeatability and inter-examiner reproducibility. These methods also included feedback sheets which served to provide feedback for students (Chapter 5).

Grades awarded from the best senior examiner who had the highest level of intra- and inter-examiner repeatability and reproducibility, respectively, were then tested against

known developed standard criteria as well as actual preparations to establish the validity of these grades (Chapter 6).

The checklist is the most common assessment method which is used within the Clinical Skills Laboratory in Universities. From five consistent methods, new checklists and reliable tools were established and tested again to demonstrate improved validity and reliability of awarded grades as well as feedback consistency (Chapter 7).

This process now requires further testing with another cohort of preparations to affirm its usefulness. The new cohort should include tooth preparations by both novice and expert operators to reflect a greater range of abilities and thus test the assessment process more rigorously.

Chapter 1 : Literature review and aims

1.1 Definition of assessment

Fish and Coles (2005) defined assessment as *"an all-embracing term for the educational activity of recognising and recording learners' achievements and their development within a specific context and in the light of the quality and scope of the education provided for them"*. It is an integral part of instruction as it determines whether or not the aims of education are being met. In other words, it is a fundamental part of the education process (Schuwirth and Van der Vleuten, 2010). According to Barr and Tagg, (1995), *"Student's assessment is at the heart of an integrated approach to student learning"*. In addition, assessment might change or modify student thinking to improve their performances in the future (Boud and Falckikov, 2007). Therefore, education is largely driven by assessment (Miller, 1990, Wass et al., 2001, Schoonheim-Klein et al., 2006, Manogue et al., 2011, Dolmons and Tigelaar, 2012, van der Vleuten et al., 2012). According to Biggs and Tang (2011), the desired competence or learning outcome should be contextualised within the taught course. It should be defined / described and qualified by teaching and assessment methods, respectively, in order to best develop student learning.

However, assessment usually has a greater impact on the students' learning compared with teaching (Boud and Falckikov, 2007). Assessment has different functions (Harlen, 2007). These functions range from i) a certification procedure leading to a pass/fail decision to ii) feedback providing comments for the student performance (Manogue et al., 2002, Boud and Falchikov, 2007, Harlen 2007). Thus, assessment and feedback are essential for the student as well as the tutor (Harden and Laidlaw, 2012).

1.2 Types of assessment

According to Taras (2008), there are two types of assessment, formative and summative assessment.

1.2.1 Formative assessment

Formative assessment is used to keep track of a student's progress through a particular course of learning. According to Gipps (1994), it *"involves using assessment information to feedback into the teaching/learning process".* In other words, it is used to present information for the students. It is not to judge the students' performances (Hyman, 1980).

Formative assessment should spring from a desire to promote learning with understanding. Students should be aware of what is expected of them and the learning environment should allow opportunities to apply their knowledge and skills. It is based on dialogue between the tutor and the student (Harlen and James, 1997). Sadler (1989) views formative assessment as a way of shaping and improving the competencies of students. In the learning process, formative assessment is feedback (Gipps, 1994). *"Teachers use feedback to make programme decisions with respect to readiness, diagnosis and remediation. Students use it to monitor the strengths and weaknesses of their performances, so that aspects associated with success or high quality can be recognised and reinforced, and unsatisfactory aspects modified or improved"* (Sadler, 1989). Therefore, it helps students to identify their strengths and weaknesses and target areas that need work and to recognize where students are struggling and address problems immediately (Brown and Cooke, 2009).

Sadler (1989) contends that even when a teacher gives valid and reliable assessments there is no guarantee that improvement in performance will follow. The important ingredient for improvement is an understanding of the level of achievement desired. The students need to know what they have to do. Consequently, they can compare their performance with the standard set down and take corrective action.

"Feedback from the teacher, which helps the student with the second of these stages, needs to be of the kind and detail which tells the student what to do to improve, the use of grades or 'good, 7/10' marking cannot do this. Grades in fact may shift attention away from the criteria and be counterproductive for formative purposes." (Gipps, 1994)

For formative assessment to act as a competency enhancer, the students should be able to have a good idea of what they should be able to achieve, similar to that of the teachers. They should be able to monitor their performance in terms of meeting these standards. Formative assessment is used to determine the level of students understanding to provide them with descriptive and informed feedback on their performance progress to encourage and guide their future learning (Epstein, 2007, McDowell et al., 2010, Manogue et al., 2011, Harden and Laidlaw, 2012).

Criterion-referenced assessment is most commonly used in formative assessment. Criterion referencing measures a students' ability by placing them along a particular skill range. Criterion-referenced assessments relate a student's score on an achievement test to a domain of knowledge. The tutor sets the level of performance which is required. It may be the total mastery of a task or it may be a minimal acceptable level. In addition, criterion-referenced assessment allows the tutor to determine students' capabilities i.e. what they can or cannot do. With criterion reference testing it is necessary to develop a set of behavioural objectives which provide:

- A clear definition of what the student should be able to do after the period of learning,
- the context in which these behaviours take place, and
- a given standard which indicates a competent level of performance.

By defining clear objectives the learning process should be enhanced (Turnbull, 1989). The main disadvantage of formative assessment which is purely criterion-referenced assessment is that it is profoundly discouraging for students who are constantly being confronted with failure. This does not matter as long as this information or feedback is used with each student consistently (Harlen and James, 1997).

Thus, the aim of 'formative assessment' is to monitor student learning to provide ongoing feedback that can be used by instructors to improve their teaching and by students to improve their learning. 'Formative assessment' refers to a wide variety of methods that teachers use to conduct evaluations of student comprehension, learning needs, and academic progress during a course (Orsmond et al., 2000). Feedback is most commonly provided for students as a formative assessment of their performance.

Feedback of formative assessment:

It is clear that assessment is an integral component of the learning process by providing the right and constant feedback on where students should be going. The teacher can support, guide and motivate the students to study (Earl, 2012). Feedback should be provided for the students in a correct way in order to prevent them repeating the same mistakes (Neher et al., 1992, Biggs and Tang, 2011).

There are many definitions of feedback but most of these definitions have a lack of consistency (Van de Ridder et al., 2008). According to Ramaprasad (1983), it is information about the gap between actual performance level and the reference level, which is subsequently used to change that gap. Hattie and Timperley (2007) defined feedback as *"information provided by an agent (e.g. teacher, peer, book, parent, self, experience) regarding aspects of one's performance or understanding. A teacher or parent can provide corrective information. A peer can provide an alternative strategy.*

A book can provide information to clarify ideas. A parent can provide encouragement, and a learner can look up the answer to evaluate the correctness of response". According to Brown and Cooke (2009), *"Feedback is an essential part of the learning process. Feedback can be positive or negative, constructive or destructive, minimal or in depth".* Boud and Molloy (2013) had the following definition, Feedback is *"a process whereby learners obtain information about their work in order to appreciate the similarities and differences between the appropriate standard for any given work, and the qualities of the work itself, in order to generate improved work".*

From the previous definitions, the authors did not mention that students who received feedback should take action for learning to happen (Sadler, 2010). Furthermore, they did not impact on the quality of the feedback information (Wingate, 2010, Nicol et al., 2014) or the level of student engagement (Evans, 2013). Therefore, feedback today not only provides the student with strong or weak points about their performance (Bloxham, 2009), it should also play an active role during and after the course by improving the capacity of the student to better manage future learning (Carless, 2006, Nicol, 2010). Feedback is a powerful tool that, if it used intelligently, enhances and develops learners' skills (Brown and Cooke, 2009).

According to the National Students Survey (NSS) in the UK (Unistats 2014), Medical and Dental students showed consistently lower satisfaction scores for feedback compared to Education students and global scores taken across a whole range of students. Education students were used as a 'standard' given that the teachers on an education course would be expected to provide ideal student feedback. At three time-points over the period 2007 to 2013, there was a gradual improvement in all three questions from the NSS which related to feedback satisfaction for Medicine and Dentistry although it consistently remained below both that for an Education degree course and also the global scores (Table 1.1). These results supported the conclusions of Branch and Paranjape (2002) as well as Boud and Molloy (2013). The work of these authors demonstrated the students complain, they did not receive any feedback from their tutor (Branch and Paranjape, 2002), or any feedback provided was not enough (Boud and Molloy, 2013).

Table 1.1 Percentage of National Students Survey of UK higher education full time student's satisfaction ('mostly agree' and 'definitely agree') to three questions related to feedback (numbers 7, 8 and 9) at three different years (2007, 2010 and 2013), for student from medicine and dentistry, education, and the overall score

National Students Survey Respondent Satisfaction				
Question		**Medicine and Dentistry**	**Education**	**Global score**
7. Feedback on my work has been prompt	2007	39 %	56 %	53 %
	2010	40 %	64 %	58 %
	2013	59 %	73 %	67 %
8. I have received detailed comments on my work	2007	31 %	72 %	59 %
	2010	33 %	74 %	62 %
	2013	51 %	81 %	72 %
9. Feedback on my work has helped me clarify things I did not understand	2007	38 %	60 %	53 %
	2010	41 %	64 %	57 %
	2013	54 %	73 %	67 %

Source: Higher Education Funding Council for England (2011 page.42) and Unistats (2014).

Boud and Molloy (2013) highlighted three assumptions that are essential in order to understand and provide clear feedback for the student to improve their learning.

1. Feedback will improve student learning and develop expertise, if the student plays an active role by giving them chance to construct their own knowledge after receiving feedback (Salder, 2010). Received feedback is analysed and discussed with the tutor and other students / teachers (Price et al., 2011). The student can connect this feedback with any existing prior knowledge (Carless et al., 2010, Nicol et al., 2014). According to Salder (1983), the feed forward concept was introduced to enhance the student performance by using tutor

comments. Salder (1989) and Hattie and Timperley (2007) developed this concept.

Salder (1989) developed a three-phase model to move feedback forward. The first phase is that the student needs to know the standard, aims and criteria to complete desired degree of performance. This phase is called 'feed up'. The second phase is that the student distinguishes the difference between his/her current performance and how that relates to the standard, aims and criteria by process of self-assessment. This phase is called 'feedback'. The third phase is that the student understands and then designs a method to reduce any gap between his/her current performance and standard, aims and criteria of course in order to make a better progress. This phase is called 'feed forward'. Similar phases were modified to improve student performance in the clinic but this concept was called, 'feed-forward interview' (Kluger van Dijk, 2010).

By providing students with this three-phase model to move feedback forward, the students will improve their performance and be able to monitor their own work to become self-regulated learners (Nicol and Macfarlane-Dick, 2006).

2. Students need to take action based on received comments from their tutor. There are different ways of providing these comments. According to Ivers et al., (2012), who completed a Cochrane Collaboration review on this subject, feedback on professional practice and healthcare outcomes is most effective when it is provided verbally and in writing with clear aims and an action plan. Higgins et al., (2002) focused on Business and Humanities students' understanding of feedback. They highlighted that the students need more explanation about their mistakes. Therefore, the tutor should engage students in feedback (Evans, 2013) by helping them understand the meaning of the feedback (Nicol, 2010, Orsmond et al., 2013). In addition, the tutors should pay more

attention when they provide comments for the students and target them differently to high and low achieving students (Orsmond et al., 2005).

3. Generalised feedback cannot be provided to all university students because they are variable and diverse group coming from a wide range of educational experiences (Boud and Molloy, 2013). Therefore, for each student and situation should have specific feedback and it is essential to ensure that the feedback message is targeted at the right student and at the appropriate level (Hattie and Timperley, 2007).

The impact of feedback for students is variable (Eva et al., 2012). At the end of a course, summative assessment is indicated to determine whether or not the student passes to the next year.

1.2.2 Summative assessment

Summative assessment is the second version of assessment which is defined as a learning evaluation in order to confirm what students know for basing decision making or certification aims (Sadler, 2005). It usually takes place at the end of a course and is designed to find out whether the instructional objectives of the course have been adequately met. With summative assessment the student is usually assigned a grade or a mark. The intention is to discover 'what has been learnt' and is the student 'fit for purpose' when the course of study is complete (Light et al., 2009, Harden and Laidlaw, 2012). The most common example of summative assessment is the examination at the end of a term or course which is used to determine whether or not students progress into the next year or pass the course. Therefore, summative assessment is also an essential part of education (Harlen, 2007).

The disadvantage of summative assessment is that the students develop a strategy to pass the examination rather than identify their strong and weak points. Thus, passing the examination will become more important for student than gaining knowledge (Biggs and Tang, 2011). In order to address this problem, Black and William (1998) commenced with a review of classroom formative assessment. Other authors have also highlighted formative assessment and how it impacts students' learning (Boud and Falchikov, 2007). Thus, formative assessment has been used to develop and improve students' learning (Rolfe and McPherson, 1995, Light et al., 2009). Some other authors suggested mixing formative and summative assessment in one single assessment (Schuwirth and Van der Vleuten, 2010, Harden and Laidlaw, 2012). On the other hand, there are problems with this approach as students can use strategies to get through the examination process without revealing what they do not know (Biggs and Tang, 2011).

Summative assessment may be either criterion-referenced or norm-referenced. Norm referencing is when the tutor describes the students' performance in terms of their position in the group. It is designed to indicate whether the test-taker did better or worse than other students who took the same test. In other words, this type is used to rank the student's performance and comparing the scores of the students with each other. In addition to the norm-referenced assessment, criterion-referenced is used according to how well it ranks students from high achievers to low. Furthermore, it can be used as a feedback to improve student's performance (Glaser, 1963, Bond, 1996). Norm-referenced has come under attack because it traditionally has focused on low level and basic skills of the students (Bond, 1996). Furthermore, it takes no account of the differing skill levels of individual cohorts of students.

In order to identify which student passes and who fails, standard setting must be determined. It is required before assessment methods can be used for examinations (summative assessment) (Beard, 2005).

Standard setting of summative assessments

In order to assess whether students have acquired relevant skills, a valid and reliable assessment should be developed that employs an appropriate standard setting (Taylor et al., 2013). For Dentistry, this will ensure students who pass this assessment are 'patient ready' and can undertake the basic dental procedures safely and adequately, while on the other hand students who fail will need to retake the module for they might potentially jeopardize patient safety.

As a result of this, and the fact that a particular assessment may act as a 'gateway' to practising dentistry on real patients; carefully assigned and fair pass marks are necessary. However, establishing a consensus on the appropriate pass mark is not an easy task in view of the complexity in evaluating such an assessment (Taylor et al., 2013).

Traditionally, tooth preparations are evaluated subjectively by tutors using a visual inspection method (Taylor et al., 2013). Such a method is often better accompanied by other analytical methods such as using a checklist that is effective in determining whether the minimum requirement for the skill has been met (Goepferd and Kerber, 1980). However, checklists can easily fail to identify 'borderline students' which, in turn, might lead to unfair evaluation. This is mostly attributed to assessor bias and misinterpretation of the checklist (Feil, 1982). Thus, in order to avoid such a problem, a standard has to be set to determine the minimum passing grade that will separate the students who deserve to get promoted to the next level from those who do not. This will indicate whether an assessment performance is good enough for its designated purpose (Puryer and O'Sullivan, 2015).

Several standards have been developed and set for dental clinical assessments (Cizek and Bunch, 2007). These standards can be classified into two groups, 'relative' and 'absolute' (Livingston and Zieky, 1982, Ben-David, 2000).

“Relative standards are expressed in terms of the performance of the cohort taking the assessment. Students will pass or fail depending upon how well they perform relative to other students taking the assessment ... This type of standard is appropriate for assessments intended to select a certain number or percentage of students” (Puryer and O’Sullivan, 2015). Effectively, this is norm referencing referred to previously.

Absolute standards are more commonly used in dental schools. *“Absolute standards are expressed in terms of the performance of students against the test material and do not compare the performance of one student with others taking the test”* (Puryer and O’Sullivan, 2015). Absolute standards are used to identify the level of students’ knowledge or clinical skills for a particular aim, such as graduation from dental school.

Assessment takers will pass or fail depending on their clinical skills and how adequately they meet the requirements of, for example, an ideal tooth preparation regardless of the performance of other students. Therefore, all students potentially could pass or fail. For credible absolute standard setting to be achieved, one or more standard setting techniques should also be used (Puryer and O'Sullivan, 2015).

An absolute standard setting can be achieved using two techniques, ‘test-centred’ and ‘examinee-centred’ (Case and Swanson, 1998). In the test-centred technique, panel staff members make estimations of how they perceive students would fulfil the minimum requirements of a successful tooth preparation. Consequently, a cut-off mark is discussed and decided, below which students will not be considered competent to do the skill and therefore they will need to retake the assessment. Yet, it might be difficult to reach a consensus on a definitive cut-off mark due to differences of expert opinion

(Livingston and Zieky, 1982). Examples of the test-centred technique are Angoff and Ebel methods.

For this to be achieved, the so called modified Ebel's method can play a central role in providing the desired setting. This looks at the relevance and importance of each step in the skill to be assessed via categorizing each step into groups such as: essential, important or indicated (Case and Swanson, 1998). Moreover, the characteristics that the prepared cavity needs to possess in order for it to be considered ‘ready to be filled’ can be used as a guide in this case. In other words, if the number of the total characteristics is, let's say 15 (5 essential, 5 important and 5 indicated); the student is expected to achieve at least 3 essential, 3 important and 2 indicated in order to pass the assessment.

Applying an appropriate standard setting requires not only the full-time staff to be involved but also part-time staff and sometimes the students themselves (Puryer and O'Sullivan, 2015). The staff members chosen need to possess thorough academic knowledge and understanding of the skill that is being assessed and they also need to be familiar with the students and the evaluation process. However, only a few staff members may be qualified to serve as members of the panel. A standard setting cannot be conveniently achieved with a limited number of experts for the process might be greatly influenced by one or more experts who possess too rigid or too flexible standards (hawk versus dove bias); therefore, a panel of more than 5 staff members is usually recommended (Livingston and Zieky, 1982, Fowell et al., 2008).

This test-centred technique can be used solely or in conjunction with an examinee-centred one where expert staff members determine an actual borderline group rather than a hypothetical one. Due to the fact that test-centred methods possess a hypothetical nature, supplemental information about the actual performance of real assessment-takers is highly advisable which can be achieved by implementing an examinee-centred method into the setting (Livingston and Zieky, 1982). This will ensure that the suggested pass/fail mark has served its purpose.

For this to be implemented, borderline regression (an examinee-centred) method can be used where another panel of experts grade the performance of the assessment-takers using a subjective score based upon how well students performed overall (i.e. global score) (Smee and Blackmore, 2001, Schoonheim-Klein et al., 2009). The global score should be independent of the numerical score adopted previously in the aforementioned modified Ebel's method. Such global score is usually comprised of 4 grade descriptors, namely; 'good', 'pass', 'Borderline' and 'fail' (Puryer and O'Sullivan, 2015).

Tooth preparations that are not good enough to be considered as a pass but at the same time not bad enough to be considered a fail are given the 'borderline grade'. Subsequently, the global scores are collected along with the assessment's original grades and are then plotted graphically to compute the statistical linear regression using a statistical software package (Smee and Blackmore, 2001, Schoonheim-Klein et al., 2009). Doing so will generate a cut-off pass mark which will, in turn, indicate whether the original standard setting assigned for the assessment is appropriate or not. Nonetheless, the borderline linear regression method has been proved to provide a high level of credibility and reliability even if used solely (Kramer et al., 2003). Therefore, it is used to determine an actual (not hypothetical) cut-off point. The example of borderline method is usually used to determine the cut-off point for Objective Structured clinical examination (OSCE) to assess the performance of undergraduate students (Kilminster and Roberts, 2004).

It is worth mentioning that setting a standard based on a hypothetical borderline student's performance via a test-centred method is usually time consuming, for the experts are required to meticulously set the desired standard, while actual observation via an examinee-centred method is usually more time efficient. This is due to the fact that it can be undertaken simultaneously throughout the duration of the assessment (Case and Swanson, 1998). On the other hand, applying an examinee-centred setting can be a complex process (Kramer et al., 2003). Borderline regression method in particular, requires an advanced level of statistical calculations which, in many instances, necessitates the collaboration of a statistician.

Both of the aforementioned techniques share a common weak point: they both require judgment that possesses a subjective nature (Zieky et al., 2006, Cizek and Bunch, 2007, Nichols et al., 2010). However, no unified approach exists that can objectively determine the ideal cavity preparation (Taylor et al., 2013). In addition, the mere determination of a cut-off mark remains to be, by far, a subjective process (Zieky et al., 2006, Puryer and O'Sullivan, 2015).

In order to overcome the potential limitations of the aforementioned standards, the same setting method can be repeated by the inclusion of more experts as panel members or, if feasible, ask different experts to repeat the procedure. This will determine the reliability of the assessment that can be also calculated using certain statistical procedures (Puryer and O'Sullivan, 2015).

The General Dental Council (GDC) has stated that several dentistry assessments in the UK appeared to happen at a very basic level of standard setting (The General Dental Council, 2013). This is alarming in view of how crucial these assessments can be for both dental schools and students.

In summary, all dental schools need to seek to ensure that valid and reliable standard settings are applied to their assessments which, if accomplished, are very likely to enhance their education outcomes.

1.3 Criteria for assessment

From the previous, both formative and summative assessments are important. They are widely used in UK Dental Schools in order to provide feedback and scores for students to reflect and develop their performances. In order to evaluate or develop a successful assessment there are a number of criteria that should be considered. The assessment should be valid, reliable, cost-effective, acceptable and have educational impact (Turnbull et al., 1998, Van der Vleuten, 1996).

The form and setting up of a successful assessment method is not easy. Regulation of education and training programmes demands different strengths of assessment methods (The General Dental Council, 2013). Therefore, the widely accepted criteria used to evaluate the strengths of a given assessment method (Watson et al., 2014) have been proposed by van der Vleuten (1996). These criteria include validity, reliability, cost and feasibility, acceptability, and education impaction. In addition, an assessment method with these criteria might motivate the student to learn (Turnbull et al., 1998, Van der Vleuten, 1996).

There are several different assessment methods described in published research literature on medical and dental education (Manogue et al., 2001, Manogue et al., 2002, Epstein, 2007, Albino et al., 2008, Kramer et al., 2009, Manogue et al., 2011, Norcini et al., 2011, Taylor et al., 2013). 'Glance and grade', 'checklist', 'point scales', and 'computer-assisted' (or software) are the most common ways to provide assessment of tooth preparations in clinical dentistry. Most of these studies focused on one to two elements of the assessment criteria rather than all points; for example, validity and reliability only.

Before reviewing some studies which are related to assessment in dentistry, there is information that should be considered before developing an assessment method. The examiners should understand the outcomes to be evaluated and need for a blend of assessment methodologies (Manogue et al., 2011). Therefore, no single methodology can test all levels of competence and performance described in Miller's Pyramid of layers (Figure 1.1). Miller (1990) proposed an outline for assessing competence in the form of a pyramid. The lower two parts assess the knowledge of the student and the upper two parts assess the competence (Miller, 1990, Epstein and Hundert, 2002, Carr, 2006, Chadwick and Holsgrove, 2009, Davies et al., 2009, Schuwirth and van der Vleuten, 2010). George Miller created this framework to assess the knowledge, attitudes and skills of the students. It was explained how the students in professions such as medicine and dentistry develop such skills. In other words, clinical professionals are often concerned not just with knowledge acquisition, but achievement of the skills and their application.

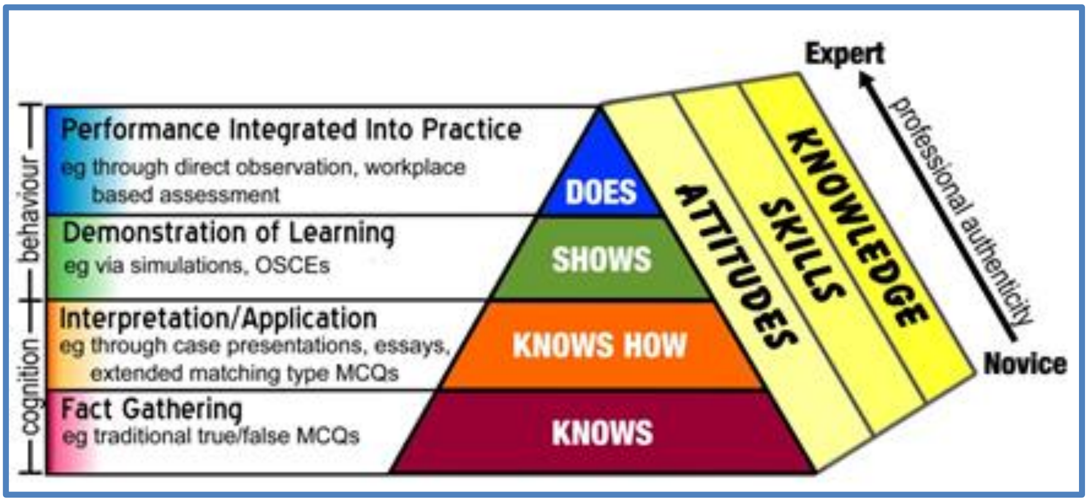


Figure 1.1 Miller's pyramid
Based on work by Miller G.E. the assessment of clinical skills/Competence/Performance; Academic Medicine 1990; 65(9);63-67 Adapted by Drs. R. Mehay and Burns, UK (January 2009)

In the pyramid, the base is used to assess the knowledge 'Knows' of the student. The next stage is 'Knows how'. It is used to assess how the students can apply the knowledge which they have. These two stages are only used to assess the cognitive domain (i.e. test of the knowledge of student). Inexperienced students usually sit in this area. The next stage of Miller's pyramid assessed is how the student can apply the knowledge. At the top of pyramid, assessing performance of the student is the function of this stage. The upper two stages assess competence and performance of the student. The cognition area should have high correlation with the performance area. A student who knows how to do something does not necessarily mean that the student will do that in practise. It is essential that the students do what they know in practise otherwise, there is no point learning it. Therefore, if the tutor wants students to apply their learning to clinical practise, the tutor needs to use assessment methods that will motivate them to progress through the 'shows and does' areas (i.e. performance) (Miller, 1990).

Thereafter, evaluating and developing an assessment can be taken forward by considering five criteria of assessment. The following pages will review these criteria based around several assessment methods which are used in dentistry.

1.3.1 Validity of assessment:

Validity *"is the extent to which the competence that the assessment claims to measure is actually being measured"* (Schuwirth and van der Vleuten, 2010). DeVon et al., (2007) and may also be defined as, *"the ability of the instrument to measure the attributes of the construct under study".* In addition, Lynn (1986) defined the validity as *"a crucial factor in the selection or application of an instrument, for validity is the extent to which that instrument measures what it is intended to measure".*

Most universities use specific assessment formats to assess undergraduate and postgraduate students. These assessment methods should be valid in order to provide more accurate feedback and guide the students to achieve the outcomes of the course.

Types of validity

To determine validity for any type of assessment, certain types of the validity must be considered; content-related, criterion-related and construct-related validity (Messick, 1995). According to Lynn (1986), *"although over 35 terms may be used to connote types of validity ... only three types are in common usage today - content, criterion-related, and construct validity"*.

These concepts of validity are used to establish the overall validity of a given assessment method. Table 1.2 shows the classifications of the validity according to Lynn (1986) and Messick (1995) with appropriate questions, developed by the author, which may be asked to help determine such validity.

Table 1.2 Types of validity

Types of validity *according to Lynn (1986) and Messick (1995)*
Content-related validity Does the assessment method include (all) the right item(s)? *Face validity (sub-type of content validity)* Does the assessment method seem to be generally (lay person's opinion) correct?
Criterion-related validity Does the assessment method accurately predict or forecast? *Criterion-concurrent validity* *Criterion-predictive validity*
Construct-related validity Does the assessment method measure the correct construct? *Convergent validity* Does assessment method with other related measures? *Discriminant validity* Does assessment method discriminate among un-related measures?

DeVon et al., (2007) preferred another classification of validity, originally suggested by Trochim (2006), and stated that *"all types of validity fall under the broad heading of construct but content and face validity are termed translational (as in translation of the construct);* (whereas) *concurrent, predictive, convergent and discriminant are types of criterion validity"* (Figure 1.2).

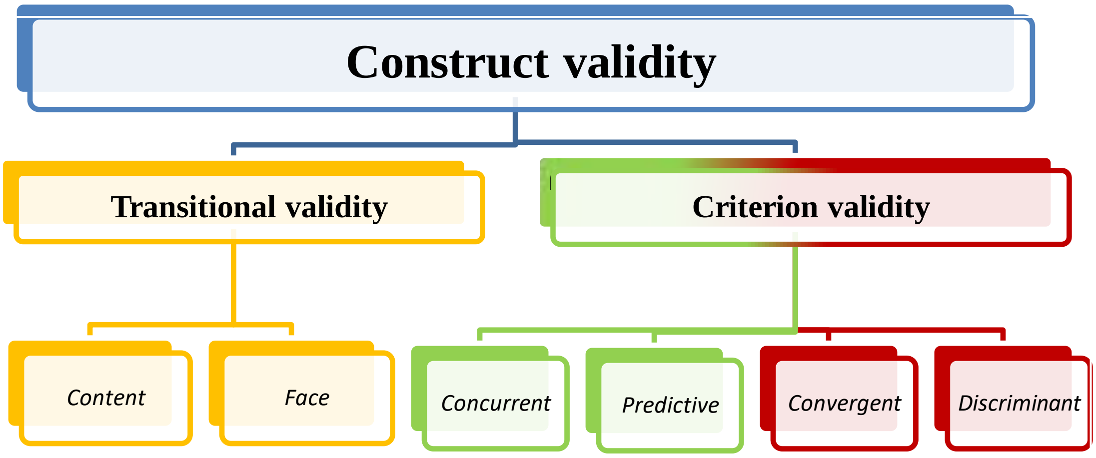


Figure 1.2 Type of validity assembled by the author based on the work of Trochim, W. (2006)

<u>Content validity:</u>

Content validity is defined as *"the degree to which an instrument has an appropriate sample of items for the construct being measured"* (Polit and Beck, 2004).

For assessment methods, content validity is whether or not the assessment method for a given test accurately reflects the whole testable domain. In other words, this assessment should reflect the objective domain which the student has achieved. Thus, content validity is not determined by assessment format but by content of the assessment (Schuwirth and van Der Vleuten, 2010). If the assessment method (e.g. Checklist) and/or outcomes were not relevant to the learning objectives, the examiners can modify or change the content of the assessment method (Streiner and Norman, 2008). Therefore, content validity should be used in order to measure whether the assessment method sufficiently covers the area it is intended to cover. This type of validity was assessed only through the ratings of experts. According to Lynn (1986), there is a two-stage process to determine the content validity for a given assessment method.

The first stage is a review of the literature to identify the content of the assessment tool (e.g. content of the checklist). This stage is called, ‘The Development Stage’. In this stage, the literature review and opinion of examiners are essential steps to collect the items and their components of the assessment method to format a matrix which is also called a ‘blueprint’. This is usually used to determine the items or categories appropriate to the assessment method (Crossley et al., 2002). From the blueprint, a Content Validity Index (CVI) is created (Lynn, 1986, Netemeyer et al., 2003, Polit and Beck, 2006, DeVon et al., 2007). The Content Validity Index (CVI) is used in order to establish, revise, delete or substitute the items and their components of the assessment tool (e.g. checklist). According to Lynn (1986), there are two types of CVIs. The first type involves the Content Validity Index of individual Items (I-CVI) and the second involves the Content Validity Index of the overall Scale (S-CVI) (Lynn, 1986, Polit and Beck, 2006).

The second stage requires a panel of examiners for ‘The Judgement-Quantification Stage’ to determine content validity. Examiners work independently in order to evaluate the Content Validity Index for Items (I-CVI) by rating items according to their relevance (Lynn 1986, Berk, 1990, Polit and Beck, 2006). According to Lynn (1986), a four-point scale: 1 = not relevant, 2 = somewhat relevant, 3 = relevant and, 4 = very relevant, should be used for rating items (I-CVI) in order to determine whether the items should be used or excluded. To recognize the agreement which can be inflated by chance factors, Lynn (1986) recommended that if the number of examiners who asked to rate the items was less than five, all the examiners must agree on the content validity for their rating. Items can be exchanged or modified, if the examiners feel they are not, or somewhat, relevant (Lynn, 1986).

Content validity is calculated by counting the results of the examiners based on the degree to which the examiners agree on the relevance of the items. Items should be ranked 3 = relevant or 4 = extremely relevant by examiners to be finally selected as an item of the new assessment method (Lynn, 1986, Polit and Beck, 2006, DeVon et al., 2007, Sirajudeen et al., 2012).

According to Lynn (1986), Content Validity Index for Scales S-CVI is also used to establish *"the proportion of the total items judged content valid"* (Lynn, 1986); *"the proportion of items on an instrument that achieved a rating of 3 or 4 by the content experts"* (Beck and Gable, 2001). In other words, S-CVI is computed as the number of items given a rating of either 3 or 4 by the experts, divided by the total number of the items on the instrument (Lynn, 1986, Polit and Beck, 2006).

According to Polit and Beck (2006), excellent I-CVI for three to five experts is 1.00 (100% agreement at the item level), while the level of the agreement for six to ten experts is a minimum 0.78 (78% agreement at the item level). In addition, they suggested that the agreement, between experts who have judged the items to be relevant, for S-CVI should be 0.90 (90%) or higher. This requires clearly defined and relevant items, carefully-selected experts (Davis, 1992), and clear instructions to the experts about the underlying constructs and the rating task (Lynn, 1986).

Ahmed et al., (2016) used a class II amalgam cavity preparation procedure assessment rubric in order to define;

- the level of student performance and
- the associated point value (i.e. clinically acceptable / unacceptable),

for each criterion within the rubric. The criteria of the assessment were adopted from Sturdevant's Art and Science of Operative Dentistry. Following a Microsoft PowerPoint presentation, eight examiners took part in a detailed discussion of;

- the 13 components of the class II amalgam cavity preparation and
- the specific criteria defining levels of student performance

for each component listed on the assessment rubric form. They were provided with the chance to ask questions throughout the calibration session. The same examiners then assessed 32 class II amalgam cavity preparations which were prepared by first year dental students as a part of the preclinical operative dentistry course. These preparations were assessed using the assessment rubric form with a periodontal probe. Reliability among the examiners improved after calibration. In addition, the objective use of a periodontal probe increased reliability.

Thus, content validity of an assessment method is usually established by content review, which should be undertaken by panel of examiners (experts) within the domain being examined (Beanland et al., 1999). Although the content validity is important to the design of the assessment method, it is not the only method used to determine validity of the assessment method (Strainer and Norman, 2008). Face validity is sometimes required.

Face validity:

Face validity is a sub-type of content validity (Beanland et al., 1999). It is defined as a *"validity conferred by the lay person's acceptance that procedure, statement, or instrument appears to be sound or relevant… face validity includes validity by assumption (a non-statistical assessment of the logical tie between the elements or items of an instrument and its purpose) and validity by definition (the determination by one or more content experts that the elements or items of an instrument represent the content domain being assessed)"* (Lynn, 1986). According to Streiner and Norman (2008), *"Face validity simply indicates whether, on the face of it, the instrument appears to be assessing the desired qualities".*

Face validity is not quantifiable method like content validity. It is the least scientific of all measures of validity (Lynn, 1986). It is only used to determine that the structure of assessment method represents a subjective decision based on a review of measure itself by one or more experts (Streiner and Norman, 2008). Thus, editorial review and pilot studies are always used to establish face validity for the assessment method. From these reviews and pilot studies, face validity is used to confirm understandability, clarity of content, consistency of assessment method and covering of learning objectives, by asking the panel of examiners (Downing and Haladyna, 2004).

Although content and face validity are important, they are not appropriate to confirm the overall validity of the assessment method, as they do not provide any evidence from the assessment of grades or scores (Streiner and Norman, 2008). Therefore, construct validity was introduced by Cronbach and Meehl (1955) to address this issue.

<u>Construct-related validity:</u>

Construct validity is used to evaluate whether or not the assessment method (construct) measures the domain of knowledge and skills being assessed. According to Crossley et al., (2002), "dental students' skills" would be a construct that might be expected to be better at the end of the course than that at the beginning of the course. Assessment grades or scores will confirm this improvement. "*A typical example for construct validity is that more intelligent students can learn faster, have superior memory skills and* (are) *better able to solve problems than less intelligent students"* (Schuwirth and van der Vleuten, 2010). Using this as an example, the students who produced more errors should have lower scores than the students who produced more correct answers (Norman and Eva, 2010). Therefore, the assessment form and its outcome should indicate this information for the examiner. Thus criteria are the essential part in the assessment form to provide scores for the students.

Several examples in the literature describe an important aspect of these assessments which is the development of explicit and meaningful criteria within a checklist (Haj-Ali and Feil, 2006, Ahmed et al., 2016). Development, application, and validation of successful assessment structures are dependent on clear and meaningful criteria for the assessment method. Knight (1997) recommended that criteria should be;

i) valid: individually, collectively, and non-compensatory (independent of one another), and
ii) reliable: all criteria should be clearly described and the levels of performance clearly defined within a matrix format (Section 1.3.2).

From recommendations of Knight, construct validity of criteria and their levels of performance are important for calibration. Knight (1997) reviewed original evaluation methods to include very specific levels of performance (descriptors) and limited these levels to;

- excellent,
- clinically acceptable and
- criterion not met,

to improve the training of students and teachers, as well as examiner agreement. Knight (1997) recommended assessment methods with calibrated criteria can be used for students to evaluate their performance in the clinic. The student must know exactly what it is that is to be achieved in order to perform a designated procedure. Additionally, to improve the students' performance, the students should receive feedback comparing their performance to an ideal (Knight, 1997).

Based on these recommendations, Haj-Ali and Feil (2006) developed grade forms with relevant criteria for each of three levels (ideal, acceptable, standard not met) for preclinical Class II amalgam cavity preparations. They used an assessment method (i.e. checklist with periodontal probe), for three rounds of assessment, by nine preclinical operative laboratory examiners, who individually evaluated ten prepared class II amalgam cavity preparations. The first round of assessment was without any calibration training. The second round of assessment was immediately after calibration, and the third round of assessment was ten weeks later with no further calibration or training. They concluded that calibration of the examiners by using clear and understandable criteria for preclinical students was essential to improve the agreement among the examiners and provide accurate scores that reflect that assessment (Haj-Ali and Feil, 2006).

The criteria may be structured into a matrix format with standards clearly specified and each criterion expanded into written statements describing different degrees of quality. This type of assessment allows examiners to specify criteria related to each step or feature in a clinical performance task and define each level of accomplishment on a scale. Licari et al., (2008) published an excellent guide for developing assessment forms for both preclinical and clinical performance and reconfirmed the importance of clearly defined criteria in a well-organized assessment form. They suggested that consistent terminology and a standardized format for all assessment forms were important both for the student and the clinical examiners. Appropriate organization of the evaluation forms can facilitate provision of specific feedback and support active participation of the student (Licari et al., 2008). Thus, construct validity is important.

There are two types of construct validity: convergent validity and discriminant validity.

Convergent validity

Convergent validity is type of construct validity. It is a comparison between the results of an assessment method with an established method (i.e. gold standard method) administrated at the same time (Campbell and Fiske, 1959, Shuttleworth, 2009). These methods should have the same parameters (Bastien et al., 2001). Measures of constructs that theoretically should be related to each other are, in fact, observed to be related to each other. The correlation between two different methods evaluating the same attributes should be high. For example, two different evaluations of a class II amalgam cavity preparation, perhaps one evaluation being shorter and easier to administer, should give the same overall result.

Discriminant validity

Discriminant or divergent validity refers to a measurement method's ability to vary indirectly with a measure of an opposite construct. Campbell and Fiske (1959) introduced this type of validity. Measures of constructs that theoretically should not be related to each other are, in fact, now observed to be related to each other. It indicated that outcome of assessment method does not correlate with other method's outcome presumed to measure conceptually dissimilar constructs (Campbell and Fiske, 1959). For example, assessment method of class II amalgam preparation is not highly correlated with other assessment methods designed to assess different types of preparation (e.g. full crown preparation).

Criterion-related validity

Criterion validity evidence involves the correlation between the test and a criterion variable (or variables) taken as representative of the construct. In other words, it compares the test with other measures or outcomes (e.g. criteria) already held to be valid (Streiner and Norman, 2008). For example, if there is high degree of correlation between the criterion variable and the grades on the assessment method which was commonly used to assess, this would be evidence of criterion validity (DeVon et al., 2007). Statistically, a Spearman correlation is the most commonly used test for categorical data while a Pearson correlation test is used for non-categorical data (DeVon et al., 2007).

Concurrent validity

If the assessment data (e.g. grades) and criterion data are collected at the same time, this is referred to as concurrent validity evidence. In other words, concurrent validity is a comparison between the results of an assessment (i.e. grades) with established examination administered at the same time (Shuttleworth, 2009). These assessment methods should be assessed using the same task (Bastien et al., 2001) and analysed using a simple correlation (e.g. Spearman correlation).

Predictive validity

Predictive validity is the extent to which a score on a scale or assessment predicts scores for the same criteria measure using a recognised standard. If the assessment data (i.e. grades) are collected first in order to predict criterion data collected at a later point in time, then this is referred to as predictive validity evidence (Cronbach and Meehl, 1955). This type of validity is used to determine the degree to which a test grade can expect how well the student will do in the future.

Predictive validity and concurrent validity are generally measured as a correlation between an assessment score and some criterion measure. These types of Criterion-related validity are most commonly used for an Objective Structured Clinical Examination (OSCE) (Brown et al., 1999, Gerrow et al., 2003). Brown et al., (1999) used criterion-related validity for an OSCE which consisted of 17 stations in conservation, periodontology and prosthetics to assess clinical competence and to provide feedback to students. The conclusion was that the OSCE was intrinsically valid and a better predictor of performance in the final examination than either a concurrent 4th year examination or Advanced-level university entry grades.

Gerrow et al., (2003), evaluated the concurrent validity of the National Dental Examining Board of Canada (NDEB) Written Examination and the OSCE by correlating students' scores with their performance in the final year of the Doctor of Dental Surgery (D.D.S) or Doctor of Dental Medicine (D.M.D) program. The subjects of this study were the 2317 students at nine Canadian dental schools who completed parts one and two of NDEB examinations between 1995 and 2000. The findings indicated positive correlations between students' examination scores and final year results. In addition, the conclusion of this study supported the concurrent validity of both NDEB examinations.

It is essential to remember that some assessment methods might be reliable but not valid. In addition, an assessment method cannot be valid unless it is reliable (Beanland et al., 1999, Polit and Hungler, 1999).

1.3.2 Reliability of assessment

Reliability relates to the extent to which examiners can consistently distinguish between different items on a measurement scale. Reliability is one of criteria for assessment. Dr. Sue Hegyvary, editor of the journal of Nursing Scholarship, commented that *"Validity and reliability are basic requirements for research. Good articles include such* *information but others do not, to the detriment of those articles, because the findings are not credible unless the data are credible".* Beanland et al., (1999) defined reliability as the degree to which an assessment method produces consistently the same results with repeated administration.

According to DeVon et al., (2007), reliability is divided into two main types, 'stability reliability' (i.e. test-retest reliability) and 'equivalence reliability' (i.e. alternative or parallel-forms reliability and Coefficient Alpha). Weiner (2007) demonstrated that the reliability can also be classified into i) test-retest reliability, ii) internal consistency reliability (i.e. split-half reliability and coefficient Alpha), and iii) inter-rater (observatory) reliability.

From the previous classifications, types of reliability are: test-retest, internal consistency, alternative or parallel forms, and inter-rater reliability, most commonly used.

Type of reliability

<u>Test-retest</u>

Stability (test-retest) reliability is used to estimate the consistency of the same test on two occasions using the same examiners, teeth and environment. Thus, this type is used to measure reliability across different points in time (DeVon et al., 2007, Waltz et al., 2010). It is also called intra-rater reliability (Rankin and Stokes, 1998). The exact nature of the test applied will depend on the data being evaluated (DeVon et al., 2007). The amount of time allowed between measures is critical. If the same objective is measured twice, the correlation between the two observations will depend in part by how much time elapses between the two measurement occasions. Two weeks to one month is the generally accepted time interval for retesting (Waltz et al., 2010). According to Polit and Beck (2004), the memory reactivity effects impact on the test-retest measures of reliability. Respondents' memories tend to decline as the time between tests lengthens (Polit and Beck, 2004). Thus, leaving reasonable time between assessment sessions is important.

Coefficient alpha

Coefficient (Cronbach's) alpha is also called equivalence reliability (DeVon et al., 2007). Coefficient alpha is the most commonly used statistic to estimate internal consistency reliability. This type is most commonly used to determine the internal consistency of questionnaires (Brink and Wood, 1998, Polit and Beck, 2004). Internal consistency measures how well the items on instrument fit together theoretically. This type of reliability is used for one test administration (Ferketich, 1990, Waltz et al., 2010). This uses to compute one split-half reliability and then randomly divide the items into another set of split halves and re-compute, and keep doing this until all possible split half estimates of reliability are achieved. Thus, coefficient alpha is mathematically equivalent to the average of all possible split-half estimates. All split-half estimates are calculated from the same sample. If the items are not correlated, coefficient alpha value is low (Nunnally and Bernstein, 1994).

In general, a coefficient alpha of 0.70 is acceptable for new scales (DeVellis, 2003). On the other hand, Nunnally and Bernstein (1994) recommended that a reliability coefficient of 0.80 is adequate, but went on to state, *"if important decisions are made with respect to specific test scores, a reliability of 0.90 is the bare minimum and a reliability of 0.95 should be considered the desirable standard"*. Bland and Altman (1997) also recommended that the coefficient alpha should be minimally 0.90 with an ideal value of 0.95. Cicchetti (1994) suggested the following reliability (coefficient alpha) guidelines for clinical significance: [reliability (< 0.70) is unacceptable, reliability (≥ 0.70 - < 0.80) is fair, reliability (≥ 0.80 - < 0.90) is good, and reliability (≥ 0.90) is excellent].

Alternative and parallel forms

Alternative or parallel forms reliability was developed by Hubley and Wagner (2004) in a study to examine two different forms of the Multidimensional Health Locus of Control Scale (MHLCS). It is also a type of equivalence reliability (DeVon et al., 2007). In alternative and parallel forms reliability, two parallel forms have to be created. These two test forms address the same construct. Both of them are administered to the same sample of people or students. The correlation between the two parallel forms is the estimate of reliability. In other words, alternative and parallel forms are used to determine reliability of scores from two test forms, each with different items to assess the same concepts. These two test forms must assess the same phenomenon and have scores with approximate means, variance and alpha coefficients (DeVellis, 2003, Waltz et al., 2010). According to Brink and Wood, (1998), some authors suggested a correlation between test forms at least 0.80 is acceptable as an alternative or parallel form of reliability.

Inter-rater reliability

This type of reliability is used to compare two or more of the observers/raters at a point in time in order to determine whether two observers are being consistent in their observations. In other words, inter-rater reliability gives a score of how much homogeneity, or consensus, there is in the ratings given by raters (Gwet, 2014). The guidelines developed by Cicchetti and Sparrow (1981) represented a simplified version of those introduced by Landis and Koch (1977). The guidelines state that, when the reliability coefficient is below 0.40 and 0.59, the level of clinical significance is fair;

when it is between 0.60 and 0.74, the level of clinical significance is good; and when it is between 0.75 and 1.00, the level of clinical significance is excellent.

Reliability versus agreement

Agreement is defined as the degree to which an assessment method produces the same consequence with repeated administration (Gisev et al., 2013). From definitions of reliability and agreement, they are different concepts and are measured differently (Tinsley and Weiss, 1975). Estimating both of them is a common objective of many research studies (Shrout, 1998, de Vet et al., 2006, Kottner et al., 2011, Gisev et al., 2013). According to Tinsley and Weiss (1975) and Kottner et al., (2011), there is often confusion between them in the literature because they have different concepts. Kottner et al., (2011), suggested that the *"reliability and agreement are not fixed properties of measurement tools but, rather, are the product of interactions between the tools, the subjects/objects, and the context of the assessment. Reliability and agreement estimates are affected by various sources of variability in the measurement setting (e.g. rater and sample characteristics, type of instrument, administration process) and the statistical approach (e.g. assumptions concerning measurement level, statistical model)".* Thus, the conclusion of a given study is only interpretable if the measurement setting is clearly described and explained both statistically and graphically. They also reported that the reliability and agreement should be calculated differently (Kottner et al., 2011). On the other hand, Caro et al., (1979) demonstrated that the reliability takes into account the amount of agreement that could be expected to occur through the chance. Thus, the difference between the concepts of reliability and agreement is not always clear in the published literature, even in the hands of experts. According to Gisev et al., (2013), there is an argument in the statistical literature about the application and relevance of the different tests to estimate reliability and agreement although this difference is not substantial.

Measurement of reliability/agreement

From the previous paragraphs, the selection of a suitable statistical test for reliability and agreement is important and dependant on the research question. The choice of test needs to be justified, bearing in mind the context and purpose of the study and ease of calculation and interpretation of the result. In addition, different types of reliability and agreement require different tests.

Weiner (2007) suggested that inter-rater reliability/agreement can be estimated using percentage of overall agreement and Kappa statistical test. For example, Ahmed et al., (2016) used average percentage agreement among the eight examiners for the pilot study to estimate inter-examiner agreement. According to Jakobsson and Westergren (2005) and Gisev et al., (2013), percentage agreement cannot be selected to estimate the level of reliability/agreement for two nominal or ordinal datasets because it does not consider the agreement expected by chance. Therefore, some studies preferred to use Kappa tests to estimate the agreement (Gisev et al., 2013). Weighted and un-weighted Kappa tests can be used to determine the agreement or reliability, if the number of examiners/data-sets is only two. Both of them consider percentage of agreement and percentage of agreement expected by chance. The difference between them is the un-weighted Kappa test does not take account of the degree of disagreement. Thus, zero weight is given to all disagreement values. On the other hand, weighted Kappa gives different weights for values which do not agree (Jakobsson and Westergren, 2005).

In addition to previous tests, a correlation between the different measurements is also used. These tests are Spearman correlation for categorical data and Pearson correlation for non-categorical data. These two tests have same concepts (Field, 2013). Gratton et al., (2016) used Spearman correlation to determine whether the excellent scores awarded from faculty members were also awarded by digital tooth preparation evaluation technology (E4D). According to Safrit (1976), Pearson correlation has at least three limitations in order to estimate reliability. First, the aim of the Pearson's correlation is to determine the relationship between two variables. Theoretically, this type of test is not appropriate to apply for the correlation of two measures from the same variable, such as the test and the retest scores for a concept. Second, it is difficult to determine the test-to-test variation when multiple tests are administered. For example, if a concept is measured three times repeatedly, three scores are obtained. Traditionally, three correlation coefficients for every two of these three scores would be calculated and examined. However, the correlation coefficient of all three scores cannot be generated at the same time. Intra-class correlations resolve this problem when three scores are examined simultaneously. Third, the Pearson's product-moment correlation cannot detect the existence of a systematic error (Yen and Lo, 2002). In order to determine the systematic error, a t-test can be used (Houston, 1983).

There are many studies that have used a t-test to measure reliability and systematic errors of non-categorical (continuous) data (Houston, 1983, DeVon et al., 2007). For example, Kateeb et al., (2016) used t-test to assess the accuracy of grades generated from visual inspection when compared to the digital grading system in order to estimate the reliability of the device and to detect systematic errors of two series of measurements (Kateeb et al., 2016). In fact, systematic errors between two series of measurements may arise over a period of time if an examiner's measuring method changes with experience. One series of measurements may be changed systematically from a series made at two separate occasions (Houston, 1983). According to Houston (1983), a t-test can only be used to measure these errors. He also recommended that the minimum sample number that can be used to detect systematic errors is 25 objects (Houston, 1983). Thus, a t-test cannot be used in order to estimate the reliability overall.

In addition, Sharaf et al., (2007) and Sherwood and Douglas (2014) preferred to use Wilcoxon signed rank test (categorical data) and Friedman test to estimate intra- and inter-examiner reliability (categorical or non-categorical data respectively). The Wilcoxon signed-rank test is a non-parametric statistical hypothesis test used when comparing two related samples, matched samples, or repeated measurements on a single sample to assess whether their population mean ranks differ. It can be used as an alternative to the paired t-test when the population cannot be assumed to be normally distributed, the data is categorical and the sample size is small (Sawilowsky and Blair, 1992, Meek et al., 2007). In this case, the Wilcoxon signed-rank test is more powerful than the t-test (Meek et al., 2007). However, this test has limitation when the difference between the groups is zero, and the observations are discarded (Pratt, 1959). The Friedman test is another statistical test used to estimate inter-examiner reliability. The Friedman test is also a non-parametric statistical test. It is used to test for differences between more than two groups or one group on three or more different occasions when the dependent variable being measured is ordinal or continuous data. The Friedman test is derived from ranks with no tie. In other words, the test is used by converting the original results to ranks in order to find the differences between groups (Gibbons and Chakraborti, 2011). Pearson's correlation, Spearman's correlation, t-test, Wilcoxon signed-rank test and the Freidman test are not proper tests to determine reliability or agreement. Some of tests compare the means of data between groups (e.g. t-test) or compare the means of rank (e.g. Wilcoxon signed-rank test) (Field, 2013).

The most common statistical test to determine intra- and inter-examiner reliability/agreement is the intra-class correlation (ICC). Intra-class correlation test (ICC) can be used for both categorical and non-categorical data to estimate reliability.

This type of test is better suited to determine the direction any differences between datasets may take (Van Stralen et al., 2008). If there are two or more datasets then ICC is an adequate method to measure the agreement and reliability among examiners (Jakobsson and Westergren, 2005, Gisev et al., 2013). There are three different models which can be used for the ICC. The particular model used is dependent on the nature of the study (Shrout and Fleiss, 1979, Rankin and Stokes, 1998). The most important interpretation of an intra-class correlation is that it is a measure of the proportion of a variance that is attributable to items measured or judged (McGraw and Wong, 1996). Intra-class correlation with absolute agreement can be used to assess the agreement while intra-class correlation with consistency can be used to assess the reliability. For example, Kateeb et al., (2016) used ICC to estimate agreement among the four examiners who assessed ninety-six teeth that were prepared for a ceramo-metal crown, and between the examiners and the digital grading software.

The analysis of reliability for measurements sometimes is not sufficient using only intra-class correlation if the data is continuous values. Intra-class correlation (ICC) can produce misleading results. For instance, the value of ICC may be low, if the sample is homogeneous (Atkinson and Nevill, 1998, Rankin and Stokes, 1998, Bland and Altman, 1999, Hopkins 2000, Lexell and Downham, 2005). Therefore, a Bland and Altman plot can be used with ICC in order to determine the agreement and systematic error. A Bland and Altman plot is a graphical representation used to estimate agreement and systematic errors between two continuous data (Bland and Altman, 1999).

For example, Seo et al., (2014), used Bland and Altman plots to evaluate the overall agreements among different measuring methods (drawing protractor, digital protractor and Computer-Aided Design CAD) for the abutment convergence angle of plastic right maxillary canines. These teeth were prepared for metal-ceramic crowns. Thereafter, intra-class correlation coefficient (ICC) was used to evaluate the reliabilities for the three different methods (Seo et al., 2014).

From Tables 1.3 and 1.4, the reliability/agreement is usually calculated using the intra-class correlation coefficient (ICC) for nominal, ordinal, ratio or continuous data, while the kappa statistic for only nominal or ordinal variables (Kottner et al., 2011, Gisev et al., 2013). From Table 1.4, the agreement between two nominal or ordinal datasets is usually calculated with the proportions of agreement while the agreement between more than two nominal or ordinal datasets is usually estimated with ICC. ICC is also used to evaluate the agreement between two or more continuous datasets (Kottner et al., 2011). A Bland and Altman plot is used to determine the limits of agreement and systemic differences for two (continuous) measurements or two different methods (Rankin and Stokes, 1998, Bland and Altman, 1999, Kottner et al., 2011, Giesv et al., 2013).

Table 1.3 lists frequently applied statistical approaches which are arranged to estimate reliability/agreement of one examiner on two separate occasions or for two examiners on one occasion (Gisev et al., 2013).

Table 1.3 Examples of inter- and intra-examiner indices suitable for use with various types of data

	Level of measurement					
	Nominal/categorical data		*Ordinal data*		*Interval and ratio data*	
Inter- or intra-examiner indices	2 examiners/ 2 occasions	>2 examiners/ >2 occasions	2 examiners/ 2 occasions	>2 examiners/ >2 occasions	2 examiners/ 2 occasions	>2 examiners/ >2 occasions
	Cohen's Kappa ICC Weighted Kappa	Fleiss Kappa ICC	Weighted Kappa ICC	Kendall coefficient of concordance ICC	Bland and Altman plots ICC	ICC

Source (Gisev et al., 2013).

Table 1.4 lists frequently applied statistical approaches which are arranged to indicate which tests are used to determine reliability and which tests are used to determine agreement (Kottner et al., 2011).

Table 1.4 Statistical methods for analysing inter-rater/intra-rater reliability and agreement studies

Level of measurement	Reliability measurement	Agreement measurement
Nominal data	Kappa statistics	Proportions of agreement
Ordinal data	Intra-class correlation (ICC) Matrix of kappa coefficients Weighted kappa	Proportions of agreement
Continuous data	Intra-class correlation (ICC)	Proportions of agreement (ranges) Standard errors of measurement (SEM) Coefficients of variation (c.v.) Bland Altman plots

Source (Kottner et al., 2011).

Intra- and inter-examiner reliability/agreement

Intra-examiner reliability or agreement describes the consistency or agreement of a single examiner in grading the same sample on two different occasions (Dhuru et al., 1978). Inter-examiner reliability or agreement estimates the degree of consistency or agreement among the examiners when they assess the performance of the same group of students on the same task (Brown et al., 1999).

Most of the studies were focussed on inter-examiner reliability/agreement more than intra-examiner reliability/agreement. Lilley et al., (1968) and Deranleau et al., (1983) concluded that intra-examiner variability is less than inter-examiner reliability/agreement. Therefore, the majority of the studies focused more on the inter-examiner reliability/agreement than intra-examiner reliability/agreement.

The following few pages will concentrate on the impact of the tools, rating scale and examiner ability on the reliability and agreement.

Reliability/agreement according to the type of assessment tools

Several studies tried to improve the agreement in order to assess performance of the students in pre-clinical simulations by using different methods which could be broadly categorised as: the 'glance and grade', 'checklist', and 'computer software with devices' (e.g. Prepassiant and Opto-Electronic devices) (Taylor et al., 2013).

Manogue et al., (2001) reported that the 'glance and grade' method is used widely for assessment in dental education because it is recommended by assessors. This method indicates a subjective global assessment of the student performance. Several studies compared the 'glance and grade' method which provide a global mark with analytical methods which used defined criteria (e.g. Checklist and Checklist with specific criteria methods) to evaluate dental preparations (Taylor et al., 2013). Vann et al., (1983) compared the 'glance and grade' method with the checklist and the checklist with specific criteria methods to clarify the work of Goepfred and Herber (1980). Goepfred and Herber (1980) developed and tested the efficiency of an analytical system for evaluating class II cavity preparation on primary teeth. They used an analytical system (e.g. checklist) to reduce the subjectivity of clinical assessment and introduced objective measures to improve examiner agreement. The result of Goepfred and Herber (1980) was that there was improvement in intra- and inter-examiner agreement. The findings of Goepfred and Herber (1980) were not tested for statistical significance. Vann et al., (1983) also used a checklist with criteria, a similar number of examiners, the same grades and descriptions, but the findings were different. They found that there was no method which enhances inter-examiner agreement, although the checklist with criteria did improve intra-examiner agreement. Fuller in 1972 compared the 'glance and grade' method with the use of preparation models and a checklist for 67 class II cavity preparations evaluated by eight examiners. The correlation between these methods demonstrated that there were significant differences between them (Fuller, 1972).

Sharaf et al., (2007) compared the 'glance and grade' method and the checklist with criterion method for 240 cavity preparations evaluated by three examiners. The findings of this study were that there was statistically significant inter-examiner variability in 87% of cases using both assessment methods. They also concluded that the level of intra-examiner variability for three examiners was not statistically significant for most of the preparations using the glance and grade method, and criteria and checklist method. They evaluated different types of cavity design on plastic primary teeth completed by thirty dental students in paediatric dentistry preclinical laboratory sessions. For the first evaluation, each examiner graded the preparation using the glance and grade method. After three days, the preparation was re-evaluated again using the same method to measure intra-examiner variability. The third evaluation was performed blindly and graded using the criteria and checklist method with an explorer to verify cavity form and dimensions. The results indicated that the problem of inter-examiner reliability still exists (Sharaf et al., 2007). Although the reliability was not always improved by using the checklist, Sherwood and Douglas in 2014 concluded that the 'glance and grade' method of assessment should be changed to objective checklist criteria scoring method which decreases the examiner variability.

Sherwood and Douglas (2014) recommended that preclinical operative work of students be assessed by objective checklist criteria scoring and it should be provided after sufficient training and calibration sessions to reduce examiner variability. The study assessed 41 undergraduate students who were in second year of study, preparing two class II disto-occlusal amalgam cavity preparations performed on plastic typodont left lower first molar and left upper first molar teeth. These students performed the cavity preparations after a one-hour didactic lecture class and a one-hour demonstration session on the class II disto-occlusal preparation. The preparations were assessed by four, blinded and independent examiners using two methods of scoring; glance and grade method and objective checklist scoring method with explorer and mouth mirror. One week after the first evaluation, the preparations were again assessed by the same examiners for a second time using the same methods of scoring. Intra- and inter-examiner reliability were calculated using Wilcoxon signed rank test and Friedman test, respectively. Thus, the results of this study demonstrated that the glance and grade method was more unreliable with both intra- and inter-examiner consistency being poor compared with objective checklist criteria scoring (Sherwood and Douglas, 2014).

According to Taylor et al., (2013), *"Given the subjectivity associated with human assessors, attention has been given to developing more objective methods of assessment"*. In order to improve reliability, Schiff et al., (1975) invented an instrument called the 'pulpal floor measuring instrument'. It was used to evaluate the depth, smoothness and flatness of the pulpal floor. They concluded that inter-examiner reliability (retest reliability) of this study ranged from 0.81 to 0.99 by using the instrument, while a retest reliability coefficient as determined by subjective instructor rating ranged from 0.66 to 0.89. So, inter-examiner reliability was improved from 0.81 to 0.99, compared to 0.66 to 0.89 for intra-examiner reliability (Schiff et al., 1975). Similarly, Haj-Ali and Feil (2006) and Ahmed et al., (2016) concluded that inter-examiner reliability was improved by using a periodontal probe as an assessment instrument along with a checklist. From the last three studies, theses authors created checklists according to the dimensions of a tooth preparation.

In the last two decades, new assessment methods have appeared to assess dental preparations objectively such as computer assisted learning or computer assisted simulation system (Pollard and Davenport, 1993, Grigg and Stephens, 1998). Examples of these devices include, E4D compare software (Renne et al., 2013), DentSim by Denx (Rose et al., 1999, Welk et al., 2008), Virtual Reality Dental Training System by Novint

(Buchanan, 2001, Jasinevicius et al., 2004), PreAssistant by Kavo (Arnetzl and Dornhofer, 2004, Kournetas et al., 2004, Cardoso et al., 2006) and the Cavity Preparation Skill Evaluation System (CPSES) (Zou et al., 2016). All of these devices have utilised the dimensions of a given tooth preparation.

According to Taylor et al., (2013), *"the PreAssistant (Kavo, Germany) is a scanner, designed to objectively assess typodont teeth. It scans model teeth by photographing them from different angles and light projectors".* The software provides visual 3D images of tooth preparation. This device can be used to compare tooth preparations visually or by calculating the dimensions of tooth preparations. The PreAssistant device cannot assess surface roughness and continuity of the finish line. In addition, it can only provide a series of measurements rather than overall assessment of tooth preparation (Taylor et al., 2013).

Kournetas et al., (2004) arranged a pilot study to assess the reliability of the PreAssistant device. They scanned four full crown preparation teeth and four unprepared teeth many times both with and without changing position of the tooth. The aim of the pilot study was to determine the minimum magnitude difference between images that can be detected by the human eye. The authors concluded that 100 µm (0.1mm) to 200 µm (0.2mm) can be detected by using this device. In addition, these eight teeth were scanned at six different angles (i.e. six planes) to assess the reliability of the device. They concluded that the repeatability of the measurements (intra-examiner reliability) was more accurate than reproducibility (Inter-examiner reliability). According to Kournetas et al., (2004), the tooth can be mounted in different positions by using this device. Thus, the variations were expected. The authors suggested that the mean accuracy of the device was 89 µm (<0.1 mm) which was accepted for education aims (Kournetas et al., 2004). Cardoso et al., (2006) also used this device to compare an ideal full ceramic crown preparation on tooth 36 by the faculty member with the same design and tooth preparation completed by a student. Twenty five tooth preparations were assessed using a visual method by the examiners and a PreAssistant device. A 0-20 grading scale was used to assess the tooth preparations. The grades were provided by the examiner while the device provided a data sheet containing the preparation analysis and feedback which was then given to the student. They concluded that the device can assess only 70% objectively while 30% of the assessment can be evaluated by the assessor visually. This conclusion is supported by the study of Arnetzl and Dornhofer (2004). According to Cardoso et al., (2006), the creation of a mathematical formula to assess a tooth preparation using the PreAssistant was difficult. Although this device provided objective assessment for the students, Cardoso et al., (2006) did not recommend using this device alone for the final classification of the student's performance.

There is another scanner with software, called 'E4D Compare software', which is used to reduce subjective assessment. According to Renne et al., (2013), 50 maxillary right first molar teeth were prepared to receive all-ceramic crowns as a practical examination by the students. The preparations were graded on a 0-100 scale by three experienced and calibrated faculty members. A gold standard preparation was selected and scanned according to the examiner assessment from the 50 tooth preparations. By using E4D Compare software, the remaining 49 tooth preparations were scanned and compared with the gold standard preparation. In this study, the authors selected 300 μm (0.3mm) as an acceptable range that the student preparations could vary from the ideal. In addition, they provided two methods to assess tooth preparations. In the first method the examiners provided grades for the 50 preparations in comparison with the gold standard preparation visually. The second method was comparison of 3D images of the 49 tooth preparations with the 3D image of the gold standard preparation by using E4D Compare software. Renne et al., (2013) concluded that although this type of software reduced the subjectivity and provided accurate and consistent assessment for the students according to the gold standard, more research needs to be done to reduce the subjectivity of selection of the gold standard preparation.

DentSim (Image Navigation Ltd, USA) is another device which has been used to teach crown preparations and endodontic access cavities. This system is used with a phantom head simulation unit. Online visual tracking of a preparation, real time feedback and evaluation are given to the student by using this system. It provides a simulated clinical environment for the students (Taylor et al., 2013). Welk et al., (2004) suggested that training on dental preparation using DentSim system will enhance student skills. Jasinevicius et al., (2004) supported suggestion of Welk et al., (2004). They concluded that the student who trained on the DentSim system received higher scores than the students who trained on traditional phantom heads. Although this device can provide objective tracking of a tooth preparation, the final assessment needs a member of staff which still leaves the problem of subjectivity associated with the assessment.

Zou et al., (2016) used the 'Cavity Preparation Skill Evaluation System' (CPSES) to provide an objective and accurate measurement for the class I cavity preparation thus avoiding human errors. It is a three-dimensional laser scanning tool with an image-processing system. According to Zou et al., (2016), > 90% of students considered that CPSES created a realistic simulation and provided appropriate guidance as well as targeted and objective recommendations; the system could consistently and reliably scan a student's tooth preparation and compare it with a theoretical ideal to provide feedback and objective feedback. Furthermore, it helped students better understand the desirable parameters of occlusal cavity preparation, encouraged student's self-paced learning and development of independent practise.

All of these assessment devices provided higher reliability than traditional assessment methods. The final assessment still needs an examiner's opinion which is part of subjectivity associated with the assessment (Cardoso et al., 2006). In addition, validity and agreement are impacted potentially, if an assessment tool with poor reliability is used. For example, global assessment with scoring preclinical dental procedures does not include description of the level of performance (AlHumaid et al., 2016).

Reliability/agreement according to the nature of the rating scale:

Although the reliability and agreement of a checklist method is better than the 'glance and grade' method, the variability of reliability and agreement still exists. Brown (1930) suggested that the rating scale must qualify as follows: a) it must be valid, b) it must be objective and c) it must be reliable; when given two or more times to the same group. Some studies tried to enhance the reliability and/or agreement through changing the nature of the rating scale within the different methods. These following studies tried to enhance inter-examiner reliability/agreement by changing the number of categories in a rating scale.

Natkin and Guild (1967) reported that the use of a nine-point rating scale decreased the inter-examiner reliability (55%). Each occasion consisted of testing ten randomly-selected student projects followed by a discussion of the criteria established for those projects which related to endodontic procedures. The authors of this study felt that the reason for the variability was due to *"instructors struggling to recognise errors and assigning an appropriate level of severity to them"* (Natkin and Guild, 1967).

Houpt and Kress (1973) compared three different rating scales for checklists: a two-point scale with two specific points (incorrect/correct), a five-point scale with upper and lower limits specified and five-point scale with detailed descriptions of each level. They concluded that the use of a two-point rating scale was found to have more inter-examiner agreement than the use of a five-point rating scale. They discovered that a five-point rating scale with descriptions would be of more value in teaching.

Hinkelman and Long (1973) used two different rating scales; a two-point rating scale (pass and fail) and a three-point rating scale (‘no improvement necessary’, ‘clinically acceptable’ and ‘clinically un-acceptable and un-correctable’). They concluded that a two-point rating scale slightly improved inter-examiner agreement compared with a three-point rating scale. However, the authors recommended using three-point rating scale because it was more useful for ranking according to ability (Hinkelman and Long, 1973).

Gaines et al., (1974) also compared two types of checklists; the first consisted of six assessment items each scoring 0 to 5. The second checklist contained six assessment items with objective statement for each score in each item. The study involved seven examiners with unstated levels of experience, evaluating only eight preparations. The second checklist demonstrated improved inter-examiner agreement (0.56) while the first checklist did not (0.26) (Gaines et al., 1974).

Deranleau et al., (1983) concluded that there was no difference in agreement between checklists with two- and three-point scales for evaluation of class II mesio-occlusal cavity preparation and porcelain jacket crown wax ups using five examiners.

Helft et al., (1987) used a five point rating scale to evaluate the marginal adaption and thickness of cemented crown on extracted teeth. The agreement was also poor. Helft et al., (1987), suggested the reason might be a poor rating scale which had no objective criteria. To improve inter-examiner correlation, definitions for each category of the rating scale is clearly important to reduce miss-interpretations, measurement error, and variance (Helft et al., 1987).

Guenzel et al., (1995) used a three-point scale to determine whether or not rubric organization provides more information as a feedback for students. They classified scale points into 'super', 'clinically-acceptable' and, 'standard not met' for preclinical amalgam cavity preparations. Their conclusion was that the rubric assisted the student to acquire information, practise and receive feedback in instructional tasks (Guenzel et al., 1995) and improvement in inter-examiner agreement (Haj-Ali and Feil, 2006).

According to Haj-Ali and Feil (2006), they suggested that before starting an assessment session, the development of grading system is important. They developed a 'grade form' for class II amalgam cavity preparation according to Knight's recommendations from a publication in 1997. Haj-Ali and Feil (2006) divided the criteria into three main categories related to class II amalgam cavity. They suggested that calibration is more successful when the number of categories on the rating scale is limited. Because of that each criteria of category had three levels of outcome which were specifically described as: 'ideal preparation', 'acceptable preparation' and 'standards not met'. Thirty class II cavities were assessed by nine examiners on three different occasions. They concluded that, with training, 'grade form', inter-examiner reliability was improved (Haj-Ali and Feil, 2006).

The number of categories on a rating scale has impact on the examiner reliability and agreement. Two-point rating scales (pass and fail) could be used to provide a certificate for students as evidence of passing the course (Houpt and Kress, 1973) while more than two-point rating scale provides more useful information for the students and improves students' skills (Hinkelman and Long, 1973).

From the previous paragraphs, the optimal number of rating scale points for maximized operational feedback instructions to students is from 3 to 5 points (Fernandez, 1967, Houpt and Kress, 1973, Lindvall and Nitko, 1975). However, increasing the number of scoring levels induces discrimination problems, and it decreases the accuracy of scoring (Goepferd and Kerber, 1980). Thus, no definite conclusions can be made with respect to the optimal number of scoring categories. In addition, using an unclear and poorly designed rating scale might impact on the reliability and agreement of the assessment result as well as examiner ability to assess (Helft et al., 1987).

According to Hauser and Bowen (2009), *"Rubrics and rating scales designed with specifics related to each step in a preclinical skill performance can be valuable tools for both the learner and the evaluator."* In addition, using standardised rubrics or rating scales as a training and calibration for the examiner might enhance the agreement (Garland and Newell, 2009). Knight in 1997 also suggested that preclinical dental students learned first to recognise defining features of the skill to be learned. These criteria should be included in rubric form and/or rating scales.

<u>Reliability/agreement according to examiner ability to assess</u>

There are other reasons reported for reliability/agreement variability which include:

- The number of training sessions (Natkin and Guild, 1967, Lilley et al., 1968, Hinkleman and Long, 1973, Houpt and Kress, 1973, Geopfred and Kerber, 1980, Scruggs et al., 1989, Haj-Ali and Feil, 2006, Ahmed et al., 2016), and,
- Examiner experience (Lilley et al., 1968, Fuller, 1972, Houpt and Kress, 1973, Jenkins et al., 1996).

Even-though the study of Natkin and Guild (1967) failed to demonstrate a correlation between the examiner experience and agreement scores, they reported a statistically significant increase in inter-examiner reliability with assessment of endodontic procedures after training sessions (Natkin and Guild, 1967).

In addition, Lilley et al., (1968) provided 37 amalgam preparations for three examiners to assess these preparations in different stages on two different occasions. They provided training to discuss criterion description for each score and to categorise fault for the examiners before commencing assessment sessions. They concluded that the intra-examiner reliability was high for the most experienced examiner. Although inter-examiner agreement of the scores related to the cavity preparation stage was slightly improved, it was not improved for any of the other stages. They found a number of pass-fail differences existed between the examiners after the training session (Lilley et al., 1968).

Fuller (1972) tried to determine the effect of the examiner experience on the intra-examiner coefficients using Spearman's rank order correlation coefficient. Instructors with less experience had lower intra-examiner coefficient than instructors with the more experience as examiners, 0.472 and 0.831, respectively. In addition, there is no evidence that reliability for raters on ratings of the same product can be enhanced by training (Fuller, 1972).

Houpt and Kress (1973) provided cavity preparations for senior staff to assess them with immediate feedback of the correct score. Senior staff trained on the assessment before commencing scoring session. The result was that the training enhanced total scores agreement among the examiners but it was not between criteria. They also highlighted the impact of experience on the reliability. They demonstrated that experience improves intra- and inter-examiner agreement (Houpt and Kress, 1973).

Hinkleman and Long (1973) also reported that there was no or little impact of training on the examiner agreement and consistency. They found that inter-examiner reliability was enhanced by using a two-point scoring system over a three-point scale. They failed to determine any impact of examiner experience on the reliability. On the other hand, they concluded that the examiners who had less experience preferred the two-point scoring system while more-experienced examiners preferred the three-point scale to recognise perfect work (Hinkleman and Long, 1973). In addition, Patridge and Mast (1978) also found there was little or no effect of the examiners' training to evaluate the preparations. This result supports the study result of Hinkleman and Long (1973).

According to Geopfred and Kerber (1980), the examiners used 'glance and grade' method for annual faculty training with an analytical checklist to assess class II preparations on primary teeth. The agreement among the examiners using the checklist was improved following the training session. Geopfred and Kerber (1980) reported that the training might not be the main reason for this result. This may be due to repetition of assessing the same teeth. However, the effect of training to improve inter- and intra-reliability of scoring tooth preparations was uncertain, since such evaluations also included qualitative judgements.

The method which is used to calibrate examiners may be more important than the experience of the calibrator when planning a training sessions. Scruggs et al., (1989) concluded this result when they tried to determine the difference between calibrated and non-calibrated examiners. An analytical checklist was provided for a group calibrated by an expert and a group calibrated by a non-expert to assess dental sealants. The authors found that the examiners who were calibrated by the expert actually decreased inter-examiner reliability after training session (Scruggs et al., 1989).

Jenkins et al., (1996) tried to identify the impact of the experience of the examiner on inter-examiner agreement. Assessment of class II amalgam cavity preparations was considered in their study. They provided 13-point grading system for examiners to evaluate the preparations. They found that variability of up to seven marks was noted. In addition, the level of pass-fail differences was unrelated to the experience of the examiner (Jenkins et al., 1996).

According to Haj-Ali and Feil (2006), the calibration with gold standard scores is part of examiner training. They suggested that before starting assessment sessions, development of a grading system is important. They developed a 'grade form' for class II amalgam cavity preparation according to Knight recommendations in 1997. Two instructors assessed a random sample of 30 class II amalgam cavities to create gold standard scores for the following three tests. Nine examiners assessed the same 30 preparations on three different occasions: prior to calibration training, immediately following training and the third occasion was ten weeks later. They reported that calibration with gold standard scores assisted the examiners to improve inter-examiner reliability (Haj-Ali and Feil, 2006).

Ahmed et al., (2016), used a similar method to conclude that inter-examiner reliability may be enhanced through a focused process of faculty calibration. They provided 32 class II amalgam cavities to be assessed by eight examiners at two different occasions. On the first occasion, each examiner was asked to assess 32 tooth preparations immediately after calibration session. On the second occasion, the examiners were asked to assess the same tooth preparations after an average time interval of six months without a further calibration session. Inter- and intra-examiner agreement were analysed using Weighted Kappa and McNemar analysis. They concluded that inter-examiner reliability was improved after the calibration session while intra- and inter-examiner reliability was decreased at six months. Therefore, Ahmed et al., (2016) suggested that more frequent calibration sessions may enhance the level of reliability among the course faculty.

In general, most of the studies tried to improve the agreement and reliability. Unfortunately, a number of studies have found that agreement and/or reliability for assessment in dentistry and related disciplines is variable and sometimes low. In addition, some studies used inappropriate statistical tests to determine the agreement and reliability.

In relation to both validity and reliability, the data which can be used to determine reliability and validity are important to indicate how useful the assessment is to evaluate student performance. DeVon et al., (2007) recommended that content validity of an assessment method can be indicated by using the process proposed by Lynn (1986) but the number of underlying dimensions, the number of items and the theoretical framework should also be considered. In addition, the scores of an assessment method should be correlated to related criteria at the same, or different, times for the same objects in order to confirm criterion-related validity. For reliability and agreement, test-retest or intra-rater reliability/agreement procedure should also be performed if stable characteristics are to be measured on two separate occasions (DeVon et al., 2007). Finally, inter-rater reliability/agreement tests should be used to determine the reliability among raters (Weiner, 2007).

The concept of the validity and reliability is explained in Figure 1.3. For example, if someone aims a gun at a target and fires three rounds, the gun is reliable if all three hit the target. The gun is unreliable if one round went above the target, the next round below and the last above again. Figure 1.3 shows the four combinations possible between validity and reliability.

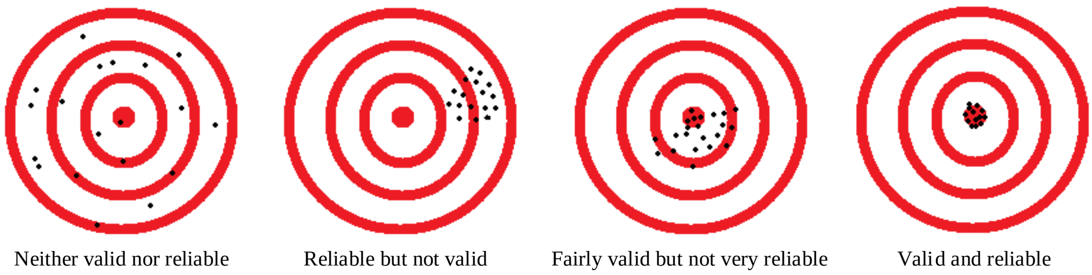


Figure 1.3 Diagram illustrate the possible combinations of validity and reliability (Bolarinwa, 2015)

1.3.3 Cost-effectiveness of assessment

Valid and reliable assessment methods are definitely costly. It takes time to design the assessment, take the assessment and score the assessment as well as staff training (Bourisicot et al., 2011). Investing in the assessment is investment in teaching and learning (van der Vleuten, 1996).

From the previous paragraphs, the ‘glance and grade’ method did not always improve examiner reliability or agreement. Although it is cheap, it does not produce effective teaching. To reduce the subjectivity and examiner scoring variability, Sherwood and Douglas (2014) recommended using objective checklist criteria scoring methods to evaluate preclinical operative work of students. This type of assessment is not expensive. Therefore, the checklist is commonly used in dental education (Taylor et al., 2013).

According to Brown (1930), using the checklists with subjective statement (i.e. criteria and their level of performance) will not enhance agreement. Gaines et al., (1974) compared two checklists; the first consisted of subjective statements and the second checklist contained objective statements. They concluded that the agreement among the examiners using the second (objective) checklist had better agreement than the first checklist (Gaines, 1974). This means that if a checklist has more objective criteria, it will provide more beneficial feedback for instructional purpose (Sherwood and Douglas, 2014).

In addition, there are studies which have utilised computer software and devices to reduce errors produced by human assessment as well as to provide consistent, accurate and objective assessment for students (Pollard and Davenport, 1993, Grigg and Stephens, 1998, Rose et al., 1999, Buchanan, 2001, Cardoso et al., 2006, Renne et al., 2013, Kateeb et al., 2016). These software and devices used the components of a checklist. An example is the Kavo PreAssistant system. This device provides valid and reliable data for most studies by utilising some of anatomical feature measurements of full crown preparations. It provided 70% of the feedback and grades for dental students. The remaining features were assessed subjectively. According to Cardoso et al., (2006), there are many merits of using the Kavo PreAssistant system and these include: i) providing better objective assessment and feedback than visual evaluation; ii) the software provides more details and detects mistakes easily, and iii) it produced fewer problems than visual evaluation in calibration of different examiners. On the other hand, there are also disadvantages which include: i) 30% of feedback is provided subjectively, ii) the creation of a mathematical formula to assess a full ceramic crown preparation when using Kavo PrepAssistant system (software version 1.05) was difficult, iii) the manufacturers and programmers must be aware of real needs of pre-clinical teaching schools, iv) it takes long time to scan and assess one model (>2 minutes per model), and v) the cost of this machine is high (Cardoso et al., 2006).

Knight (2007) suggested assessment methods might be still good value, if the cost of the assessment method might be not overtake teaching costs. Schuwirth and van der Vleuten (2010) suggested that *"a pre-requisite for a cost-effective assessment programme is an explicit description of its goals, both in terms of what is to be assessed and how it is to be assessed. Only then can an evaluation be conducted into whether the programme is optimally cost-effective"*.

1.3.4 Acceptability of assessment

Acceptability of the assessment method is another requirement for accurate results and progress (Hays et al., 2002). Even the best assessment method is useless if the teachers and students do not accept it and are not familiar with it (Schuwirth and van der Vleuten, 2010, Norcini et al, 2011). If the assessment method is accepted by students and teachers, this will guide the students to the objectives of the course as well as provide them fair feedback (Norman et al., 1991). Therefore, it is important to clarify the aim and content of the assessment method to achieve acceptance. For example, a clearly defined rating scale and a checklist are essential components to provide the student with accurate feedback (Taylor et al., 2013, Ahmed et al., 2016). According to Houpt and Kress (1973), examiners were more reliable in their judgments with two-point scale, while the five-point scale with description was more beneficial for instructional purposes (Houpt and Kress, 1973). Hinkelman and Long (1973) and Deranleau et al., (1983) concluded the same result. These three studies also indicated that a two-point scale provided greater examiner agreement level but they highlighted that it provided little information as a feedback for the students. If the student did not receive proper and accurate feedback from the teacher, this might impact, negatively, on the students' learning by focusing only on pass and fail scores. Therefore, Hinkelman and Long (1973) and Deranleau et al., (1983) preferred a three-point scale as it provided better comments for the students.

In addition to the number of categories on a rating scale, a description of the criteria and the levels of performance are also important (Knight, 1997). Paskins et al., (2010) evaluated the use of criteria which could be simply assessed and had limited chance for mis-interpretation, based on a checklist formulated to evaluate the management of simulated respiratory and cardiac emergencies. Two examiners used the checklist and demonstrated a high agreement value among the examiners (>0.9). Because the checklist provided clear information for the examiners, it also provided reproducible outcomes. The authors concluded that the checklist list was valid (Paskins et al., 2010). Helft et al., (1987) and Paskins et al., (2010) suggested that defined criteria will reduce variance between examiners. Therefore, cheap, clear, valid and reliable checklists will provide acceptable and accurate assessments for students.

In conclusion, acceptability of the assessment method for the teachers and students is fundamental to provide instructional advantage in teaching and determining components of the student's performance.

1.3.5 Educational impact of assessment

According to Schuwirth and Van der Vleuten (2010) the idiom, "*'Students don't do what you expect, student do what you inspect' epitomises the educational impact of assessment"*. Thus, the assessment has a major effect on student's learning. In addition, *"the driving influence of assessment is a powerful tool to ensure that students learn what, and how, teacher wants them to learn"* (Schuwirth and van der Vleuten, 2010).

From these sentences, development of an assessment method should be considered in relation to content of what the examiner wishes to assess as well as the format of the assessment method. Therefore, 'Constructive alignment' between the educational and assessment aims must be considered when an assessment method is developed (Biggs and Tang, 2011). If they are not aligned, the assessment aims will not prevail (Van der Vleuten, 1996). If the examiner provides valid, reliable and acceptable assessment methods, the examiner can guide the students' learning and skills in a positive way.

From the five criteria of assessment and the results of previous studies, the assessment methods used in Clinical Skills Laboratories are widely variable. The 'glance and grade' method and 'objective checklist' are the most common assessment tools which are used (Vann et al., 1983, Manogue et al., 2001). Additionally, they were used as the fundamental elements for computer assisted systems. Therefore, the valid and reliable checklist is selected in this study because it is the core of the assessment at the clinical and preclinical courses.

1.4 Checklist for assessment

Conservative and fixed prosthodontic concepts and techniques are introduced to second and third year dental students through participation in a clinical skills laboratory courses at Dundee Dental School. The most common assessment methods used are the 'glance and grade' and checklist methods. Meckenzie in 1973 suggested that the 'glance and grade' assessment should be supplemented with other forms of assessment, such as a checklist, to improve reliability (Mackenzie, 1973). In addition, Sherwood and Douglas (2014) concluded that the 'glance and grade' method alone is not accurate to assess students' performance. Thus, the 'glance and grade' should be associated with the checklist to provide more accurate assessment and feedback for undergraduate dental students.

The checklist is used to acquire and develop students' clinical skills through the hands-on training on phantom heads by providing feedback. Dental students, during their clinical skills laboratory module, are expected to develop several skills prior to dealing with real patients. Tooth preparation upon plastic teeth which are attached to a phantom head is one of the most basic elements in the clinical skills laboratory. Most universities utilise checklists to assess the students' performance on plastic teeth during or at the end of the course (Kramer et al., 2009). Additionally, the purpose of using the checklist is to guide the students through the aims of the course by providing valid, reliable, acceptable, cheap and effective feedback and grades.

The contents of a checklist have been utilised to establish several computer-assisted systems. According to Rosenberg et al., (2003), these systems have proven to be as effective as other methods of teaching. Thus, development of a checklist is the core of many teaching and assessment methods. In order to establish a computer-assisted learning system, the dimensions of the features of a tooth preparation are utilised (Cardoso et al., 2006, Esser et al., 2006). According to Cardoso et al., (2006) and Esser et al., (2006), their system used the measurements of full crown preparations to provide objective assessment for students (e.g. Kavo PrepAssisstant device).

The dimensions of tooth preparations were also utilised in most of the studies which focused on enhancing calibration between examiners (Albino et al., 2008). Therefore, dimensions are important to develop effective assessment methods (Haj-Ali and Feil, 2006).

In addition to dimensions of tooth preparations, Knight (1997) suggested some recommendations in order to calibrate the examiners. He established the prescriptions for the grading system and also specified several elements that define acceptable criteria used in the grading context. He suggested that the criteria should have the following properties:

1. have individual and collective validity,
2. be independent of each other,
3. be sequenced based on the natural order of the procedure itself,
4. be capable of being objectively tested and,
5. provide clearly defined levels of performance which are called ‘descriptors’ for each criterion.

In addition to the previous points, calibration is more successful when the number of points on a rating scale is limited and when the determination of a rating for a particular criterion has been operationalized (e.g. a grade of 3 is given if the occlusal depth of the class I amalgam cavity is less than 1.5mm, defined as ‘Shallow’ and estimated by using a periodontal probe) (Knight, 1997).

Haj-Ali and Feil (2006) and Ahmed et al., (2016) utilised the dimensions of tooth preparation to improve reliability, these studies did not state that the reliability or agreement with a valid gold standard had been achieved. In addition, there are other studies which concluded that a checklist did not always provide consistent results by improving criteria, rating scales and calibration (Taylor et al., 2013, Ahmed et al., 2016). The reason for this may be due to there being no purely objective assessment. A purely objective checklist is very difficult to create. Some features of tooth preparations are not measurable. Therefore, both objective and subjective criteria must be selected. According to Cardoso et al., 2006, Kavo PrepAssistant device provides 70% of assessment objectively through evaluation of tooth preparation dimensions. To improve objective components of assessment, some authors utilised instruments as assessment tools along with the checklists to produce more accurate assessment and feedback for the students (Haj-Ali and Feil, 2006, Ahmed et al., 2016). Thus, ideal/acceptable measurements or features from the body of available literature should be considered before assessing student preparations.

In this study, the class II amalgam cavity preparation and the full veneer gold crown preparation were selected as part of a key assessment in the clinical skills laboratory. In order to assess the student performance more objectively, ideal or acceptable feature measurements for these types of preparations were searched and selected from the available literature. Some of these measurements were selected from several studies (Haj-Ali and Feil, 2006, Park et al., 2009, Ahmed et al., 2016) and are shown in Table 1.5 and Table 1.6. The data in these tables demonstrate the most common ideal or acceptable measurements mentioned in the available literature relating to protocols for class II amalgam cavity and full veneer gold shell crown preparation. The dimensions and components of these types of tooth preparation are utilised in this study and divided into objective and subjective features (Chapter 6). The specific features of each type of tooth preparation are now discussed.

1.4.1 Class II amalgam preparation features

In 1908, Black defined interproximal caries as a lesion which occur in the area located slightly gingival to the point of contact between the posterior teeth (Black, 1914). Treatment of interproximal caries usually is limited to two choices: non-invasive (preventive) or invasive (restorative). Class II amalgam cavity preparation is invasive preparation technique. The shape of class II amalgam cavity has been changed considerably over the years. These modifications are strongly influenced by changes in caries prevalence, the introduction of fluoride, the use of more efficient cutting tools, and the development of dental material and techniques. The more modern preparations were developed at a time when the caries prevalence had started to decrease and when more efficient rotating cutting instruments and material had become available (Jokstad, 1992, Söderholm et al., 1998).

It is very difficult to identify the most ideal and/or accepted dimensions for class II amalgam cavity for posterior teeth because these dimensions i) are different from one study to another (i.e. principles of preparation), ii) depend on the shape of the tooth (i.e. molar or premolar) and, iii) depend on the type of tooth materials (i.e. natural or artificial tooth).

Table 1.5 illustrates the most common ideal dimensions for each feature of class II amalgam cavity preparation which have been collected from a narrative search of the literature and protocols.

Table 1.5 Measurements for features of class II amalgam cavity preparations

Summary of isthmus width (mm) found in literature and protocols of dental institutions			
Study/Protocol	*Year*	*Tooth type*	*Isthmus width or ratio*
G.V. Black	1914	Posterior teeth	1/3 of intercuspal distance
Gabel	1951	Molar	1/8 of intercuspal distance
Rodda	1972	Posterior teeth	¼ of intercuspal distance (~1.00 mm – 1.50 mm)
Jokstad and Mjör	1987	Posterior teeth	>1.00 mm and < 2:5 of intercuspal distance
Haj-Ali and Feil	2006	Posterior teeth	1.00 – 1.50 mm
Park et al.,	2009	Posterior teeth	1.00 – 2.00 mm
Akpata et al.,	2013	Molar	¼ to 1/8 of intercuspal distance
Ahmed et al.,	2016	Molar	1.00 mm and 1/3 of intercuspal distance
Summary of occlusal depth (mm) found in literature and protocols of dental institutions			
Study/Protocol	*Year*	*Tooth type*	*Occlusal depth*
G.V. Black	1914	Posterior teeth	2.00 mm
Jokstad and Mjör	1987	Posterior teeth	2.00 mm
Roberson et al.,	2002	Posterior teeth	1.50 – 2.00 mm
Haj-Ali and Feil	2006	Posterior teeth	1.50 – 2.00 mm
THE CONSORTIUM, Overseas Registration Examination, Part 2.	2012	Upper posterior tooth	~2.00 mm
Akpata et al.,	2013	Molar	1.50 – 2.00 mm
Summary of box depth (mesio-distal direction) (mm) found in literature and protocols of dental institutions			
Study/Protocol	*Year*	*Tooth type*	*Box depth*
Jokstad and Mjör	1987	Posterior teeth	1.00 – 1.50 mm
Baum et al.,	1995	Posterior teeth	1.20 mm
Roberson et al.,	2002	Posterior teeth	~1.00 mm (0.80 mm)
Haj-Ali and Feil	2006	Posterior teeth	1.00 – 1.50 mm
Park et al.,	2009	Posterior teeth	1.00 mm
Hilton et al.,	2013	Posterior teeth	1.50 mm
Quizlet	2017	Posterior teeth	1.20 – 1.50 mm
Summary of box depth (distance from marginal ridge to gingival floor) (mm) found in literature and protocols of dental institutions			
Study/Protocol	*Year*	*Tooth type*	*Box depth gingivally* Distance from marginal ridge to gingival floor
Jokstad and Mjör	1987	Posterior teeth	4.00 mm
Jokstad	1989	Premolar	3.20 – 3.60 mm
Baum et al.,	1995	Posterior teeth	3.00 – 4.00 mm or 1.00 to 2.00 mm below contact point
Quizlet	2017	Posterior teeth	3.50 – 4.00 mm
Summary of box width (Bucco-lingual/palatal or proximal extension of the box) (mm) found in literature and protocols of dental institutions			
Study/Protocol	*Year*	*Tooth type*	*Box width*
Jokstad and Mjör	1987	Molar	Bucco-lingual extension= >1:5 mm - <3:5*
Haj-Ali and Feil	2006	Posterior teeth	0.50 mm from adjacent tooth in buccal, lingual, and gingival
THE CONSORTIUM, Overseas Registration Examination, Part 2.	2012	Upper posterior tooth	Proximal contact clearance = ~0.50 mm
Ahmed et al.,	2016	Molar	Proximal contact clearance = 0.50 mm in all directions (buccal, lingual/palatal, and gingival)
Assessment of Clinical Skills	2016	Posterior tooth	Proximal and/or gingival margins clear adjacent teeth 0.50 mm or less
Summary of pulpal axial wall length (mm) found in literature and protocols of dental institutions			
Study/Protocol	*Year*	*Tooth type*	*Pulpal axial wall length*
Jokstad and Mjör	1987	Molar	< 2.00 mm
Satwik and Neelakantan	2016	Right lower molar tooth	~ 1.00 mm
Quizlet	2017	Molar teeth	1.50 – < 2.00 mm

Summary of marginal ridge width measurement (mm) found in literature and protocols of dental institutions			
Study/Protocol	*Year*	*Tooth type*	*Box width*
Roberson et al.,	2002	Premolar	1.6 mm
Haj-Ali and Feil	2006	Unknown Posterior teeth	≥ 1.00 mm
Satwik and Neelakantan	2016	Right lower molar tooth	2 mm

*Assess the bucco-lingual extension relative to the interproximal circumference.

Evaluation of the class II amalgam cavity preparation must be based on a definition of ideal cavity design (measurement) to assess students' outcomes. From Table 1.5, there is no one specific ideal class II cavity to compare with student's cavity. So, assessment of students' performance is usually dependent on those ideal cavity preparation features and the measurements which have been chosen by staff members in order to assess student's work more objectively. Of course cavity preparation on patients is led by the extent and spread of caries. These dimensions only test a student's dexterity and ability to cut stereotypical shapes. Thus, it is a test of the student's ability to do what they are taught to do. It is worth noting that different levels of precision are afforded in the various measurements of the ideal cavity reviewed in Table 1.5 as implied by the number of significant figures quoted. Is it not reasonable to expect students and examiners to be capable of measuring to 1/10th of a millimetre (mm)?

1.4.2 Full veneer gold shell crown preparation features:

The full veneer gold shell crown preparation is used to preserve the remaining tooth structure, for example after endodontic treatment (Ricketts and Bartlett, 2011). In the middle of the 20th century, indirect fabrication techniques for prostheses were developed after the creation of more accurate impression materials. These materials are used to take an impression of prepared tooth, and then cast to prepare a model of the tooth. The crown is constructed in wax on the model then cast in gold/metal. The development of dental instruments also made preparation of a tooth a much simpler and less time consuming procedure (Marra, 1970).

At the same time, several critical features of the full veneer gold shell crown preparation have been discussed in order to provide guidelines to develop tooth preparations and restorations (Goodacre et al., 2001) to preserve tooth structure (Diego, 1996). Goodacre et al., (2001) and Tiu et al., (2015) reviewed the features of preparations for fixed prostheses along with their historical basis and measuring method. Unfortunately, few studies have stressed the features of an ideal crown preparation (Goodacre et al., 2001). In addition, some studies have tried to analyse tooth preparation and describe features which were important for longevity of the prostheses (Goodacre et al., 2001, Tiu et al., 2015).

The ideal properties of a full metal crown preparation are variable from study to study but most of them agreed that gold/metal crown requires minimal tooth reduction. In addition, there is no substantial evidence to confirm recommended reduction of axial and occlusal surfaces (Goodacre et al., 2001).

The following Table shows the most acceptable dimensions for full metal crown preparations (Table 1.6).

Table 1.6 Measurements for features of full gold crown preparations

Summary of total occlusal convergence found in the literature and protocols of dental institutions			
Study/Protocol	*Year*	*Tooth type*	*Total occlusal convergence angle*
Prothero	1923	Molar	2° - 5° on one side (TOC = 4° – 10°)
Jørgensen	1955	Posterior teeth	2° – 5° on one side (TOC = 4° – 10°)
Tylman and Malone	1978	Molar	3° - 14°
Noonan and Goldfogel	1991	Molar	15.7° – 19.2°
Wilson and Chan	1994	Molar	6° - 12°
Goodacre et al.,	2001	Molar	10° - 20°
Ayad et al.,	2005	Posterior teeth	14.1° – 19.8°
O'Sullivan	2005	Molar	6° - 10°
Smith and Howe	2007	Molar	7.5° - 12°
THE CONSORTIUM, Overseas Registration Examination, Part 2.	2012	Posterior teeth	6° - 20°
Yoon et al.,	2014	First molar teeth	6.3° – 16.9°

Summary of occlusal reduction (functional cusps) found in literature			
Study/Protocol	*Year*	*Tooth type*	*Occlusal reduction (functional cusps)*
Blair et al.,	2002	Posterior teeth	1.50 mm
Rosenstiel et al.,	2006	Molar	≥ 1.50 mm
Ricketts and Bartlett	2011	Posterior teeth	< 2.00 mm
Summary of occlusal reduction (non-functional cusps) found in literature			
Study/Protocol	*Year*	*Tooth type*	*Occlusal reduction (non-functional cusps)*
Blair et al.,	2002	Posterior teeth	1.00 mm
Rosenstiel et al.,	2006	Molar	≥ 1.00 mm
Ricketts and Bartlett	2011	Posterior teeth	< 2.00 mm
Summary of finish line depth found in literature and protocols of dental institutions			
Study/Protocol	*Year*	*Tooth type*	*Finish line depth*
Blair et al.,	2002	Posterior teeth	>0.30 to ≤ 1.00 mm
Rosenstiel et al.,	2006	Posterior teeth	≥ 0.50 mm
THE CONSORTIUM, Overseas Registration Examination, Part 2.	2012	Posterior teeth	0.50 – 0.75 mm
Assessment of Clinical Skills	2016	Posterior teeth	0.50 mm
Summary of finish line location (mm) found in literature and protocols of dental institutions			
Study/Protocol	*Year*	*Tooth type*	*Finish line location*
Goodacre et al.,	2001	Posterior teeth	Supra-gingival without specific measurement
THE CONSORTIUM, Overseas Registration Examination, Part 2.	2012	Posterior teeth	~ 1.00 mm
Assessment of Clinical Skills	2016	Posterior teeth	0.50 – 1.00 mm
Summary of axial reduction (mm) found in literature and protocols of dental institutions			
Study/Protocol	*Year*	*Tooth type*	*Axial reduction*
Goodacre et al.,	2001	Molar	≥ 0.50 mm or The height of tooth preparation ≥ 4.00 mm
O'Sullivan	2005	Posterior teeth	The height of tooth preparation ≥ 4.00 mm
Assessment of Clinical Skills	2016	Posterior teeth	≥ 0.50 - < 1.50 mm

From Table 1.6, there are no specific dimensions for full crown preparations which can be used as ideal measurements to compare with students' preparations. So, the assessment of students' performance is also dependent on the ideal preparation features and the measurements which have been chosen by staff members in order to assess students' work more objectively.

1.5 Conclusions from literature review

The assessment of students during practice placements is a process that is vitally important to their development throughout the course. From a student viewpoint, education is mostly motivated by assessment. Universities are mainly involved with ‘fitness for award’ – whether the student accomplished the appropriate level and breadth of learning to be granted a degree.

Most of the studies have concluded that the ‘glance and grade’ method alone was not valid and reliable assessment method (Vann et al., 1983, Manogue et al., 2001, Sherwood and Douglas 2014). On the other hand, the majority of studies found that using a computer-assisted system did enhance the reliability of an assessment (Arnetzl and Dornhofer, 2004, Kournetas et al., 2004, Cardosa et al., 2006). The core of electronic assessment was comprised of the components of a checklist. Thus, using the checklist in assessment becomes more common than the ‘glance and grade’ method, and computer-assist system.

Dental examiners and dental students are expected to make clinical judgments based on their education. Therefore, validity and reliability measurements of clinical practice assessment must be founded in addition to ‘fitness for practice’. There are studies which have focused on active steps to enhance the levels of consistency, agreement or calibration among faculty without concentrating on the validity (Scruggs et al., 1989, Haj-Ali and Feil, 2006). A lack of validity or reliability of an assessment method may lead to confusion and frustration among the students. For example, there has been no study which discussed clearly the way to establish the content validity for a checklist when used as an assessment tool for clinical work. This issue has demonstrated a deficiency in dental education. In addition, the subsequent development of a grade from an assessment system adds to its complexity (Taylor et al., 2013).

In dental schools, much of the articles and literature reviews on evaluation have considered methods to enhance calibration among assessors in the preclinical laboratory and clinic courses. The calibration needs gold standard examiners or scores. The gold standard examiners or scores were not established clearly in most studies. For example, some studies selected the highest qualified or experienced examiner as a gold standard

(Curtis et al., 2008, Cho et al., 2010, Mays and Levine, 2014, Tuncer et al., 2015, Alhumaid et al., 2016). In the literature, the examiner with highest qualification or the greatest experience was not always the examiner who produced the most reliable result (Hinkleman and Long, 1973, Patridge and Mast, 1978, Scruggs et al., 1989, Jenkins et al., 1996). Therefore, the gold standard examiner or scores selection is a very difficult and has not been discussed clearly.

The design of scoring criteria was also mentioned in several studies to evaluate quality of student's performance, such as variety of ranking factors, variety of assessors or tutors for tests, techniques for creating ranking requirements, the level of details that should be involved in the scale, and how student's assessment is calculated (Taylor et al., 2013). Several studies concluded that objective and clearly defined criteria should be selected (Brown, 1930, Gaines et al., 1974, Helft et al., 1987, Sherwood and Douglas, 2014). Taylor et al., (2013) reviewed that several studies have focused on the number of points in a rating scale. If the number of points in the rating scale is two, they provide higher agreement than three or more points in the rating scale. On the other hand, three points in the rating scale gives more instructional advantage in teaching and determining components of a student's performance than two-point scale (Taylor et al., 2013).

In addition, the overall shape of a tooth preparation is important to assist the student's evaluation of their performance. In other words, the criteria and their descriptors should be established according to the geometric shape of the tooth preparation. These objective criteria may be used to analyse student's preparation in detail and mistakes can be detected (Cardoso et al., 2006).

In order to reduce the variation between examiners, some studies selected a specific tool with a checklist to enhance the reliability of judgements among the examiners; for example, a periodontal probe with a checklist (Haj-Ali and Feil, 2006, Ahmed et al, 2016).

Determining the maximum number of mistakes which can be accepted to pass the student was also important. Therefore, absolute standard setting must be considered (Puryer and O'sullivan 2015). Indeed, establishment of standard setting for the checklist which is used to assess performance on the dental clinic was also not mentioned in the literature.

To create or develop a new assessment method, judgments of the current examiners should be evaluated to determine the weaknesses and strengths of the examiners' ability to assess the students' performance.

1.6 Thesis outline following literature review

In this thesis, the first four chapters were designed to evaluate the type and quality of feedback and assessment in the clinical skills laboratory at Dundee Dental School. These studies focused on conservative and fixed prosthodontic concepts and techniques which are introduced to second and third year dental students through participation in a clinical skills laboratory courses. Five different methods were provided for the examiners to assess tooth preparations on different occasions. Examiners teaching these courses in clinical skills laboratory are expected to provide reliable formative and summative feedback on the student performance.

One chapter was also designed to evaluate the validity of the grades awarded by the examiners. While examiner repeatability is important, it is equally important that examiner grading reflects what is truly known about the object (tooth preparation). Thus, it is equally important that the grading is valid.

The checklist is the most common assessment method which is used within the Clinical Skills Laboratory in many universities. Although the checklist provides feedback for students to raise the level of student achievement and skills, it has also problems which might discourage them from learning. Examples of these problems are the scaling system used, the use of standard setting and the definitions of criteria which will be assessed. Therefore, new checklists with reliable assessment tools were established in order to identify if this method of assessment will provide valid outcomes and improve feedback and assessment reliability. Finally, checklists with clear descriptor and reliable assessment tools were evaluated in the Chapter 7 of this thesis.

1.7 General aims and null hypotheses of the thesis

The aims of this study were:

- To evaluate the assessment and feedback which was provided by examiner(s) at Dundee Dental School, and
- To improve the agreement and consistency of both grades-awarded and feedback-given, respectively, by examiners in clinical skills laboratory at Dundee Dental School.

Null hypotheses:

The examiner(s) at Dundee Dental School did not provide valid and reliable grades and feedback for dental students in clinical skills laboratory at Dundee Dental School.

It was not possible to change (improve) the validity and reliability of the grades and feedback for dental students in clinical skills laboratory at Dundee Dental School

Chapter 2 : Evaluating current assessment and feedback methods at Dundee Dental School in the Clinical Skills Laboratory

Chapters

Chapter 2	**Evaluating current assessment and feedback methods at Dundee Dental School in the Clinical Skills Laboratory**
	Aims: **The aims of this Chapter were to identify, in relation to the class II amalgam cavity preparation and the full veneer gold shell crown preparation undertaken in the Clinical Skills Laboratory at Dundee Dental School:** **1. the types and quality of feedback and assessment provided for students and,** **2. the disadvantages associated with current feedback.**
Chapter 3	**Independent evaluation of student class II amalgam cavity and full gold shell-crown preparations by senior academic staff**
	Aim: The aim of this study was to evaluate inter-examiner agreement between a convenience sample of three senior academic staff members, who have more than twenty years of clinical and teaching experience each, when evaluating operative procedures (class II amalgam cavity and full veneer gold shell crown preparation) on plastic teeth, from a sample year-cohort of dental students, at Dundee Dental School.
Chapter 4	**Evaluation of selected student class II amalgam cavity and full gold shell-crown preparations by senior academic staff and additional teaching staff**
	Aims: 1. To develop a sub-set of class II amalgam cavity, and full veneer gold shell crown, preparations. 2. To determine intra-examiner agreement for a group of senior academic staff when evaluating a sub-set of class II amalgam cavity preparations and full-veneer gold shell crown preparations. 3. To study a large group of additional teaching staff at Dundee Dental School in order to determine intra- and inter-examiner agreement when evaluating a sub-set of class II amalgam cavity preparations and full-veneer gold shell crown preparations.
Chapter 5	**Evaluation of selected student class II amalgam cavity and full gold shell crown preparations by using specific additional tools and feedback sheets**
	Aims: 1. To identify and develop specific additional tools to assist with evaluation of tooth preparations. 2. To determine examiner agreement for both class II amalgam cavity and full veneer gold shell crown preparations using the specific additional tools and feedback sheets for both grades awarded and repeatability and reproducibility of detailed feedback provided by senior academic staff.
Chapter 6	**Identification of gold standard (representative grades) for 26 class II amalgam cavities and 30 full gold shell-crown preparations**
	Aims: These objectives can be set out as a number of steps, each of which had a specific aim which is expanded within the methodology of this chapter. Thus, the two aims of this chapter were: 1. to identify the best senior academic staff examiner who had the highest correlation between grades awarded with the number of negative points identified; 2. to identify those tooth preparations which were awarded a grade by the best examiner in order to: • determine the specific feature measurements of this sub-group of grade tooth preparations and: • compare the objective class II amalgam cavity or full veneer gold shell crown preparation dimensions (SAMFs) recorded by the researcher and those subjective tooth preparation evaluations (SAFs) for similar teeth by the senior examiner with dimensions presented in the dental literature and subsequent calibration with the grades awarded by the best examiner.
Chapter 7	**Identification of reliable tools and development of a new checklist to assess class II amalgam cavities and full veneer gold shell crown preparations**
	Aims: 1. To identify a reliable specific additional measurement tool. 2. To develop two reliable new checklists (New Checklist CII Preparation and New Checklist Gold shell Crown Preparation). 3. To determine intra- and inter-examiner agreement using the grades and consistency of the New Checklists (Class II Preparation and Gold Shell Crown Preparation) and compare these data with that from previously used 'Gray and Mhanni feedback sheets' to ascertain if the agreement and consistency of the New Checklists have been improved.
Chapter 8	**General conclusions, recommendation, and further studies**

2.1 Introduction

Conservative and fixed prosthodontic concepts and techniques are introduced to second and third year dental students through participation in a clinical skills laboratory courses at Dundee Dental School. These courses contain a didactic portion during which learn the theoretical aspects of conservation and fixed prosthodontics. Throughout the UK, students rely on different interpretations by different Dental Schools of conservation and fixed prosthodontic methods multiple for application, reinforcement and enhancement of theoretical principles during a simultaneous laboratory portion of the courses. School examiners teaching the courses are expected to provide consistent formative and summative feedback of the student performance. Low reliability of examiners may lead to student confusion and frustration (Haj-Ali and Feil, 2006). Therefore, the maxim "Assessment drives learning" (Miller, 1990, Wass et al., 2001, Manogue et al., 2011, Norcini et al., 2011, Vander Vleuten et al., 2012), can be used to understand the problem provide at least a partial solution. In summary, student learning to determine the clear nature of the problem, analyse the problem in detail and the skills to solve the problem, are essential to motivate the student to learn.

This thesis is focused on the assessment and feedback of student performance in order to identify the most common problem. In further chapters, clarification and further definition the problem will be reported clearly. Then, researcher will conclude whether or not the problem can ultimately be solved.

If the problem is defined poorly or not completely understood, it is much more difficult to solve than a problem which has been clearly defined and analysed. The way a problem is worded and understood has a huge impact on the number, quality, and type of proposed solutions.

An initial approach for this thesis was to evaluate the Course Guides for both Conservation and Fixed Prosthodontics to ascertain if there were clear in their definition of the problems the students faced and the feedback that would be given.

2.2 Course guide for Conservation at Dundee Dental School

The aims of the Conservation course are to teach, “the prevention, diagnosis and progression of dental caries, its removal when required by means of conservative cavity preparation and the restoration of prepared cavities with restorative materials” (e.g. amalgam or composite resin).

One of the main and essential methods to teach the students is the use of the clinical skills laboratory. Attendance at the clinical skills laboratory is compulsory. The concepts of cavity preparation will be taught by preparing cavities, to eradicate simulated caries, in the plastic teeth of the phantom head patient. Demonstrators provide advice and feedback on cavities which have been prepared by students. Later in the course, student will be assessed formally by means of prescribed tests. The students’ performance in these tests will be a major factor determining exemption from the Professional Degree Practical Examination.

2.2.1 Assessment in Conservation

The assessment is also important to ensure safe treatment of patients in the clinical part of the dental course. Formative and summative assessments are involved in this course. Formative assessment is a range of formal and informal assessment procedures conducted by examiners during the learning process to modify or change teaching and learning activities to improve student performance (Crooks, 2001). Summative assessment is indicated to assess students’ performances at the end of the course (Harlen and James, 1997). Therefore, teachers may assess student individual progress, and thereby ascertain what the student is capable of independently using a series of prescribed tests.

All prescribed tests are marked on the University 1-5 scale where 5 is a good pass, 4 is a bare pass, and 3 is a marginal fail. A grade of 1 is automatically obtained if either the wrong tooth or cavity is cut.

2.3 Course guide for Fixed Prosthodontics at Dundee Dental School

This course focuses on learning how to restore form and function for teeth with the use of fixed prosthodontics (bridges and laboratory fabrications [e.g., single unit crowns and veneers]).

Preparations will be carried out on plastic (Typodent) teeth mounted in model jaws. This will overcome many of the practical difficulties associated with the use of extracted natural teeth.

Three examiners will collectively mark all of students' preparations at the end of the course to give student a grade which is consistent with the entire year. Students will only be allowed to carry out such procedures on patients when the student has demonstrated competence in the Clinical Skills Laboratory and in a practical examination at the course completion.

2.3.1 Assessment in Fixed Prosthodontics

Formative assessment is also involved in this course. Before each tooth preparation, students should describe briefly the key design characteristics. Following the procedure student should reflect on the difficulties they experienced and how they propose to address these. These comments should be written in the student's logbook.

In addition to mini-assignments and logbook completion, there is a session called, 'Class Practical Examination'. This is the summative assessment used at the end of a learning activity to make decision to provide the final judgement (Harlen and James, 1997). During this examination, the student will be expected to prepare a full coverage metal crown preparation which, in this thesis, is called a, 'Full Veneer Gold Shell Crown Preparation' (FVGSC), on an upper posterior tooth.

2.4 Evaluation of assessment method at Dundee Dental School

No studies have analysed carefully each of the components of the evaluation system being used at Dundee Dental School, in order to identify specific areas of low agreement and reliability, and then taken targeted steps to improve agreement and reliability. This study will focus on the feedback and assessment at Dundee Dental School to determine whether or not these parameters, in relation to student performance, are consistent and fair.

2.5 Aim and null Hypothesis

Aim

The aims of this part of the study were to identify, in relation to the class II amalgam cavity preparation and the full veneer gold shell crown preparation undertaken in the Clinical Skills Laboratory at Dundee Dental School:

- the types and quality of feedback and assessment provided for students and,
- the disadvantages associated with current feedback.

Null Hypothesis:

Feedback and assessment from the tutors, in relation to the class II amalgam cavity preparation and the full veneer gold shell crown preparation undertaken in the Clinical Skills Laboratory at Dundee Dental School, did not motivate the students to learn.

2.6 Material and Methods

In order to determine the type and quality of feedback and assessment provided for dental students in relation to the class II amalgam cavity preparation and the full veneer gold shell crown preparation undertaken in the Clinical Skills Laboratory at Dundee Dental School, the following steps were applied:

- One cohort of students was selected who completed the Class II amalgam cavity preparation in the academic year 2013/2014 and then subsequently completed the full veneer gold shell crown preparation in the academic year 2014/2015. These sessions were important because they formed the main part of a final practical exam for second and third year BDS at Dundee Dental School.
- The researcher (AM) attended several sessions in the clinical skills laboratory.
- The type of feedback and the type of assessment were both recorded through observation of staff interaction with students.
- The quality of the feedback was requested from each student, by answering two questions (see Figure 2.1), and a short interview with the researcher (AM).
- The quality of the assessment was through data collection and evaluation in the subsequent chapters.

> - What is the type of feedback which encourages you to learn at Dundee Dental School in the clinical laboratory skills?
> - What is the disadvantage of the current feedback and assessment at Dundee Dental School?

Figure 2.1 Two open questions to determine the best method to provide feedback for student and the quality of feedback at Dundee Dental School

2.7 Results

From Dundee Dental School Programme Handbooks for, 'Management of Dental Caries Course', and, 'Fixed Prosthodontics and Laboratory Fabricated Restorations', the School offered a variety of ways to learn in the clinical skills laboratory. These included:

1. a short introduction as a review of the topics (including videos and live demonstrations) for the students before commencing the practice session,
2. subjective formative assessment/feedback provided verbally by tutor after the students finish their tooth preparations and,
3. Mini-assignments from the students before starting practice session in the fixed prosthodontics course only and used as a formative assessment.

2.7.1 Type of feedback and assessment at Dundee Dental School

Only oral <u>feedback</u> was provided. Formative <u>assessment</u> was provided verbally while, at the end of the course, summative <u>assessment</u> was recorded as either a grade on a 1-5 scale or a simple pass/fail grade.

2.7.2 The quality of feedback at Dundee Dental School

In relation to the quality of the feedback both the researcher observations and student comments were collected.

a. The quality of feedback according to the researcher observations

Students were divided into three groups of between 25 and 30. During the practice session, most of the students found that class II amalgam cavity preparation and full veneer gold shell crown preparation were very difficult to make by using a high speed handpiece. For class II amalgam cavity preparations in 2013/2014, using only one fissure diamond bur to prepare a class II cavity without touching the adjacent plastic tooth is hard for them. In addition, most of the students asked about the depth and width of the proximal box in relation to the dimensions of the bur. Most of the class II cavities which were prepared by students had a large irregular proximal box. Furthermore, the majority of the adjacent teeth were damaged.

For full veneer gold shell crown preparations in 2014/2015, the tutors showed videos that demonstrated full veneer gold shell crown preparation for the lower first molar tooth, using a high speed handpiece with a round-ended, chamfer, diamond bur for the same cohort of students. The video showed the student how to create a pre-operative silicone impression to form the provisional crown. A second pre-operative silicone impression, in this thesis called an impression index, was made. The impression index was sectioned in a bucco-lingual direction at long axis of tooth being prepared. The impression index allowed the operator to evaluate the amount of tooth reduction after tooth preparation. The tutor did not provide the amount of reduction for the students in relation to the impression index or other instrument to allow them to evaluate their works. The students commenced preparation of the first molar tooth after watching the demonstration. The researcher found that the students struggled to prepare an ideal full veneer gold shell crown preparation, especially in relation to the axial distal wall of the lower first molar.

b. The quality of feedback according to the student comments

The students who attended Clinical Skills Laboratory sessions in 2013/2014 were involved in this part of study. The total number of students who answered the questionnaire was 45 out of 69 (65%) students. The excluded students were omitted because they did not answer either of the questions, their writing was illegible or the answers given did not relate to the questions asked. Figure 2.2 illustrated the most preferable type of feedback was tutor's feedback (93.4%). They found this type of feedback crucial to initiate self-improvement. Furthermore, Figure 2.2 showed a small minority of dental students favoured peer feedback from others (e.g. friends, books, or tutors).

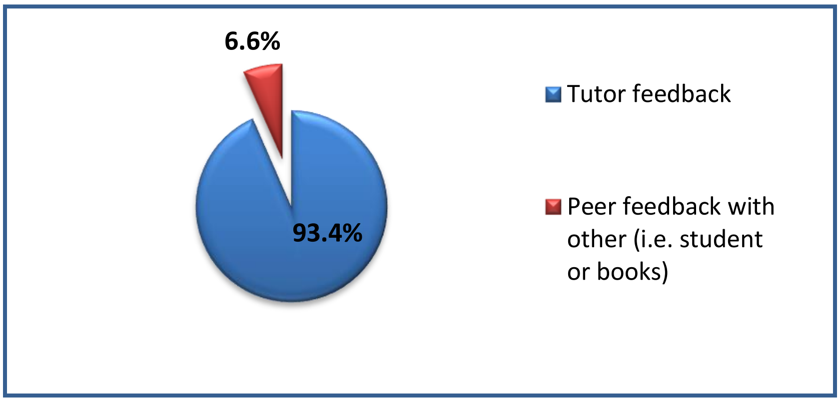


Figure 2.2 Pie chart to illustrate the preference for type of feedback expressed by the students from a questionnaire

The disadvantages of currently received feedback were that the students complained:

1. it was not prompt
2. it was not clear and,
3. it did not contain objective evaluations such as acceptable dimensions or features

These results are supported by the National Students Survey (NSS) in the UK (High Education Funding Council for England, 2011 Page 8-14) and (Unistats 2014) (see Chapter 1).

From these observations and comments, the following question was raised, "does the tutor provide the students with valid and reliable feedback and assessment?"

2.8 Discussion

- Dundee Dental Students struggle to prepare teeth because there are no specific dimensions which were provided to the students before commencing preparation. Feedback provides comments for the students and it also provides a plan for students to learn from it and enhances their performance (Boud and Molloy, 2013).

- The best feedback method which was favoured by the students was from the tutors who were considered to have more experience. Although the students preferred the feedback from the tutor, the tutor did not provide satisfactory comments for them. Most of the student complaints were that they did not receive enough feedback to both compare and enhance their performance in relation to according to an ideal or acceptable standard. In addition, most of tutors' feedback was generally verbal. These comments from the students are supported by the studies of Branch and Paranjape (2002) and Boud and Molloy (2013). Therefore, Bound and Molloy (2013) highlighted three assumptions that are essential to understand and enhance feedback to prompt students' performance (see Chapter 1).

    1. Feedback helps students to identify their strengths and weaknesses and target areas that need work.
    2. Student needs to take action on received comments from tutor.
    3. Feedback cannot be provided for all students by same way. The way of providing of these comments are important.

2.9 Conclusion

According to the observations of the researcher and students' comments, the null hypothesis was accepted. Feedback from the tutor was preferred by Dental Students at Dundee Dental School in order to improve their performances.

The conclusions of this chapter were based on an opinion of the researcher (AM) and students' comments. In the next few chapters (i.e. Chapter 3, 4, 5 and 6), researcher (AM) will try to find evidence for these conclusions by evaluating the quality of the assessment and consistency of feedback at Dundee Dental School for work undertaken in the Clinical Skills Laboratory.

Chapter 3：Independent evaluation of student class II amalgam cavities and full gold shell-crown preparations by senior academic staff

	Chapters
Chapter 2	**Evaluating current assessment and feedback methods at Dundee Dental School in the Clinical Skills Laboratory** **Aims:** The aims of this Chapter were to identify, in relation to the class II amalgam cavity preparation and the full veneer gold shell crown preparation undertaken in the Clinical Skills Laboratory at Dundee Dental School: 1. the types and quality of feedback and assessment provided for students and, 2. the disadvantages associated with current feedback.
Chapter 3	**Independent evaluation of student class II amalgam cavity and full gold shell-crown preparations by senior academic staff** **Aim:** **The aim of this study was to evaluate inter-examiner agreement between a convenience sample of three senior academic staff members, who have more than twenty years of clinical and teaching experience each, when evaluating operative procedures (class II amalgam cavity and full veneer gold shell crown preparation) on plastic teeth, from a sample year-cohort of dental students, at Dundee Dental School.**
Chapter 4	**Evaluation of selected student class II amalgam cavity and full gold shell-crown preparations by senior academic staff and additional teaching staff** **Aims:** 1. To develop a sub-set of class II amalgam cavity, and full veneer gold shell crown, preparations. 2. To determine intra-examiner agreement for a group of senior academic staff when evaluating a sub-set of class II amalgam cavity preparations and full-veneer gold shell crown preparations. 3. To study a large group of additional teaching staff at Dundee Dental School in order to determine intra- and inter-examiner agreement when evaluating a sub-set of class II amalgam cavity preparations and full-veneer gold shell crown preparations.
Chapter 5	**Evaluation of selected student class II amalgam cavity and full gold shell crown preparations by using specific additional tools and feedback sheets** **Aims:** 1. To identify and develop specific additional tools to assist with evaluation of tooth preparations. 2. To determine examiner agreement for both class II amalgam cavity and full veneer gold shell crown preparations using the specific additional tools and feedback sheets for both grades awarded and repeatability and reproducibility of detailed feedback provided by senior academic staff.
Chapter 6	**Identification of gold standard (representative grades) for 26 class II amalgam cavities and 30 full gold shell-crown preparations** **Aims:** These objectives can be set out as a number of steps, each of which had a specific aim which is expanded within the methodology of this chapter. Thus, the two aims of this chapter were: 1. to identify the best senior academic staff examiner who had the highest correlation between grades awarded with the number of negative points identified; 2. to identify those tooth preparations which were awarded a grade by the best examiner in order to: • determine the specific feature measurements of this sub-group of grade tooth preparations and: • compare the objective class II amalgam cavity or full veneer gold shell crown preparation dimensions (SAMFs) recorded by the researcher and those subjective tooth preparation evaluations (SAFs) for similar teeth by the senior examiner with dimensions presented in the dental literature and subsequent calibration with the grades awarded by the best examiner.
Chapter 7	**Identification of reliable tools and development of a new checklist to assess class II amalgam cavities and full veneer gold shell crown preparations** **Aims:** 1. To identify a reliable specific additional measurement tool. 2. To develop two reliable new checklists (New Checklist CII Preparation and New Checklist Gold shell Crown Preparation). 3. To determine intra- and inter-examiner agreement using the grades and consistency of the New Checklists (Class II Preparation and Gold Shell Crown Preparation) and compare these data with that from previously used 'Gray and Mhanni feedback sheets' to ascertain if the agreement and consistency of the New Checklists have been improved.
Chapter 8	**General conclusions, recommendation, and further studies**

3.1 Introduction

In recent years, there has been an increasing interest in measuring examiner agreement and reliability, both in dentistry and other academic disciplines. This is in order to provide reliable and accurate feedback for students as well as accurate evaluation of student work for progression within a given course. Most of the feedback is provided by teachers after evaluating student work. Therefore, in dentistry, reliable and accurate feedback for the student plays an important role in order to achieve a higher level of performance in preclinical dental education before student starts treating patients on the clinic. Unfortunately, a number of studies have found that agreement and/or reliability for assessment in dentistry and related disciplines is variable and low (Lilley et al., 1968, Fuller, 1972, Salvendy et al., 1973, Myers, 1977, Vann et al., 1983, Jenkins et al., 1998, Sharaf et al., 2007). In addition, studies typically conclude that intra-examiner variability is less than inter-examiner variability (Lilley et al., 1968, Deranleau et al., 1983, Sharaf et al., 2007). In other words, examiners agree with themselves more than they agree with each other. However, the way in which teaching in many UK Dental Schools is often arranged means that several examiners may evaluate the work of different students within a particular year group. Thus, measuring agreement amongst examiners was selected in this part of study to evaluate the assessment system in Operative Dentistry at Dundee Dental School.

Some of the first clinical feedback that dental students receive is during the clinical skills laboratory when they prepare teeth to receive, for example, a Class II amalgam or a full veneer gold shell crown. Therefore, preparing a class II amalgam cavity for second year students (2013/2014) and a full veneer gold shell crown preparation for third year students (2014/2015) in the clinical skills laboratory in Dundee Dental School were chosen for this work.

3.2 Aim and null hypothesis

Aim:

The aim of this study was to evaluate inter-examiner agreement between a convenience sample of three senior academic staff members, who have more than twenty years of clinical and teaching experience each, when evaluating operative procedures (class II amalgam cavity and full veneer gold shell crown preparation) on plastic teeth, from a sample year-cohort of dental students, at Dundee Dental School.

Null hypothesis:

There was no inter-examiner agreement for a convenience sample of three senior academic staff members who evaluated Class II amalgam cavity preparations and full veneer gold-crown preparations within a sample year-cohort of dental students at Dundee Dental School.

3.3 Material and methods

3.3.1 The sample year-cohort

The year selected was a convenience sample of the dental students working in the clinical skills laboratory at the time of this period of post-graduate study. This sample year completed the preparation of a class II amalgam cavity in the second year of their BDS course and a full veneer gold shell crown preparation in the third year of their Bachelor of Dental Surgery (BDS) course. Thus, the cohort of students was the same for each tooth preparation. These students were required to complete each preparation within 120 minutes.

3.3.2 Selection of examiners

The tooth preparations were assessed by up to three different senior academic staff members who spend much of their time evaluating student tooth preparations and have done so for over 20 years. The senior academic staff members were a convenience sample who also supervised the author's postgraduate studies.

3.3.3 Selection of tooth preparations

a. Preparation of class II amalgam cavities

Following didactic theoretical and practical teaching and practice of class II amalgam cavity preparation, students worked in the clinical skills laboratory with plastic teeth (Frasaco GmbH, Germany), fixed in Typodont models and mounted in a phantom head simulator (A-dec incorporation – Nuneaton, Liberty Way). The procedure that was evaluated was a mesio-occlusal cavity, for restoration using dental amalgam, in a maxillary second premolar. These preparations were made under examination conditions, on up to two occasions, which was part of the course evaluation. Students used a diamond fissure bur (UnoDent Ltd, No. BDM541) in a high-speed hand-piece only to perform these procedures (Figure 3.1).

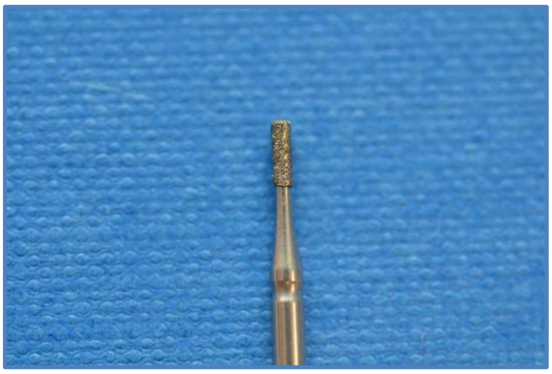


Figure 3.1 Picture of a diamond fissure bur

The resulting cavities were coded with a random number taken from a book of raffle tickets so the examiners were blinded to which student completed each cavity.

Conditions and methods for examination of class II amalgam cavities

All cavities evaluated were situated within an intact Typodont arch from a phantom head by three senior academic staff as examiners. They were examined under ambient lighting conditions, out of the phantom head and in an office environment. The method of cavity assessment was at the discretion of each examiner with one proviso: Examiners were asked to evaluate, on one occasion, all tooth preparations using a personally-devised, four- or five-point scale. This magnitude of each scale was to ensure that evaluations were comparable with currently used methods to determine student progression within the course. The scales ranged from the most acceptable cavity to the least acceptable cavity and are detailed in Table 3.1.

Table 3.1 Grading system for three senior academic staff as examiners for class II amalgam cavity preparation

Examiner 1 (personally-devised scale)	
Grade	**Descriptions**
A	This cavity is good to place amalgam immediately.
B	Roughly finished but can place amalgam (acceptable)
C	Modification required to place amalgam (Not acceptable)
D	No amount of modification would make acceptable (Not acceptable)
E	Wrong tooth

Examiner 2 (personally-devised scale)				
	Pass		***Fail***	
Evaluation	**A**	**B**	**C**	**D**
Will the cavity work to retain a class II amalgam?	Yes	Yes	No	No
Is the cavity underprepared and redeemable?	No	Yes or No	Yes	No
Is the cavity over-prepared?	No	No	No	Yes
Is there damage to the adjacent tooth?	No	No	Yes	Yes

Examiner 3 (personally-devised scale)	
Grade	**Description**
A	• Occlusal key and box undercut, • Adequate depth of occlusal key (2-3mm) • Key - key shaped with resistance form • Isthmus not too narrow > 1.5 - 2 mm • Floor of box beneath contact point • No unsupported enamel at floor of box - or generally • Appropriate cavo-surface angle for amalgam • No acute line angles (i.e. GV Black type box) • No damage to adjacent tooth • Not necessary to clear contact point buccally and lingually provided matrix band can be placed
B	• Occlusal key and box undercut, • Adequate depth of occlusal key (2-3mm) • Floor of box beneath contact point • No unsupported enamel at floor of box - or generally • Appropriate cavo-surface angle for amalgam • No acute line angles (i.e. GV Black type box) • Accept minor damage to adjacent tooth provided candidate knew how to deal with it • Not necessary to clear contact point buccally and lingually provided matrix band can be placed
C	• Occlusal key or box not undercut, • Adequate depth of occlusal key (2-3mm) • Floor of box beneath contact point • Unsupported enamel at floor of box - or generally • Appropriate cavo-surface angle for amalgam • Acute line angles (i.e. GV Black type box) • Minor damage to adjacent tooth but candidate did not know how to deal with it • Clear contact point buccally and lingually beyond expected for minimal cavity
D	• Occlusal key and box not undercut, • Inadequate depth of occlusal key (< 2-3mm) • Occlusal key - no resistance form • Floor of box not beneath contact point • Unsupported enamel at floor of box - or generally • Inappropriate cavo-surface angle for amalgam • Acute line angles throughout cavity (i.e. GV Black type box) • Marked damage to adjacent tooth • Cavity clearing contact point buccally and lingually with wide margin

b. Preparation of full veneer gold shell crown

The student cohort used for the evaluation of the class II amalgam cavity in the second year of their BDS course subsequently progressed to the third year of the BDS course where they prepared plastic teeth (Frasaco GmbH, Germany) for a full veneer gold shell crown on the upper first molar tooth within a phantom head. The teeth were prepared using tapered high-speed diamond with a rounded tip (Chamfer) burs (No.8856.314.016 TPS2, Komet Dental, UK) as part of their normal instruction for this part of their BDS course (Figure 3.2). The resulting crown preparations were coded with a random number taken from a book of raffle tickets so the examiners were blinded to which student completed each cavity.

Figure 3.2 Picture of a tapered high-speed diamond with a rounded tip (Chamfer) bur

<u>Conditions and methods for examination of full veneer gold shell crown preparations</u>

All crown preparations were evaluated by two senior academic staff (also used to assess the Class II cavities) as examiners. Crown preparations were situated within an intact Typodont arch, removed from the phantom head, and were examined under ambient lighting conditions in an office environment. The method of cavity assessment was at the discretion of each examiner with one provision: examiners were asked to evaluate, on one occasion, all tooth preparations using a four-point scale. This magnitude of scale was to ensure that evaluation was comparable with currently used methods to determine student progression within the course. The scales ranged from the most acceptable cavity to the least acceptable cavity and are detailed in Table 3.2.

Table 3.2 Grading system of senior academic staff as examiners for full veneer gold shell crown preparation

Examiner 1 (personally-devised scale)	
Grade	**Description**
A	No damage to adjacent teeth, appropriate taper, minimal chamfer all around, smooth flowing outlines, appropriate occlusal reduction and functional cusp bevel, margins mesial and distal below contact points
B	1-2 of the following which would still render the preparation fit for purpose: Minor damage to adjacent teeth, increased taper, chamfer too deep, irregular outline, too much occlusal reduction, inappropriate functional cusp bevel, margins at contact point
C	More than three of the following which would render the preparation unfit for purpose as it stood – but correctable with advice: Minor to moderate damage to adjacent teeth, increased taper, chamfer too deep, irregular outline, too much occlusal reduction, inappropriate functional cusp bevel, margins at contact point
D	Extensive damage to adjacent teeth.

Examiner 2 (personally-devised scale)	
Grade	**Description**
A	Would work, not over-tapered and no damage to adjacent tooth
B	Would work, not over-tapered but minor damage to adjacent tooth
C	Undercut and damage to adjacent tooth
D	Over-tapered and damage to adjacent tooth

3.3.4 Statistical analysis to determine inter-examiner agreement for senior academic staff examiners

For ordinal and nominal data there is no requirement to determine the distribution of results prior to calculation of inter-examiner agreement. For two examiners only, Cohen's un-weighted Kappa score was determined whereas for more than two examiners, the intra-class correlation was chosen (Gisev et al., 2013). Alphabetical grades A, B, C, D and E for each examiner were changed into numerical values 5, 4, 3, 2 and 1 respectively before commencing statistical analysis by using SPSS package.

The data were entered into a statistical package (SPSS version 22) to determine inter-examiner agreement using the two way mixed intra-class correlation test (ICC type 3) with absolute agreement for the class II amalgam cavity and Cohen's un-weighted kappa for the full veneer gold shell crown preparation (see section 3.4 'Results'). The reasons for selecting specific intra-class correlation and Cohen's un-weighted Kappa are explained in the section 3.5 'Discussion'.

3.4 Results

3.4.1 Selection of examiners

All three examiners evaluated the class II amalgam cavity preparations. Only two of these examiners evaluated the full veneer gold shell crown preparations.

3.4.2 Selection of tooth preparations

a. Class II amalgam cavity preparation

Seventy second year dental students completed a total of 118 cavities. Most students completed two cavities but some, who, according to course requirements, had completed their first cavity to an acceptable standard, were not required to complete a second cavity.

b. Full veneer gold shell crown preparation

Seventy third year students each completed a full veneer gold shell crown preparation in a plastic tooth mounted in a phantom head.

3.4.3 Inter-examiner agreement for senior academic staff examiners

a. Class II amalgam cavity preparation

By using Intra-class correlation test for the class II amalgam cavity, Table 3.3 demonstrated that there was low agreement between the three examiners who assessed the class II amalgam cavities [single measurement ($ICC_{3,1}$ = 0.518, 95% CI = 0.415 – 0.613). The relatively wide 95% confidence interval would also support this poor agreement.

Table 3.3 Intra-class Correlation Coefficient for senior academic staff as examiners for class II amalgam cavity preparation

	Intraclass Correlation	95% Confidence Interval		Significance ($p \leq 0.05$)
		Lower Bound	Upper Bound	
Single Measures	0.518	0.415	0.613	0.000

b. Full veneer gold shell crown preparation

By using Cohen's un-weighted kappa test for the full veneer gold shell crown preparation, Table 3.4 shows the results for the Cohen's un-weighted kappa agreement for the full veneer gold shell crown preparation by two senior academic staff examiners.

Table 3.4 Cohen's un-weighted Kappa (inter-examiner agreement) for senior academic staff members assessment of full veneer gold shell crowns

Kappa value	**Level of agreement**	**Significance ($p \leq 0.05$)**
0.089	Slight	0.215

Cohen's un-weighted Kappa agreement test was utilised because the number of senior assessors was two. The Kappa value was low and demonstrated that there was no significant agreement between examiners (K= 0.089). Once again agreement between examiners would support the overall conclusion from Kappa values that the agreement among examiners was low. This is discussed further later in this chapter.

3.5 Discussion

The class II amalgam cavity preparation and the full veneer gold shell crown preparation were selected in this study because they are part of a preclinical examination at Dundee Dental School. These teeth were prepared by students who have undergone same level of training experience to reduce the bias in tooth preparation's quality. Moreover, the senior academic teachers who were involved were a convenience sample of evaluators (examiners) considered to be representative of Dundee Dental

School staff. Well-designed research studies must therefore include procedures that measure agreement among the various data collectors.

According to Gisev et al., (2013), inter-examiner agreement can be tested by having two examiners or more. Three senior academic staff as examiners were included in class II amalgam cavity preparation and two examiners for the full veneer gold shell crown preparation.

Some common ways which are used to calculate agreement are percentage agreement, Cohn's un-weighted Kappa, weighted Kappa and intra-class correlation (ICC). Percentage agreement can be selected if the agreement expected by chance is not considered. Weighted and un-weighted Kappa tests are used to determine intra- and inter-examiner agreement, if the number of examiners is two. Both of them consider percentage agreement and percentage of agreement expected by chance. The difference between them is un-weighted Kappa test does not take account of the degree of disagreement. Zero weight is given to all disagreement values. On the other hand, weighted Kappa gives different weights for disagreement values (Jakobsson and Westergren, 2005, Gisev et al., 2013). A Kappa ratio of 1 indicates perfect agreement whereas 0 indicates poor agreement (Cohen, 1968, Landis, Koch, 1977, Jakobsson and Westergren, 2005, Sim and Wright, 2005).

In this part of study, Cohen's un-weighted Kappa agreement was selected to evaluate the agreement between two examiners (inter-examiner agreement) for evaluation of the full veneer gold shell crown preparation. The Cohen's un-weighted kappa test was selected because the preparations to be assessed, the categories of evaluation and the examiners were all independent. In addition, the result of the assessment is an important decision for the student who may either fail or pass and provides consistent feedback for students. Cohen's un-weighted Kappa was selected in this study, although there are several studies preferring the use of a weighted Kappa agreement (Kraemer et al., 2002, Gisev et al., 2013). Cohen's un-weighted Kappa will throw away the ordering information and treat the data as if it is composed of discrete categories to calculate absolute agreement among examiners (i.e. nominal data). Consequently, using an un-weighted kappa on categorical data would tend to result in a lower agreement score than using weighted kappa.

Landis and Koch (1977) have proposed the following as standards for strength of agreement for Cohen's kappa coefficient:

< 0 = poor,

0 –0.20 = slight,

0.21 – 0.40 = fair,

0.41 – 0.60 = moderate,

0.61 – 0.80 = substantial, and

0.81–1.00 = almost perfect.

If there are more than two examiners, who assess the tooth preparations, intra-class correlation ICC is an adequate method to measure inter-examiner agreement (Jakobsson and Westergren, 2005, Gisev et al., 2013). Intra-class correlations (ICCs) are used to measurement data homogeneity. The most essential interpretation of an intra-class correlation is that it is a measure of the proportion of a variance that is attributable to items measurement (i.e. tooth preparations) and examiners (McGrow and Wong, 1996). An intra-class correlation ratio of 1 indicates perfect agreement with no measurement error, whilst 0 indicates no agreement (Shrout and Fleiss, 1979, Fleiss, 1986).

There are three models for the intra-class correlation ICC. These are, 'One-way random (1)', 'Two-way random (2)' and, 'Two-way mixed (3)', models. For this study the two-way mixed model (3) with absolute agreement (A) was selected because each preparation was assessed by each examiner and the three examiners were the only examiners (convenience sample) included in this particular study (Shrout and Fleiss, 1979, Rankin and Stokes, 1998).

The most important interpretation of an intra-class correlation is that it is a measure of the proportion of a variance [i.e. variance of mean square for items (tooth preparations), mean square for judges (examiners) and mean square error (residual mean square)] that is attributable to items of measurement (tooth preparations) and judges (examiners) (McGrow and Wong, 1996). The mean squares from different sources of variance, was obtained by ANOVA and are used in the intra-class correlation ICC (Shourt and Fleiss, 1979).

The items in this study were tooth preparations (i.e. class II amalgam cavity preparation and full veneer gold shell crown preparation) and judges were senior academic staff examiners. According to Rankin and Strokes (1998), when each item (tooth preparation) is rated by the same examiners, who are the only examiners of interest in the study, a two-way mixed type (type 3) of intra-class correlation should be considered (Renkin and Stokes, 1998). In addition, the agreement will reflect the accuracy of score which was given by specific examiners but cannot be applied generally to any other examiners. Therefore, a two-way mixed model (3) of intra-class correlation with absolute agreement (A) was selected.

Intra-class correlation (ICC) reports two measures. These are average measure (k) and single measure (1). Average measure (k) is rarely calculated or reported in the literature and essentially describes the agreement of two or more examiners for a whole dataset. Average measurement is utilised when the unit of analysis is the mean measurement obtained either from more than one measurement or from more than one examiner. Of greater interest is the inter-examiner agreement for individual items (tooth preparation) within the whole dataset. However, single measures of intra-class correlation are much more informative (Shrout and Fleiss, 1979, McGrow and Wong, 1996). For this, the single measure (1) parameter is more useful and more widely reported.

For example, if three examiners have four categories they can each award to 26 class II amalgam cavities; it is possible for all examiners to award different categories to all cavities. However, if all examiners award each category with the same frequency across the dataset then the average measure will be perfect whereas there may be absolutely no agreement between examiners on any single item (cavity) within that dataset. As such the single measure intra-class correlation ICC $_{(3,1)}$ will be very low or zero. Thus it is the single measure intra-class correlation ICC $_{(3,1)}$ parameter which is the more important and clinically relevant of the two (Rankin and Stokes, 1998, Yen and Lo, 2002).

In this part of study, the intra-class correlation (ICC) of interest was for 'single measure' (1) because, in reality, the examiner only assessed the preparation once at one period of time (Yen and Lo, 2002). So, single measure of intra-class correlation $_{(3,1)}$ with absolute agreement (A) was selected to determine inter-examiner agreement amongst three senior academic staff as examiners.

There are few consensuses in the research literature to define a 'sufficient' intra-class correlation (ICC) score. While 0.70 would be sufficient for a measure used for research purposes, some researchers advocate a value of 0.80 or 0.90 as a minimum when using scores for making important decisions about the health of an individual (Bland and Altman, 1997, Nunnally and Bernstein, 1994, Polit and Beck, 2006). In this study, the decision is essentially that of fail or pass. Therefore, an intra-class correlation for single measure of 0.80 to 0.90 is probably a minimal acceptable standard. Clearly, the intra-class correlation of 0.518 achieved in this study, for the evaluation of the class II amalgam cavity preparation, is low.

A test result for Intra-class correlation and Cohen's Kappa tests (calculated from the null hypothesis and the sample) are called statistically significant if it is deemed unlikely to have occurred by chance, assuming the truth of the null hypothesis. A statistically significant result, when a probability (p-value) is less than a threshold (0.05 level of significance), justifies the rejection of the null hypothesis, but only if the *a priori* probability of the null hypothesis is not high.

Tables 3.3 and 3.4 showed that the agreement among senior academic staff was low for both class II amalgam cavity preparation and full gold-crown preparation. This finding supports the results of Lilley et al., (1968), Fuller (1972), Salvendy et al., (1973), Jenkins et al., (1998), Quinn et al., (2003) and Paskins et al., (2010).

Some authors have suggested that the variability of agreement might be due to examiner experience, internal examiner bias, interpretation and design of rating scales, knowledge or training reason (Natkin and Guild, 1967, Fuller, 1972, Houpt and Kress, 1973, Helft et al., 1987).

In general, researchers struggle because, the greater the consequences resulting from the evaluation, the greater the need for high inter-examiner agreement (Nunnally and Bernstein, 1994, LeBreton and Sentor, 2008).

3.6 Conclusions

Amongst senior academic staff;

1 the level of inter-examiner agreement for the class II amalgam cavity preparation was low and,

2 the level of inter-examiner agreement for the full gold shell-crown preparation assessment, was slight.

From the result, null hypothesis is accepted. There was no inter-examiner agreement for a convenience sample of three senior academic staff members who evaluated Class II amalgam cavity preparations and full veneer gold-crown preparations within a sample year-cohort of dental students at Dundee Dental School. These results indicate that the problem of inadequate inter-examiner agreement exists at Dundee Dental School within a small group of experienced teachers when they each use their own devised and applied grading systems. Furthermore, there is data lacking for intra-examiner agreement and this element of examiner evaluation should also be evaluated.

Evaluations of student tooth preparations occur on the clinics many times each day by experienced practitioners who teach students. Such practitioners come from varied backgrounds and it is essential to determine what agreement exists within this much larger cohort of examiners in order to establish the extent of any potential problem existing within evaluation of student clinical work.

In order to make the above evaluations possible a smaller number of tooth preparations are required as the repeated evaluation of sample sizes of 118 class II cavity preparations and 70 full veneer gold shell crown preparations is un-wieldy.

Chapter 4 : Evaluation of selected student class II amalgam cavities and full gold shell-crown preparations by senior academic staff and additional teaching staff

Chapters	
Chapter 2	**Evaluating current assessment and feedback methods at Dundee Dental School in the Clinical Skills Laboratory** **Aims:** The aims of this Chapter were to identify, in relation to the class II amalgam cavity preparation and the full veneer gold shell crown preparation undertaken in the Clinical Skills Laboratory at Dundee Dental School: 4. the types and quality of feedback and assessment provided for students and, 5. the disadvantages associated with current feedback.
Chapter 3	**Independent evaluation of student class II amalgam cavity and full gold shell-crown preparations by senior academic staff** **Aim:** The aim of this study was to evaluate inter-examiner agreement between a convenience sample of three senior academic staff members, who have more than twenty years of clinical and teaching experience each, when evaluating operative procedures (class II amalgam cavity and full veneer gold shell crown preparation) on plastic teeth, from a sample year-cohort of dental students, at Dundee Dental School.
Chapter 4	**Evaluation of selected student class II amalgam cavity and full gold shell-crown preparations by senior academic staff and additional teaching staff** **Aims:** **1. To develop a sub-set of class II amalgam cavity, and full veneer gold shell crown, preparations.** **2. To determine intra-examiner agreement for a group of senior academic staff when evaluating a sub-set of class II amalgam cavity preparations and full-veneer gold shell crown preparations.** **3. To study a large group of additional teaching staff at Dundee Dental School in order to determine intra- and inter-examiner agreement when evaluating a sub-set of class II amalgam cavity preparations and full-veneer gold shell crown preparations.**
Chapter 5	**Evaluation of selected student class II amalgam cavity and full gold shell crown preparations by using specific additional tools and feedback sheets** **Aims:** 1. To identify and develop specific additional tools to assist with evaluation of tooth preparations. 2. To determine examiner agreement for both class II amalgam cavity and full veneer gold shell crown preparations using the specific additional tools and feedback sheets for both grades awarded and repeatability and reproducibility of detailed feedback provided by senior academic staff.
Chapter 6	**Identification of gold standard (representative grades) for 26 class II amalgam cavities and 30 full gold shell-crown preparations** **Aims:** These objectives can be set out as a number of steps, each of which had a specific aim which is expanded within the methodology of this chapter. Thus, the two aims of this chapter were: 1. to identify the best senior academic staff examiner who had the highest correlation between grades awarded with the number of negative points identified; 2. to identify those tooth preparations which were awarded a grade by the best examiner in order to: • determine the specific feature measurements of this sub-group of grade tooth preparations and: • compare the objective class II amalgam cavity or full veneer gold shell crown preparation dimensions (SAMFs) recorded by the researcher and those subjective tooth preparation evaluations (SAFs) for similar teeth by the senior examiner with dimensions presented in the dental literature and subsequent calibration with the grades awarded by the best examiner.
Chapter 7	**Identification of reliable tools and development of a new checklist to assess class II amalgam cavities and full veneer gold shell crown preparations** **Aims:** 1. To identify a reliable specific additional measurement tool. 2. To develop two reliable new checklists (New Checklist CII Preparation and New Checklist Gold shell Crown Preparation). 3. To determine intra- and inter-examiner agreement using the grades and consistency of the New Checklists (Class II Preparation and Gold Shell Crown Preparation) and compare these data with that from previously used 'Gray and Mhanni feedback sheets' to ascertain if the agreement and consistency of the New Checklists have been improved.
Chapter 8	**General conclusions, recommendation, and further studies**

4.1 Introduction

Traditionally, students who have learned in a clinical skills laboratory have received a grade from their teacher for their ability to prepare teeth to receive direct or indirect restorations. Feedback may also be given which generally reflects the teacher's personal preferences (Yates et al., 1976). Pre-clinical and clinical performances are difficult to evaluate accurately and consistently, because the evaluation of competent student performance by the examiner consists of subjective and objective skills (Helft et al., 1987).

Although the stressful nature of clinical evaluation, both for students (McKay, 1978) and for faculty (Seymour, 1989), is recognized, and various methods of preclinical and clinical evaluation have been described (Wooley, 1977), many studies have been published to establish the reliability and accuracy of observation as a tool for evaluating dental student performances (Hinkelman and Long, 1973, Houpt and Kress, 1973). Most of these studies concluded that teachers frequently have widely different evaluation procedures and criteria; most of which are subjective. Subjective feedback may inadequately inform the student of specific strengths and weakness on which to focus in order to improve their learning.

In addition, establishing agreement among Faculty members has not been an easy undertaking (Natkin and Guild, 1967, Houpt and Kress, 1973, Gaines et al., 1974). Surveys have shown that inconsistent Faculty evaluation is a significant source of discouragement and the major reason for the student decision to, “do just enough to get by” (Natkin and Guild, 1967).

The reasons for Faculty variability were; assessor experience, examiner bias, design of the rating scales as well as training and severity of standards set by each instructor (Natkin and Guild, 1967, Fuller, 1972, Houpt and Kress, 1973, Helft et al., 1987).

According to Duncan, if feedback from the teacher is reliable and consistent, learning tends to be predictable and efficient. Conversely, when feedback is inconsistent, learning tends to be unpredictable and inefficient. Therefore, efforts have been made to improve the reliability of faculty evaluations of student performance (Duncan, 1979).

This study aimed to evaluate intra-examiner agreement for the previously-described group of senior academic staff. In addition, inter and intra-examiner agreement for a large group of additional teaching staff who regularly evaluate student work within a clinical environment at Dundee Dental School was determined.

4.2 Aims and null hypotheses

Aims

1 To develop a sub-set of class II amalgam cavity, and full veneer gold shell crown, preparations.

2 To determine intra-examiner agreement for a group of senior academic staff when evaluating a sub-set of class II amalgam cavity preparations and full-veneer gold shell crown preparations.

3 To study a large group of additional teaching staff at Dundee Dental School in order to determine intra- and inter-examiner agreement when evaluating a sub-set of class II amalgam cavity preparations and full-veneer gold shell crown preparations.

Null hypotheses

There is no intra-examiner agreement by senior staff for evaluation of a sub-group of class II amalgam cavities and full veneer gold shell crown preparations Furthermore, there is no intra and inter-examiner agreement for the evaluation of a sub-group of class II amalgam cavities and full veneer gold shell crown preparations, by additional teaching staff, who undertake the daily grading of preclinical procedures at Dundee Dental School.

4.3 Material and Methods

4.3.1 Panel of examiners

The three senior academic staff members who assessed 118 class II amalgam cavity preparations in the previous part of the study were also selected as panel of examiners for this part of the study. In addition, a large group of additional teaching staff was recruited as part of two teaching days in 2014 and 2015 at Dundee Dental School was also selected for this part of the study. These additional staff comprised Outreach tutors as well as lecturing staff within Dundee Dental School.

4.3.2 Development of a sub-set of class II amalgam cavities and full veneer gold shell crown preparations

a. Development of sub-set of class II amalgam cavity preparations

Previously, three senior teachers assessed 118 class II amalgam cavity preparations to determine inter-examiner agreement (see Chapter 3). The results showed poor agreement. Only 26 (22%) class II amalgam cavities demonstrated complete agreement between all three senior academic staff. This convenience sample, sub-set of 26 class II amalgam cavities was chosen to determine intra-examiner agreement of the senior academic staff as well as inter and intra-examiner agreement of the larger group of examiners (additional teaching staff). This sub-set of cavities was selected according to a particular characteristic. This type of sample is called stratified random sample (Greenhalgh, 2014). It is a random sample but the target is first stratified according to particular characteristics, the cavities with total agreed grades between the senior examiners were selected (n = 26 cavities). This smaller sub-set of cavities was considered to be a more manageable size for evaluation by a larger group of additional teaching staff instead of the original sample of 118 cavities. Furthermore, the proposed method for evaluation and re-evaluation by the additional teaching staff had to be completed within the same working day due to time and travel constraints of busy staff, often coming from outwith Dundee and with other clinical work commitments.

This sub-set of class II amalgam cavities (n=26) was used in this Chapter (i.e. to determine intra- and inter-examiner agreement), Chapter 5, Chapter 6 and Chapter 7.

b. Development of sub-set of full veneer gold shell crown preparations

For full gold shell crown preparation, two senior academic staff (the third was unavailable) evaluated 70 student-preparations to determine inter-examiner agreement (see Chapter 3). There was agreement for only 11 (16%) full veneer gold shell crown preparations. This was considered to be insufficient for further evaluation. In order to develop a sample size comparable with that used for the class II amalgam cavity evaluation, a sample size of 30 preparations was suggested. The additional full veneer gold shell crown preparations were sought from those where disagreement between the two senior academic staff was least (i.e. only one grade difference between examiners).

These sub-set of full veneer gold shell crown preparations (n=30) was used in this Chapter (i.e. to determine intra- and inter-examiner agreement), Chapter 5, Chapter 6 and Chapter 7.

4.3.3 Conditions and methods for examination of the class II amalgam cavity and full veneer gold shell crown preparation

a. Senior academic staff – class II amalgam cavity:

Previously, three senior teachers assessed 118 class II amalgam cavity preparations to determine inter-examiner agreement as an occasion one evaluation (see Chapter 3). The same three senior academic staff members assessed 26 class II amalgam cavity preparations as an occasion two evaluation. The cavities were examined in ambient (room) light with teeth set in an anatomical position within a phantom-head dental arch. The dental arch was examined out of the phantom head and could be held within the hand of the examiner and rotated as required. No measuring devices were used for the evaluation of class II amalgam cavity preparations. All evaluations were completed using visual assessment only and the personal preference of evaluation scale stated by each examiner.

A second evaluation, using the same criteria outlined in Chapter 3 (see Table 3.1), was undertaken, by the same three examiners, at least one week after the first evaluation. This was known as the 'Occasion two' evaluation.

b. Senior academic staff – full veneer gold shell crown

For a second time, due to examiner availability only two of the above senior academic staff members were available to assess the sub-set of 30 full metal crown preparations. For full gold shell-crown preparation, seventy (n = 70) full veneer gold shell crown preparations were assessed by two senior examiners as an occasion one evaluation (see Chapter 3). Crown preparations were also examined in ambient (room) light with teeth set in an anatomical position within a phantom-head dental arch. The dental arch was examined out of the phantom head and could be held within the hand of the examiner and rotated as required. No measuring devices were used for the evaluation of the full veneer gold shell crown preparations. All evaluations were completed using visual assessment only and the personal preference of evaluation scale stated by each examiner.

Thirty full veneer gold shell crown preparations were chosen as a sub-set (see section 4.3.2) from the first assessment ‘Occasion one’ of two senior examiners. A second evaluation, using the same criteria outlined in Chapter 3 (see Table 3.2), was undertaken, by the same two examiners, at least one week after the first evaluation. This was known as the ‘Occasion two’ evaluation.

c. Additional teaching staff – class II amalgam cavity

Additional academic staff members (n = 34) were recruited as part of two teaching days at Dundee Dental School in 2014. All additional teaching staff examiners were asked to rate the 26 class II amalgam cavities on two occasions, which were at the start and end of the first training day, using a five-point scale (from A to E) with which they were familiar with from the undergraduate clinic. The descriptions for this rating scale are shown in Table 4.1.

On each occasion, preparations were examined in ambient (room) light with teeth set in an anatomical position within a phantom-head dental arch. The dental arch was examined out of the phantom head and could be held within the hand of the examiner and rotated as required. No measuring devices were used. Tooth preparations were scored on paper on two occasions by additional academic staff examiners. After the first occasion, the same additional academic staff examiners re-scored similar tooth preparations after 4 hours.

Table 4.1 Grading system for class II amalgam cavity preparation at clinical skills laboratory

Clinical Grade	Description
A	Completes clinical task to high standard with no intervention/advice from staff
B	Can complete clinical task to clinically acceptable standard with minimum intervention/advice from staff
C	Requires intervention/advice from staff to complete clinical task to clinically acceptable standard
D	Very limited or no understanding of procedure. Requires significant intervention/support to complete clinical task to clinically acceptable standard. Makes errors in carrying out clinical procedures. Poor performance
E	Student observes procedure/intervention, but takes no active role

d. Additional teaching staff – full veneer gold shell crown

Additional academic staff members (n = 30) were recruited as part of two teaching days at Dundee Dental School in 2015. All additional teaching staff examiners were asked to rate the 30 full veneer gold shell crown preparations on two occasions, which were at the start and end of the first training day, using a five-point scale (from A to E) with which they were familiar with from the undergraduate clinic. The evaluation criteria used for evaluation of the full veneer gold shell crown preparation are described in Table 4.2.

On each occasion, preparations were examined in ambient (room) light with teeth set in an anatomical position within a phantom-head dental arch. The dental arch was examined out of the phantom head and could be held within the hand of the examiner and rotated as required. No measuring devices were used. Tooth preparations were scored on paper on two occasions by additional academic staff examiners. After the first occasion, the same additional academic staff examiners re-scored similar tooth preparations after 4 hours.

Table 4.2 Grading system for full gold shell-crown preparation at clinical skills laboratory

Grade	Description
A	No damage to adjacent teeth, appropriate taper, minimal chamfer all around, smooth flowing outlines, appropriate occlusal reduction and functional cusp bevel, margins mesial and distal below contact points
B	1-2 of the following which would still render the preparation fit for purpose: Minor damage to adjacent teeth, increased taper, chamfer too deep, irregular outline, too much occlusal reduction, inappropriate functional cusp bevel, margins at contact point
C	More than three of the following which would render the preparation unfit for purpose as it stood – but correctable with advice: Minor to moderate damage to adjacent teeth, increased taper, chamfer too deep, irregular outline, too much occlusal reduction, inappropriate functional cusp bevel, margins at contact point
D	Extensive damage to adjacent teeth
E	Wrong prepared tooth

4.3.4 Statistical analysis to determine intra-examiner agreement for senior academic staff and intra- and inter-examiner agreement for additional teaching staff

a. Intra-examiner agreement for senior academic staff

Assessment data of senior academic staff was entered into a statistical package (SPSS version 22) to determine intra-examiner agreement using the Cohen's un-weighted Kappa test for the class II amalgam cavity and for the full veneer gold shell-crown preparation. Alphabetical grades A, B, C, D and E for each examiner were changed into numerical values 5, 4, 3, 2 and 1 respectively before commencing statistical analysis by using SPSS package.

Inter-examiner agreement for senior academic staff members was determined for both the class II amalgam cavity and full gold shell-crown preparations in Chapter 3.

b. Intra-examiner agreement for additional academic staff

Assessment data of senior academic staff was entered into a statistical package (SPSS version 22) to determine intra-examiner agreement using the Cohen's un-weighted Kappa test for the class II amalgam cavity and for the full veneer gold shell crown preparation. Alphabetical grades A, B, C, D and E for each examiner were also changed into numerical values 5, 4, 3, 2 and 1 respectively before commencing statistical analysis by using SPSS package.

c. Inter-examiner agreement for additional academic staff

In this part of the study, inter-examiner agreement for additional academic staff was determined by using intra-class correlation ICC $_{(3,1)}$ (SPSS version 22). Single measures were calculated for each occasion of evaluations.

Intra-class correlation compares the variances of different mean square of items [(n = 26 class II cavities and n = 30 full veneer gold shell crown preparations] and judges (additional group of teaching staff) for both occasion one and occasion two evaluations to calculate the degree of absolute agreement between examiners. The additional teaching staff examiners can therefore only agree to award one of five categories which were expressed as a numerical value (5, 4, 3, 2 or 1).

4.4 Results

4.4.1 Panel of examiners

Three senior academic staff repeated the examination of the class II amalgam cavity whereas only two senior academic staff repeated the full veneer gold shell crown preparations.

In 2014, 34 examiners attended a teaching day in order to evaluate the class II amalgam cavities, whereas 30 examiners who attended the same annual training event in 2015 evaluated the full gold shell-crown preparations. These examiners were called additional teaching staff in this study (Figure 4.1).

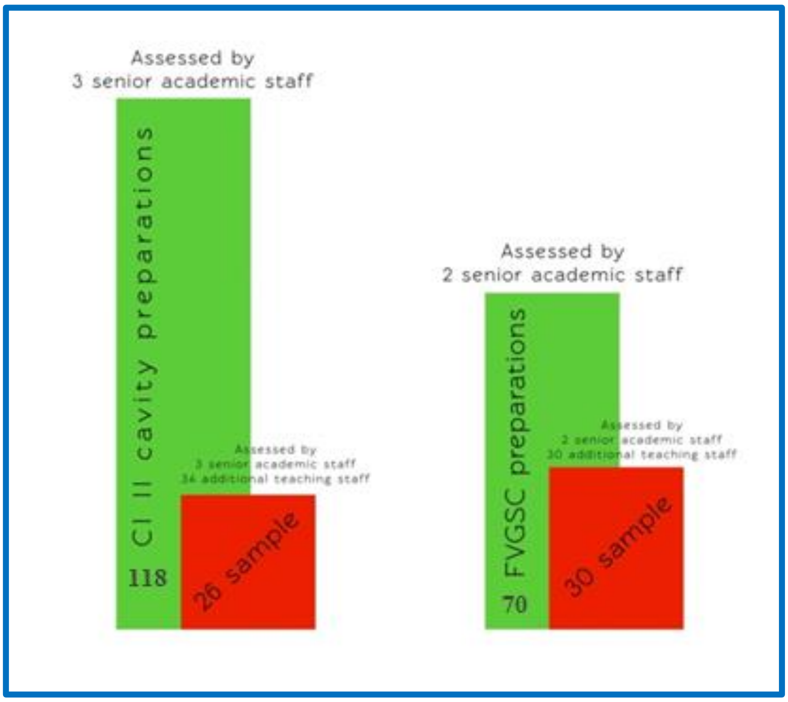


Figure 4.1 Diagram illustrate the numbers of preparations and samples (Stratified random sample) for class II amalgam cavity preparation (Cl II) and full veneer gold shell crown preparation (FVGSC) which were assessed by senior academic staff and additional teaching staff

4.4.2 Development of a sub-sets of Class II amalgam cavity and full veneer gold shell crown preparation

In 2014, a subset of 26 class II amalgam cavity preparations was selected where there was absolute agreement between senior academic staff as examiners. Ten cavities were described as ideal, three cavities were acceptable, six cavities were correctable and seven cavities were unacceptable. In 2015, a subset of 30 full veneer gold shell crown preparations was also selected where there was absolute agreement between senior academic staff as examiners for 11 preparations of which, five preparations were ideal, four preparations were acceptable and two preparations which were correctable. The remaining 19 full veneer gold shell crown preparations were selected on the basis of best agreement. For these preparations five were ideal, nine were acceptable and five were correctable. Best agreement was on the basis of only one grade difference between examiners.

4.4.3 Result of statistical analysis

a. Intra-examiner agreement for senior academic staff when evaluating a sub-set of class II amalgam cavity preparations

Intra-examiner agreement was determined for each member of senior academic staff individually using Cohen's un-weighted Kappa test. The results are displayed in Table 4.3 and show that each individual senior academic staff member only demonstrated fair intra-examiner agreement (Landis and Koch, 1977).

Table 4.3 Cohen's un-weighted Kappa intra-examiner agreement of senior academic staff as examiners of class II amalgam cavities

Examiner	Kappa value	Level of agreement	Significance ($p \leq 0.05$)
1	0.224	Fair	0.019
2	0.228	Fair	0.041
3	0.283	Fair	0.008

b. Intra-examiner agreement for senior academic staff when evaluating a sub-set of full veneer gold shell crown preparations

The Cohen's un-weighted Kappa value was determined for each examiner individually. The results are displayed in Table 4.4 and demonstrate that agreement was either fair or moderate (Landis and Koch, 1977).

Table 4.4 Cohen's un-weighted Intra-examiner Kappa agreement of senior academic staff as examiners of full veneer gold shell crown preparations

Examiner	Kappa value	Level of agreement	Significance ($p \leq 0.05$)
1	0.352	Fair	0.000
2	0.597	Moderate	0.000

c. Intra-examiner agreement for additional teaching staff when evaluating a sub-set of class II amalgam cavity preparations

Cohen's un-weighted Kappa intra-examiner agreement was determined for each individual member of additional teaching staff. The results are displayed in Table 4.5 and demonstrate a range of intra-examiner Kappa scores for members of the additional teaching staff.

In terms of percentage, 35.3% of the additional teaching staff demonstrated slight intra-examiner agreement (Kappa value = 0.00 to 0.20), 44.1% demonstrated fair agreement (Kappa value = 0.21 to 0.40) and 20.6% demonstrated moderate (Kappa value = 0.40 to 0.60) intra-examiner agreement (Table 4.6, Figure 4.2).

Table 4.5 Un-weighted Kappa intra-examiner agreement for additional teaching staff as examiners of class II amalgam cavities

Examiner code	Kappa value	Level of agreement	Significance ($p \leq 0.05$)	Examiner code	Kappa value	Level of agreement	Significance ($p \leq 0.05$)
1	0.232	Fair	0.067	**32**	0.15	Slight	0.232
2	0.473	Moderate	0.000	**33**	0.456	Moderate	0.000
3	0.319	Fair	0.006	**34**	0.443	Moderate	0.000
4	0.361	Fair	0.001	**35**	0.0211	Fair	0.044
5	0.573	Moderate	0.000	**36**	0.184	Slight	0.084
7	0.487	Moderate	0.000	**37**	0.254	Fair	0.015
8	0.052	Slight	0.669	**38**	0.116	Slight	0.217
10	0.321	Fair	0.003	**39**	0.022	Slight	0.849
13	0.346	Fair	0.004	**40**	0.111	Slight	0.273
24	0.424	Moderate	0.001	**41**	0.268	Fair	0.015
25	0.387	Fair	0.002	**42**	0.284	Fair	0.011
26	0.257	Fair	0.036	**43**	0.237	Fair	0.154
27	0.053	Slight	0.679	**44**	0.103	Slight	0.228
28	0.323	Fair	0.026	**45**	0.068	Slight	0.534
29	0.035	Slight	0.769	**46**	0.458	Moderate	0.000
30	0.067	Slight	0.614	**47**	0.251	Fair	0.027
31	0.265	Fair	0.031	**50**	0.165	Slight	0.112

The highlighted values represent statistically significant differences.

Table 4.6 Number and percentage of additional teaching staff and level of intra-examiner agreement as examiners of class II amalgam cavities

Examiner	Slight (k = 0.00-0.20)	Fair (k = 0.21-0.40)	Moderate (k = 0.41-0.60)
Total number of examiners	12	15	7
Percentage	35.29%	44.12%	20.59%

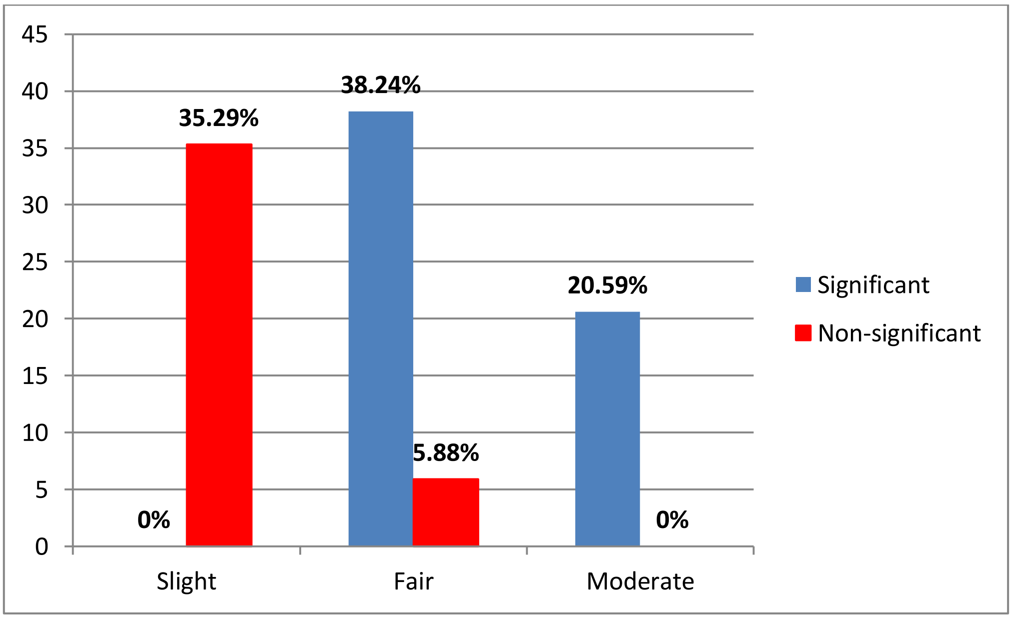


Figure 4.2 Bar chart of the percentage significant and non-significant intra-examiner agreement of additional teaching staff as examiners of class II amalgam cavities

d. Intra-examiner agreement for additional teaching staff when evaluating a sub-set of full veneer gold shell crown preparations

Intra-examiner agreement was then determined for each individual member of additional teaching staff by using Cohen's un-weighted Kappa test. The results are displayed in Table 4.7 and demonstrate a range of intra-examiner Kappa scores for members of the additional teaching staff.

In terms of percentage, 3.3% of the additional teaching staff demonstrated poor intra-examiner agreement, 33.3% demonstrated slight agreement, 23.3% demonstrated fair intra-examiner agreement and 40% demonstrated moderate intra-examiner agreement (Table 4.8 and Figure 4.3).

Table 4.7 Un-weighted intra-examiner kappa agreement for additional teaching staff as examiners of full veneer gold shell crown preparations

Examiner code	Kappa value	Level of agreement	Significance (p≤0.05)	Examiner code	Kappa value	Level of agreement	Significance (p≤0.05)
1	0.150	Slight	0.207	**18**	0.279	Fair	0.027
2	0.470	Moderate	0.000	**19**	0.459	Moderate	0.001
3	0.377	Fair	0.001	**20**	0.452	Moderate	0.000
4	0.190	Slight	0.208	**21**	0.506	Moderate	0.000
5	-0.024	Poor	0.814	**22**	0.450	Moderate	0.000
7	0.427	Moderate	0.000	**24**	0.125	Slight	0.363
8	0.592	Moderate	0.000	**26**	0.586	Moderate	0.000
9	0.249	Fair	0.039	**27**	0.006	Slight	0.949
10	0.331	Fair	0.002	**28**	0.177	Slight	0.107
12	0.143	Slight	0.222	**31**	0.124	Slight	0.219
13	0.546	Moderate	0.000	**32**	0.431	Moderate	0.005
14	0.330	Fair	0.005	**33**	0.412	Moderate	0.000
15	0.110	Slight	0.361	**34**	0.027	Slight	0.814
16	0.308	Fair	0.023	**35**	0.607	Moderate	0.000
17	0.163	Slight	0.208	**37**	0.258	Fair	0.018

The highlighted values represent statistically significant differences.

Table 4.8 Number and percentage of additional teaching staff and level of intra-examiner agreement as examiners of class II amalgam cavities

Examiner	Poor (K < 0.00)	Slight (K = 0.00-0.20)	Fair (K = 0.21-0.40)	Moderate (K = 0.41-0.60)
Total number of examiners	1	10	7	12
Percentage	3.34%	33.33%	23.33%	40%

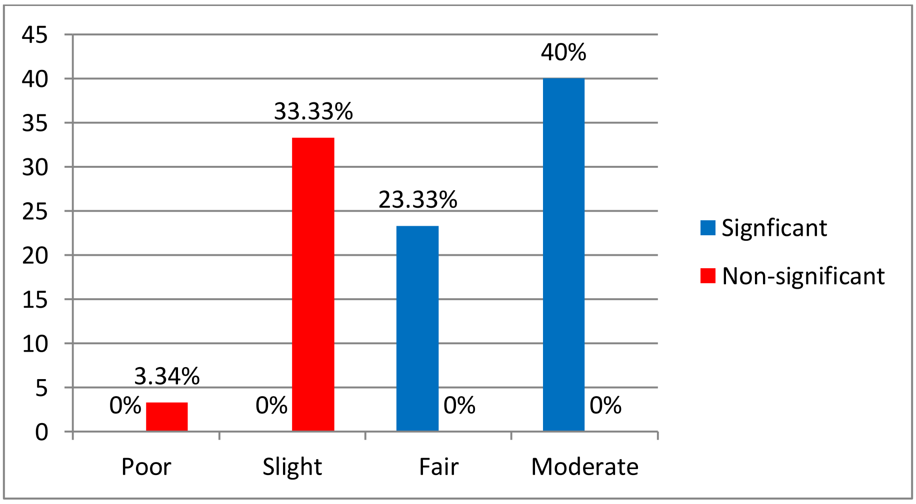


Figure 4.3 Bar chart of the percentage significant and non-significant intra-examiner agreement of additional teaching staff as examiners of full veneer gold shell crown preparations

e. Inter-examiner agreement for additional teaching staff when evaluating a sub-set of class II amalgam cavity preparations

The results for the intra-class correlation coefficient (ICC) for the evaluation of class II amalgam cavity preparations by additional teaching staff are shown in Table 4.9.

Table 4.9 Intra-class Correlation Coefficient for a large group of additional teaching staff for class II amalgam cavity preparation assessment on two occasions

Occasion	Intra-class Correlation Single measure	95% Confidence Interval		Significance ($p \leq 0.05$)
		Lower Bound	Upper Bound	
1	0.413	0.294	0.579	0.000
2	0.463	0.339	0.627	0.000

As shown in Table 4.9, inter-examiner agreement of the additional teaching staff who assessed the class II amalgam cavities was low (single measure of ICC= 0.413 in the occasion 1 and single measure of ICC= 0.463 in the occasion 2).

f. Intra-examiner agreement for additional teaching staff when evaluating a sub-set of full veneer gold shell crown preparations

As shown in Table 4.10, single measures of intra-class correlation demonstrated there was *poor* agreement between additional teaching staff (single measures of ICC = 0.370) in the occasion 1 and (single measures of ICC = 0.429) in the occasion 2. Thus, the intra-class correlation value of additional group of teaching staff in the occasion 2 was marginally better than in the occasion 1.

Table 4.10 Intra-class Correlation Coefficient for a large group of additional teaching staff for full gold shell-crown preparation assessment on two occasions.

Occasion	Intra-class Correlation Single measure	95% Confidence Interval		Significance ($p \leq 0.05$)
		Lower Bound	Upper Bound	
1	0.370	0.268	0.509	0.000
2	0.429	0.322	0.569	0.000

4.5 Discussion

4.5.1 Development of a sub-set of class II amalgam cavities and full veneer gold shell crown preparations

The convenience samples of class II cavity preparations and full veneer gold shell crown preparations were chosen for three reasons. First, they were samples where there was best agreement by senior academic staff and second, the spread of grades within each sample represented a spectrum of student achievement from ideal to unacceptable. The third reason was the simple practicality of having a sample size suitable for evaluation by a large group of examiners within a reasonable time period. Similar sample sizes have been chosen by other authors evaluating similar outcomes (Goepfred and Kerber, 1980, Helft et al., 1987, Cardoso et al., 2006, Seo et al., 2014).

4.5.2 Intra-examiner agreement for senior academic staff when evaluating a sub-set of class II amalgam cavity preparations and full-veneer gold shell crown preparations.

The agreement and disagreement for each examiner who assessed both class II amalgam cavity, and full veneer gold shell crown, preparations was analysed by determining a Kappa value. Evaluating intra-examiner is a common objective of many medical and dental research studies. The concept of examiner agreement is fundamental to the design and evaluation of feedback or assessment instruments which are used in different institutes. There are different statistical tests that exist to measure agreement. There is often confusion about which test is most appropriate to use. This may lead to incomplete and inconsistent reporting of results, if the test is not appropriate (Kottner et al., 2011, Gisev et al., 2013).

In this part of study, the information to consider when selecting an intra-examiner agreement test was that the number of examinations was a maximum of two. In addition and despite the fact that different scales were used, the data were both ordinal (categorised from best to worst) and categorical (pass or fail). The Cohen's un-weighted Kappa test can be applied to data that are not normally distributed, but it is best suited to a close-ended ordinal scale, such as the five point scales used in this study (Cohen, 1960). Cohen also argued that the weighted Kappa should be used particularly if the variables have more categories than binary (more than yes and no), because the distance from agreement should also be taken into consideration (Cohen, 1968).

In this part of the study, evaluation of examiner agreement using Cohen's un-weighted Kappa score is acceptable. Evaluation of ordinal data only would normally require the use of a weighted Kappa score while the use of categorical data would require the use of an un-weighted Kappa score (Kottner et al., 2011, Gisev et al., 2013). However, because the candidate needs to know whether they have passed or failed and then, in addition, what score they achieved and what the feedback should be provided, Cohen's un-weighted Kappa scores were calculated as nominal categories for all data (Cohen, 1960) (see Chapter 3). Therefore, Cohen's un-weighted Kappa test was selected to evaluate intra-examiner agreement for a group of three senior academic staff who assessed 26 class II cavity preparations. The value of un-weighted Kappa was positive but it was low. Kappa value was only fair for three senior academic staff (Table 4.3). The p-value was less than 0.05 which is significant, thus, the null hypothesis that there is no intra-examiner agreement for senior examiners evaluating class II cavities is accepted.

Intra-examiner agreement for two senior academic staff who assessed 30 full veneer gold shell crown preparations demonstrated that the Kappa values were positive and statistically significant for both examiners although one examiner had fair-agreement while the other had moderate-agreement, bordering on good-agreement (Table 4.4). This variation is not good for students as feedback. Feedback should have better agreement. Thus, the results support the null hypothesis that there is no intra-examiner agreement for senior examiners evaluating full veneer gold shell crown preparations and, therefore, the null hypothesis is accepted.

These results were supported by many other authors (Natkin and Guild, 1967, Fuller, 1972, Houpt and Kress, 1973, Helft et al., 1987). The reasons for agreement variability might be due to examiner experience, internal examiner bias, interpretation and design of rating scales, knowledge or training reason (Natkin and Guild, 1967, Fuller, 1972, Houpt and Kress, 1973, Helft et al., 1987).

4.5.3 Intra- and inter-examiner agreement for additional teaching staff when evaluating a sub-set of class II amalgam cavity preparations and full veneer gold shell crown preparations.

In order to determine the intra- and inter-examiner agreement of additional teaching staff, who are often experienced practitioners who come from varied backgrounds, it is essential to evaluate what agreement exists for these staff at Dundee Dental School. Furthermore, it is important to establish the extent of any potential problem existing within evaluation of student clinical work.

The Cohen's un-weighted Kappa test was applied to calculate intra-examiner agreement for class II amalgam cavity preparations and full veneer gold shell crown preparations in the same way as for senior academic staff. For inter-examiner agreement, intra-class correlation ICC $_{(1,3)}$ was also calculated for class II amalgam cavity preparations and full veneer gold shell crown preparations in the same way as for senior academic staff (see section 4.4.3).

Cohen's Un-weighted Kappa values of intra-examiner agreement for the group of additional teaching staff who assessed 26 class II cavity preparations were positive. Seven out of 34 examiners had moderate agreement while the rest were slight or poor (Table 4.6). These results demonstrated a lack of agreement between the occasion one and occasion two for most examiners. Moreover, the p-value was less than 0.05 thus, the null hypothesis that there is no intra-examiner agreement for additional teaching staffs, when evaluating class II cavities, is accepted.

For full veneer gold shell crown preparation, the majority of un-weighted Kappa values for each examiner of additional teaching staff were positive. Twelve out 30 examiners had moderate agreement while the rest were slight or poor (Table 4.8). These results demonstrated a lack of agreement between the occasion one and occasion two for most examiners. Moreover, the p-value was less than 0.05 thus, the null hypothesis that there is no intra-examiner agreement for additional teaching staffs, when evaluating full veneer gold shell crown preparations, is accepted.

Single measures values $ICC_{(1,3)}$ for both class II amalgam cavity preparation and full veneer gold shell crown preparation which assessed by additional teaching staff were poor to moderate (Tables 4.9 and 4.10).

The results of this part of study support the conclusion that significant intra and inter-examiner variation occurs in evaluating student's preparation by using 'glance and grading' method (Sherwood and Douglas, 2014).

According to Sherwood and Douglas, (2014), the traditional glance and grading method should be changed to an objective checklist criteria scoring method which reduces the examiner variability and seeks to improve the examiner reliability especially with regard to lesser experienced faculty members.

In the current study, intra- and inter-examiner variability of individual appeared during assessment because a rating scale was used which did not define criteria and their levels of performance objectively and clearly. Providing a definitive level of performance for each criterion provides a greater consistency and agreement of evaluation (Gaines et al., 1974, Goepfred and Herber, 1980, Knight, 1997). In addition, the rating system which was used in this part of study had a five-point rating scale (Tables 4.1 and 4.2) for additional teaching staff which was poorly defined for each component of the tooth preparation (e.g. depth of occlusal cavity).

Knight (1997) demonstrated many requirements for criteria of assessment method and scaling system in order to reduce variability of agreement and reliability. Criteria must be valid individually, collectively, and non-compensatory (independent of one another). Moreover, criteria must be reliable. All criteria must clearly define with specifically described tests and levels of performance. Therefore, the level of each scale must be confined to those that can be clearly outlined to provide accurate feedback for the student (Knight, 1997).

4.6 Conclusions

From the results, there is moderate to poor intra- and inter-examiner agreement for both senior academic and additional teaching staff who undertake the daily grading of preclinical procedures at Dundee Dental School. The level of intra- and inter-examiner agreement was disappointing for both senior academic staff and the large group of additional teaching staff. These results indicate that the problem of inadequate inter- and intra-examiner agreement exists at Dundee Dental School.

In addition, the result of Chapter 3 and 4 supported the conclusion of the Chapter 2. "Although the students preferred the feedback from the tutor, the tutor did not provide satisfied comments for them. Most of the student complaints were that they did not receive enough feedback to both compare and enhance their performance in relation to according to an ideal or acceptable standard" (see Chapter 2).

In order to assess whether students acquired a dental skill, a valid and reliable assessment should be developed to improve examiner agreement and feedback to student learners in a clinical skills laboratory (Taylor et al., 2013). The same could surely be said for such skills on the clinic and with real patients.

There are several studies which have tried to improve intra- and inter-examiner agreement (Salvendy et al., 1973, Geopferd and Kerber, 1980, Knight, 1997, Jenkins et al., 1998, Haj-Ali and Feil, 2006, Kateeb et al., 2016), using different assessment methods. The most successful method, the focus of the next chapter (Chapter 5), was the development of a feedback sheet to improve intra- and inter-examiner agreement.

Chapter 5 : Evaluation of selected student class II amalgam cavities and full gold shell crown preparations by using specific additional tools and feedback sheets

Chapters	
Chapter 2	**Evaluating current assessment and feedback methods at Dundee Dental School in the Clinical Skills Laboratory** **Aims:** The aims of this Chapter were to identify, in relation to the class II amalgam cavity preparation and the full veneer gold shell crown preparation undertaken in the Clinical Skills Laboratory at Dundee Dental School: 1. the types and quality of feedback and assessment provided for students and, 2. the disadvantages associated with current feedback.
Chapter 3	**Independent evaluation of student class II amalgam cavity and full gold shell-crown preparations by senior academic staff** **Aim:** The aim of this study was to evaluate inter-examiner agreement between a convenience sample of three senior academic staff members, who have more than twenty years of clinical and teaching experience each, when evaluating operative procedures (class II amalgam cavity and full veneer gold shell crown preparation) on plastic teeth, from a sample year-cohort of dental students, at Dundee Dental School.
Chapter 4	**Evaluation of selected student class II amalgam cavity and full gold shell-crown preparations by senior academic staff and additional teaching staff** **Aims:** 1. To develop a sub-set of class II amalgam cavity, and full veneer gold shell crown, preparations. 2. To determine intra-examiner agreement for a group of senior academic staff when evaluating a sub-set of class II amalgam cavity preparations and full-veneer gold shell crown preparations. 3. To study a large group of additional teaching staff at Dundee Dental School in order to determine intra- and inter-examiner agreement when evaluating a sub-set of class II amalgam cavity preparations and full-veneer gold shell crown preparations.
Chapter 5	**Evaluation of selected student class II amalgam cavity and full gold shell crown preparations by using specific additional tools and feedback sheets** **Aims:** **1. To identify and develop specific additional tools to assist with evaluation of tooth preparations.** **2. To determine examiner agreement for both class II amalgam cavity and full veneer gold shell crown preparations using the specific additional tools and feedback sheets for both grades awarded and repeatability and reproducibility of detailed feedback provided by senior academic staff.**
Chapter 6	**Identification of gold standard (representative grades) for 26 class II amalgam cavities and 30 full gold shell-crown preparations** **Aims:** These objectives can be set out as a number of steps, each of which had a specific aim which is expanded within the methodology of this chapter. Thus, the two aims of this chapter were: 1. to identify the best senior academic staff examiner who had the highest correlation between grades awarded with the number of negative points identified; 2. to identify those tooth preparations which were awarded a grade by the best examiner in order to: • determine the specific feature measurements of this sub-group of grade tooth preparations and: • compare the objective class II amalgam cavity or full veneer gold shell crown preparation dimensions (SAMFs) recorded by the researcher and those subjective tooth preparation evaluations (SAFs) for similar teeth by the senior examiner with dimensions presented in the dental literature and subsequent calibration with the grades awarded by the best examiner.
Chapter 7	**Identification of reliable tools and development of a new checklist to assess class II amalgam cavities and full veneer gold shell crown preparations** **Aims:** 1. To identify a reliable specific additional measurement tool. 2. To develop two reliable new checklists (New Checklist CII Preparation and New Checklist Gold shell Crown Preparation). 3. To determine intra- and inter-examiner agreement using the grades and consistency of the New Checklists (Class II Preparation and Gold Shell Crown Preparation) and compare these data with that from previously used 'Gray and Mhanni feedback sheets' to ascertain if the agreement and consistency of the New Checklists have been improved.
Chapter 8	**General conclusions, recommendation, and further studies**

5.1 Introduction

The previous two chapters have demonstrated poor intra- and inter-examiner agreement to rank tooth preparations using visual examination for class II amalgam cavities and full veneer gold shell crown preparations, completed by a convenience sample of undergraduate dental students and examined by senior academic staff who used their own criteria for assessment. Effectively, this means that undergraduate students could have received inconsistent grading of their work and inconsistent feedback leading to poor understanding of the processes and theory required to successfully complete these types of tooth preparation. It was, therefore, important to find a way of improving this agreement for the benefit of students, staff and the educational practises within the Department. Clearly, such a method should be valid and reliable, as discussed previously in this thesis, and facilitate student learning through feedback (formative assessment) in order to determine whether the student has acquired the relevant skills (summative feedback).

Taylor et al., (2013) suggested that a valid and reliable assessment should be developed that employs an appropriate standard setting. This type of assessment acts as the gateway to practicing dentistry on real patients. Traditionally, tooth preparations are evaluated subjectively by visual inspection, helped sometimes by measuring tools, for example, a feedback sheet (Taylor et al., 2013). The validity and reliability of the feedback sheet is improved when it is accompanied by additional analytical methods which are objective and effective to determine whether the minimum requirement for the skill is met (Goepferd and Kerber, 1980, Sherwood and Douglas, 2014). However, evaluation by using a feedback sheet, without the accompaniment of additional analytical methods, is often subject to assessor bias and misinterpretation (Feil, 1982).

This part of the study attempted three things which are listed below.

1. To identify additional tools for the evaluation of tooth preparations.
2. Using these additional tools, could we improve the intra- and inter-examiner agreement of senior academic staff for the grading of class II amalgam cavities and full veneer gold shell crown preparations?
3. Although the ability to determine repeatedly a grade for a tooth preparation is important, it is undoubtedly more useful for the student to be informed consistently why their tooth preparation was awarded a particular grade thus making feedback consistent.

The following chapter is therefore set out by defining the aims, the hypotheses and identifying additional tools for the evaluation of tooth preparations. There follows an extensive section describing the methodology for evaluation of these tools to determine inter and intra-examiner repeatability of grades awarded and feedback given. The chapter ends with a summary.

5.2 Aim and Null hypothesis

Aims of this part of study were:

- To identify and develop specific additional tools to assist with the evaluation of class II amalgam cavity and full veneer gold shell crown preparations (see section 5.3).
- To determine examiner agreement using the specific additional tools and feedback sheets for:
    - grades awarded (see section 5.4) and
    - repeatability and reproducibility of detailed feedback provided by senior academic staff (see section 5.4).

Null Hypothesis:

The use of specific additional tools (including the development of a feedback sheet) to rank and provide feedback for tooth preparations by students does not improve intra- and inter-examiner agreement of senior academic staff.

5.3 Selection of specific additional tools and development feedback sheets

5.3.1 Introduction

a. Selection of specific additional tools

Dentists are creatures of habit. Once they find a method that works for them they often do not change it without very good reason (Suvinen et al., 1998). It is, therefore, very important that when students are first exposed to a procedure there is careful selection of the tools that are recommended for the task. It, therefore, seems sensible to utilise dental tools which students used during tooth preparations in a phantom head to evaluate their work.

b. Development of feedback sheet

In addition to the specific additional tools, a feedback sheet should be utilised to assess the student performance accurately. Sherwood and Douglas (2014) concluded that *"This study calls for change in evaluation of preclinical operative work of students from traditional glance and grading method to objective checklist criteria scoring method which reduces the examiner variability, this method of scoring has to be introduced after sufficient training and calibration sessions for improving the examiner reliability especially with regard to lesser experienced faculty members"*. Therefore, feedback sheets for class II amalgam cavity preparation and full veneer gold shell crown preparation were established.

5.3.2 Material and methods

a. Tools for more objective measurement of tooth preparations

With regard to the preparation and restoration of a class II amalgam cavity, students at Dundee Dental School are advised to cut the whole preparation using a high-speed diamond fissure bur (BDM541, UnoDent LTd: cutting length = 4.00mm, diameter = 1.00mm) as shown in Figures 3.1 and 5.1. Once the cavity has been completed it must be large enough to accept an amalgam condenser (Amalgam plugger, Blacks # 0/1, DE Healthcare Products: length = 2.50mm, diameter = 1.00mm) with which to pack the amalgam restorative material appropriately (Figure 5.1). Both the bur and the amalgam condenser can be used to effectively measure the dimensions of a class II amalgam preparation in a phantom head tooth.

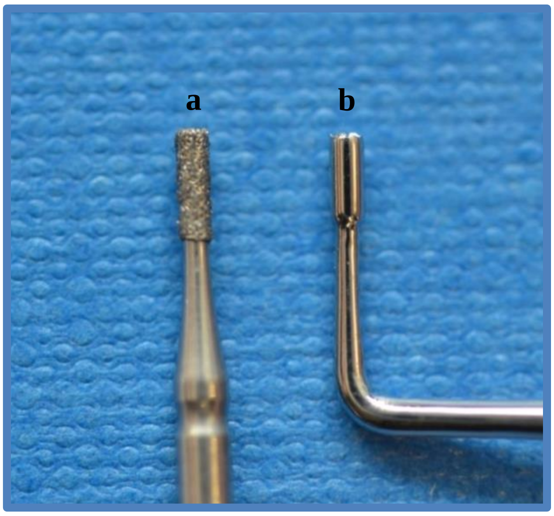


Figure 5.1 Picture of specific additional tools for Class II amalgam cavity evaluation
a. Diamond fissure bur **b.** Amalgam condenser

With regard to the preparation of a full veneer gold shell crown preparation at Dundee Dental School, the use of pre-operative silicone indices is recommended along with the use of a tapered high-speed diamond bur with a rounded tip (No.8856.314.016 TPS2, Komet Dental, UK) to create a tapered preparation with a chamfered finish line (Figures 3.2 and 5.2). The pre-operative indices can be used to evaluate the amount of tooth reduction in three dimensions (Figure 5.2). This is possible, to an extent, with tapered high-speed diamond with a rounded tip (Chamfer) bur although it is easily used to evaluate the quality of the finish line and the presence or absence of undercuts (Blair et al., 2002, Shillingburg et al., 2012).

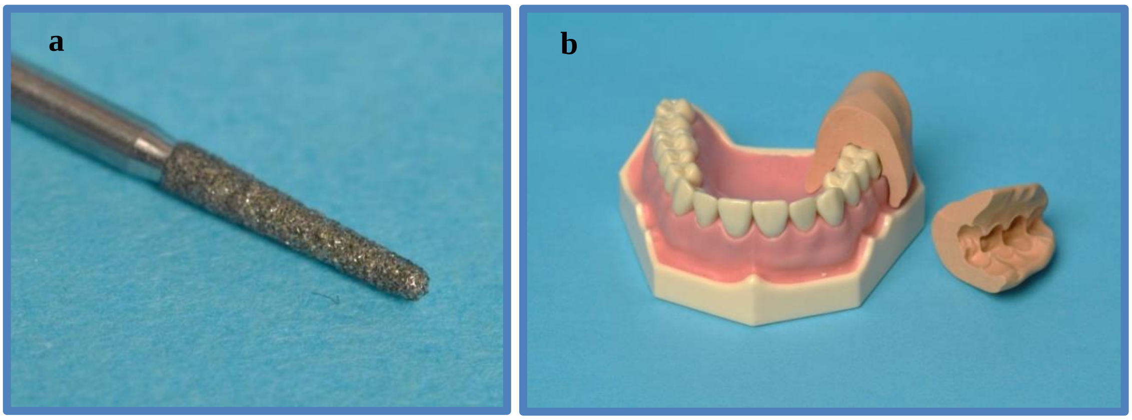


Figure 5.2 Picture of specific additional tools for Full veneer gold shell crown preparation evaluation
(**a.** Tapered diamond bur with a rounded tip, **b.** Pre-operative sectional indices)

For the selection of specific additional tools it, therefore, seems sensible to utilise the amalgam condenser and the diamond fissure bur to evaluate a class II amalgam cavity prepared in a phantom head tooth. In the same way, tapered high-speed diamond with a rounded tip (Chamfer) bur and the silicone indices are sensible specific additional tools with which to evaluate the full veneer gold shell crown preparation in a phantom head tooth.

b. The feedback sheet development

In addition to the specific additional tools described above, a feedback sheet was used for both preparation types. The feedback sheet for the class II amalgam cavity was generously provided by Dr Gordon Gray from the University of Bristol Dental School and was one he had used for several years to evaluate the phantom head tooth preparations by Bristol Dental Students (Gray 2016, personal communication). This feedback sheet is subsequently referred to as the 'Gray feedback sheet' in this thesis. No such feedback sheet existed for the full veneer gold shell crown preparation at Dundee Dental School and, therefore, a feedback sheet for this purpose was developed. This feedback sheet is subsequently referred to as the 'Mhanni feedback sheet' in this thesis.

In order to connect the concept of a feedback sheet with identifiable and measurable parameters, it is generally considered that determination of the content and criterion validity are fundamental steps for validation of this type of assessment method (Lynn, 1986, Wynd et al., 2003).

Content validity was used in order to measure whether the feedback sheet sufficiently covers the area, it is intend to cover. Content of feedback sheet should be reflected in all areas in which it is intended to cover. This type of validity was assessed only through the experts' ratings. According to Lynn (1986), there is a two-step process in order to determine the content validity for, in this case, a feedback sheet.

For content validity, the first stage is a review of the literature to identify the content of the feedback sheet. This stage is called, 'The Development Stage'. In this stage, the literature review and opinion of examiners are essential steps to collect the items (categories) and their criteria of the feedback sheet which will also include instructions for use and the grading system. Content Validity Index for Items (I-CVI) is created from the first stage for the second stage (Table 5.1) (Lynn, 1986, Netemeyer et al., 2003, Polit and Beck, 2006, DeVon et al., 2007). Content validity index (CVI) information was used in order to guide the researcher in revising, deleting or substituting items and their criteria of the feedback sheet.

The second stage requires a panel of examiners (i.e. senior academic staff) for 'The Judgement-Quantification Stage' to determine content validity. Examiners work independently in order to evaluate the Content Validity Index for Items (I-CVI) by rating items and criteria according to their relevance (Lynn, 1986, Berk, 1990, Polit and

Beck, 2006). According to Lynn (1986), a four point scale: 1 = not relevant, 2 = somewhat relevant, 3 = relevant and, 4 = very relevant, should be used for rating items and/or criteria (I-CVI) in order to determine whether the items and/or criteria should be used or excluded. Items and/or criteria can be exchanged or modified, if the examiners feel they are not or somewhat relevant. Content validity is calculated by counting the results of the examiners based on the degree to which the examiners agree on the relevance of the items and criteria. Content Validity Index for Scale (S-CVI) is also calculated as a number of items giving a rating of either 3 or 4 by the experts, divided by the total number of the items on the instrument - that is also, the proportion in agreement about relevance (Lynn, 1986).

Items and/or criteria should be ranked 3 = relevant or 4 = extremely relevant by all examiners in order to be included as an item and/or criterion in the new feedback sheet (Lynn, 1986, Polit and Beck, 2006, DeVon et al., 2007, Sirajudeen et al., 2012).

Criterion validity is also used, as a type of validation, to identify the relationship between negative points of feedback sheet with grades awarded. A negative point is given for each feature of the tooth preparation that is not acceptable according to the assessment of senior examiners. The strength of correlation between negative points and grades supports the extent to which the feedback sheet provides accurate feedback to the students (Waltz et al., 2010).

<u>Determination of the validity of class II amalgam cavity preparation using the 'Gray feedback sheet'</u>

The 'Gray feedback sheet' has been used for many years by the University of Bristol and it was considered sensible to establish 'Criterion validity' using a single examiner at Dundee University. The following paragraphs describe how this was undertaken.

Two hundred and forty two mesio-occlusal cavities (i.e. class II amalgam cavities) in plastic upper premolar teeth were completed by a cohort of second year dental students within the Operative Skills Laboratory under examination conditions in 2014/2015. Each student within the cohort completed up to four different cavities on a different days. The cavities were coded with a random number taken from a book of raffle tickets so the evaluator was blinded to which student completed each cavity (Appendix 1).

One senior academic member of staff (RGC) initially assigned each cavity to an interim grading category based upon his initial gut reaction. Thereafter, a more detailed assessment of the cavities ensued with adjustment of the interim grade being made, where appropriate, based upon the visually apparent "goodness-of-fit" to the category.

Next, for each cavity, the 'Gray feedback sheet' was completed by one examiner (RGC) and scored by awarding one mark for each error (negative point). An error was defined by any of the criteria appearing in the two right-hand-side columns in Figure 5.3. Moderate/severe damage of the adjacent tooth automatically incurred a maximum eleven marks. The, 'Number of negative points', was recorded in Appendix 1.

Finally, where the feedback sheet did not justify the initial grade awarded, there was an opportunity for further adjusting the final grade. The, 'Final grade', was recorded in Appendix 1.

For criterion validity, determining the relationship between final grades and the number of negative points was essential in order to ensure this feedback sheet provided accurate or reasonable advice for the student's work. This relationship was determined by calculation of the correlation of the number of negative points from the 'Gray feedback sheet' with grading ultimately awarded by one of the senior academic staff (RGC).

A Spearman correlation coefficient was calculated to explore the relationship of 'Final Grade' awarded to the 'Number of negative points', by using the statistical Prism package (Version 6.0, GraphPad software, USA).

Management of Dental Caries Course
Feedback and Marking Sheet

CANDIDATE
Test
Date
Class II

BOX			
Outline	OK	Rough/irregular	
Position	OK	Too far B/P	
Depth gingivally	OK	Too deep	Too shallow not clear G
B-P width	OK	Too wide	Narrow not clear B/P
M-D depth	OK	Too deep pulpo-axial wall	Too shallow pulpal- axial wall
Unsupported enamel	None	Yes – B/P/G	
Retention form	OK	Parallel walls	Divergent walls

OCCLUSAL			
Lock	OK	Rough/irregular	Not follow fissure
Depth	OK	Too deep	Too shallow
Width	OK	Too wide	Too narrow

DAMAGE ADJACENT TOOTH			FAIL
	NONE	Minor	Mod/Severe

Score errors as 1 mark each with a FAIL incurring 11marks

Figure 5.3 Schematic representation of the original 'Gray feedback sheet'

<u>Development and determination of the validity of full veneer gold shell crown preparation 'Mhanni feedback sheet'</u>

A new feedback sheet was proposed to evaluate the full veneer gold shell crown preparation. For content validity, the two-stage process comprising the, 'Development stage' and the, 'Judgement-quantification stage' was required to determine the content validity of the proposed feedback sheet (Martuza, 1977, Lynn, 1986, Polit and Beck, 2006, DeVon et al., 2007). This new feedback sheet for the full veneer gold shell crown preparation was called the, 'Mhanni feedback sheet'.

For the, 'Development stage' three elements were used. These elements were; a) comments from additional teaching staff, b) the grading systems used by three senior academic staff and, c) information provided by a narrative review of the literature in the section 1.4.2 (Polit and Beck, 2006, Sirajudeen et al., 2012).

A large group of additional teaching staff (n = 30) was recruited as part of a teaching day in November 2015 at Dundee Dental School. A collection of 30 full veneer gold shell crown preparations was identified in Chapter 4. The additional teaching staff were asked to rate the same collection of full veneer gold shell crown preparations identified in Chapter 4, on two occasions. Furthermore, the additional teaching staff were asked to provide their comments for each preparation. From, a) the additional teaching staff comments and b) the components of the grading systems devised by each of the senior academic staff and c) the narrative literature review, the initial categories (items) and criteria of a feedback sheet for the full veneer gold shell crown preparation was constructed.

For the, 'Judgment-Quantification Stage', a four point scale, range from 1 = not relevant to 4 = extremely relevant, was used, by senior academic staff, to rate items and/or criteria to determine whether such items and/or criteria should be used or excluded, exchanged or modified. Only items and/or criteria which were ranked as 3 = relevant or 4 = very relevant by all senior academic staff were selected for the final content of this additional and new feedback sheet for the full veneer gold shell crown preparation. The results of this process can be found in section 5.3.3.

Having devised the 'Mhanni feedback sheet' it was necessary to determine the criterion validity using a single examiner (DNJR) at Dundee University. The following paragraphs describe how this was undertaken.

Seventy full veneer gold shell crown preparation in plastic upper first molar teeth were completed by a cohort of third year dental students within the Operative Skills Laboratory under examination conditions in 2014/2015. Each student within the cohort completed one preparation. The preparations were coded with a random number so the evaluator was blinded to which student completed each preparation (Appendix 2).

One senior academic member of staff (DNJR) initially assigned each preparation to an interim grading category based upon his initial gut reaction. Thereafter, a more detailed assessment of the preparations ensued with adjustment of the interim grade being made, where appropriate, based upon the visually apparent "goodness-of-fit" to the category.

Next, for each preparation, the 'Mhanni feedback sheet' was completed and scored by awarding one mark for each error (negative point). An error was defined by any of the criteria appearing in the two right-hand-side column in Figure 5.4. Moderate to severe damage of the adjacent teeth and destructive shape of bucco-lingual and proximal convergence of prepared tooth automatically incurred a maximum 20 marks. The, 'Number of negative points', was recorded in Appendix 2.

Finally, where the 'Mhanni feedback sheet' did not justify the initial grade awarded, there was an opportunity for further adjusting the final grade. The, 'Final grade', was recorded in Appendix 2.

FEEDBACK SHEET FOR GOLD CROWN

Model Number: *Grade:*

Occlusal surface			
Occlusal reduction	Adequate	Under-prepared	Over-prepared
Contour of occlusal preparation	Yes (follow the contour tooth surfaces)	No	
Axial surface(s)			
Buccal reduction	Adequate	Under-prepared	Over-prepared
Lingual reduction	Adequate	Under-prepared	Over-prepared
Mesial reduction	Adequate	Under-prepared	Over-prepared
Distal reduction	Adequate	Under-prepared	Over-prepared
Undercuts	No	Yes	Yes – more than one surface
Bucco-lingual convergence	Adequate (Convergence)	Improper convergence*	No (destructive shape)** FAIL
Proximal convergence	Adequate (Convergence)	Improper convergence*	No (destructive shape)** FAIL
Contour of preparation	Follow tooth surfaces contour	One not follow	More than one not follow
Contact area with adjacent teeth	Cleared	Yes – one side	Not clear
Functional cusps			
Functional cusp bevel reduction	Adequate	Not - symmetrical	No
Location of functional bevel	Adequate	No	
Finish line			
Chamfer finish line	Yes	No	
Level of finish line to gingival margin	At gingival (at or above by 0.5mm)	Supra-gingival (more than 0.5mm)	Subgingival and/or Supraginigival
Depth of finish line all around	Even	Uneven	Deep

Final preparation			
Texture of final preparation except margin	Adequate	Rough (irregular) Sharp edges	
Texture of margin	Smooth and well define	Rough (irregular)	
Adjacent teeth damage			
Mesial tooth	No damage	Minor damage	Moderate/severe damage FAIL
Distal tooth	No damage	Minor damage	Moderate/severe damage FAIL

**Improper convergence = one wall is taper or two walls are parallel.*

***Destructive shape = over-prepared, occlusal wider than cervical or tapered tooth.*

(Negative points) out of 20 = Percentage score = % Grade

Score errors as 1 mark each with a FAIL incurring 20 marks

Figure 5.4 Schematic representation of the developed 'Mhanni feedback sheet'

For criterion validity, determining the relationship between final grades and the number of negative points was essential in order to ensure the 'Mhanni feedback sheet' provided accurate or reasonable advice for the student's work. This relationship was determined by calculation of the correlation of the number of negative points from the 'Mhanni feedback sheet' with grading ultimately awarded by one of the senior academic staff (DNJR).

A Spearman correlation coefficient was calculated to explore the relationship of 'Final Grade' awarded to the 'Number of negative points', by using the statistical Prism package (Version 6.0, GraphPad software, USA).

5.3.3 Results

a. Selection of specific additional tools and validity of 'Gray feedback sheet'

From the previous, seems sensible to select the amalgam condenser and the diamond fissure bur to evaluate a class II amalgam cavity prepared in a phantom head tooth. In addition, validity of 'Gray feedback sheet' had already been undertaken by the University of Bristol. In this part of study, criterion validity was undertaken using a single examiner (RGC) at Dundee University.

For criterion validity, Appendix 1 shows the grade awarded and number of negative points awarded. The ticket number was used to identify the student. It is included here as a reference for possible future further work. The data are plotted in Figure 5.5. For the whole dataset, a Spearman correlation analysis showed a high negative correlation (r = -0.83) between the grade awarded and the number of faults found. It can be said that the test had a high degree of validation support, and it can be used as a selection tool (DeVon et al., 2007).

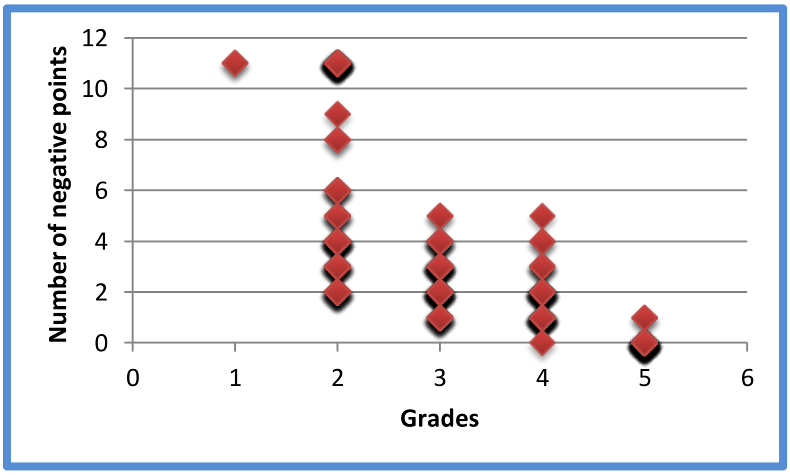


Figure 5.5 Scatter plot of the grades and the number of faults (negative points) awarded from the 'Gray feedback sheet'

<u>All tests Spearman r = -0.83 (95% Confidence Interval's -0.79 - -0.87)</u>

These data indicate good correlation with between the 'Gray feedback sheet' and a previously used ranking scale. The advantage of the 'Gray feedback sheet' for the students was the ability to provide more detailed feedback to the student.

b. Selection of specific additional tools and validity of 'Mhanni feedback sheet'

Tapered high-speed diamond with a rounded tip (Chamfer) bur and the silicone indices are sensible specific additional tools with which to evaluate the full veneer gold shell crown preparation in a phantom head tooth.

To develop a feedback sheet for full veneer gold shell crown preparations, the researcher (AM) defined the construct of interest and dimensions by searching the literature and seeking expert opinions (Martuza, 1977, Lynn, 1986, Netemeyer et al., 2003, Polit and Beck, 2006, DeVon et al., 2007). Three senior academic staff examiners were then asked to review the potential scale items and validate that they are appropriate indicators of construct (Martuza, 1977, Lynn, 1986, Schultz and Whitney, 2005, Waltz et al., 2010).

According to Lynn (1986), the researcher computed content validity index for items of full veneer gold shell crown preparation (I-CVI) are shown in Table 5.1. Content validity was calculated for each criterion under 6 items. Individual criteria were evaluated by three senior academic staff examiners. A four points scale ranging, 1 = not relevant to 4 = extremely relevant, was used for determining whether the item and their criteria should retained or rejected. Each senior academic staff examiner was asked to rate each scale item in terms of its relevance in order to determine the underlying construct. For each criterion, the I-CVI was computed as the number of senior examiners giving a rating of either 3 or 4, divided by the total number of senior academic staff examiners (n = 3). To recognize the agreement which can be inflated by chance factors, Lynn (1986) recommended that if the number of examiners who asked to rate the items was less than five, all the examiners must agree on the content validity for their rating. All items were rated as 3 (relevant) or 4 (extremely relevant), were retained (Lynn, 1986). For example, I-CVI for three senior examiners should be given a rating of either 3 or 4 to retain the item and their criteria. Content Validity Index for

Scale (S-CVI) is also computed as a number of items giving a rating of either 3 or 4 by the experts, divided by the total number of the items on the instrument - that is also, the proportion in agreement about relevance. According to Table 5.1, the proportion in agreement about relevance for three senior examiners was totally agreed (100%).

The Table 5.1 demonstrates that all items of 'Mhanni feedback sheet' and their criteria were accepted from three senior academic staff examiners and retained.

Table 5.1 Content validity index (CVI) and relevance of senior examiners for items and criteria of full veneer gold shell crown preparation

No.	*Criteria*	*Examiner 1 Relevancy*	*Examiner 2 Relevancy*	*Examiner 3 Relevancy*	*I-CVI*
Item 1: Occlusal surface					
1	Occlusal reduction	4	4	4	Relevant
2	Contour of occlusal preparation	4	3	3	Relevant
Item 2: Axial surface(s)					
1	Buccal reduction	3	4	3	Relevant
2	Lingual reduction	3	4	3	Relevant
3	Mesial reduction	4	4	3	Relevant
4	Distal reduction	4	4	3	Relevant
5	Undercuts	4	3	4	Relevant
6	Bucco-lingual convergence	4	3	3	Relevant
7	Proximal convergence	4	3	3	Relevant
8	Contact area with adjacent teeth	4	4	4	Relevant
9	Contour of preparation is follow tooth surfaces contour	4	3	3	Relevant
Item 3: Functional cusps					
1	Functional cusp bevel reduction	4	3	3	Relevant
2	Location of functional bevel	4	3	3	Relevant
Item 4: Finish line					
1	Type of finish line	3	3	3	Relevant
2	Level of finish line related to gingival margin	4	3	3	Relevant
3	Depth of finish line	4	4	3	Relevant
Item 5: Texture of tooth preparation					
1	Texture of final preparation except margin	3	3	3	Relevant
2	Texture of margin preparation	4	3	3	Relevant
Item 6: Adjacent teeth damage					
1	the mesial tooth	4	4	4	Relevant
2	The distal tooth	4	4	4	Relevant
				S-CVI	100%

(A four points scale, 1 = not relevant, 2 = somewhat relevant, 3 = relevant and 4 = extremely relevant, was used for determining whether items should retained or rejected)

For criterion validity, Appendix 2 shows the grade awarded and the number of negative points awarded for the assessment by one examiner (DJNR). The coded number was used to identify the model. Correlation analyses showed good agreement between the grade awarded and the number of faults found (Figure 5.6).

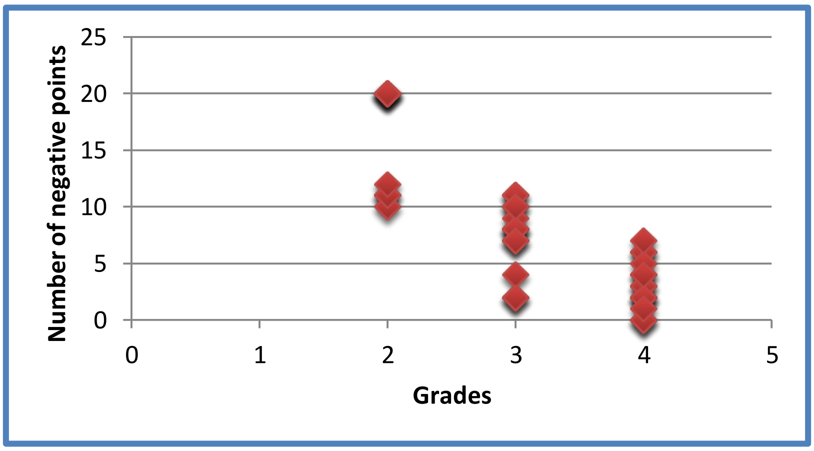


Figure 5.6 Scatter plot of the grades and the number of faults (negative points) awarded from the 'Mhanni feedback sheet'

Spearman correlation (r) was -0.82 (95% Confidence Interval = -0.72 to -0.89). This data indicate good correlation between the 'Mhanni feedback sheet' and a previously used ranking scale. The advantage of the 'Mhanni feedback sheet' for the students was the ability to provide more detailed feedback to the student.

5.4 The ability of specific additional tools and feedback sheets to improve intra-and inter-examiner agreement

5.4.1 Introduction

Additional tools and feedback sheets have been selected and developed, for the evaluation by senior academic staff, for both class II amalgam cavity and full veneer gold shell crown preparations, created by undergraduate students in an Operative Skills

Laboratory. These tools include burs, an amalgam condenser, pre-operative silicone indices and two valid feedback sheets, one for each type of tooth preparation. The next stage is to evaluate the ability of these tools and feedback sheets to improve intra-and inter-examiner agreement and feedback.

Agreement involves several possible parameters. The overall grade awarded for each preparation by an examiner is an obvious parameter for which to establish agreement. However, the feedback sheets each describe multiple different negative aspects (points) of the student preparations. Feedback sheets also identify positive aspects to tooth preparation but these are generally simply described as '**OK or Adequate**'. Generally speaking, it is much easier for an examiner to focus on those aspects of a tooth preparation which should be improved: negative aspects.

When an examiner identifies a positive or negative aspect of a tooth preparation, the **category**, the **criteria** and the **level of performance (descriptors)** are all identified (see Figures 5.5 and 5.6). For example, in relation to a class II amalgam cavity preparation, the occlusal (**category**) depth (**criterion**) of the cavity may have been prepared too shallow (**descriptor**) which is redeemable (**level of performance**).

The feedback sheet requires the examiner to make a judgement for each criterion of each category of a tooth preparation. For example, the examiner is required to make a judgement of the depth (criterion) of the occlusal part (category) of a class II amalgam cavity preparation. The quality of that judgment can only be selected at the level of performance. To continue the example, the depth (criterion) of the occlusal part (category) of a class II amalgam cavity preparation may be OK (positive point), too shallow (negative point) or too deep (negative point).

Both the number of negative points and also the actual description of each point (at criteria level) are important in providing accurate and consistent student feedback. For example, if three examiners evaluate a class II amalgam cavity and each examiner notes three negative points using the feedback sheet, it is obviously important that the negative points noted are consistent for all examiners and not, for example, a different three negative points for each examiner. Furthermore, the quality of the feedback as well as its consistency can also be evaluated. Thus, the following question is posed; is feedback accurate/repeatable at the criterion level (e.g. the depth of the occlusal part of your cavity is too shallow)?

Thus, agreement using the feedback sheets should evaluate:

a) the grades awarded by the examiners,
b) the number of negative points identified by the examiners and,
c) the sameness/consistency of those negative (and positive) points identified at the level of performance for each criterion.

5.4.2 Material and methods

a. Improving intra- and inter-examiner agreement for Class II amalgam cavity preparation by using additional tools and 'Gray feedback sheet'

From the previous results in this chapter, the protocol was developed and tested using senior academic staff. From previous work outlined in this thesis (see Chapter 4), 26 selected class II cavities were evaluated on two separate occasions a minimum of one week apart. As stated previously, these cavities were representative of samples of students work, drawn from tests completed by the second year Bachelor of Dental Surgery (BDS) students, during the academic year of 2013-2014. All possible grades were including in this sample. These were the same cavities used in the staff training day (see Chapter 4).

All grading was performed on the bench top by three senior academic staff (RGC, DNJR, and AFH) who followed a sequence of five grading stages using techniques outlined in the next paragraph. After each grading stage, the researcher (AM) noted the numbers of the cavities allocated to each grade as well as the identifying code number for each tooth. All grading stages were performed at the same sitting. All cavities were graded using the scale in Table 5.2.

Table 5.2 The grading scale for Class II amalgam cavity preparation

Grade	Criteria
1- FAIL	Wrong tooth/cavity cut
2- FAIL	Major safety concerns of such a nature as to render the cavity beyond redemption.
3- FAIL	Cause for concern that although not catastrophic in nature, indicate lack of control/understanding that cumulatively render the cavity unsatisfactory.
4- PASS	No deficient areas but performance lacks fine headpiece control that does not compromise patient safety. No iatrogenic damage to the adjacent tooth, other than to superficial enamel, due to the proximity of the contact point.
5- EXEMPT	No deficient areas and fine headpiece control demonstrated by virtue of degree of superior cavity finish. No iatrogenic damage to the adjacent tooth.

Grading was accomplished in five stages:

Stage 1: Each examiner initially visually inspected (eyeballed) the cavities and allocated them to an interim grading category (Table 5.2) placing the models on the bench thereafter to form groupings of cavities according to the interim grade awarded. The researcher (AM) noted these down and the time taken for this process.

Stage 2: Each examiner looked through the groupings of stage 1 and adjusted the cavity grades to ensure that, within the interim grade grouping of cavities, there were no outliers. Outlying cavities were moved to a more appropriate grouping, based on the grading scale in Table 5.2. The researcher (AM) noted down these movements and the time taken for this process.

Stage 3: Keeping the cavity groupings the same as stage 2. Each cavity was reassessed by inserting a small round amalgam condenser to ensure that the width of the occlusal part of the cavity was sufficiently wide for amalgam placement and was as not grossly over-prepared (Figure 5.1b). If the condenser did not fit, or demonstrated the occlusal part of the cavity was over-prepared, then the examiner had an opportunity to reassign the grade awarded for the cavity based on the grading scale in Table 5.2. The researcher (AM) noted down those cavities where the grade was reassigned and the time taken for this process.

Stage 4: Keeping the cavity groupings the same as stage 3, a single use fissure diamond bur (BDM541, UnoDent LTd) with working length is 4.00 mm, and 1.00 mm for diameter (Figure 5.1a), was inserted into the cavities to confirm or otherwise:

- That the cavity occlusal floor was at the correct depth - which was defined as 1.50 to 2.00 mm – i.e. less or half the length of the working end of the diamond fissure bur.
- That the gingival floor of the box was at the correct gingival level – which was defined as one full length of the working end of the diamond bur from the original occlusal surface.
- That the width of the box, when the bur was placed at the contact point, was correct and not overcut – which was defined as a width of three diamond burs with the central one coincident with the central portion of the contact point.

If the bur did not fit, or demonstrated the cavity was over prepared, then the examiner had an opportunity to reassign the grade awarded for the cavity based on the scale in Table 5.2. The researcher (AM) noted down those cavities where the grade was reassigned and the time taken for this process.

Stage 5: A ‘Gray feedback sheet’ was completed. On this occasion, both the bur and the condenser could be used to help with the evaluations required by the ‘Gray feedback sheet’ (Figure 5.3). The ‘Number of negative points’ was evaluated by each examiner who, once again, had the opportunity to change the grade awarded for the cavity. The researcher noted down these changes and the ‘Final grade’ could now be awarded for each cavity by each examiner. The researcher (AM) also noted the time taken for this process.

b. Improving intra- and inter-examiner agreement for full veneer gold shell crown preparation by using additional tools and ‘Mhanni feedback sheet’

The protocol was also developed and tested using three senior academic staff for the full veneer gold shell crown. From previous work outlined in Chapter 4, thirty selected full veneer gold shell crown preparations were evaluated on two separate occasions a minimum of one week apart. As stated previously, full veneer gold shell crown preparations were representative of samples of students work, drawn from a test completed by the third BDS students, during the academic year of 2014-2016. All possible grades were including in this sample. These were the same full veneer gold shell crown preparations used in the staff training day (see Chapter 4).

All grading was performed on the bench top by three senior academic staff (RGC, DNJR, and AFH) who followed a sequence of five grading stages using techniques outlined in the next paragraph. After each grading stage, the researcher (AM) noted the numbers of the preparations allocated to each grade as well as the identifying code number for each tooth. All grading stages were performed at the same sitting. All full veneer gold shell crown preparations were graded using the scale in Table 4.2 (see Chapter 4).

Grading was accomplished in five stages:

Stage 1: Each examiner initially visually inspected (eyeballed) the full veneer gold shell crown preparations and allocated them to an interim grading category (Table 4.2) placing the models on the bench thereafter to form groupings of full veneer gold shell crown preparations according to the interim grade awarded. For example, Examiner establishes whether or not the contact areas with adjacent teeth were cleared. The researcher (AM) noted number of grade down for each tooth preparation and the time taken for this process.

Stage 2: Each examiner looked through the groupings of stage 1 and adjusted the full veneer gold shell crown preparation grades to ensure that, within the interim grade collection of preparations, there were no outliers. Outlying full veneer gold shell crown preparations were moved to a more appropriate grouping, based on criteria in Table 4.2. The researcher (AM) noted down these movements and the time taken for this process.

Stage 3: Keeping the full veneer gold shell crown preparation groupings the same as stage 2, each prepared tooth was reassessed using a tapered high-speed diamond with a rounded tip (No.8856.314.016 TPS2, Komet Dental, UK) used to create a tapered preparation with a chamfered finish line (length of working part of the bur is 8.00 mm, and diameter is 1.00 mm at the tip and 1.60 mm at the top of working part) (Figures 3.2 and 5.2a). This bur was used to confirm:

- That the finish line of full veneer gold shell crown preparation was at the correct depth - which was defined as 0.50 mm – i.e. half thickness of chamfer diamond bur at the tip.
- That presence or absence of undercuts on the axial walls by holding the bur parallel to the axial wall of the prepared teeth.

If a half-thickness of the chamfer diamond bur tip did not fit on the finish line, this demonstrated that the finish line was under- or over- prepared. The same bur was used to establish any undercuts on the axial walls. The examiner had a chance to reassign the grade awarded for the full veneer gold shell crown preparation based on the scale in Table 4.2. The researcher (AM) noted down those full veneer gold shell crown preparations where the grade was reassigned and the time taken for this process.

Stage 4: Keeping the full veneer gold shell crown preparation groupings the same as stage 3, sectional silicone putty indices were used which involving at least one tooth either side of the full veneer gold shell crown preparation (Figure 5.2b). Each full veneer gold shell crown preparation was reassessed by using two pre-operative sectional indices. The two indices were sectioned at 90 degrees to confirm that the occlusal and axillary contours of the preparation, as well as the preparation convergences were correct.

If one feature did not fit, then the examiner had an opportunity to reassign the grade awarded for the full veneer gold shell crown preparation based on the grading scale in Table 4.2. The researcher (AM) noted down those full veneer gold shell crown preparation where the grade was reassigned and the time taken for this process.

Stage 5: A ‘Mhanni feedback sheet’ was completed. On this occasion, both the bur and the pre-operative sectional indices could be used to help with the evaluations required by the ‘Mhanni feedback sheet’ (Figure 5.4). The ‘Number of negative points’ was evaluated by each examiner who, once again, had the opportunity to change the grade awarded for the preparation. The researcher noted down these changes and the ‘Final grade’ could now be awarded for each full veneer gold shell crown preparation by each examiner. The researcher (AM) also noted the time taken for this process.

c. Assessing the repeatability and reproducibility of detailed feedback from the perspective of, “consistency of message”, to a learner

Class II amalgam cavity preparation ‘Gray feedback sheet’

The ‘Gray feedback sheet’ has several features. A set of criteria constitute descriptors of various categories within the overall assessment of a class II cavity. Each criterion is assessed and four levels of performance can be applied. These levels were designated, satisfactory, redeemable, unmodifiable and irredeemable fail. Not all levels of performance could be applied to all criteria. For example, irredeemable fail could only be applied when there was moderate/severe damage to the adjacent tooth.

Thus, using the ‘Gray feedback sheet’, the evaluation criteria were divided into three main categories; Box (proximal part), Occlusal (occlusal preparation) and Damage (to the adjacent tooth). For each criterion, four levels of performance were specifically described: i), ‘satisfactory’, ii), ‘redeemable’, iii), ‘unmodifiable’ and iv) ‘irredeemable fail’. These changes form a ‘Modified Gray feedback sheet’ (Figure 5.7).

Management of Dental Caries Course Feedback Sheet					
Category	Criteria	**Level of performance (Descriptors)**			
		Satisfactory	*Redeemable*	*Unmodifiable*	*Irredeemable fail*
Box	Outline	OK		Rough/ irregular	
	Position	OK		Too far B/P	
	Depth gingivally	OK	Too shallow not clear G	Too deep	
	B-P width	OK	Narrow not clear B/P	Too wide	
	M-D depth	OK	Too shallow pulpal- axial wall	Too deep pulpoaxial wall	
	Unsupported enamel	NONE	Yes – B/P/G		
	Retention form	Ok	Parallel walls	Divergent walls	
Occlusal	Lock	OK	Rough/irregular	Not follow fissure	
	Depth	OK	Too shallow	Too deep	
	Width	OK	Too narrow	Too wide	
Damage adjacent tooth		NONE	Minor		Moderate/ Severe

Figure 5.7 Schematic representation of the 'Modified Gray feedback sheet'

Each level of performance was coded 1 = satisfactory, 2 = redeemable, 3 = unmodifiable and 4 = irredeemable fail. Based on class II cavity evaluation on two occasions by three examiners each using the 'Gray feedback sheet', agreement for each criterion according to the level of performance was determined for each examiner (repeatability) using SPSS (version 22). Inter-examiner agreement (reproducibility) at the level of performance for each criterion of the ‘Gray feedback sheet’ was also determined by using SPSS (version 22).

Full veneer gold shell crown preparation ‘Mhanni feedback sheet’

The ‘Mhanni feedback sheet’ has several features. A set of criteria constitute descriptors of various categories within the overall assessment of a full veneer gold shell crown preparation. Each criterion is assessed and four levels of performance can be applied. These levels of performance were designated, satisfactory, redeemable, unmodifiable and irredeemable fail. Not all levels could be applied to all criteria. For example, irredeemable fail could only be applied when there was moderate/severe damage to the adjacent tooth.

Thus, using the feedback sheet, the evaluation criteria were divided into six main categories; Occlusal surface, Axial surfaces, Functional cusps, Finish line, Final preparation, and Damage to the adjacent teeth. For each criterion, four levels of performance were specifically described: i), ‘satisfactory’, ii), ‘redeemable’, iii), ‘unmodifiable’ and iv) ‘irredeemable fail’ (Figure 5.8).

Each level of performance was coded 1 = satisfactory, 2 = redeemable, 3 = unmodifiable and 4 = irredeemable fail. Based on evaluation of full veneer gold shell crown preparation on two occasions by three examiners each using the ‘Mhanni feedback sheet’, agreement for each criterion according to the level of performance was determined for each examiner (repeatability) using SPSS (version 22). Inter-examiner agreement (reproducibility) at the level of performance for each criterion of the ‘Mhanni feedback sheet’ was also determined by using SPSS (version 22).

Feedback Sheet for full veneer gold shell crown preparation					
Category	Criteria	**Level of performance (Descriptors)**			
		Satisfactory	*Redeemable*	*Unmodifiable*	*Irredeemable fail*
Occlusal surface	Occlusal reduction	Adequate	Under-prepared	Over-prepared	
	Contour of occlusal preparation	Yes (follow the contour tooth surfaces)		No	
Axial surface(s)	Buccal reduction	Adequate	Under-prepared	Over-prepared	
	Lingual reduction	Adequate	Under-prepared	Over-prepared	
	Mesial reduction	Adequate	Under-prepared	Over-prepared	
	Distal reduction	Adequate	Under-prepared	Over-prepared	
	Undercuts	No	Yes	Yes – more than one surface	
	Bucco-lingual convergence	Adequate (Convergence)	Improper convergence		No (destructive shape)
	Proximal convergence	Adequate (Convergence)	Improper convergence		No (destructive shape)
	Contour of preparation	Follow tooth surfaces contour	One not follow	More than one not follow	
	Contact area with adjacent teeth	Cleared	Yes – one side	Not clear	
Functional cusps	Functional cusp bevel reduction	Adequate	Not - symmetrical	No	
	Location of functional bevel	Adequate		No	

Category	Criteria	Level of performance (Descriptors)			
		Satisfactory	*Redeemable*	*Unmodifiable*	*Irredeemable fail*
Finish line	Chamfer finish line	Yes		No	
	Level of finish line to gingival margin	At gingival (at or above by 0.5mm)	Supra-gingival (more than 0.5mm)	Subgingival and/or Supra-ginigival	
	Depth of finish line all around	Even	Uneven	Deep	
Final preparation	Texture of final preparation except margin	Adequate	Rough (irregular) Sharp edges		
	Texture of margin	Smooth and well define	Rough (irregular)		
Adjacent teeth damage	Mesial tooth	No damage	Minor damage		Moderate/ severe damage
	Distal tooth	No damage	Minor damage		Moderate/ severe damage

Figure 5.8 Schematic representation of the 'Modified Mhanni Feedback sheet'

5.4.3 Statistical analysis

a. Assessing the repeatability and reproducibility of grades and the number of negative points by using additional tools and feedback sheets

For each examiner, un-weighted Cohen's Kappa statistic was computed, using SPSS (Version 22, IBM Corporation, USA), to determine the intra-examiner agreement between each of the two occasions for each stage of grading and, at stage 5, the number of negative points awarded. The intra-class correlation was used to determine inter-examiner agreement, for each of the five stages, between senior academic staff and on each of the two occasions when they performed the evaluation (see section 5.4.4).

b. Assessing the repeatability and reproducibility of detailed feedback from the perspective of, "consistency of message", to a learner.

Each level of performance was coded and computed by using SPSS (Version 22, IBM Corporation, USA). Based on evaluation of class II amalgam cavity preparation and full veneer gold shell crown preparation on two occasions by three senior examiners each using the feedback sheet, intra-examiner agreement (repeatability) for each criterion according to the level of performance was determined using un-weighted Cohen's Kappa test. The intra-class correlation measurement (single measure) was used to determine inter-examiner agreement (reproducibility) at the level of performance for each criterion of the 'Gray and Mhanni feedback sheets' among senior examiners (see section 5.4.4).

5.4.4 Results

a. Assessing the repeatability and reproducibility of grades and the number of negative points by using additional tools and feedback sheets

<u>Intra-examiner agreement for class II amalgam cavity preparation according to grades for each stage</u>

Tables 5.3, 5.4 and 5.5 show, for examiners 1, 2, and 3, respectively, the grades awarded for each cavity at each stage of the grading process on each occasion of grading. These are colour coded for each stage of grading to indicate agreement and disagreement (Figure 5.9). The agreement and disagreement percentages were also illustrated for each stage of grading process in these tables. The tables also show the number of negative points awarded from the 'Gray feedback sheet' on each occasion and by the same colour coding convention indicates agreement and disagreement. The

number of times from the first and second grading occasions, where the same final conclusion was reached, is also summarised for each examiner in these tables.

Colour code	Description
(dark green)	Agreement – Pass
(light green)	Agreement – Fail
(red)	Disagreement - in borderline
(pink)	Disagreement - Pass or Fail

Figure 5.9 Description of the four colour codes to describe agreement between examiners

Table 5.3 Examiner 1 grades awarded for each class II amalgam cavity at each stage of the grading process on each occasion of grading (with agreement and disagreement percentages), and the number of negative points awarded from the 'Gray feedback sheet'

Model Number	Eyeball		Confirm Eyeball		Condenser		Bur		**Gray feedback sheet**		Negative Points	
	First	Second	First	Second	First	Second	First	Second	**First**	**Second**	First	Second
5	2	4	2	2	2	2	2	2	**2**	**2**	2	2
8	3	3	3	3	3	3	3	3	**3**	**3**	3	2
15	2	3	4	4	4	4	4	4	**4**	**4**	1	2
16	3	3	4	3	4	3	4	3	**4**	**4**	1	1
36	5	4	5	5	5	5	5	5	**5**	**5**	0	0
39	4	3	4	4	4	4	4	4	**5**	**5**	0	0
40	4	3	3	3	3	3	4	3	**4**	**5**	1	0
41	2	4	3	4	3	4	4	4	**5**	**5**	0	2
43	4	4	3	4	3	4	3	3	**5**	**3**	0	2
46	5	4	5	5	5	5	5	5	**5**	**5**	0	0
53	3	4	3	4	3	4	3	4	**3**	**2**	2	1
54	3	4	3	5	3	5	3	5	**3**	**5**	1	0
57	2	2	2	2	2	2	2	2	**3**	**2**	4	11
62	3	3	3	3	3	3	3	3	**3**	**3**	2	5
73	2	2	2	2	2	2	2	2	**2**	**2**	11	11
78	3	3	3	3	3	3	3	3	**3**	**3**	2	2
80	3	3	3	4	3	4	3	4	**3**	**4**	2	3
83	3	4	4	4	4	4	3	4	**5**	**5**	0	0
85	3	3	3	3	3	3	3	2	**3**	**3**	2	3
87	4	5	4	5	4	5	4	5	**5**	**5**	0	0
88	2	2	2	3	2	3	2	2	**2**	**2**	5	5
94	2	2	2	2	2	2	2	2	**2**	**2**	11	11
109	3	3	3	4	3	4	3	4	**3**	**5**	1	0
111	2	3	3	2	3	2	3	2	**2**	**2**	3	4
120	5	4	5	4	5	4	5	4	**5**	**4**	1	1
138	2	2	2	2	2	2	2	2	**2**	**2**	11	11
Disagreement	13 (50%)		11 (42%)		11 (42%)		11 (42%)		**8 (31%)**			
Agreement	13 (50%)		15 (58%)		15 (58%)		15 (58%)		**18 (69%)**			

Table 5.4 Examiner 2 grades awarded for each class II amalgam cavity at each stage of the grading process on each occasion of grading (with agreement and disagreement percentages), and the number of negative points awarded from the 'Gray feedback sheet'

Model Number	Eyeball		Confirm Eyeball		Condenser		Bur		**Gray feedback sheet**		Negative Points	
	First	Second	First	Second	First	Second	First	Second	**First**	Second	First	Second
5	4	3	4	3	4	3	4	3	**4**	**3**	2	3
8	3	4	3	3	3	3	3	3	**3**	**3**	3	3
15	4	4	4	3	4	3	4	3	**4**	**3**	2	2
16	5	4	5	4	5	4	5	4	**5**	**4**	2	3
36	4	5	5	5	5	5	5	5	**5**	**5**	0	1
39	4	4	4	4	3	4	4	4	**4**	**4**	1	2
40	3	3	3	3	3	3	3	4	**4**	**4**	2	2
41	3	3	3	3	3	3	3	3	**3**	**3**	3	5
43	3	3	3	3	3	3	3	3	**3**	**3**	3	2
46	4	5	4	5	4	5	4	5	**4**	**5**	2	0
53	3	3	3	3	3	3	3	3	**3**	**3**	5	4
54	4	3	3	3	3	3	3	3	**3**	**3**	4	3
57	3	3	3	3	3	4	4	4	**4**	**4**	5	3
62	3	3	3	3	3	3	3	3	**3**	**3**	3	3
73	2	2	2	2	2	2	2	2	**2**	**2**	11	11
78	3	4	3	3	3	3	3	3	**3**	**3**	4	3
80	3	4	4	4	3	4	3	4	**4**	**4**	2	4
83	3	3	3	4	3	4	3	4	**3**	**4**	6	2
85	3	3	3	3	3	3	3	3	**3**	**3**	4	5
87	5	5	4	5	4	5	4	5	**4**	**5**	1	1
88	3	3	3	3	3	3	3	3	**3**	**3**	5	6
94	3	3	3	3	3	3	3	3	**3**	**3**	3	5
109	4	3	4	3	3	3	3	3	**3**	**3**	2	2
111	3	3	3	3	3	3	3	3	**3**	**3**	4	4
120	5	4	5	4	5	4	5	4	**5**	**5**	0	2
138	2	2	2	2	2	2	2	2	**2**	**2**	11	11
Disagreement	10 (39%)		8 (31%)		11 (42%)		9 (35%)		**6 (23%)**			
Agreement	16 (62%)		18 (69%)		15 (58%)		17 (65%)		**20 (77%)**			

Table 5.5 Examiner 3 grades awarded for each class II amalgam cavity at each stage of the grading process on each occasion of grading (with agreement and disagreement percentages), and the number of negative points awarded from the 'Gray feedback sheet'

Model Number	Eyeball		Confirm Eyeball		Condenser		Bur		**Gray feedback sheet**		Negative Points	
	First	Second	First	Second	First	Second	First	Second	**First**	**Second**	First	Second
5	4	4	3	3	3	3	3	2	**3**	**2**	2	3
8	3	4	3	3	3	3	2	3	**2**	**3**	3	3
15	3	4	3	4	3	4	4	4	**4**	**4**	2	1
16	2	3	3	3	2	3	2	3	**2**	**3**	3	3
36	4	5	5	5	5	5	5	5	**5**	**5**	0	0
39	5	4	5	5	5	5	5	5	**5**	**5**	0	0
40	4	5	5	5	5	5	4	4	**4**	**4**	2	2
41	3	3	3	3	3	3	3	3	**3**	**3**	3	4
43	4	3	3	3	3	3	3	3	**3**	**3**	2	2
46	3	5	3	5	3	5	3	5	**5**	**5**	0	0
53	3	3	3	3	3	3	3	3	**2**	**3**	5	4
54	4	4	4	4	4	4	3	4	**4**	**4**	1	2
57	2	3	3	3	2	3	2	3	**2**	**2**	3	3
62	2	3	3	3	3	3	3	3	**3**	**3**	2	3
73	2	2	3	2	2	2	2	2	**2**	**2**	11	11
78	3	4	3	4	3	4	3	4	**3**	**4**	2	2
80	4	4	4	4	4	4	4	4	**4**	**4**	1	2
83	4	5	4	5	4	5	4	5	**5**	**5**	0	0
85	2	3	3	3	2	3	2	3	**2**	**2**	5	5
87	3	3	3	4	3	4	4	4	**4**	**4**	1	1
88	2	2	3	2	2	2	2	2	**2**	**2**	5	6
94	3	3	3	2	2	2	2	2	**2**	**2**	11	11
109	4	5	4	5	4	5	4	5	**5**	**5**	0	0
111	2	2	3	2	2	2	2	2	**2**	**2**	5	5
120	4	4	4	4	4	4	4	4	**4**	**4**	1	2
138	2	2	3	2	2	2	2	2	**2**	**2**	11	11
Disagreement	14 (54%)		11 (42%)		8 (31%)		10 (38%)		**5 (19%)**			
Agreement	12 (46%)		15 (58%)		18 (69%)		16 (62%)		**21 (81%)**			

Each type of assessment stage had its own time to complete. Table 5.6 shows time taken to assess class II amalgam cavity preparation for each stage and for each examiner.

Table 5.6 Time spent (in seconds) to assess class II amalgam cavity preparation

Stage	**Occasion number**	**Examiner 1**	**Examiner 2**	**Examiner 3**	**Average time spent per stage (sec)**	**Average time spent per model (sec)**
		Time spent for evaluation (seconds)				
Eyeball	1	360	133	300	264	10
Eyeball	2	300	180	300	260	10
						***Average* 10**
Confirm eyeball	1	300	180	240	240	9
Confirm eyeball	2	360	90	240	230	9
						***Average* 9**
Condenser	1	180	83	180	148	6
Condenser	2	180	180	180	180	7
						***Average* 7**
Bur	1	360	240	360	320	12
Bur	2	240	180	240	220	9
						***Average* 11**
Gray feed-back sheet	1	1380	1500	1440	1440	55
Gray feed-back sheet	2	1800	2400	1800	2000	77
						***Average* 66**

Tables 5.7, 5.8 and 5.9 show for each examiner and stage of grading the Un-weighted Kappa statistic as calculated in SPSS to assess repeatability.

Landis and Koch (1977) have proposed the following values as standards for the strength of agreement as assessed by the Un-weighted Kappa coefficient:

- ≤ 0 = poor,
- 0.01–0.20 = slight,
- 0.21–0.40 = fair,
- 0.41–0.60 = moderate,
- 0.61–0.80 = substantial and,
- 0.81–1.00 = almost perfect.

Table 5.7 Measurement of intra-examiner (Kappa) agreement for the class II amalgam cavity preparations for examiner 1 according to the grades awarded

Type of assessment stage	Kappa value	Significance ($p \leq 0.05$)
Eyeball	0.299	0.009
Confirm eyeball	0.427	0.000
Condenser	0.427	0.000
Bur	0.430	0.000
Gray feedback sheet	0.583	0.000

Table 5.8 Measurement of intra-examiner (Kappa) agreement for the class II amalgam cavity preparations for examiner 2 according to grades awarded

Type of assessment stage	Kappa value	Significance ($p \leq 0.05$)
Eyeball	0.378	0.004
Confirm eyeball	0.485	0.000
Condenser	0.323	0.011
Bur	0.431	0.001
Gray feedback	0.637	0.000

Table 5.9 Measurement of intra-examiner (Kappa) agreement for the class II amalgam cavity preparations for examiner 3 according to grades awarded

Type of assessment stage	Kappa value	Significance ($p \leq 0.05$)
Eyeball	0.265	0.019
Confirm eyeball	0.386	0.000
Condenser	0.530	0.000
Bur	0.479	0.000
Gray feedback sheet	0.739	0.000

It is apparent that the final stage of assessment ‘Gray feedback sheet’ achieves the highest agreement for each of three examiners while the use of the bur improves intra-examiner agreement for two of the three examiners. In addition, it shows that examiners 3 and 2 ultimately demonstrate substantial agreement after all stages have been completed whereas examiner 2 displays moderate agreement. All examiners show improvement in repeatability through the stages of grading.

Inter-examiner agreement for class II amalgam cavity preparation according to grades for each stage

Table 5.10 Inter-examiner agreement for the class II amalgam cavity preparation for each stage and occasion according to grades

Inter-examiner agreement for Class II amalgam cavity preparation					
Stage	*Occasion*	*Number of examiners*	*Single measurement ICC*	*95 % of CI*	*Best single measurement if examiner deleted*
Eyeball	1	3	0.458	0.225 – 0.676	0.529 if Examiner 2 is excluded
Eyeball	2	3	0.470	0.237 – 0.686	0.509 if Examiner 3 is excluded
Confirm eyeball	1	3	0.511	0.281 – 0.716	0.705 if Examiner 3 is excluded
Confirm eyeball	2	3	0.672	0.476 – 0.822	0.717 if Examiner 2 is excluded
Condenser	1	3	0.714	0.536 – 0.847	0.718 if Examiner 3 is excluded
Condenser	2	3	0.706	0.520 – 0.843	0.717 if Examiner 2 is excluded
Bur	1	3	0.560	0.339 – 0.749	0.653 if Examiner 3 is excluded
Bur	2	3	0.700	0.514 – 0.839	0.817 if Examiner 2 is excluded
Gray feedback sheet	1	3	0.540	0.313 – 0.736	0.657 if Examiner 2 is excluded
Gray feedback sheet	2	3	0.692	0.503 – 0.834	0.855 if Examiner 2 is excluded

The highlighted values represent the highest inter-examiner agreement.

Table 5.10 shows the agreement between senior academic staff in using different methods of grading for evaluation of 26 class II amalgam cavity preparations, indicating the occasion two assessment for each stage was better than the occasion one. In addition, condenser and bur stages produced the best inter-examiner agreement in comparing with the other stages. In fact, there was marginally – very small difference between all stages. Table 5.10 also shows that bur and ‘Gray feedback sheet’ stages produced excellent agreement (ICC > 0.80) between some senior examiners, if one of senior academic staff (examiner 2) was excluded. From the previous data, examiner 2 failed to demonstrate sufficient inter-examine agreement.

Intra-examiner agreement for class II amalgam cavity preparation according to negative points 'Gray feedback sheet' for each stage

Intra-examiner agreement (Kappa) for the number of negative points awarded by each senior academic staff examiner on each of two occasions is shown in Table 5.11. The highest value was for examiner 3 while the lowest value was for examiner 2.

Table 5.11 Intra-examiner agreement for the class II amalgam cavity preparations according to the number of negative points for each examiner using 'Gray feedback sheet'

Examiners	Kappa Value	Significance ($p \leq 0.05$)
Examiner 1	0.382	0.000
Examiner 2	0.211	0.013
Examiner 3	0.589	0.000

Inter-examiner agreement for class II amalgam cavity preparation according to negative points 'Gray feedback sheet' for each stage and occasion

The inter-examiner agreement was evaluated using intra-class correlation (ICC). As displayed in Table 5.12, there was good agreement among three senior academic staff, (0.785) and (0.802), for occasion one and two respectively. By process of elimination, this table also shows that the best agreement according to the number of negative points was for examiner 3 who was the only examiner not excluded when the best single measurement was determined.

Table 5.12 Inter-examiner agreement among senior academic staff who assessed class II amalgam cavity preparations according to negative points for each occasion using the 'Gray feedback sheet'

Occasion	**Intra-class correlation** single measure	**95% confidence interval**		**Significance ($p\leq0.05$)**	**Best single measurement if examiner deleted**
		Lower bound	Upper bound		
1	0.785	0.635	0.888	0.000	0.924 if Examiner 2 is excluded
2	0.802	0.662	0.897	0.000	0.872 if Examiner 1 is excluded

<u>Intra-examiner agreement for full veneer gold shell crown preparation according to grades for each stage</u>

Tables 5.13, 5.14 and 5.15 show, for examiners 1, 2, and 3, respectively, the grades awarded for each preparation at each stage of the grading process on each occasion of grading. These tables were coded by different colours which were described in this section 5.4.4 (Figure 5.9). These tables also shows the number of negative points awarded from the 'Mhanni feedback sheet' on each occasion and by the same colour coding convention indicates agreement and disagreement. The total of first and second grading occasions where the same final conclusion was reached is also summarised for each examiner in these tables.

Table 5.13 Examiner 1 grades awarded for each full veneer gold shell crown preparation at each stage of the grading process on each occasion of grading (with agreement and disagreement percentages), and the number of negative points awarded from the 'Mhanni feedback sheet'

Model Number	Eyeball		Confirm Eyeball		Bur		Impression index		**Mhanni feedback sheet**		Negative Points	
	First	Second	First	Second	First	Second	First	Second	**First**	**Second**	First	Second
1	4	4	4	3	4	4	3	4	3	4	5	2
3	4	4	4	4	5	3	5	4	4	4	5	3
4	3	4	3	3	3	3	3	3	3	3	9	10
5	3	4	3	4	3	4	3	4	3	5	5	0
7	4	4	4	4	4	4	4	5	4	5	2	1
13	3	3	3	3	3	3	4	3	3	3	6	7
14	5	5	5	5	5	5	5	5	5	5	0	0
18	3	3	3	3	3	3	3	3	3	3	6	8
20	5	5	5	4	5	5	5	5	5	5	0	0
21	5	5	5	5	5	5	4	5	5	5	0	0
25	4	4	4	4	4	4	4	4	3	4	6	4
26	2	2	2	2	2	2	2	2	2	2	20	20
29	3	3	4	3	3	4	3	3	3	3	9	8
31	4	5	4	4	5	5	4	4	5	4	1	2
51	4	4	4	4	5	5	4	4	5	5	0	1
52	2	3	4	4	4	2	2	2	4	3	1	2
54	4	3	4	4	3	3	4	3	3	3	8	8
57	5	5	5	5	5	4	4	4	5	5	0	2
58	3	3	3	3	3	3	3	3	3	3	7	8
59	2	2	2	2	2	2	2	2	2	2	20	20
60	4	3	4	3	4	3	3	4	3	3	9	6
63	4	3	4	3	3	3	4	3	3	3	8	12
67	4	4	4	4	4	5	4	5	4	5	3	3
69	5	4	5	4	5	4	5	5	4	4	3	5
70	3	4	4	4	4	4	4	4	3	3	7	5
71	5	5	5	5	5	4	3	5	5	5	1	2
73	5	4	5	3	4	3	4	3	4	3	1	6
74	4	4	3	4	4	3	4	3	3	4	5	8
78	4	5	5	5	5	5	4	5	5	5	0	0
88	3	3	3	3	3	3	3	3	2	2	10	11
Disagreement	11 (37%)		9 (30%)		11 (37%)		14 (47%)		**9 (30%)**			
Agreement	19 (63%)		21 (70%)		19 (63%)		16 (53%)		**21 (70%)**			

Table 5.14 Examiner 2 grades awarded for each full veneer gold shell crown preparation at each stage of the grading process on each occasion of grading (with agreement and disagreement percentages), and the number of negative points awarded from the 'Mhanni feedback sheet'

Model Number	Eyeball		Confirm Eyeball		Bur		Impression index		**Mhanni feedback sheet**		Negative Points	
	First	Second	First	Second	First	Second	First	Second	**First**	**Second**	First	Second
1	4	3	4	3	4	3	3	3	3	3	8	5
3	4	5	4	4	4	4	4	4	3	3	6	7
4	4	4	4	3	4	3	4	3	3	3	9	20
5	3	4	3	4	3	4	3	4	3	4	9	7
7	5	4	5	4	5	4	4	4	4	4	3	2
13	3	3	3	3	3	3	4	3	4	3	6	8
14	4	4	4	4	4	4	4	4	4	4	5	3
18	3	3	3	3	3	3	3	3	3	3	11	10
20	5	4	5	4	5	4	4	4	3	4	8	2
21	4	4	4	4	4	4	4	4	4	4	6	4
25	5	5	5	5	5	5	4	4	4	5	5	1
26	3	3	3	3	3	3	3	3	3	2	8	20
29	4	4	4	4	4	4	4	4	4	4	5	4
31	4	5	4	4	4	4	4	4	4	4	4	2
51	5	4	5	4	5	4	4	4	4	4	3	6
52	4	4	4	4	4	4	4	4	3	4	11	4
54	3	3	3	3	3	3	4	3	3	3	10	7
57	5	5	5	5	5	5	4	4	4	5	5	0
58	4	4	3	4	3	4	3	3	3	3	9	9
59	3	3	3	2	3	2	3	2	2	2	20	20
60	3	3	3	3	3	3	3	3	3	3	9	8
63	3	4	4	4	4	4	4	4	3	4	8	6
67	4	3	4	3	4	3	4	4	4	4	4	3
69	4	5	4	4	4	4	4	4	4	4	3	2
70	4	4	4	4	3	4	3	4	3	4	7	4
71	5	5	4	5	4	5	4	4	4	5	2	1
73	5	5	4	4	4	4	4	4	4	4	2	3
74	4	4	4	4	4	4	4	4	3	4	9	5
78	4	4	4	4	4	4	4	4	4	4	3	2
88	3	3	3	3	3	3	3	3	2	2	13	10
Disagreement	10 (30%)		10 (30%)		11 (37%)		6 (20%)		**11 (37%)**			
Agreement	20 (70%)		20 (70%)		19 (63%)		24 (80%)		**19 (63%)**			

Table 5.15 Examiner 3 grades awarded for each full veneer gold shell crown preparation at each stage of the grading process on each occasion of grading (with agreement and disagreement percentages), and the number of negative points awarded from the 'Mhanni feedback sheet'

Model number	Eyeball		Confirm Eyeball		Bur		Impression index		**Mhanni feedback sheet**		Negative Points	
	First	Second	First	Second	First	Second	First	Second	**First**	**Second**	First	Second
1	5	3	4	3	3	3	3	3	3	3	4	9
3	4	3	2	3	2	3	3	3	3	4	5	4
4	2	2	2	2	2	2	2	2	2	2	20	20
5	4	3	2	3	2	3	3	3	3	3	7	15
7	4	5	4	5	4	5	3	4	3	4	2	4
13	2	3	2	3	2	3	2	3	2	4	10	6
14	4	3	4	3	4	5	3	4	3	3	4	4
18	3	3	3	3	3	3	3	3	3	3	9	10
20	4	5	5	5	4	5	4	4	4	3	1	3
21	3	3	3	3	3	3	3	3	3	3	2	4
25	4	5	4	5	3	4	3	3	3	3	6	8
26	2	3	2	4	2	3	2	3	2	2	20	20
29	3	3	3	3	3	3	3	3	3	3	8	15
31	2	3	2	3	2	3	2	3	2	3	20	9
51	4	4	4	4	3	4	3	3	3	3	3	9
52	3	4	3	4	3	4	3	4	3	4	3	5
54	4	3	4	3	2	3	3	3	3	3	6	6
57	4	3	4	4	4	4	3	3	3	3	5	7
58	4	4	3	4	3	2	3	2	3	2	7	15
59	2	3	2	2	3	2	2	2	2	2	20	20
60	4	3	3	4	3	4	3	4	3	3	8	7
63	4	4	3	4	3	4	3	4	3	4	7	8
67	3	3	3	3	3	3	3	3	3	3	2	9
69	4	3	4	3	4	3	3	3	3	3	3	6
70	4	5	4	5	3	3	3	3	3	3	4	7
71	5	4	5	4	5	4	3	4	3	3	3	3
73	4	3	4	4	4	4	3	4	3	3	8	11
74	4	4	4	4	4	4	3	4	3	4	4	1
78	3	3	3	3	3	3	3	3	3	3	5	12
88	3	3	2	3	2	3	2	3	2	3	11	10
Disagreement	19 (63%)		18 (60%)		19 (63%)		13 (43%)		**10 (30%)**			
Agreement	11 (37%)		12 (40%)		11 (37%)		17 (57%)		**20 (70%)**			

Each type of assessment stage had its own time to complete. Table 5.16 shows the time taken for each stage for each examiner.

Table 5.16 Time spent (in seconds) to assess full veneer gold shell crown preparation

Stage	**Occasion number**	**Examiner 1**	**Examiner 2**	**Examiner 3**	**Average time spent per stage (sec)**	**Average time spent per model (sec)**
		Time spent for evaluation (sec)				
Eyeball	1	600	480	600	560	19
Eyeball	2	600	420	420	480	16
						***Average* 18**
Confirm eyeball	1	480	300	360	380	13
Confirm eyeball	2	600	600	360	520	17
						***Average* 15**
Bur	1	600	480	360	480	16
Bur	2	720	420	360	500	17
						***Average* 17**
Impression index	1	1200	1020	420	880	29
Impression index	2	870	1800	420	1030	34
						***Average* 32**
Mhanni feed-back sheet	1	2700	2520	3120	2780	93
Mhanni feed-back sheet	2	2520	3000	3600	3040	101
						***Average* 97**

Tables 5.17, 5.18, and 5.19 show intra-examiner agreement for each examiner and stage of grading, using Un-weighted Kappa statistic test (SPSS).

Table 5.17 Measurement of intra-examiner (Kappa) agreement for the full veneer gold shell crown preparations for examiner 1 according to grades awarded

Type of assessment stage	Kappa value	Significance (p≤0.05)
Eyeball	0.475	0.000
Confirm eyeball	0.562	0.000
Bur	0.488	0.000
Impression index	0.347	0.001
Mhanni feedback sheet	0.573	0.000

Table 5.18 Measurement of intra-examiner (Kappa) agreement for the full veneer gold shell crown preparations for examiner 2 according to grades awarded

Type of assessment stage	Kappa value	Significance (p≤0.05)
Eyeball	0.477	0.000
Confirm eyeball	0.436	0.001
Bur	0.389	0.003
Impression index	0.549	0.001
Mhanni feedback sheet	0.409	0.001

Table 5.19 Measurement of intra-examiner (Kappa) agreement for the full veneer gold shell crown preparations for examiner 3 according to grades awarded

Type of assessment stage	Kappa value	Significance (p≤0.05)
Eyeball	0.134	0.140
Confirm eyeball	0.156	0.142
Bur	0.050	0.659
Impression index	0.150	0.166
Mhanni feedback sheet	0.268	0.032

The highlighted values represent statistically significant agreement differences.

It is apparent that the final stage of assessment 'Mhanni feedback sheet' achieves the highest agreement for the most of senior examiners. In addition, it shows that examiner 3 demonstrates poor agreement whereas examiners 1 and 2 display fair and moderate agreement.

Inter-examiner agreement for full veneer gold shell crown preparation according to grades for each stage

Table 5.20 Inter-examiner agreement for the full veneer gold shell crown preparation for each stage and occasion according to grades

Inter-examiner agreement for full veneer gold shell crown preparation					
Stage	*Occasion*	*Number of examiners*	*Single measurement ICC*	*95% of CI*	*Best single measurement if one examiner is excluded*
Eyeball	1	3	0.501	0.287 – 0.695	0.559 if Examiner 2 excluded
Eyeball	2	3	0.298	0.086 – 0.529	0.600 if Examiner 3 excluded
Confirm eyeball	1	3	0.510	0.241 – 0.751	0.594 if Examiner 3 excluded
Confirm eyeball	2	3	0.460	0.240 – 0.665	0.686 if Examiner 3 excluded
Bur	1	3	0.438	0.220 – 0.647	0.618 if Examiner 3 excluded
Bur	2	3	0.384	0.165 – 0.605	0.459 if Examiner 1 excluded
Impression index	1	3	0.252	0.023 – 0.501	0.498 if Examiner 3 excluded
Impression index	2	3	0.352	0.134 – 0.577	0.497 if Examiner 3 excluded
Mhanni feedback sheet	1	3	0.342	0.099 – 0.580	0.601 if Examiner 3 excluded
Mhanni feedback sheet	2	3	0.375	0.145 – 0.600	0.693 if Examiner 3 excluded

The highlighted values represent the highest inter-examiner agreement.

Table 5.20 showed that assessment of occasion two for each stage was not always better than occasion one. The highest levels of agreement in occasion one of eyeball and confirm eyeball stages, which produced poor to moderate inter-examiner agreement. On the other hand, the lowest level was in impression index stage occasion one. Furthermore, examiner 3 was the worst examiner in the most stages in relation to other examiners. Therefore, if examiner 3 was excluded, the inter-examiner agreement was improved. The best inter-examiner agreement stage if examiner 3 is excluded was for the ‘Mhanni feedback sheet’ stage (occasion two).

Intra-examiner agreement for full veneer gold shell crown preparation according to the number of negative points 'Mhanni feedback sheet'

Table 5.21 summarises the level of intra-examiner agreement of negative point number which was given by each examiner on two different occasions by using Un-weighted Kappa test. Data illustrates that the highest Kappa value was for examiner 1 whereas examiner 2 was the lowest value.

Table 5.21 Intra-examiner agreement for the full veneer gold shell crown preparation for each examiner according to the number of negative points

Senior examiners	Kappa Value	Significance ($p \leq 0.05$)
Examiner 1	0.197	0.002
Examiner 2	-0.006	0.904
Examiner 3	0.135	0.008

The highlighted values represent statistically significant agreement differences.

Inter-examiner agreement for full veneer gold shell crown preparation according to the number of negative points 'Mhanni feedback sheet'

Table 5.22 identifies the level of inter-examiner agreement for the number of negative points in occasion one was lower than in the occasion two. By process of elimination, this table also shows that the best agreement according to the number of negative points was for examiner 1 who was the only examiner not deleted when the best single measurement was determined.

Table 5.22 Inter-examiner agreement for the full veneer gold shell crown preparation according to the number of negative points for each examiner using 'Mhanni feedback sheet'

Occasion	**Intra-class correlation** single measure	**95% confidence interval**		**Significance ($p \leq 0.05$)**	**Best single measurement if examiner deleted**
		Lower bound	Upper bound		
1	0.562	0.356	0.739	0.000	0.643 if Examiner 2 is excluded
2	0.647	0.669	0.927	0.000	0.801 if Examiner 3 is excluded

b. Assessing the repeatability and reproducibility of detailed feedback from the perspective of, "consistency of message", to a learner

Intra-examiner repeatability according to criteria of 'Gray feedback sheet'

Table 5.23 summarises the Un-weighted Kappa scores for intra-examiner agreement at the level of performance for each criterion of the 'Gray feedback sheet' for each senior examiner. Examiner 1 demonstrated almost perfect intra-examiner agreement for 4/11 criteria of the, 'Gray feedback sheet' and substantial agreement for a further 4/11 criteria. Examiner 2 demonstrated substantial intra-examiner agreement for 1/11 criteria of the, 'Gray feedback sheet'. Examiner 3 demonstrated almost perfect intra-examiner agreement for 4/11 criteria of the, 'Gray feedback sheet' and substantial agreement for a further 3/11 criteria.

There was no substantial intra-examiner agreement by all three examiners for any criterion of the, 'Gray feedback sheet'. For examiners 1 and 3, almost perfect intra-examiner agreement was observed for 'occlusal width' of the, 'Gray feedback sheet'. For examiners 1 and 2, substantial intra-examiner agreement was observed for 'mesio-distal depth of the box' of the 'Gray feedback sheet'.

All other intra-examiner agreement varied between poor and moderate. Indeed, the lowest level of intra-examiner agreement for criteria of the, 'Gray feedback sheet' which have not specific features or measurements only for, ‘box outline’, ‘unsupported enamel’, ‘retention form’ and ‘occlusal lock’.

These results are not encouraging and are explored further in the discussion section of this chapter.

Table 5.23 Intra-examiner agreement and percent of agreement between three senior examiners according to criteria of the ‘Gray feedback sheet’ on two occasions

	Examiner 1			**Examiner 2**			**Examiner 3**		
	Kappa agreement	Significance ($p \leq 0.05$)	%	Kappa agreement	Significance ($p \leq 0.05$)	%	Kappa agreement	Significance ($p \leq 0.05$)	%
Box									
Box outline	-0.054	0.768	88	0.469	0.005	92	0.532	0.006	81
Position of box	**1.000**	**0.000**	100	0.458	0.019	92	0.435	0.018	81
Depth gingivally	0.644	0.000	85	0.514	0.000	69	0.864	0.000	92
Bucco-lingual width	0.733	0.000	85	0.438	0.001	65	0.832	0.000	92
Mesio-distal depth	0.679	0.000	88	0.680	0.000	88	0.873	0.000	96
Unsupported enamel	**1.000**	**0.000**	100	0.336	0.085	85	0.649	0.000	96
Retention Form	**1.000**	**0.000**	100	-0.002	0.985	35	0.798	0.000	92
Occlusal									
Occlusal lock	0.618	0.000	85	0.347	0.043	81	-0.020	0.838	92
Occlusal depth	0.567	0.004	85	0.359	0.034	65	0.693	0.000	88
Occlusal width	**1.000**	**0.000**	100	0.361	0.006	88	**1.000**	**0.000**	100
Damage to adjacent tooth	0.601	0.000	81	0.401	0.008	73	0.869	0.000	92

The high-lighted values represent statistically significant agreement differences.

Inter-examiner reproducibility according to criteria of 'Gray feedback sheet'

Table 5.24 summarises the intra-class correlation measurements (single measures) to determine inter-examiner agreement at the level of performance for each criterion of the 'Gray feedback sheet' among senior examiners.

There was moderate to substantial inter-examiner agreement for the width of the box preparation in both bucco-lingual and mesio-distal dimensions as well as the occluso-gingival depth of the box. Each of these features was evaluated using specific additional tools such as a bur or an amalgam condenser.

In general, there was only poor or slight inter-examiner agreement between the levels of performance that describe criteria for which there is no specific tool for measurement, such as, box outline, position of the box, unsupported enamel, and occlusal lock. There was also poor inter-examiner agreement for occlusal width for which the amalgam condenser was used to help with evaluation.

Table 5.24 Inter-examiner agreement (single measures) and confidence interval for criteria of the 'Gray feedback sheet' among three senior examiners for each of two occasions

	Occasion 1				**Occasion 2**			
	Intra-class correlation	95% confident interval		Significance ($p \leq 0.05$)	Intra-class correlation	95% confident interval		Significance ($p \leq 0.05$)
	(Single measures)	*lower*	*upper*		*(Single measures)*	*lower*	*upper*	
Box								
Box outline	0.239	0.024	0.491	0.013	0.160	-0.039	0.414	0.063
Position of box	0.306	0.077	0.554	0.004	0.057	-0.114	0.299	0.275
Depth gingivally	0.604	0.378	0.781	0.000	0.634	0.429	0.798	0.000
Bucco-lingual width	0.627	0.416	0.794	0.000	0.654	0.448	0.812	0.000
Mesio-distal depth	0.669	0.477	0.820	0.000	0.752	0.588	0.869	0.000
Unsupporte d enamel	0.235	0.011	0.493	0.021	0.161	-0.052	0.432	0.073
Retention Form	0.281	0.061	0.529	0.003	0.173	-0.012	0.413	0.007
Occlusal								
Occlusal lock	0.162	-0.048	0.423	0.072	0.456	0.223	0.675	0.000
Occlusal depth	0.313	0.086	0.558	0.001	0.533	0.308	0.731	0.000
Occlusal width	0.000	-0.175	0.251	0.485	-0.027	-0.210	0.233	0.579
Damage to adjacent tooth	0.340	0.108	0.584	0.002	0.311	0.074	0.563	0.005

The highlighted values represent statistically significant agreement differences.

Intra-examiner repeatability according to criteria of 'Mhanni feedback sheet'

Table 5.25 summarises the Un-weighted Kappa scores for intra-examiner agreement at the level of performance for each criterion of the 'Mhanni feedback sheet' for each senior examiner. Examiner 1 demonstrated almost perfect intra-examiner agreement for 3/20 criteria of the, 'Mhanni feedback sheet' and substantial agreement for a further 8/20 criteria. Examiner 2 demonstrated substantial intra-examiner agreement for 1/20 criteria of the, 'Mhanni feedback sheet'. Examiner 3 demonstrated substantial intra-examiner agreement for 3/20 criteria of the, ‘Mhanni feedback sheet’.

There was no substantial intra-examiner agreement by all three examiners for any criterion of the, ‘Mhanni feedback sheet’. For examiners 1 and 3, substantial intra-examiner agreement was observed for both the, ‘lingual reduction’ and the, ‘clear contact area with adjacent teeth’ criteria of the, ‘Mhanni feedback sheet’.

All other intra-examiner agreement varied between poor and moderate. Indeed, all examiners exhibited poor to fair agreement only for, ‘contour of preparation’, ‘depth of finish line all around’ and ‘texture of final preparation’.

These results are not encouraging and are explored further in the discussion section of this chapter.

Table 5.25 Intra-examiner agreement and percent of agreement between three senior examiners according to the criteria of the ‘Mhanni feedback sheet’ on two occasions

	Examiner 1			Examiner 2			Examiner 3		
	Kappa agreement	Significance ($p \leq 0.05$)	%	Kappa agreement	Significance ($p \leq 0.05$)	%	Kappa agreement	Significance ($p \leq 0.05$)	%
Occlusal surface									
Occlusal reduction	0.355	0.006	63	0.651	0.000	80	0.312	0.019	63
Contour of occlusal preparation	0.783	0.000	90	0.242	0.047	83	0.328	0.053	70
Axial surface(s)									
Buccal reduction	0.638	0.000	83	0.373	0.002	63	0.475	0.000	77
Lingual reduction	0.706	0.000	83	0.205	0.122	60	0.625	0.000	90
Mesial reduction	0.779	0.000	90	0.299	0.077	63	0.252	0.101	63
Distal reduction	0.591	0.000	80	0.359	0.037	70	0.327	0.019	67
Undercuts	0.651	0.000	97	0.359	0.033	83	0.474	0.020	93
Bucco-lingual convergence	0.815	0.000	93	0.415	0.005	73	0.204	0.056	80
Proximal convergence	0.143	0.414	80	0.493	0.006	77	-0.061	0.611	77
Contour of preparation	0.239	0.024	77	0.189	0.167	67	0.076	0.118	50
Contact area with adjacent teeth	0.839	0.000	97	0.520	0.004	87	0.651	0.000	97

	Examiner 1			Examiner 2			Examiner 3		
	Kappa agreement	Significance (p≤0.05)	%	Kappa agreement	Significance (p≤0.05)	%	Kappa agreement	Significance (p≤0.05)	%
Functional cusps									
Functional cusp bevel reduction	0.489	0.000	80	0.040	0.629	33	-0.252	0.038	30
Location of functional bevel	0.467	0.009	77	0.025	0.540	30	-0.171	0.249	47
Finish line									
Chamfer finish line	0.488	0.007	77	0.129	0.150	70	0.670	0.000	83
Level of finish line to gingival margin	0.667	0.000	90	0.314	0.084	77	-0.062	0.575	73
Depth of finish line all around	0.139	0.331	57	0.264	0.032	57	0.089	0.527	47
Final preparation									
Texture of final preparation except margin	0.268	0.114	87	0.368	0.041	73	0.153	0.256	67
Texture of margin	**1.000**	**0.000**	100	-0.053	0.735	87	0.048	0.788	73
Adjacent teeth damage									
Mesial tooth	0.783	0.000	97	0.453	0.009	83	0.492	0.003	83
Distal tooth	0.786	0.000	90	0.241	0.099	67	0.335	0.020	63

The high-lighted values represent statistically significant agreement differences.

<u>Inter-examiner reproducibility according to criteria of ‘Mhanni feedback sheet’:</u>

Table 5.26 summarises the intra-class correlation measurements (single measures) to determine inter-examiner agreement at the level of performance for each criterion of the ‘Mhanni feedback sheet’ among senior examiners. These data were the same data generated to determine intra-examiner agreement which had to take place over two occasions of evaluation. Thus, inter-examiner agreement for each occasion could be determined and there are some interesting comparisons between the two occasions.

The best inter-examiner agreement in any occasion of evaluation was for the criterion of, ‘distal tooth’ for the category of ‘Damage to adjacent teeth’ (ICC = 0.728) within the occasion two of evaluation. This was the only incidence of substantial agreement. There was moderate inter-examiner agreement for five criteria but this was spread across occasion one and occasion two. All other inter-examiner agreement was between poor and fair. For 11/20 criteria detailed in the, ‘Mhanni feedback sheet’, data from occasion one evaluation demonstrated better inter-examiner agreement than data from occasion two evaluation. These results are not encouraging and are explored further in the discussion section of this chapter.

Table 5.26 Inter-examiner agreement (single measures) and confidence interval for the criteria of the 'Mhanni feedback sheet' among three senior examiners for each of two occasions

	Occasion 1				**Occasion 2**			
	Intra-class correlation	95% confident interval		Significance (p≤0.05)	Intra-class correlation	95% confident interval		Significance (p≤0.05)
	(Single measures)	*lower*	*upper*		*(Single measures)*	*lower*	*upper*	
Occlusal surface								
Occlusal reduction	0.231	0.019	0.474	0.017	0.281	0.071	0.513	0.003
Contour of occlusal preparation	0.220	0.010	0.463	0.020	0.022	-0.118	-0.227	0.386
Axial surface(s)								
Buccal reduction	0.250	0.048	0.483	0.005	0.327	0.111	0.556	0.001
Lingual reduction	0.262	0.048	0.500	0.008	0.250	0.048	0.483	0.005
Mesial reduction	0.358	0.141	0.581	0.000	0.400	0.179	0.618	0.000
Distal reduction	0.294	o.o72	0.531	0.004	0.265	0.040	0.509	0.010
Undercuts	0.460	0.237	0.667	0.000	0.213	0.012	0.450	0.018
Bucco-lingual convergence	0.221	0.026	0.452	0.008	0.314	0.087	0.551	0.003
Proximal convergence	0.094	-0.063	0.310	0.113	0.046	-0.144	0.294	0.325
Contour of preparation	0.147	-0.063	0.399	0.091	-0.008	-0.176	0.227	0.515
Contact area with adjacent teeth	0.451	0.231	0.658	0.000	0.566	0.360	0.741	0.000
Functional cusps								
Functional cusp bevel reduction	0.107	-0.030	0.302	0.035	0.028	-0.118	0.238	0.364
Location of functional bevel	0.018	-0.108	0.209	0.393	-0.010	-0.150	0.198	0.529
Finish line								
Chamfer finish line	0.266	0.059	0.498	0.002	0.557	0.350	0.735	0.000
Level of finish line to gingival margin	0.559	0.349	0.738	0.000	0.158	-0.043	0.402	0.067
Depth of finish line all around	0.373	0.155	0.595	0.000	0.226	0.018	0.466	0.016
Final preparation								
Texture of final preparation except margin	0.132	-0.060	0.373	0.098	0.141	-0.059	0.383	0.087
Texture of margin	0.482	0.267	0.681	0.000	0.237	0.027	0.476	0.013
Damage to adjacent teeth								
Mesial tooth	0.259	0.044	0.498	0.009	0.393	0.174	0.611	0.000
Distal tooth	0.552	0.344	0.732	0.000	0.728	0.567	0.848	0.000

The high-lighted values represent statistically significant agreement differences.

5.5 Discussion

Three senior examiners evaluated class II amalgam cavities and full veneer gold shell crown preparations on two occasions at least one week apart.

On each occasion, each examiner

- followed a common, cumulative, five-stage scheme of evaluation using common criteria (including feedback sheets 'Gray and Mhanni') to determine a grade for each type of preparation and
- used common feedback sheets 'Gray and Mhanni' alone to provide feedback comments, based around negative points, for each type of preparation.

This complex evaluation can be broken down logically as follows (Figure 5.10):

1) **From the evaluation of the class II amalgam cavity preparation,**
   a) the grades awarded could be used to determine
      i) intra-examiner agreement and
      ii) inter-examiner agreement,
   b) the number of negative points awarded could be used to determine
      i) intra-examiner agreement and
      ii) inter-examiner agreement ,
   c) and the consistency of negative points identified could be used to determine
      i) intra-examiner agreement and
      ii) inter-examiner agreement.

2) **From the evaluation of the full veneer gold shell crown preparation,**

   a) the grades awarded could be used to determine

      i) intra-examiner agreement and

      ii) inter-examiner agreement,

   b) the number of negative points awarded could be used to determine

      i) intra-examiner agreement and

      ii) inter-examiner agreement.

   c) and the consistency of negative points identified could be used to determine

      i) intra-examiner agreement and

      ii) inter-examiner agreement.

Figure 5.10 Outline of the principle findings from evaluation of the class II amalgam preparation and the full veneer gold shell crown preparation

From Figure 5.10, the principle findings from the complex evaluation which was broken down were:

5.5.1 Evaluation of class II amalgam cavity preparations.

(1ai) Intra-examiner agreement according to grades awarded improved through the cumulative stages of grading for all examiners (Tables 5.7 – 5.9). The 'Gray Feedback sheet' achieved the highest agreement followed by the diamond fissure bur assessment stage while traditional visual assessment (eyeball stage) was the lowest.

(1aii) On the other hand, inter-examiner agreement for class II amalgam cavity preparation did not improve through the cumulative stages of grading for both occasion one and occasion two. However, occasion two, for each stage, was better than occasion one of assessment. Furthermore, the amalgam condenser and bur stages produced better inter-examiner agreement compared with other stages. Once again, traditional visual assessment (eyeball stage) had the lowest inter-examiner agreement. If examiner 2 was excluded, the bur and ‘Gray feedback sheet’ stages produced better inter-examiner agreement, ICC = 0.817 and 0.855, respectively (Table 5.10).

(1bi) Intra-examiner agreement for the number of negative points according to the ‘Gray feedback sheet’ demonstrated the highest level of agreement for examiner 3 while examiner 2 was the lowest, Kappa 0.589 and 0.211, respectively (Table 5.11).

(1bii) Generally, the inter-examiner agreement among senior academic staff was good but became better if examiner 2 was excluded in occasion one and if examiner 1 was excluded in occasion two 0.924 and 0.872, respectively (Table 5.12).

From the results of intra- and inter-examiner agreement according to the number of negative points identified, the examiner who had the best agreement was Examiner 3.

(1ci and 1cii) For assessing the repeatability and reproducibility of detailed comments from the ‘Gray feedback sheet’, the overall consistency of feedback to the student was low from the perspective of both intra- and inter-examiner agreement (Tables 5.23 and 5.24). However, there were elements from the feedback sheet that achieved higher consistency. Generally, these elements were those where some objective evaluation of the class II amalgam cavity was possible, for example, intra-examiner and inter-examiner agreement for the mesio-distal depth of the box preparation (Tables 5.23 and 5.24).

5.5.2 Evaluation of full veneer gold shell crown preparations.

(2ai) In contrast to the class II amalgam cavity preparation, intra-examiner agreement of the grades awarded for the full veneer crown preparation did not always improve through the cumulative stages of grading. The 'Mhanni Feedback sheet' achieved the highest intra-examiner agreement for the greatest proportion of senior examiners followed by traditional visual assessment stages (Tables 5.17 to 5.19). The lowest level of agreement was for the bur assessment stage. Moreover, the highest level of intra-examiner agreement according to grades was for <u>examiner 1</u>.

(2aii) Inter-examiner agreement for full veneer gold shell crown preparation was also not improved through the cumulative stages of grading for occasion one and occasion two except for two stages; the use of the impression index and the use of the 'Mhanni feedback sheet' (Table 5.20). Examiner 3 was the worst assessor in the most stages. If Examiner 3 was excluded, the inter-examiner agreement improved and the highest level of inter-examiner agreement was traditional visual (confirm eyeball) and 'Mhanni feedback sheet' stages ICC = 0.686 and 0.693, respectively (Table 5.20).

(2bi) For intra-examiner agreement for full veneer gold shell crown preparation according to the number of negative points, the highest level agreement was for Examiner 1 while Examiner 2 was the lowest level (Table 5.21).

(2bii) On the other hand, the level of inter-examiner agreement according to number of negative points in occasion one was lower than in occasion two (Table 5.22).

(2ci and 2cii) Generally, for assessing the repeatability and reproducibility of detailed comments from the 'Mhanni feedback sheet', the overall repeatability and reproducibility of feedback to the student was low from the perspective of both intra- and inter-examiner agreement for some senior examiners (Tables 5.25 and 5.26). However, there were elements from the feedback sheet that achieved higher consistency. These elements were those where some objective evaluation of the full veneer gold shell crown preparation was possible, for example, intra-examiner and inter-examiner agreement for the contact area with adjacent teeth.

"Assessment drives learning" (Wass et al., 2001), therefore, the primary purpose of this part of study was to develop and validate feedback sheets for class II amalgam cavity and full veneer gold shell crown preparation assessments in a clinical skills laboratory. The 'Mhanni feedback sheet' was developed and validated using methods described by Lynn (1986), DeVon et al., (2007) and Sirajudeen et al., (2012). The results presented in this chapter provided the first step to evaluate the utility of this approach and subsequently, the impetus to create revised checklists.

Intra-examiner agreement for class II amalgam cavity and full veneer gold shell crown preparation through the cumulative stages of grading for occasion one and occasion two was improved by using specific additional tools. On the other hand, inter-examiner agreement was low for full veneer gold shell crown preparation and better for class II amalgam cavity preparation. Even-though inter-examiner agreement was low, there was improvement from occasion one to occasion two by using a diamond fissure bur and 'Gray feedback sheet' for class II amalgam cavity preparation and the impression index and 'Mhanni feedback sheet' for full veneer gold shell crown preparation. From Tables 5.19 and 5.20, it is clear that the examiner 3 was not familiar with full veneer gold shell crown preparation assessment.

There are several studies supporting the findings in this part of the study. Intra-examiner agreement was better than inter-examiner agreement of grades for tooth preparations in studies by Lilley et al., (1968), Fuller, (1972), Salvendy et al., (1973), Deranleau et al., (1983), Jenkins et al., (1998) and Sharaf et al., (2007). The same result for examiner agreement was also reported by Vann et al., (1983) and Sherwood and Douglas (2014), when they compared visual assessment with a checklist. For inter-examiner agreement, Vann et al., (1983) and Sharaf et al., (2007) reported that there was no method for improving inter-examiner agreement. This result was also concluded by other studies (Goepferd and Kerber, 1980, Satterthwaite and Grey, 2008).

For negative points, intra-examiner agreement was better in the 'Gray feedback sheet' than in the 'Mhanni feedback sheet'. This was because the number of criteria and the design of the rating scales were different between the two sheets. In addition, inter-examiner agreement for negative points from 'Gray or Mhanni feedback sheets' was better than intra-examiner agreement. Helft et al., (1987) reported that, even though all examiners who used a similar scaling system and checklist, there was disagreement among them. It might have arisen because the rating scale system lacked objective criteria (Helft et al., 1986, O'Donnell et al., 2011, Alhumaid et al., 2016). Thus, rating system and levels of performance for each criterion should be provided clearly for the examiners to assess students' work accurately. In addition, Feil in 1982 concluded that criteria of checklist and/or scaling system give chance for examiner to provide specific grade according to their own interpretations (Feil, 1982). Other reasons suggested for such disagreement were examiner experience, internal rater bias, and training (Lilley et al., 1968, Houpt and Kress, 1973, Helft et al., 1987, Sharaf et al., 2007, Alhumaid et al., 2016).

The grading system and negative points from the feedback sheet were not defined objectively by using additional tools in this part of study. Therefore, inter-examiner agreement was low. This result was supported by the work of Ganies (1974), Helft et al., (1987) and O'Donnell et al., (2011) reports. Although inter-examiner agreement was low in this part of study, most of the occasion two of assessment stages for class II amalgam cavity preparation and some for full veneer gold shell crown were better than the occasion one assessment stage. According to Lillely et al., (1968), occasion one may represent a training session for the assessor in preparation for occasion two. Therefore, inter-examiner agreement for the second time became better (Lillely et al., 1968).

Halft et al., (1987) and Knight (1997) suggested that the clearly defined grading systems and levels of criteria of feedback sheets provide less scope for interpretation by examiners. In this part of the study, consistency in assessment feedback among senior examiners was poor. It is speculated this was because most senior examiners did not use the specific additional tools correctly. For example, examiners did not always use the bur alongside the 'Gray feedback sheet' to record their grades. Goepferd and Kerber (1980) observed, if an examiner used specific additional tools properly, the consistency of feedback sheet was improved. They also used a, 'glance and grade' system, a set of criteria and a checklist. They reported that there was improvement in both intra- and inter-examiner agreement using these methods. In contrast, Vann et al., (1983) used the same grades and descriptors as Goepferd and Kerber (1980), and reported that there was no improvement in inter-examiner agreement with the use of criterion and checklists after comparison with global assessment (Vann et al., 1983). Taylor et al., (2013) suggested a possible reason for this difference in results was because Goepferd and Kerber (1980) used a, 'glance and grade' system first and then repeated this with the additional use of criterion and checklist while Vann et al., (1983) compared global assessment with criterion and checklist only. On the other hand, Sherwood and Douglas in 2014 suggested that preclinical operative work of students be assessed by objective checklist criteria scoring rather than use glance and grade method and the checklist should be introduced after training and calibration sessions to decrease examiner inconsistency. In many teaching institutions, the, 'glance and grade' method is still used. Schiff et al., (1975) tried to reduce subjective assessment by using a tool called a, 'Pulpal floor measuring instrument'. Even-though this tool provided some advantages, it was not suitable to assess all features of the cavity. Therefore, in this study, the researcher (AM) tried to find the best tools which can be used to assess tooth preparation in order to improve intra- and inter-examiner agreement for grades and consistency of feedback.

5.6 Conclusion

According to grade:

For class II amalgam cavity preparation, intra-examiner agreement according to grades awarded improved through the cumulative stages of grading for all senior examiners. The 'Gray Feedback sheet' achieved the highest level of agreement. In contrast to the class II amalgam cavity preparation, intra-examiner agreement of the grades awarded for the full veneer crown preparation did not always improve through the cumulative stages of grading. The ‘Mhanni Feedback sheet’ achieved the highest intra-examiner agreement for the greatest proportion of senior examiners.

On the other side, inter-examiner agreement for class II amalgam cavity preparation did not improve through the cumulative stages of grading for both occasion one and occasion two. However, occasion two, for each stage, was better than occasion one of assessment. Furthermore, the amalgam condenser and bur stages produced better inter-examiner agreement compared with other stages. In contrast to the class II amalgam cavity preparation, inter-examiner agreement for full veneer gold shell crown preparation was also not improved through the cumulative stages of grading for occasion one and occasion two except for two stages; the use of the impression index and the use of the ‘Mhanni feedback sheet’.

According to negative points of feedback sheet:

The use of a feedback sheet to assess class II amalgam cavity and full veneer gold shell crown preparations did not improve intra-examiner agreement of senior academic staff, while inter-examiner agreement was better.

Intra-examiner agreement for the number of negative points demonstrated the highest and the lowest level of agreement for senior academic staff examiners. Inter-examiner agreement for the number of negative points among senior academic staff was moderate. Furthermore, Occasion one for senior academic staff examiners produced closely similar level of inter-examiner agreement according to number of negative points with occasion two.

According to consistency of feedback sheet:

Consistency is a very essential part to provide reliable and fair feedback for student's work. Repeatability for each examiner (intra-examiner agreement) was better than among examiners (inter-examiner agreement). Especially for the feature which has specific measurement. Therefore, the use of specific additional tools (including the development of a feedback sheets) to assess tooth preparations improved intra and inter-examiner agreement of senior academic staff for some features which have specific measurement, for example, depth or width of the feature.

All in all, feedback sheets which were used did not always provide repeatable and reproducible judgment for the student. The reason for this might be attributed to assessor bias and misinterpretation. Therefore, valid, clear descriptions and reliable feedback sheet are essential to judge student work. Hence an important question is, "How can researchers know whether their repeatability or reproducibility relates to valid observations without gold standard data?" In other words, although the researcher may be able to prove the assessment method repeatability and internal consistency, and, therefore reliability, the assessment method itself may not be valid (see Chapter 6).

Chapter 6 : Development of a standard (representative grades) for 26 class II amalgam cavities and 30 full gold shell-crown preparations

Chapters	
Chapter 2	**Evaluating current assessment and feedback methods at Dundee Dental School in the Clinical Skills Laboratory** **Aims:** The aims of this Chapter were to identify, in relation to the class II amalgam cavity preparation and the full veneer gold shell crown preparation undertaken in the Clinical Skills Laboratory at Dundee Dental School: 1. the types and quality of feedback and assessment provided for students and, 2. the disadvantages associated with current feedback.
Chapter 3	**Independent evaluation of student class II amalgam cavity and full gold shell-crown preparations by senior academic staff** **Aim:** The aim of this study was to evaluate inter-examiner agreement between a convenience sample of three senior academic staff members, who have more than twenty years of clinical and teaching experience each, when evaluating operative procedures (class II amalgam cavity and full veneer gold shell crown preparation) on plastic teeth, from a sample year-cohort of dental students, at Dundee Dental School.
Chapter 4	**Evaluation of selected student class II amalgam cavity and full gold shell-crown preparations by senior academic staff and additional teaching staff** **Aims:** 1. To develop a sub-set of class II amalgam cavity, and full veneer gold shell crown, preparations. 2. To determine intra-examiner agreement for a group of senior academic staff when evaluating a sub-set of class II amalgam cavity preparations and full-veneer gold shell crown preparations. 3. To study a large group of additional teaching staff at Dundee Dental School in order to determine intra- and inter-examiner agreement when evaluating a sub-set of class II amalgam cavity preparations and full-veneer gold shell crown preparations.
Chapter 5	**Evaluation of selected student class II amalgam cavity and full gold shell crown preparations by using specific additional tools and feedback sheets** **Aims:** 1. To identify and develop specific additional tools to assist with evaluation of tooth preparations. 2. To determine examiner agreement for both class II amalgam cavity and full veneer gold shell crown preparations using the specific additional tools and feedback sheets for both grades awarded and repeatability and reproducibility of detailed feedback provided by senior academic staff.
Chapter 6	**Identification of gold standard (representative grades) for 26 class II amalgam cavities and 30 full gold shell-crown preparations** **Aims:** **These objectives can be set out as a number of steps, each of which had a specific aim which is expanded within the methodology of this chapter. Thus, the two aims of this chapter were:** **1. to identify the best senior academic staff examiner who had the highest correlation between grades awarded with the number of negative points identified;** **2. to identify those tooth preparations which were awarded a grade by the best examiner in order to:** • **determine the specific feature measurements of this sub-group of grade tooth preparations and:** • **compare the objective class II amalgam cavity or full veneer gold shell crown preparation dimensions (SAMFs) recorded by the researcher and those subjective tooth preparation evaluations (SAFs) for similar teeth by the senior examiner with dimensions presented in the dental literature and subsequent calibration with the grades awarded by the best examiner.**
Chapter 7	**Identification of reliable tools and development of a new checklist to assess class II amalgam cavities and full veneer gold shell crown preparations** **Aims:** 1. To identify a reliable specific additional measurement tool. 2. To develop two reliable new checklists (New Checklist CII Preparation and New Checklist Gold shell Crown Preparation). 3. To determine intra- and inter-examiner agreement using the grades and consistency of the New Checklists (Class II Preparation and Gold Shell Crown Preparation) and compare these data with that from previously used 'Gray and Mhanni feedback sheets' to ascertain if the agreement and consistency of the New Checklists have been improved.
Chapter 8	**General conclusions, recommendation, and further studies**

6.1 Introduction

Assessment of tooth preparations within the clinical skills laboratory is one of the most common subjective judgments used to evaluate student performance at Dundee Dental School. Subjective judgment has a major impact on the student attitude and motivation (Sherwood and Douglas, 2014). Therefore, validity, repeatability and reproducibility of the assessment of tooth preparations should be focused to provide objective elements of judgment in preference to subjective elements of judgment (Streiner and Norman, 2008). By focussing on objective judgment, the motivation of students to learn and acquire new skills might, therefore, improve.

The results gathered in Chapter 5 concluded that there was an improvement in intra-examiner agreement to rank tooth preparations using specific additional tools (e.g. amalgam condensers, burs and indices) for class II amalgam cavities and full veneer gold shell crown preparations, while inter-examiner agreement demonstrated no such improvement. These results support the conclusions of Haj-Ali and Feil (2006), who concluded that intra-examiner repeatability was greater than inter-examiner reproducibility. It was also apparent from Chapter 5 that the use of a feedback sheet, to determine the number of negative points when assessing tooth preparations, failed to improve intra-examiner agreement of senior academic staff, although it did enhance inter-examiner agreement. The result of Chapter 5 were summarised in Appendix 3.

In addition, the feedback sheet was used to evaluate the repeatability and reproducibility of feedback given to the student. From Chapter 5, although the overall repeatability and reproducibility of feedback to the student was low, feedback was more repeatable for individual examiners (intra-examiner repeatability) than between examiners (inter-examiner reproducibility). These results are supported by the conclusions of Sharaf et al., (2007) and Sherwood and Douglas (2014), who concluded that the use of objective checklists improved intra-examiner repeatability but failed to improve repeatability between examiners. However, other authors have concluded that checklist and performance criteria did not help improved grading intra- or inter-examiner repeatability (Houpt and Kress, 1973, Vann et al., 1983). Therefore, feedback sheets which were used in Chapter 5 did not always provide objective judgment for the student. Furthermore, the grades which were judged by using feedback sheet might not reflect the teeth preparations truly.

There is a question raised by these contrary results. This question is, "How can researchers know whether their repeatability relates to valid observations without gold standard data?"

While examiner repeatability is important, it is equally important that examiner grading reflects what is truly known about the tooth preparation. Thus, it is equally important that the grading is valid. In a recent review article, several studies did not state the method of grade calibration clearly (Mays and Branch-Mays, 2016). On the other hand, some authors have addressed the problem of; how can an evaluator select the grades which truly reflect the standard of the tooth preparation?

Selection of a grade from the examiner who had the highest specialty or greatest experience as a gold standard was one of the most common recommended methods (Curtis et al., 2008, Cho et al., 2010, Mays and Levine, 2014, Tuncer et al., 2015, Alhumaid et al., 2016). A few studies used an averaged value from all examiners (Cho et al., 2010, Callen et al., 2015). No previous study has investigated if the selection of a grade from only one expert or several experts is valid to act as a gold standard reflecting the tooth preparation truly.

From Chapter 5 in this thesis, intra- and inter-examiner agreements of the senior academic staff was variable; some combinations of senior academic staff had low agreements and some combinations of senior academic staff had a higher agreement. The same finding was evident when a feedback sheet was used in order to establish an overall grade (Appendix 3). Therefore, it was not always valid to select the average grade from several senior academic staff as a gold standard to truly reflect the tooth preparation.

In terms of what reflects the tooth preparation truly (the gold standard), only objective measurement has been shown to have good repeatability. Subjective evaluation has been used with mixed results (Sherwood and Douglas, 2014). Some aspects of tooth preparation evaluation require specialist methodology not available to the researcher (AM) of this thesis; for example, profilometry was used to evaluate surface roughness of enamel (Rao et al., 2011) and can be used to determine the proximal surface damage of adjacent teeth. Furthermore, these specialist methods are not available in many dental schools. Thus, if the subjective evaluation is to be used in order to contribute to a definition what reflects the tooth preparation truly, it would be logical to limit this to a binary response such as yes/no. Therefore, the combined use of both forms of evaluation (objective and subjective) can often exploit the advantage of each. Fundamentally, any feedback to a student is designed to help them make their own judgments in the clinic of the acceptability of the preparations they make for teeth in the oral cavity.

Objective measurements might include the convergence angle of a crown preparation or the depth of a proximal box for a class II amalgam cavity preparation. Binary subjective evaluation might include the presence or absence of damage to the proximal surfaces of adjacent teeth or clearance of the contact point with adjacent teeth. Such things can be independently verified and used to determine a gold standard reflecting what is truly known about the tooth preparation (the gold standard). In this thesis, the features of tooth preparations that were evaluated objectively were called Specific Anatomical Feature Measurements (SAFMs) and subjectively were called Specific Anatomical Features (SAFs).

If the grades of the examiners all agreed with the gold standard descriptors, all examiners would agree with each other (Haj-Ali and Feil, 2006). However, there is a fundamental problem. An agreed gold standard descriptor for tooth preparations does not exist. Thus the author was faced with the problem of developing such a standard based on the published literature which described acceptable tooth preparations. This cannot be called a 'Gold Standard' and the phrase used in this thesis will be 'Developed Standard'. Ultimately, this Developed Standard may acquire 'gold' status but this will be for others to judge.

Therefore, the objectives of this chapter were to determine:

- the senior academic staff who had the best grade agreements (see Chapter 5) had the best correlation with the number of negative points and,
- if this individual had the best agreement with a Developed Standard.

6.2 Aims and null hypothesis

Aim:

These objectives can be set out as a number of steps, each of which had a specific aim which is expanded within the methodology of this chapter. Thus, the two aims of this chapter were:

1. to identify the best senior academic staff examiner who had the highest correlation between grades awarded with the number of negative points identified;
2. to identify those tooth preparations which were awarded a grade by this examiner in order to:
    a. determine the specific feature measurements of this sub-group of grade tooth preparations and:
    b. compare the objective class II amalgam cavity or full veneer gold shell crown preparation dimensions (SAFMs) recorded by the researcher (AM) and those subjective tooth preparation evaluations (SAFs) for similar teeth by the senior examiner with dimensions presented in the dental literature and subsequent calibration with the grades awarded by this examiner.

Null Hypothesis:

Tooth preparations with all grades awarded by the best senior academic staff examiner do not agree with the passing and failing grade features for tooth preparations reported in the literature and measured objectively by the researcher (AM).

6.3 Material and methods

This part of the study was carried out on similar samples used in Chapter 5. Some statistical analysis on the grades was used to determine the best intra- and inter-examiner agreement for senior academic staff (see Chapter 5). In this Chapter, the grades which were awarded from senior staff academic staff examiners were correlated with the number of negative points in order to confirm the best senior examiner. In addition, grades awarded from the best senior examiner were compared with the passing grade (ideal and acceptable) features for tooth preparations reported in the literature in order to reject or accept the null hypothesis.

In order to compare the grades awarded from the best senior examiner with the ideal or acceptable specific feature measurements, 26 class II amalgam cavities were scanned using a 3D optical scanner and 30 full veneer gold shell crown preparations were photographed in order to determine SAFMs of prepared teeth to define the relationship between SAFMs and feedback sheet grades for the best senior examiner.

6.3.1 Identification of the best senior examiner and grades for class II amalgam cavity and full veneer gold shell crown preparations

The examiner who had the highest intra-examiner repeatability by using Kappa test and the greatest positive impact on inter-examiner reproducibility by using intra-class correlation coefficient (ICC), was calculated to identify the best senior academic staff examiner according to grade repeatability (see Chapter 5).

Statistical analysis

Spearman correlation analysis (SPSS) of the ‘Gray and Mhanni feedback sheet’ grades from the 5th assessment stage (see Chapter 5) awarded by the three senior examiners versus, the number of negative points awarded from the same feedback sheets, was used in order to confirm the best senior examiner. The best senior examiner was defined and confirmed by the strongest negative correlation between grades and the number of negative points (Tables 6.1, 6.2, and 6.3).

6.3.2 Measuring the specific anatomical feature measurements (SAFMs) for each type of tooth preparation

This section firstly identified from the literature what had been measured (see Chapter 1) and compared this information with what the researcher (AM) was able to measure from the prepared blocks of teeth containing the tooth preparations. Subsequently different methods were used to make measurements on different occasions to determine the reliability of SAFMs for each prepared tooth.

Class II amalgam cavity preparation

Based upon the dimensions of a class II amalgam cavity thought to be of clinical significance and informed by the previous literature review [see Chapter 1 (Table 1.5)], the following specific cavity features (Figure 6.1) were measured;

1. Box- depth of box gingivally,
2. Box - bucco-palatal width
    a. gingivally (floor of the box),
    b. occlusally,
3. Box floor (mesio-distal) depth,
4. Box - pulpal axial wall length,
5. Occlusal - isthmus width
    a. cavity floor,
    b. occlusally,
6. Occlusal - cavity width along a line drawn from buccal to palatal cusp tips,
7. Occlusal - cavity depth
    a. at palatal side along the line described in point 6
    b. at buccal side along the line described in point 6, and

c. at the distal extent of the occlusal preparation.

8. Marginal ridge thickness

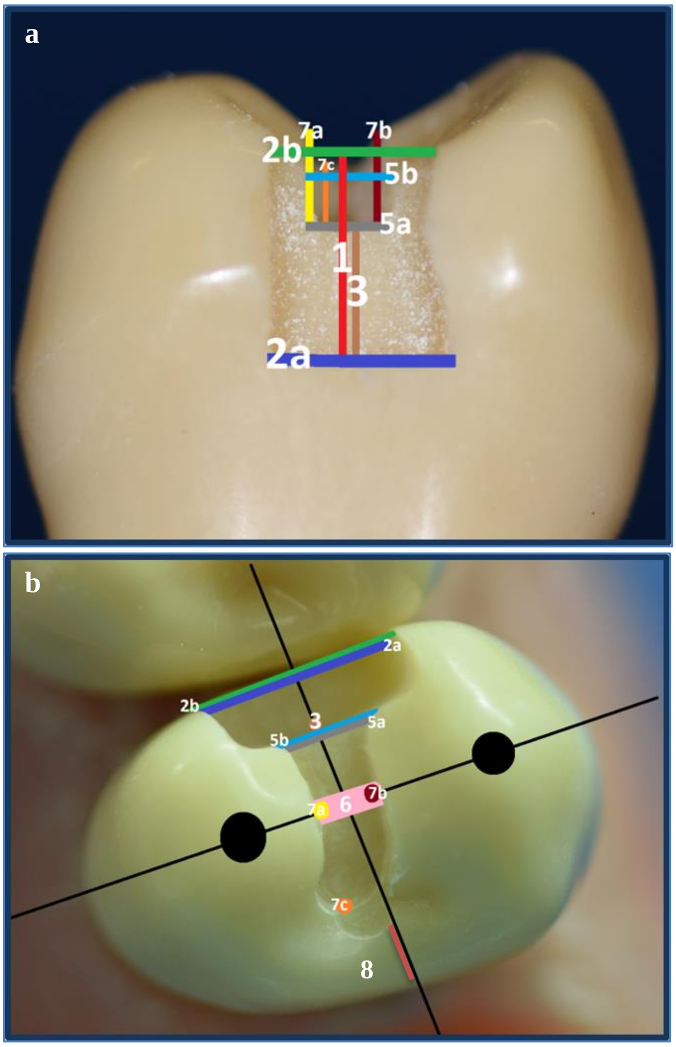


Figure 6.1 Diagram of upper second premolar cavity dimensions from a) mesial and b) occlusal views

Each class II amalgam cavity was measured by two different methods. Measurements were made directly, using a digital calliper (Mitutoyo, made in Japan) (Figure 6.2), and indirectly, from 3D digital images with a, ".STL", extension (StereoLithography).

For the indirect method, the files were acquired by scanning the cavities using an intraoral scanner (Lava^{TM} Chairside Oral Scanner, 3M, ESPE). MeshLab software (version 1.3.3) was used to measure SAFMs of the class II amalgam cavity preparations (Figure 6.3). Each direct and indirect measurement was made on two occasions, one week apart.

Figure 6.2 Photograph of the digital calliper used to measure specific features of the class II amalgam cavity for the upper second premolar

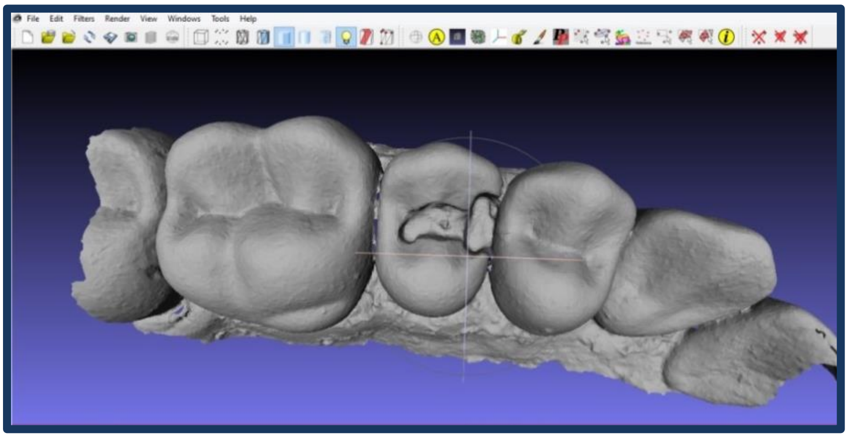


Figure 6.3 .STL image of a class II amalgam cavity for an upper second premolar analysed using MeshLab software

However, there were some specific anatomical features which could not be measured by using specific tools or methods. For example, the retention forms of the class II cavity preparation. In these cases, the specific anatomical features were abbreviated as, “SAFs”. To simplify matters, a decision was made that the evaluation of these SAFs should be binary. The binary pattern was used to reduce subjective evaluation.

For example; subjective evaluation might include the *presence* or *absence* of retention form for the proximal box or the *presence* or *absence* of damage to the adjacent tooth. Such things can be independently verified and used to determine the passing grade (acceptable) features.

To calibrate MeshLab software and the digital calliper, a range of ParaPostXP, parallel-sided, impression plastic posts (ParaPostXP Casting Technique System, Casting Introductory Kit, Coltène/Whaledent®) of 0.90, 1.00, 1.14, 1.25, 1.25, 1.40, 1.50 and 1.75mm were measured X2 by the researcher (AM). Manual measurements were performed directly by using the digital calliper. For MeshLab software, the plastic posts were scanned using an intraoral scanner (Lava™ Chairside Oral Scanner, 3M, ESPE). Thereafter, MeshLab software (version 1.3.3) was used to measure diameters of the Parapost on the monitor. Statistical analysis was carried out using Intra-class correlation (ICC) to measure the correlation between the two measuring methods in relation to diameters of the plastic post according to the manufacturer. The results demonstrated that there was a highly positive correlation [ICC = 1.00, 95% Confidence Interval (0.99 – 1.00)]. In addition, Appendix 4 demonstrates that there was highly positive correlation between the digital calliper and MeshLab measurements ($r^2 = 0.9992$, $y = 0.9978x + 0.0028$).

Full veneer gold shell crown preparation

Based upon the dimensions of a full veneer gold shell crown preparation thought to be of clinical significance, as informed by the previous literature review [see Chapter 1 (Table 1.6)], the following specific preparation features (Figures 6.4 and 6.5) were measured;

1. Total occlusal convergences,
    a. mesio-distal plane (angle)
    b. bucco-palatal plane (angle).
2. occlusal reduction,
    a. on the mesio-facial cusp
    b. on the disto-facial cusp
    c. on the bucco-mesial cusp
    d. on the palato-mesial cusp.
3. Axial reduction,
    a. on the mesial side
    b. on the distal side
    c. on the buccal side
    d. on the palatal side.

Each feature of full veneer gold shell crown preparation was measured by only one method. Measurements were made indirectly, using an ImageJ software (Version 1.47, USA), from 2D digital images. ImageJ software is a free, public-domain, Java image processing programme, which has been used as a measuring tool in several studies. According to Kerner et al., (2007), ImageJ software is a reliable and repeatable method when linear measurements have been made (Kerner et al., 2007). Therefore, another method was not required.

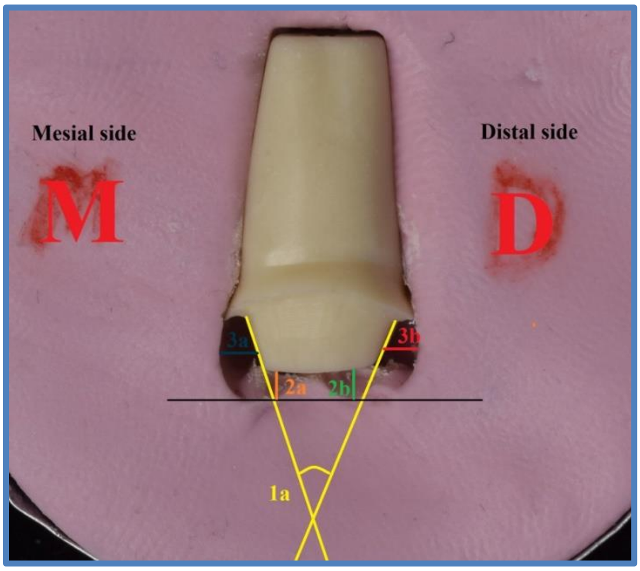


Figure 6.4 Diagram of upper first molar dimensions from buccal view

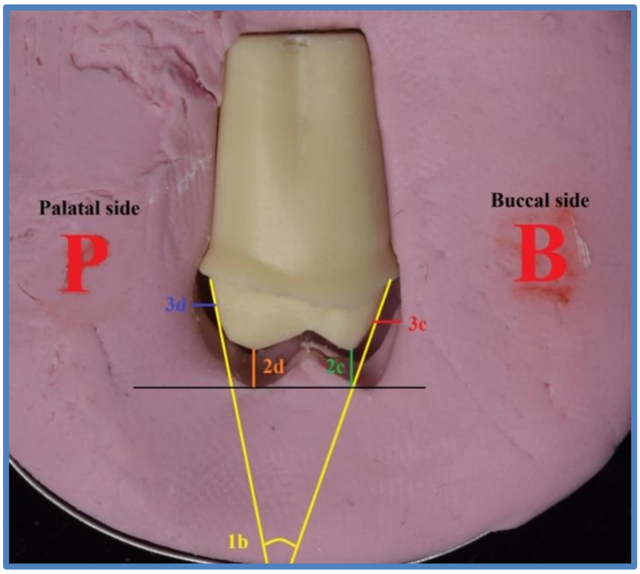


Figure 6.5 Diagram of upper first molar dimensions from mesial view

Before capturing the image, position of prepared tooth was set-up. To ensure that the positions of the Typodont teeth were stable and repeatable, custom-fit models were made by using putty impression (Lab-Putty, condensation type, Coltene, USA). Four models were made by taking impressions for distal and palatal sides of upper right and left *un-prepared* standard-sized molar teeth. These putty models were made to fit the full veneer gold shell crown molar preparation in the impression of unprepared molar tooth. Two models were used to provide mesial and buccal views of the upper right first molar tooth preparation and two models were used to provide mesial and buccal views of the upper left first molar tooth preparation (Figure 6.6). These models allowed the placement of all the prepared teeth in the same position each time. The putty models allowed the researcher (AM) to capture the prepared tooth in two different planes (i.e. buccal and mesial planes) to measure bucco-palatal, mesio-distal planes (i.e. total occlusal convergences), occlusal reduction and axial reduction of the same tooth on two occasions, one week apart.

Each full veneer gold shell crown preparation tooth was positioned in the impression model. Mesio-distal and bucco-palatal planes of prepared tooth were photographed by a digital camera.

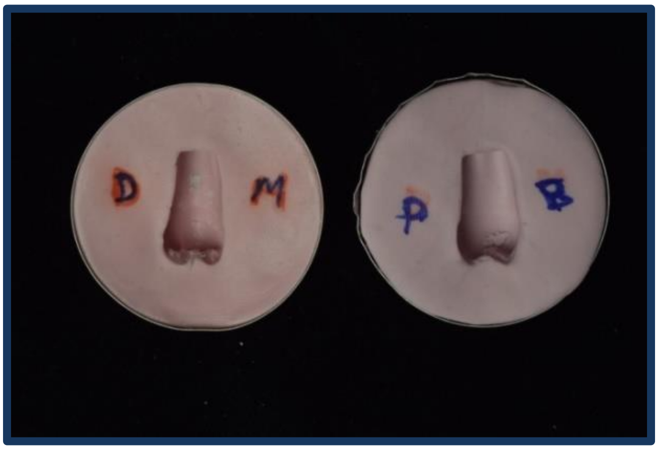


Figure 6.6 Picture of models for upper left first molar tooth
(D: Distal, M: Mesial, P: Palatal, and B: Buccal)

Images were captured with a digital single-lens reflect camera (DSLR, Nikon D3100) with a macro lens (Sigma 105 mm f/2.8 EX DG) and ring flash (Sigma MACRO EM-140 DG) which was set up on a tripod at distance of 30 cm from the table surface. A black background was set up to increase the contrast between the background and the tooth in the putty model, which allowed the axial walls and scale of the metal endo ruler (Miltex dental – Endo ruler – stainless steel, USA) to be more easily recognised on the monitor (Figure 6.7). The Endo ruler was used to calibrate scale measurements using ImageJ software (version 1.47, USA). Images of the samples were taken on the same day and under a constant light source. Images were imported into ImageJ software in order to measure specific anatomical features of the full veneer gold shell crown preparations.

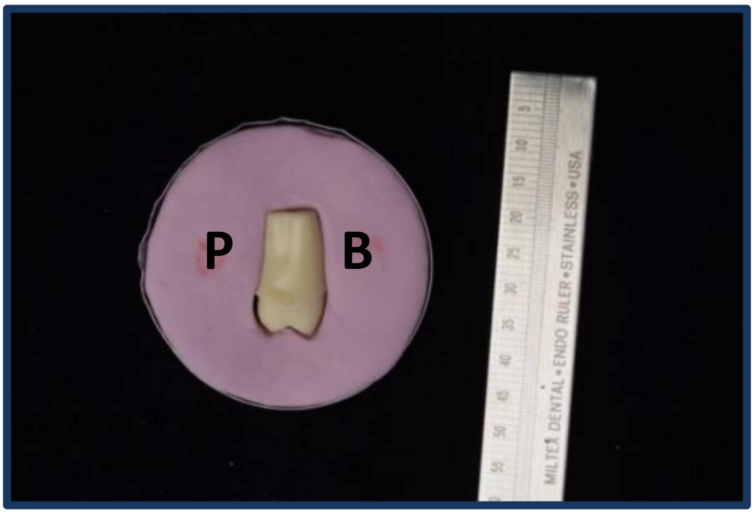


Figure 6.7 Picture of the impression model with a prepared tooth and endo ruler
(P: Palatal and B: Buccal side)

Specific anatomical features of the full veneer gold shell crown were measured indirectly on the computer screen by using ImageJ software (SAFMs). In determining total occlusal convergence (TOC) angle, a study of Yoon et al., (2014) used the gingival portion of the tooth preparation as this is largely responsible for its retention and resistance form. Therefore, total occlusal convergence (TOC) angle was measured by meeting two opposing gingival portion axial walls lines of a preparation. All features were measured on two occasions, one week apart by using ImageJ (Figure 6.8).

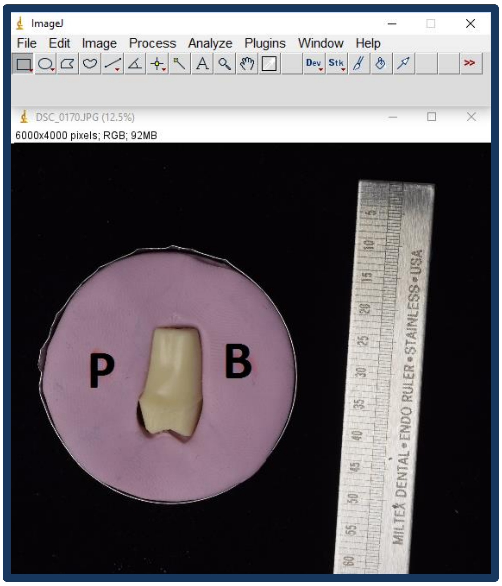


Figure 6.8 .JPG image of a full veneer gold shell crown preparation for an upper first molar analysed using ImageJ software
(P: Palatal and B: Buccal side)

However, there were some specific anatomical features (SAFs) which could not be measured by using specific tools or methods. For example, any undercuts of the full veneer gold shell crown preparation. To simplify matters, a decision was made that the evaluation of these SAFs should be binary. The binary subjective evaluation might include the *presence* or *absence* of damage to the proximal surfaces of adjacent teeth or *clearance* or *non-clearance* of the contact point with adjacent teeth. Such things can be independently verified and used to determine the passing grade features.

Statistical analysis

This part of the study used descriptive statistics to summarize the measurement (mm) of specific anatomical features with the mean and standard deviation (±SD) using one or two different methods on two occasions [see result section 6.4 – Tables (6.4 – 6.15)].

6.3.3 Identification of reliable measurements for class II amalgam cavity and full veneer gold shell crown preparations

Several studies in the literature have described and reported paired sample t-tests and an initial analysis of presented data in order to detect systematic error of two series of measurements. Systematic errors between two series measurements may arise over a period of time if an examiner's measuring method changes with experience. One series of measurements may be changed systematically from a series made at different time. Houston (1983) suggested 25 models as a minimum sample number to detect systematic error of two series of measurement (Houston, 1983). Therefore, all samples of class II amalgam cavities and full veneer gold shell crown preparations were included.

Random error may also arise as a result of changing of two measurement points for each specific anatomical feature of tooth preparation (Houston, 1983). The standard deviation is used to quantify the amount of variation (Random error) of a set of two or more measurements (Bland and Altman, 1996). The reason for this variation is that many landmarks of tooth preparation are difficult to identify, and examiner's opinion about the exact location of the point may vary at random (Taylor et al., 2013). This allows the researcher to ascertain if large differences exist between datasets but may not give information as to what direction differences may take. The main disadvantage of paired sample t-test by comparing two means of two groups is that the means of two set of measurements sometimes can be equal while the (random) differences between measurements can be huge (Chhapola et al., 2015). Therefore, intra-class correlation was used to determine the reliable measurement for each feature.

Subsequent evaluation of the Intra-class Correlation Coefficient (ICC) is also often reported. This evaluation is better suited to determine the direction of any differences between datasets may take. However, a high correlation, by using Intra-class Correlation Coefficient (ICC), does not necessarily imply that there is high agreement between two measurements for one method or for two different measuring methods. In addition, the Intra-class Correlation Coefficient (ICC) fails to provide information on the type of association between the measurements. It may show excellent correlation despite the presence of significant systematic bias. The correlation cannot distinguish between systemic or random differences in two measurements (Van Stralen et al., 2008). Therefore, Bland and Altman plots were constructed to evaluate the agreement and systemic bias between two measurements for the same method or for two different methods. A Bland and Altman plot was also used to compare a new measurement method with a gold standard (Bland and Altman, 1999).

For these data, repeatability of the SAFMs and reliability of the methods, for each dimension in Figures 6.1, 6.4 and 6.5, were assessed by calculating paired sample t-tests and the intra-class correlation using SPSS (Version 22) and a Bland and Altman plot using Medcalc software (version 12.7.0.0). By using row measurement data (see result section 6.4.3), statistical analysis was completed to determine reliability/agreement between two occasions measurements for each method.

Reliability of MeshLab, ImageJ software and measurement points for each SAFM of class II amalgam cavity and full veneer gold shell crown preparation are required. Therefore, determining repeatability between measurements made on two occasions, using MeshLab software, and comparing with measurements from digital calliper (gold standard) was essential in order to identify reliability of MeshLab measuring method and measurement points for each SAFM of the class II amalgam cavity preparation. Although the ImageJ software used was a reliable method (Kerner et al., 2007), measurement points for each SAFM of the full veneer gold shell crown preparation are also required. Therefore, determining repeatability between measurements made on two occasions, using ImageJ software, was essential in order to identify reliability of measurement points for each SAFM of the full veneer gold shell crown preparation.

To identify a reliable measuring method and measurement points for the class II amalgam cavity and measurement points for the full veneer gold shell crown preparation, the following steps were applied:

Step 1: using a paired sample t-test (SPSS) to compare two measurements means from two different occasions for each SAFM of the tooth preparation recorded by each of the measurement methods (see Tables 6.16 to 6.19) and also between measurement methods when different methods were used.

Step 2: summarising the intra-examiner reliability, using intra-class correlation (ICC), to compare two measurements from two different occasions for each SAFM of the tooth preparation recorded by each of the measurement methods (see Tables 6.20 to 6.23) and also between measurement methods when different methods were used.

Step 3: creating a Bland Altman plot for each feature using each method for both occasions and different methods, where appropriate [see Table 6.24, Appendices 5 and 6].

There were some specific anatomical features (SAFs) which could not be measured by using specific tools or methods. A decision was made that the evaluation of these SAFs should be binary. Such features can be independently verified and used to determine the passing grade features (see section 6.3.4).

6.3.4 Comparison of a) the objective class II amalgam cavity or full veneer gold veneer shell crown preparation dimensions recorded by the researcher (AM) and those subjective tooth preparation evaluations by the senior examiners exhibiting the best agreement with b) dimensions presented in the dental literature and subsequent calibration with c) the grades of the best senior examiner

The class II amalgam cavity and full veneer gold shell crown preparations both have features which can be measured objectively (referred to as Specific Anatomical Feature Measurements or SAFM) and evaluated subjectively (referred to as Specific Anatomical Features or SAF). The objective measurements have been determined using reliable methods used by the researcher (AM). The subjective evaluations have been completed by three senior examiners and, where two or three senior examiners agree, then this was accepted as the reliable subjective evaluation.

For the Class II amalgam cavity preparation, there were eight features (depth of the box, box-bucco-palatal width, box floor depth, box-pulpal axial wall length, occlusal isthmus width, occlusal cavity width, occlusal cavity depth and marginal ridge thickness) which could be measured objectively and six features which were evaluated subjectively. For the full veneer gold shell crown there were three features (occlusal convergence angle, occlusal reduction and axial reduction) which could be measured objectively and twelve features which were evaluated subjectively.

There is some data in the dental literature to describe acceptable class II amalgam cavities as well as acceptable full veneer gold shell crown preparations. A narrative search of several systematic reviews and other studies was used to identify the acceptable values of SAFMs for both types of tooth preparation in the literature (see Chapter 1). The acceptable measurements for each feature of a class II amalgam cavity and a full veneer gold shell crown preparation are variable. Therefore, the widest range of the measurements for each feature (SAFM) was determined. The range of SAFMs was called an acceptable SAFM range (see section 6.4.4). According to Ahmed et al., (2016), acceptable SAFM ranges provide more objective assessments for the student's performance and increase examiner reliability (Ahmed et al., 2016). In addition, using SAFMs of tooth preparation will reduce subjectivity of the evaluation (Tiu et al., 2014). In this part of the study, the acceptable SAFM range from the literature was reported for each feature by selecting wide range of measurements, for example, acceptable TOC angles is (3° – 20°).

After reporting acceptable SAFM ranges from the literature, reliable SAFMs for the 26 class II amalgam cavities and 30 full veneer gold shell crown preparation from one or two different methods was also reported. The reliability of SAFMs was determined by using three different statistical analysis tests (see section 6.3.3). How do these data compare with the findings described in the previous paragraphs? This is a difficult comparison and plotting both sets of objective data on a graph is helpful to decide if objective measurements agree, or do not agree, with data from the literature. For subjective evaluations, the comparison is easier. If the examiners agree there is undercut in a class II amalgam cavity and the literature says there should be undercut then there is agreement.

Knight (1997) described a method for calibration for the information described in the previous three paragraphs. Essentially, this was undertaken by converting overall evaluations of tooth preparations (both objective and subjective) to a dichotomous scale of pass (all evaluations meet the standard) or fail (at least one recommendation does not meet the standard). Once this dichotomous scale has been established it can be compared with the grades of the best examiner which have also been dichotomised to pass or fail.

This section will describe this process in a series of steps.

Step 1, Determine i), reliable average measurements (means) for each objective SAFM recorded by the researcher (AM) and ii), the binary evaluations (Yes/No), determined by the three senior examiners, for each subjective specific anatomical feature (SAF) and for both types of tooth preparation, which exhibited the most agreement.

Step 2, Compare i), the reliable mean of objective SAFMs from the most reliable method (using graphical representation) with acceptable measurements which were suggested in the literature. For subjective evaluations, the dental literature stated whether or not a feature should be present and the subjective evaluation of the examiners reported for each tooth preparation whether or not the senior examiners agreed the feature was present.

From **steps 1** and **2**, the most reliable SAFMs have been identified from the statistical analysis of reliable measuring methods using paired sample t-test, intra-class correlation and Bland and Altman plots (see result section 6.4.3). These measurements were then compared with acceptable measurements which were suggested in the literature.

To investigate the relationship of SAFMs, SAFs, and grades awarded from the best senior examiner for class II amalgam cavity and full veneer gold shell crown preparation, the last step was applied:

Step 3: Calibrating the tooth preparations scores dichotomised as pass/fail from Step 2 with dichotomised grades (converted to Pass/Fail scores) awarded by the best senior academic staff examiner.

Knight (1997) suggested that if one of SAFMs or SAF did not meet the standard recommended in the literature, the preparation model was given a fail score. If the model met all the standards described in the literature, this preparation was given a pass score (Knight, 1997). In addition, the grades of the best examiner were also converted into the pass and fail by assigning grades 1,2,3 as fail, and grades 4,5 as pass (Figure 6.9).

Calibration of the best examiner grades was performed by comparing the scores of the SAFMs and SAF with scores of grades awarded from the best senior examiner using percentage agreement and Cohen's Kappa coefficient test (SPSS). This percentage agreement and Cohen's Un-weighted Kappa test indicate that the grades awarded by the best examiner represent what can be most objectively evaluated about the tooth preparation. This is probably about as close to a Developed Standard that can be achieved.

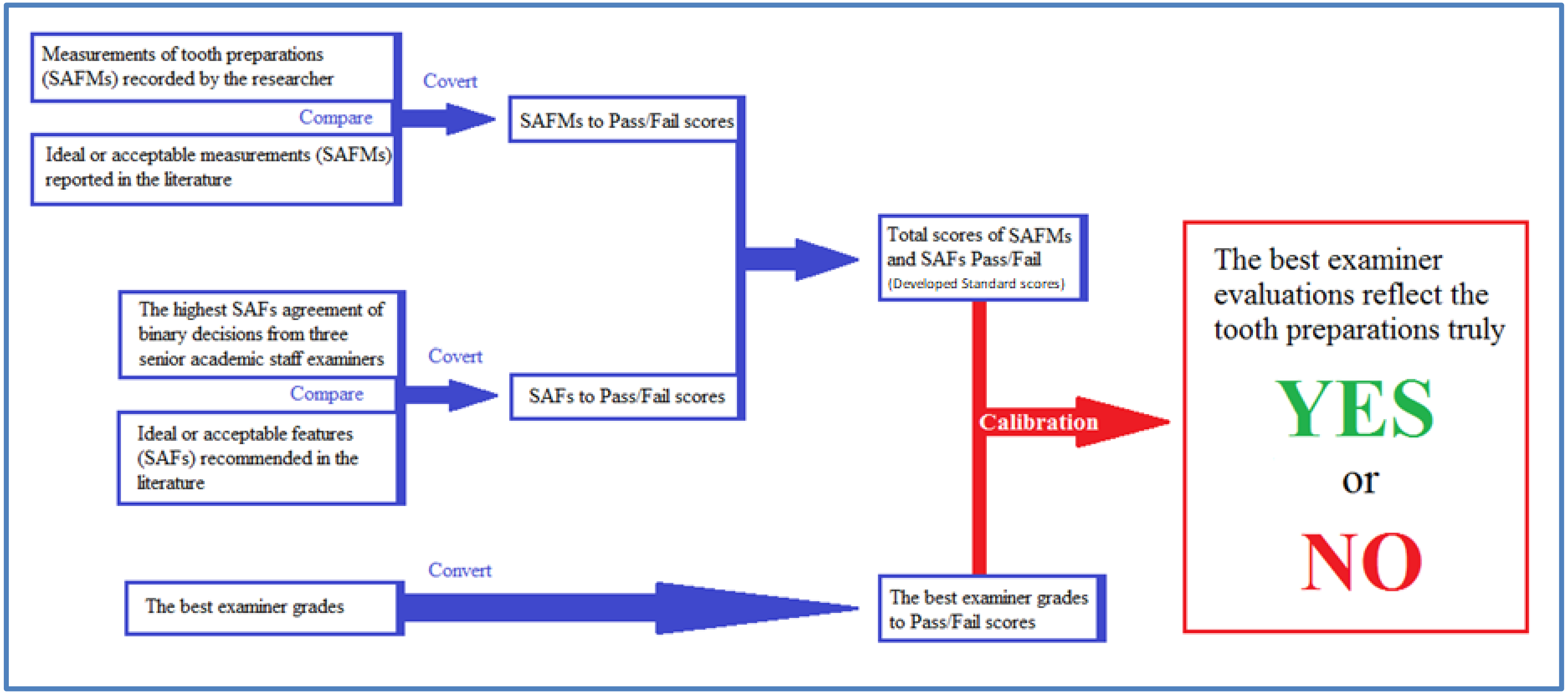


Figure 6.9 Diagram to illustrate the calibration of reliable tooth preparation dimensions recorded by the researcher (AM) and those subjective tooth preparation evaluations by the senior examiners exhibiting the best agreement with the grades of the best senior examiner

6.4 Results

6.4.1 Identification of the best senior examiner and grades for class II amalgam cavity and full veneer gold shell crown preparations

The data from Chapter 5 concluded that examiner 3 for class II amalgam cavity and examiner 1 for full veneer gold shell crown had better intra- and inter-examiner agreement than other examiners by using Cohen's Kappa agreement. In addition, Tables 5.10, 5.12, 5.20 and 5.22 in Chapter 5, identified that the best two examiners who had the highest intra-class correlation (ICC) were examiner 1 and examiner 3 for the class II amalgam cavity preparation and examiner 1 and examiner 2 for the full veneer gold shell crown preparation. For intra-examiner and inter-examiner agreement, examiner 3 and examiner 1 were the best for class II amalgam cavity and full veneer gold shell crown preparation respectively (see Chapter 5). The results of Chapter 5 were organised into a tabular form to enable comparison of performance between senior academic staff (Appendix 3).

To confirm that examiner 3 and examiner 1 were the best examiners for class II amalgam cavity and full veneer gold shell crown preparations, respectively, Spearman correlation analysis (SPSS) of the 'Gray and Mhanni feedback sheet' of 5th assessment stage grades awarded by the three senior examiners versus the number of negative points awarded from the same feedback sheets was calculated. The strongest negative correlation was selected as the best examiner and his scores were selected as the best grades (Tables 6.1, 6.2 and 6.3).

Table 6.1 summarises the correlation between the award 'Gray feedback sheet' negative points (5th stage) versus grades awarded for each examiner on each occasion and overall in order to determine the best examiner and grades for 26 class II amalgam cavities.

Table 6.1 The correlation between 'Gray feedback sheet' negative points (5th stage) versus the grades awarded by each examiner on two occasions for class II amalgam cavities

Examiner 1		
Spearman correlation	*Value*	*Significance ($p \leq 0.05$)*
1st Gray sheet grades and 1st negative points	-0.919	0.000
2nd Gray sheet grades and 2nd negative points	-0.838	0.000
1st and 2nd Gray Sheet grades with 1st and 2nd negative points (All)	-0.868	0.000
Examiner 2		
Spearman correlation	*Value*	*Significance ($p \leq 0.05$)*
1st Gray sheet grades and 1st negative points	-0.802	0.000
2nd Gray sheet grades and 2nd negative points	-0.719	0.000
1st and 2nd Gray Sheet grades with 1st and 2nd negative points (All)	-0.764	0.000
Examiner 3		
Spearman correlation	*Value*	*Significance ($p \leq 0.05$)*
1st Gray sheet grades and 1st negative points	-0.948	0.000
2nd Gray sheet grades and 2nd negative points	-0.933	0.000
1st and 2nd Gray Sheet grades with 1st and 2nd negative points (All)	-0.931	0.000

The highlighted line indicates the highest correlation.

It is clear that all senior examiners had a good Spearman correlation between negative points and grades. Although all senior academic staff examiners had between a moderate and strong level of correlation, the best examiner was examiner 3 in this respect. Occasion one of the 'Gray feedback sheet' stage assessment for examiner 3 had the strongest negative correlation between grades awarded and the number of the negative points.

For the full veneer gold shell crown preparation, Table 6.2 demonstrates the results of the Spearman correlation between the awarded 'Mhanni feedback sheet' negative points versus grades awarded for each senior examiner on each occasion and overall in order to determine the best examiner and grades for 30 full veneer gold shell crown preparations.

Table 6.2 The correlation between 'Mhanni feedback sheet' negative points (5th stage) versus the grades awarded by each examiner on two occasions for full veneer gold shell crown preparation

Examiner 1		
Spearman correlation	*Value*	*Significance ($p \leq 0.05$)*
1st Mhanni sheet grades and 1st negative points	-0.938	0.000
2nd Mhanni sheet grades and 2nd negative points	-0.877	0.000
1st and 2nd Mhanni sheet grades with 1st and 2nd negative points (All)	-0.904	0.000
Examiner 2		
Spearman correlation	*Value*	*Significance ($p \leq 0.05$)*
1st Mhanni sheet grades and 1st negative points	-0.892	0.000
2nd Mhanni sheet grades and 2nd negative points	-0.884	0.000
1st and 2nd Mhanni sheet grades with 1st and 2nd negative points (All)	-0.889	0.000
Examiner 3		
Spearman correlation	*Value*	*Significance ($p \leq 0.05$)*
1st Mhanni sheet grades and 1st negative points	-0.739	0.000
2nd Mhanni sheet grades and 2nd negative points	-0.659	0.000
1st and 2nd Mhanni sheet grades with 1st and 2nd negative points (All)	-0.617	0.000

The highlighted line indicates the highest correlation.

It was also clear that all senior academic staff examiners had also a good negative relationship between negative points and grades. Even-though examiner 1 had excellent negative correlation, occasion one for the same examiner was the best in this respect.

In addition, Table 6.3 supported results of Chapter 5. This table also displayed the correlation between the grades for all senior academic staff for both rounds together with the number of negative points awarded for both rounds together. Furthermore, it determined the correlation between the grades for the best two senior academic staff who had the best inter-examiner reproducibility for both rounds together [i.e. examiner 1 and 3 for class II amalgam cavity and examiner 1 and 2 for the full veneer gold shell crown preparation (see Chapter 5)] with the number of negative points awarded for both rounds together. Lastly, this table demonstrated the correlation between the grades for the best senior academic staff who had the best inter-examiner reproducibility for both rounds together (Table 6.3).

Table 6.3 The correlation between the grades for senior academic staff for occasion one and two with the number of negative points awarded

Spearman correlation between grades and negative points awarded ……	*Class II amalgam cavity*	*Full veneer gold shell crown*
	Value	*Value*
for **all senior academic staff** for occasion one and two	-0.835	-0.813
for **the best two inter-examiner agreement** for occasion one and two	-0.896	-0.896
for **the best single inter-examiner agreement** for occasion one and two	-0.931	-0.904

The highlighted line indicates the highest correlation.

From Tables 6.1, 6.2, 6.3 and Appendix 3, the senior examiner with the best agreement grades could be identified for each tooth preparation. Examiner 3 was the examiner with the best intra- and inter-examiner agreement for *the class II amalgam cavity preparations*, while Examiner 1 was the examiner with the best intra- and inter-examiner agreement for *the full veneer gold shell crown preparations*. It was, therefore, important to find a way of verifying that these examiners were actually the best examiners and their grades truly reflect tooth preparation.

6.4.2 Measuring the specific anatomical features measurements (SAFMs) for each type of tooth preparation

From Figure 6.1, twelve SAFMs of the class II amalgam cavity were measured objectively by using direct and indirect methods (Tables 6.4 to 6.11). Some objective SAFMs were measured in order to determine other SAFs which were mentioned in the 'Gray feedback sheet'. For example, measurement of the bucco-palatal width of the gingival floor and bucco-palatal width of the box occlusally were used to evaluate retention form of the proximal box. In addition to binary subjective evaluation, measuring other specific anatomical features (SAFs) might support subjective evaluation positively. Such things can be confirmed and used to determine the passing grade (acceptable) features.

From Figures 6.4 and 6.5, ten SAFMs for the full veneer gold shell crown preparation were determined objectively by using an indirect method (Tables 6.12 to 6.15). Some other objective specific features were also determined to evaluate other SAFs which were established in the 'Mhanni feedback sheet'. For example, measurements of mesio-distal and bucco-palatal angles were used to evaluate undercuts.

Class II amalgam cavity preparation:

The measurements (mm) of 26 class II amalgam cavity features with mean and standard deviation values using two different methods on two occasions now follow. The measuring methods were either direct using a pair of digital callipers or indirect method 'MeshLab software' analysis of 3D scans.

Tables 6.4, 6.5, 6.6, and 6.7 summarise the measurements (mm) made, using the digital calliper, of the twelve SAFM for each of the 26 class II amalgam cavities on two separate occasions.

Table 6.4 Mean and standard deviation measurements (mm), for class II amalgam cavity preparations for i) box depth gingivally, ii) bucco-palatal width gingivally and iii) bucco-palatal width occlusally using a digital calliper

Models	**Depth of box gingivally (mm)**			**Bucco-palatal width of the box floor (mm)**			**Bucco-Palatal width occlusally (mm)**		
	First occasion	Second occasion	**Mean ± standard deviation**	First occasion	Second occasion	**Mean ± standard deviation**	First occasion	Second occasion	**Mean ± standard deviation**
5	3.20	3.10	**3.15** ± 0.07	2.20	2.28	**2.24** ± 0.06	2.90	2.85	**2.73** ± 0.18
8	2.15	2.10	**2.13** ± 0.04	2.30	2.26	**2.28** ± 0.03	2.60	2.62	**2.26** ± 0.20
15	3.29	3.27	**3.28** ± 0.01	2.38	2.38	**2.38** ± 0.00	2.75	2.71	**2.48** ± 0.04
16	2.97	3.05	**3.01** ± 0.06	2.57	2.65	**2.61** ± 0.06	2.05	2.10	**2.08** ± 0.04
36	2.90	2.85	**2.88** ± 0.04	3.18	3.25	**3.22** ± 0.05	2.85	2.80	**2.83** ± 0.04
39	3.20	3.10	**3.15** ± 0.07	2.70	2.75	**2.73** ± 0.04	2.64	2.54	**2.59** ± 0.07
40	2.30	2.45	**2.38** ± 0.11	3.65	3.55	**3.60** ± 0.07	3.30	3.25	**3.28** ± 0.04
41	2.60	2.50	**2.55** ± 0.07	3.70	3.72	**3.71** ± 0.01	3.80	3.70	**3.75** ± 0.07
43	2.35	2.20	**2.28** ± 0.11	2.60	2.75	**2.68** ± 0.11	2.75	2.85	**2.80** ± 0.07
46	3.23	3.30	**3.27** ± 0.05	2.95	2.90	**2.93** ± 0.04	2.95	3.10	**3.03** ± 0.11
53	2.75	2.70	**2.73** ± 0.04	2.75	2.70	**2.73** ± 0.04	3.10	3.05	**3.08** ± 0.04
54	2.60	2.70	**2.65** ± 0.07	2.20	2.25	**2.23** ± 0.04	2.55	2.52	**2.54** ± 0.02
57	3.20	3.10	**3.15** ± 0.07	3.55	3.55	**3.55** ± 0.00	3.07	3.16	**3.12** ± 0.06
62	2.30	2.32	**2.31** ± 0.01	2.51	2.55	**2.53** ± 0.03	2.81	2.95	**2.88** ± 0.10
73	4.75	4.71	**4.73** ± 0.03	1.96	2.05	**2.01** ± 0.06	1.95	2.10	**2.03** ± 0.11
78	2.90	2.86	**2.88** ± 0.03	2.65	2.66	**2.66** ± 0.01	2.55	2.46	**2.51** ± 0.06
80	3.23	3.13	**3.18** ± 0.07	2.15	2.08	**2.12** ± 0.05	2.28	2.37	**2.33** ± 0.06
83	2.25	2.20	**2.23** ± 0.04	2.40	2.50	**2.45** ± 0.07	2.53	2.60	**2.57** ± 0.05
85	2.40	2.43	**2.42** ± 0.02	3.75	3.75	**3.75** ± 0.00	3.75	3.70	**3.73** ± 0.04
87	2.72	2.76	**2.74** ± 0.03	3.15	3.10	**3.13** ± 0.04	3.20	3.25	**3.23** ± 0.04
88	3.10	3.11	**3.11** ± 0.01	4.62	4.78	**4.70** ± 0.11	4.65	4.60	**4.63** ± 0.04
94	2.75	2.65	**2.70** ± 0.07	3.00	2.99	**3.00** ± 0.01	2.90	3.00	**2.95** ± 0.07
109	2.95	3.05	**3.00** ± 0.07	2.33	2.40	**2.37** ± 0.05	2.65	2.55	**2.60** ± 0.07
111	1.55	1.45	**1.50** ± 0.07	4.75	4.50	**4.63** ± 0.18	3.95	4.05	**4.00** ± 0.07
120	2.65	2.67	**2.66** ± 0.01	2.80	2.80	**2.80** ± 0.00	2.65	2.80	**2.73** ± 0.11
138	3.75	3.85	**3.80** ± 0.07	3.60	3.40	**3.50** ± 0.14	4.05	3.50	**3.78** ± 0.39

Table 6.5 Mean and standard deviation measurements (mm), for class II amalgam cavity preparations for i) box depth (mesio-distal), ii) pulpal axial wall length and iii) isthmus width occlusally using a digital calliper

Models	Box floor (mesio-distal) depth (mm)			Pulpal axial wall length (mm)			Isthmus width occlusally (mm)		
	First occasion	Second occasion	**Mean ± standard deviation**	First occasion	Second occasion	**Mean ± standard deviation**	First occasion	Second occasion	**Mean ± standard deviation**
5	1.95	1.80	**1.88** ± 0.11	1.75	1.60	**1.68** ± 0.11	1.30	1.27	**1.29** ± 0.02
8	1.40	1.32	**1.36** ± 0.06	1.30	1.35	**1.33** ± 0.04	1.25	1.19	**1.22** ± 0.04
15	1.45	1.35	**1.40** ± 0.07	2.05	2.00	**2.03** ± 0.04	1.20	1.24	**1.22** ± 0.03
16	1.60	1.65	**1.63** ± 0.04	2.06	2.05	**2.06** ± 0.01	1.15	1.29	**1.22** ± 0.10
36	1.45	1.31	**1.38** ± 0.10	1.34	1.38	**1.36** ± 0.03	1.30	1.38	**1.34** ± 0.06
39	1.40	1.32	**1.36** ± 0.06	2.06	2.00	**2.03** ± 0.04	1.32	1.31	**1.32** ± 0.01
40	1.25	1.22	**1.24** ± 0.02	1.10	1.12	**1.11** ± 0.01	1.38	1.55	**1.47** ± 0.12
41	1.35	1.40	**1.38** ± 0.04	1.25	1.30	**1.28** ± 0.04	2.10	2.15	**2.13** ± 0.04
43	1.54	1.52	**1.53** ± 0.01	1.35	1.40	**1.38** ± 0.04	1.35	1.40	**1.38** ± 0.04
46	1.35	1.35	**1.35** ± 0.00	1.35	1.40	**1.38** ± 0.04	1.40	1.35	**1.38** ± 0.04
53	1.14	1.16	**1.15** ± 0.01	1.20	1.25	**1.23** ± 0.04	1.48	1.40	**1.44** ± 0.06
54	0.89	0.90	**0.90** ± 0.01	1.20	1.25	**1.23** ± 0.04	1.38	1.38	**1.38** ± 0.00
57	0.90	0.82	**0.86** ± 0.06	1.05	1.10	**1.08** ± 0.04	1.11	1.11	**1.11** ± 0.00
62	1.41	1.33	**1.37** ± 0.06	1.30	1.25	**1.28** ± 0.04	1.65	1.74	**1.70** ± 0.06
73	1.60	1.50	**1.55** ± 0.07	3.75	3.90	**3.83** ± 0.11	1.38	1.46	**1.42** ± 0.06
78	1.45	1.45	**1.45** ± 0.00	1.60	1.44	**1.52** ± 0.11	1.24	1.11	**1.18** ± 0.09
80	1.20	1.21	**1.21** ± 0.01	1.70	1.80	**1.75** ± 0.07	1.78	1.75	**1.77** ± 0.02
83	1.30	1.35	**1.33** ± 0.04	1.06	1.13	**1.10** ± 0.05	1.22	1.26	**1.24** ± 0.03
85	1.22	1.35	**1.29** ± 0.09	1.30	1.26	**1.28** ± 0.03	1.55	1.45	**1.50** ± 0.07
87	1.20	1.33	**1.27** ± 0.09	1.51	1.50	**1.51** ± 0.01	1.48	1.58	**1.53** ± 0.07
88	1.45	1.31	**1.38** ± 0.10	1.94	2.10	**2.02** ± 0.11	1.48	1.35	**1.42** ± 0.09
94	0.90	0.92	**0.91** ± 0.01	0.91	0.75	**0.83** ± 0.11	3.05	3.00	**3.03** ± 0.04
109	1.30	1.40	**1.35** ± 0.07	1.75	1.70	**1.73** ± 0.04	1.20	1.20	**1.20** ± 0.00
111	2.00	1.94	**1.97** ± 0.04	0.90	0.70	**0.80** ± 0.14	1.40	1.60	**1.50** ± 0.14
120	1.45	1.35	**1.40** ± 0.07	1.05	0.91	**0.95** ± 0.10	1.40	1.45	**1.43** ± 0.04
138	1.66	1.75	**1.71** ± 0.06	2.00	2.10	**2.05** ± 0.07	1.60	1.45	**1.53** ± 0.11

Table 6.6 Mean and standard deviation measurements (mm), for class II amalgam cavity preparations for i) isthmus floor width, ii) occlusal cavity width in the middle and iii) marginal ridge thickness using a digital calliper

Models	Isthmus floor width			Occlusal cavity width (in the middle)			Marginal ridge thickness		
	First occasion	Second occasion	**Mean ± standard deviation**	First occasion	Second occasion	**Mean ± standard deviation**	First occasion	Second occasion	**Mean ± standard deviation**
5	1.22	1.30	**1.26** ± 0.06	1.17	1.25	**1.21** ± 0.06	1.40	1.44	**1.42** ± 0.03
8	1.22	1.20	**1.21** ± 0.01	1.27	1.24	**1.26** ± 0.02	1.40	1.60	**1.50** ± 0.14
15	1.10	1.15	**1.13** ± 0.04	1.17	1.24	**1.21** ± 0.05	0.95	0.95	**0.95** ± 0.00
16	1.25	1.20	**1.23** ± 0.04	1.30	1.27	**1.29** ± 0.02	1.20	1.09	**1.15** ± 0.08
36	1.28	1.33	**1.31** ± 0.04	1.13	1.10	**1.12** ± 0.02	1.05	1.17	**1.11** ± 0.08
39	1.27	1.32	**1.30** ± 0.04	1.28	1.31	**1.30** ± 0.02	1.50	1.70	**1.60** ± 0.14
40	1.36	1.47	**1.42** ± 0.08	1.30	1.31	**1.31** ± 0.01	1.40	1.40	**1.40** ± 0.00
41	2.00	1.98	**1.99** ± 0.01	1.25	1.51	**1.38** ± 0.18	1.40	1.50	**1.45** ± 0.07
43	1.25	1.23	**1.24** ± 0.01	1.10	1.13	**1.12** ± 0.02	1.21	1.10	**1.16** ± 0.08
46	1.21	1.21	**1.21** ± 0.00	1.14	1.15	**1.15** ± 0.01	1.40	1.45	**1.43** ± 0.04
53	1.15	1.13	**1.14** ± 0.01	1.10	1.20	**1.15** ± 0.07	1.51	1.54	**1.53** ± 0.02
54	1.22	1.21	**1.22** ± 0.01	1.45	1.32	**1.39** ± 0.09	1.91	1.95	**1.93** ± 0.03
57	1.10	1.05	**1.08** ± 0.04	1.00	0.99	**1.00** ± 0.01	1.55	1.58	**1.57** ± 0.02
62	1.30	1.35	**1.33** ± 0.04	1.05	1.28	**1.17** ± 0.16	1.29	1.47	**1.38** ± 0.13
73	1.30	1.25	**1.28** ± 0.04	1.25	1.13	**1.19** ± 0.08	1.80	2.00	**1.90** ± 0.14
78	1.10	1.00	**1.05** ± 0.07	1.05	1.08	**1.07** ± 0.02	0.85	1.15	**1.00** ± 0.21
80	1.50	1.41	**1.46** ± 0.06	1.15	1.10	**1.13** ± 0.04	1.80	1.99	**1.90** ± 0.13
83	1.15	1.18	**1.17** ± 0.02	1.00	1.00	**1.00** ± 0.00	1.60	1.50	**1.55** ± 0.07
85	1.30	1.40	**1.35** ± 0.07	1.20	1.24	**1.22** ± 0.03	1.53	1.55	**1.54** ± 0.01
87	1.75	1.98	**1.87** ± 0.16	1.30	1.20	**1.25** ± 0.07	1.45	0.87	**1.16** ± 0.41
88	1.40	1.35	**1.38** ± 0.04	1.10	1.10	**1.10** ± 0.00	1.35	1.44	**1.40** ± 0.06
94	1.36	1.30	**1.33** ± 0.04	1.20	1.15	**1.18** ± 0.04	1.75	1.90	**1.83** ± 0.11
109	1.21	1.35	**1.28** ± 0.10	1.15	1.00	**1.08** ± 0.11	1.35	1.20	**1.28** ± 0.11
111	1.30	1.25	**1.28** ± 0.04	1.21	1.10	**1.16** ± 0.08	1.45	1.20	**1.33** ± 0.18
120	1.24	1.20	**1.22** ± 0.03	1.35	1.20	**1.28** ± 0.11	1.50	1.30	**1.40** ± 0.14
138	1.21	1.17	**1.19** ± 0.03	1.15	1.10	**1.13** ± 0.04	0.55	0.60	**0.58** ± 0.04

Table 6.7 Mean and standard deviation measurements (mm), for class II amalgam cavity preparations for i) occlusal cavity depth at palatal, ii) buccal in the middle and iii) distal sides using a digital calliper

Models	Occlusal cavity depth (palatal side in the middle)			Occlusal cavity depth (buccal side in the middle)			Occlusal cavity depth (at distal side)		
	First occasion	Second occasion	**Mean ± standard deviation**	First occasion	Second occasion	**Mean ± standard deviation**	First occasion	Second occasion	**Mean ± standard deviation**
5	1.40	2.05	**1.73** ± 0.46	1.60	1.80	**1.70** ± 0.14	1.30	1.13	**1.22** ± 0.12
8	1.00	0.86	**0.93** ± 0.10	1.00	1.10	**1.05** ± 0.07	0.90	0.88	**0.89** ± 0.01
15	1.35	1.54	**1.45** ± 0.13	1.55	1.58	**1.57** ± 0.02	1.35	1.35	**1.35** ± 0.00
16	1.70	1.59	**1.65** ± 0.08	2.25	2.25	**2.25** ± 0.00	1.15	1.26	**1.21** ± 0.08
36	2.10	2.05	**2.08** ± 0.04	2.15	2.14	**2.15** ± 0.01	1.25	1.30	**1.28** ± 0.04
39	1.60	1.85	**1.73** ± 0.18	1.50	1.80	**1.65** ± 0.21	1.80	1.55	**1.68** ± 0.18
40	1.70	1.77	**1.74** ± 0.05	1.55	1.40	**1.48** ± 0.11	1.20	1.20	**1.20** ± 0.00
41	1.60	1.42	**1.51** ± 0.13	1.40	1.50	**1.45** ± 0.07	1.10	1.20	**1.15** ± 0.07
43	1.20	1.40	**1.30** ± 0.14	1.35	1.43	**1.39** ± 0.06	1.40	1.35	**1.38** ± 0.04
46	2.60	2.75	**2.68** ± 0.11	2.27	2.30	**2.29** ± 0.02	2.13	2.10	**2.12** ± 0.02
53	1.05	1.10	**1.08** ± 0.04	1.35	1.30	**1.33** ± 0.04	1.15	1.00	**1.08** ± 0.11
54	2.30	2.44	**2.37** ± 0.10	1.85	1.90	**1.88** ± 0.04	1.65	1.45	**1.55** ± 0.14
57	2.10	2.00	**2.05** ± 0.07	2.10	2.15	**2.13** ± 0.04	1.70	1.65	**1.68** ± 0.04
62	1.43	1.45	**1.44** ± 0.01	1.40	1.35	**1.38** ± 0.04	0.60	0.85	**0.73** ± 0.18
73	1.20	1.17	**1.19** ± 0.02	1.45	1.88	**1.67** ± 0.30	1.00	0.84	**0.92** ± 0.11
78	1.15	1.38	**1.27** ± 0.16	1.70	1.45	**1.58** ± 0.18	0.55	0.65	**0.60** ± 0.07
80	1.46	1.45	**1.46** ± 0.01	1.45	1.48	**1.47** ± 0.02	0.80	0.90	**0.85** ± 0.07
83	1.90	1.95	**1.93** ± 0.04	1.70	1.75	**1.73** ± 0.04	1.10	1.15	**1.13** ± 0.04
85	1.50	1.51	**1.51** ± 0.01	1.35	1.35	**1.35** ± 0.00	0.75	0.85	**0.80** ± 0.07
87	1.75	1.74	**1.75** ± 0.01	1.50	1.55	**1.53** ± 0.04	1.06	1.05	**1.06** ± 0.01
88	1.40	1.20	**1.30** ± 0.14	1.25	1.32	**1.29** ± 0.05	0.95	1.10	**1.03** ± 0.11
94	2.45	2.50	**2.48** ± 0.04	2.20	2.21	**2.21** ± 0.01	1.60	1.60	**1.60** ± 0.00
109	1.50	1.60	**1.55** ± 0.07	1.10	1.25	**1.18** ± 0.11	0.90	1.00	**0.95** ± 0.07
111	1.10	1.80	**1.45** ± 0.49	1.30	1.15	**1.23** ± 0.11	1.20	0.60	**0.90** ± 0.42
120	1.90	2.00	**1.95** ± 0.07	1.63	2.10	**1.87** ± 0.33	1.30	1.25	**1.28** ± 0.04
138	2.35	2.40	**2.38** ± 0.04	1.90	2.00	**1.95** ± 0.07	1.10	1.05	**1.08** ± 0.04

Tables 6.8, 6.9, 6.10 and 6.11 summarise the measurements (mm) made, using the MeshLab software, of the 12 SAFM for each of the 26 class II amalgam cavities on two occasions.

Table 6.8 Mean and standard deviation measurements (mm), for class II amalgam cavity preparations for i) box depth, ii) bucco-palatal width gingivally and iii) bucco-palatal width occlusally using MeshLab software

Models	Depth of box gingivally (mm)			Bucco-palatal width of the box floor (mm)			Bucco-Palatal width occlusally (mm)		
	First occasion	Second occasion	**Mean ± standard deviation**	First occasion	Second occasion	**Mean ± standard deviation**	First occasion	Second occasion	**Mean ± standard deviation**
5	3.10	3.17	**3.14** ± 0.05	2.44	2.46	**2.45** ± 0.01	3.01	3.00	**3.01** ± 0.01
8	2.20	2.12	**2.16** ± 0.06	2.40	2.46	**2.43** ± 0.04	2.66	2.70	**2.68** ± 0.03
15	3.35	3.34	**3.35** ± 0.01	2.35	2.40	**2.38** ± 0.04	2.74	2.72	**2.73** ± 0.01
16	2.90	3.05	**2.98** ± 0.11	2.83	2.76	**2.80** ± 0.05	2.62	2.69	**2.66** ± 0.05
36	2.85	2.78	**2.82** ± 0.05	3.16	3.10	**3.13** ± 0.04	3.01	2.94	**2.98** ± 0.05
39	3.18	3.32	**3.25** ± 0.10	2.69	2.65	**2.67** ± 0.04	2.59	2.58	**2.59** ± 0.01
40	2.41	2.45	**2.43** ± 0.03	3.46	3.50	**3.48** ± 0.04	3.24	3.21	**3.23** ± 0.02
41	2.65	2.55	**2.60** ± 0.07	3.43	3.45	**3.44** ± 0.01	3.97	4.00	**3.99** ± 0.02
43	2.33	2.46	**2.40** ± 0.09	2.74	2.70	**2.72** ± 0.03	2.97	3.00	**2.99** ± 0.02
46	3.10	3.10	**3.10** ± 0.00	2.75	2.70	**2.73** ± 0.04	3.12	3.18	**3.15** ± 0.04
53	2.72	2.73	**2.73** ± 0.01	2.65	2.70	**2.68** ± 0.04	3.38	3.40	**3.39** ± 0.01
54	2.67	2.73	**2.70** ± 0.04	2.25	2.28	**2.27** ± 0.02	2.63	2.65	**2.64** ± 0.01
57	3.25	3.24	**3.25** ± 0.01	3.60	3.57	**2.59** ± 0.02	3.10	3.15	**3.13** ± 0.04
62	2.25	2.25	**2.25** ± 0.00	2.97	2.95	**2.96** ± 0.01	3.19	3.17	**3.18** ± 0.01
73	4.55	4.50	**4.53** ± 0.04	2.05	2.07	**2.06** ± 0.01	2.29	2.35	**2.32** ± 0.04
78	2.80	2.90	**2.85** ± 0.07	2.58	2.52	**2.55** ± 0.04	2.50	2.50	**2.50** ± 0.00
80	3.17	3.11	**3.14** ± 0.04	2.20	2.24	**2.22** ± 0.03	2.47	2.56	**2.52** ± 0.06
83	2.19	2.10	**2.15** ± 0.06	2.90	2.80	**2.85** ± 0.07	3.00	2.97	**2.99** ± 0.02
85	2.50	2.45	**2.48** ± 0.04	3.65	3.70	**3.68** ± 0.04	3.74	3.70	**3.72** ± 0.03
87	2.80	2.86	**2.83** ± 0.04	3.12	3.16	**3.14** ± 0.03	3.58	3.60	**3.59** ± 0.01
88	3.30	3.30	**3.30** ± 0.00	4.63	4.73	**4.68** ± 0.07	4.85	4.90	**4.88** ± 0.04
94	2.65	2.66	**2.66** ± 0.01	3.10	2.99	**3.05** ± 0.08	3.00	3.00	**3.00** ± 0.00
109	2.75	2.70	**2.73** ± 0.04	2.68	2.70	**2.69** ± 0.01	3.10	3.05	**3.08** ± 0.04
111	1.51	1.60	**1.56** ± 0.06	4.88	4.79	**4.84** ± 0.06	4.11	4.18	**4.15** ± 0.05
120	2.47	2.50	**2.49** ± 0.02	2.70	2.60	**2.65** ± 0.07	2.93	2.84	**2.89** ± 0.06
138	3.70	3.74	**3.72** ± 0.03	3.66	3.70	**3.68** ± 0.03	4.12	4.05	**4.09** ± 0.05

Table 6.9 Mean and standard deviation measurements (mm), for class II amalgam cavity preparations for i) box depth (mesio-distal), ii) pulpal axial wall length and iii) isthmus width occlusally using MeshLab software

Models	Box floor (mesio-distal) depth (mm)				Pulpal axial wall length (mm)				Isthmus width occlusally (mm)			
	First occasion	Second occasion	**Mean ± standard deviation**		First occasion	Second occasion	**Mean ± standard deviation**		First occasion	Second occasion	**Mean ± standard deviation**	
5	1.85	1.80	**1.83**	± 0.04	1.78	1.60	**1.69**	± 0.13	1.33	1.31	**1.32**	± 0.01
8	1.38	1.45	**1.42**	± 0.05	1.40	1.42	**1.41**	± 0.01	1.38	1.35	**1.37**	± 0.02
15	1.61	1.69	**1.65**	± 0.06	2.03	2.00	**2.02**	± 0.02	1.34	1.36	**1.35**	± 0.01
16	1.62	1.54	**1.58**	± 0.06	2.04	2.00	**2.02**	± 0.03	1.23	1.35	**1.29**	± 0.08
36	1.25	1.30	**1.28**	± 0.04	1.24	1.30	**1.27**	± 0.04	1.36	1.36	**1.36**	± 0.00
39	1.10	1.05	**1.08**	± 0.04	2.10	2.05	**2.08**	± 0.04	1.31	1.38	**1.35**	± 0.05
40	1.12	1.15	**1.14**	± 0.02	1.10	1.05	**1.08**	± 0.04	1.60	1.65	**1.63**	± 0.04
41	1.50	1.54	**1.52**	± 0.03	1.23	1.15	**1.19**	± 0.06	2.20	2.35	**2.28**	± 0.11
43	1.46	1.41	**1.44**	± 0.04	1.30	1.39	**1.35**	± 0.06	1.39	1.40	**1.40**	± 0.01
46	1.35	1.30	**1.33**	± 0.04	1.31	1.21	**1.26**	± 0.07	1.40	1.50	**1.45**	± 0.07
53	1.01	1.08	**1.05**	± 0.05	1.13	1.12	**1.13**	± 0.01	1.40	1.50	**1.45**	± 0.07
54	0.91	0.94	**0.93**	± 0.02	1.15	1.21	**1.18**	± 0.04	1.50	1.43	**1.47**	± 0.05
57	0.85	0.95	**0.90**	± 0.07	1.10	1.13	**1.12**	± 0.02	1.18	1.20	**1.19**	± 0.01
62	1.18	1.10	**1.14**	± 0.06	1.20	1.12	**1.16**	± 0.06	1.70	1.70	**1.70**	± 0.00
73	1.75	1.70	**1.73**	± 0.04	3.75	3.80	**3.78**	± 0.04	1.60	1.70	**1.65**	± 0.07
78	1.33	1.35	**1.34**	± 0.01	1.62	1.70	**1.66**	± 0.06	1.21	1.29	**1.25**	± 0.06
80	1.19	1.16	**1.18**	± 0.02	1.80	1.79	**1.80**	± 0.01	1.80	1.84	**1.82**	± 0.03
83	1.27	1.30	**1.29**	± 0.02	1.05	1.03	**1.04**	± 0.01	1.34	1.38	**1.36**	± 0.03
85	1.25	1.17	**1.21**	± 0.06	1.15	1.11	**1.13**	± 0.03	1.50	1.44	**1.47**	± 0.04
87	1.33	1.25	**1.29**	± 0.06	1.32	1.31	**1.32**	± 0.01	1.99	1.90	**1.95**	± 0.06
88	1.43	1.41	**1.42**	± 0.01	2.00	1.99	**2.00**	± 0.01	1.61	1.57	**1.59**	± 0.09
94	0.90	0.93	**0.92**	± 0.02	0.85	0.86	**0.86**	± 0.01	3.04	3.10	**3.07**	± 0.04
109	1.36	1.25	**1.31**	± 0.08	1.74	1.70	**1.72**	± 0.03	1.39	1.50	**1.45**	± 0.08
111	2.00	1.91	**1.96**	± 0.06	0.86	0.96	**0.91**	± 0.07	1.60	1.51	**1.56**	± 0.06
120	1.30	1.28	**1.29**	± 0.01	0.95	1.05	**1.00**	± 0.07	1.60	1.61	**1.61**	± 0.01
138	1.75	1.77	**1.76**	± 0.01	1.95	1.90	**1.93**	± 0.04	2.10	2.00	**2.05**	± 0.07

Table 6.10 Mean and standard deviation measurements (mm), for class II amalgam cavity preparations for i) isthmus floor width, ii) occlusal cavity width in the middle and iii) marginal ridge thickness using MeshLab software

Models	Isthmus floor width			Occlusal cavity width (in the middle)			Marginal ridge thickness		
	First occasion	Second occasion	**Mean ± standard deviation**	First occasion	Second occasion	**Mean ± standard deviation**	First occasion	Second occasion	**Mean ± standard deviation**
5	1.27	1.23	**1.25** ± 0.03	1.28	1.12	**1.20** ± 0.11	1.62	1.70	**1.66** ± 0.06
8	1.18	1.25	**1.22** ± 0.05	1.12	1.15	**1.14** ± 0.02	1.82	1.65	**1.74** ± 0.12
15	1.00	1.05	**1.03** ± 0.04	1.10	1.00	**1.05** ± 0.07	1.94	1.69	**1.82** ± 0.18
16	1.20	1.29	**1.25** ± 0.06	1.12	1.10	**1.11** ± 0.01	1.28	1.41	**1.35** ± 0.09
36	1.32	1.40	**1.36** ± 0.06	1.21	1.20	**1.21** ± 0.01	1.24	1.20	**1.22** ± 0.03
39	1.22	1.20	**1.21** ± 0.01	1.10	1.13	**1.12** ± 0.02	2.42	2.28	**2.35** ± 0.10
40	1.48	1.40	**1.44** ± 0.06	1.36	1.31	**1.34** ± 0.04	1.34	1.32	**1.33** ± 0.01
41	1.85	2.00	**1.93** ± 0.11	1.31	1.26	**1.29** ± 0.04	1.58	1.51	**1.55** ± 0.05
43	1.25	1.30	**1.28** ± 0.04	1.10	1.04	**1.07** ± 0.04	1.21	1.29	**1.25** ± 0.06
46	1.20	1.18	**1.19** ± 0.01	0.90	1.10	**1.00** ± 0.14	1.38	1.35	**1.37** ± 0.02
53	1.15	1.23	**1.19** ± 0.06	0.95	0.95	**0.95** ± 0.00	1.58	1.54	**1.56** ± 0.03
54	1.16	1.11	**1.14** ± 0.04	1.22	1.29	**1.26** ± 0.05	2.04	1.53	**1.79** ± 0.36
57	1.05	0.95	**1.00** ± 0.07	0.97	0.95	**0.96** ± 0.01	1.53	1.45	**1.49** ± 0.06
62	1.47	1.40	**1.44** ± 0.05	1.28	1.30	**1.29** ± 0.01	1.47	1.55	**1.51** ± 0.06
73	1.24	1.16	**1.20** ± 0.06	1.16	1.15	**1.16** ± 0.01	1.75	1.85	**1.80** ± 0.07
78	0.85	0.95	**0.90** ± 0.07	0.91	0.88	**0.90** ± 0.02	1.14	1.07	**1.11** ± 0.05
80	1.50	1.45	**1.48** ± 0.04	1.03	1.00	**1.02** ± 0.02	2.06	2.09	**2.08** ± 0.02
83	1.10	1.13	**1.12** ± 0.02	0.92	1.00	**0.96** ± 0.06	1.88	1.64	**1.76** ± 0.17
85	1.42	1.47	**1.45** ± 0.04	1.00	1.01	**1.01** ± 0.01	1.48	1.46	**1.47** ± 0.01
87	1.90	1.87	**1.89** ± 0.02	1.20	1.32	**1.26** ± 0.08	1.07	1.09	**1.08** ± 0.01
88	1.35	1.44	**1.40** ± 0.06	0.90	1.10	**1.00** ± 0.14	1.54	1.39	**1.47** ± 0.11
94	1.32	1.30	**1.31** ± 0.01	1.28	1.20	**1.24** ± 0.06	1.80	1.77	**1.79** ± 0.02
109	1.33	1.30	**1.32** ± 0.02	0.91	0.95	**0.93** ± 0.03	1.54	1.50	**1.52** ± 0.03
111	1.31	1.22	**1.27** ± 0.06	1.35	1.16	**1.26** ± 0.13	1.42	1.51	**1.47** ± 0.06
120	1.18	1.10	**1.14** ± 0.06	1.20	1.18	**1.19** ± 0.01	1.55	1.66	**1.61** ± 0.08
138	1.15	1.13	**1.14** ± 0.01	1.10	1.06	**1.08** ± 0.03	0.61	0.60	**0.61** ± 0.01

Table 6.11 Mean and standard deviation measurements (mm), for class II amalgam cavity preparations for i) occlusal cavity depth at palatal, ii) buccal in the middle and iii) distal sides using MeshLab software

Models	**Occlusal cavity depth (palatal side in the middle)**			**Occlusal cavity depth (buccal side in the middle)**			**Occlusal cavity depth (at distal side)**		
	First occasion	Second occasion	**Mean ± Standard Deviation**	First occasion	Second occasion	**Mean ± Standard Deviation**	First occasion	Second occasion	**Mean ± Standard Deviation**
5	1.70	1.80	**1.75** ± 0.07	1.75	1.65	**1.70** ± 0.07	0.98	0.80	**0.89** ± 0.13
8	1.10	1.15	**1.13** ± 0.04	0.97	1.05	**1.01** ± 0.06	0.59	0.60	**0.60** ± 0.01
15	1.21	1.50	**1.36** ± 0.21	1.58	1.50	**1.54** ± 0.06	1.43	1.48	**1.46** ± 0.04
16	1.50	1.47	**1.49** ± 0.02	2.30	2.10	**2.20** ± 0.14	0.85	1.25	**1.05** ± 0.28
36	1.92	2.00	**1.96** ± 0.06	1.90	1.85	**1.88** ± 0.04	1.44	1.15	**1.30** ± 0.21
39	1.64	1.70	**1.67** ± 0.04	1.65	1.63	**1.64** ± 0.01	1.55	1.40	**1.48** ± 0.11
40	1.90	1.75	**1.83** ± 0.11	1.60	1.40	**1.50** ± 0.14	1.20	1.15	**1.18** ± 0.04
41	1.64	1.60	**1.62** ± 0.03	1.36	1.36	**1.36** ± 0.00	0.95	0.88	**0.92** ± 0.05
43	1.40	1.35	**1.38** ± 0.04	1.44	1.45	**1.45** ± 0.01	1.37	1.31	**1.34** ± 0.04
46	2.62	2.60	**2.61** ± 0.01	2.10	2.15	**2.13** ± 0.04	1.90	1.80	**1.85** ± 0.07
53	1.32	1.24	**1.28** ± 0.06	1.54	1.32	**1.43** ± 0.16	0.86	0.85	**0.86** ± 0.01
54	2.40	2.50	**2.45** ± 0.07	1.85	1.75	**1.80** ± 0.07	1.51	1.40	**1.46** ± 0.08
57	2.10	2.10	**2.10** ± 0.00	2.15	2.11	**2.13** ± 0.03	1.72	1.66	**1.69** ± 0.04
62	1.49	1.42	**1.46** ± 0.05	1.35	1.38	**1.37** ± 0.02	0.88	0.95	**0.92** ± 0.05
73	1.10	1.20	**1.15** ± 0.07	2.10	2.10	**2.10** ± 0.00	0.75	0.93	**0.84** ± 0.13
78	1.43	1.22	**1.33** ± 0.15	1.63	1.50	**1.57** ± 0.09	0.50	0.49	**0.50** ± 0.01
80	1.45	1.38	**1.42** ± 0.05	1.45	1.47	**1.46** ± 0.01	0.65	0.62	**0.64** ± 0.02
83	1.97	2.05	**2.01** ± 0.06	1.80	1.76	**1.78** ± 0.03	1.14	1.13	**1.14** ± 0.01
85	1.53	1.64	**1.59** ± 0.08	1.60	1.50	**1.55** ± 0.07	0.95	1.05	**1.00** ± 0.07
87	1.94	1.91	**1.93** ± 0.02	1.84	1.65	**1.75** ± 0.13	1.49	1.02	**1.26** ± 0.33
88	1.14	1.35	**1.25** ± 0.15	1.40	1.45	**1.43** ± 0.04	0.85	1.00	**0.93** ± 0.11
94	2.40	2.50	**2.45** ± 0.07	2.25	2.15	**2.20** ± 0.07	1.55	1.60	**1.58** ± 0.04
109	1.64	1.50	**1.57** ± 0.10	1.44	1.59	**1.52** ± 0.11	0.79	0.70	**0.75** ± 0.06
111	1.30	1.40	**1.35** ± 0.07	1.25	1.30	**1.28** ± 0.04	0.82	0.92	**0.87** ± 0.07
120	1.99	2.00	**2.00** ± 0.01	1.99	1.86	**1.93** ± 0.09	1.10	1.06	**1.08** ± 0.03
138	2.37	2.42	**2.40** ± 0.04	1.93	1.91	**1.92** ± 0.01	1.10	0.97	**1.04** ± 0.09

Full veneer gold shell crown preparation:

The next demonstrate the results (mm) of the ten SAFMs, along with mean and standard deviation data, determined by using an indirect method on two occasions to assess full veneer gold shell crown preparations.

Tables 6.12, 6.13, 6.14 and 6.15 summarise the ten SAFMs (mm) made, using the ImageJ software (version 1.47), of the specific anatomical features of each of the 30 full veneer gold shell crown preparations on two separate occasions.

Table 6.12 Mean and standard deviation measurements (mm), for the full veneer gold shell crown preparation for i) total occlusal convergence (i.e. proximal convergence) and ii) occlusal reduction from the buccal view using ImageJ software

Buccal view									
Models Number	Total occlusal convergence (i.e. proximal convergence)			Occlusal reduction					
				Mesial side			Distal side		
	First occasion	Second occasion	Mean ± Standard Deviation	First occasion	Second occasion	Mean ± Standard deviation	First occasion	Second occasion	Mean ± Standard deviation
1	19.54	17.16	**18.35** ±1.68	2.33	2.20	**2.27** ±0.09	2.27	2.20	**2.24** ±0.05
3	16.80	16.75	**16.78** ±0.04	1.84	1.78	**1.81** ±0.04	1.52	1.60	**1.56** ±0.06
4	16.53	16.41	**16.47** ±0.08	1.21	1.20	**1.21** ±0.01	1.16	1.18	**1.17** ±0.01
5	8.50	8.70	**8.60** ±0.14	1.87	1.78	**1.83** ±0.06	1.55	1.52	**1.54** ±0.02
7	18.00	17.60	**17.80** ±0.28	1.92	2.00	**1.96** ±0.06	1.65	1.51	**1.58** ±0.10
13	24.48	25.80	**25.14** ±0.93	1.98	1.91	**1.95** ±0.05	2.06	2.10	**2.08** ±0.03
14	26.80	27.30	**27.05** ±0.35	2.10	1.98	**2.04** ±0.08	1.72	1.80	**1.76** ±0.06
18	14.10	14.30	**14.20** ±0.14	1.00	1.10	**1.05** ±0.07	1.21	1.15	**1.18** ±0.04
20	17.20	17.10	**17.15** ±0.07	2.98	2.93	**2.96** ±0.04	2.10	2.20	**2.15** ±0.07
21	14.10	13.90	**14.00** ±0.14	2.05	1.95	**2.00** ±0.07	1.82	1.82	**1.82** ±0.00
25	18.10	17.70	**17.90** ±0.28	1.60	1.66	**1.63** ±0.04	1.61	1.68	**1.65** ±0.05
26	21.70	19.60	**20.65** ±1.48	1.32	1.36	**1.34** ±0.03	0.91	0.95	**0.93** ±0.03
29	19.35	19.80	**19.58** ±0.32	1.31	1.35	**1.33** ±0.03	1.00	1.05	**1.03** ±0.04
31	17.48	18.30	**17.89** ±0.58	2.72	2.69	**2.71** ±0.02	1.70	1.66	**1.68** ±0.03
51	16.50	17.23	**16.87** ±0.52	1.12	1.13	**1.13** ±0.01	1.10	1.12	**1.11** ±0.01
52	20.60	22.21	**21.41** ±1.14	1.73	1.79	**1.76** ±0.04	1.67	1.68	**1.68** ±0.01
54	31.49	33.20	**32.35** ±1.21	2.20	2.18	**2.19** ±0.01	1.10	1.17	**1.14** ±0.05
57	17.53	18.00	**17.77** ±0.33	1.42	1.38	**1.40** ±0.03	1.84	1.81	**1.83** ±0.02
58	15.80	16.25	**16.03** ±0.32	1.80	1.71	**1.76** ±0.06	2.00	2.10	**2.05** ±0.07
59	20.72	21.30	**21.01** ±0.41	3.02	3.10	**3.06** ±0.06	2.50	2.41	**2.46** ±0.06
60	8.00	6.80	**7.40** ±0.85	0.64	0.60	**0.62** ±0.03	0.63	0.60	**0.62** ±0.02
63	16.60	16.67	**16.64** ±0.05	3.40	3.32	**3.36** ±0.06	2.28	2.30	**2.29** ±0.01
67	19.42	18.60	**19.01** ±0.58	2.02	2.03	**2.03** ±0.01	1.08	1.15	**1.12** ±0.05
69	18.61	17.90	**18.26** ±0.50	1.82	1.85	**1.84** ±0.02	1.36	1.30	**1.33** ±0.04
70	15.57	16.10	**15.84** ±0.37	0.85	0.82	**0.84** ±0.02	0.56	0.51	**0.54** ±0.04
71	22.50	21.10	**21.80** ±0.99	1.35	1.38	**1.37** ±0.02	1.00	1.12	**1.06** ±0.08
73	17.80	18.30	**18.05** ±0.35	1.93	1.90	**1.92** ±0.02	1.41	1.35	**1.38** ±0.04
74	31.74	30.10	**30.92** ±1.16	1.38	1.40	**1.39** ±0.01	1.31	1.36	**1.34** ±0.04
78	13.20	13.10	**13.15** ±0.07	1.71	1.76	**1.74** ±0.04	1.50	1.55	**1.53** ±0.04
88	38.90	39.20	**39.05** ±0.21	2.89	2.98	**2.94** ±0.06	2.71	2.80	**2.76** ±0.06

Table 6.13 Mean and standard deviation measurements (mm), for the full veneer gold shell crown preparation for axial reduction from buccal view using ImageJ software

Models Number	Buccal view					
	Axial reduction					
	Mesial side			Distal side		
	First occasion	Second occasion	Mean ± Standard deviation	First occasion	Second occasion	Mean ± Standard deviation
1	1.66	1.57	**1.62** ±0.06	1.52	1.55	**1.54** ±0.02
3	2.90	2.69	**2.80** ±0.15	2.35	2.35	**2.35** ±0.00
4	2.12	2.20	**2.16** ±0.06	2.52	2.55	**2.54** ±0.02
5	2.15	2.27	**2.21** ±0.08	2.35	2.41	**2.38** ±0.04
7	2.48	2.48	**2.48** ±0.00	2.39	2.41	**2.40** ±0.01
13	2.41	2.42	**2.42** ±0.01	2.48	2.48	**2.48** ±0.00
14	3.49	3.45	**3.47** ±0.03	3.75	3.70	**3.73** ±0.04
18	1.93	1.86	**1.90** ±0.05	2.28	2.20	**2.24** ±0.06
20	2.57	2.45	**2.51** ±0.08	2.17	1.95	**2.06** ±0.16
21	1.96	1.94	**1.95** ±0.01	2.08	2.10	**2.09** ±0.01
25	1.66	1.72	**1.69** ±0.04	2.80	2.83	**2.82** ±0.02
26	1.85	2.00	**1.93** ±0.11	1.96	1.80	**1.88** ±0.11
29	2.54	2.35	**2.45** ±0.13	1.45	1.45	**1.45** ±0.00
31	2.35	2.36	**2.36** ±0.01	1.45	1.43	**1.44** ±0.01
51	1.80	1.79	**1.80** ±0.01	2.44	2.49	**2.47** ±0.04
52	2.50	2.62	**2.56** ±0.08	3.00	2.90	**2.95** ±0.07
54	2.43	2.39	**2.41** ±0.03	3.31	3.22	**3.27** ±0.06
57	1.86	1.73	**1.80** ±0.09	2.58	2.42	**2.50** ±0.11
58	1.90	1.93	**1.92** ±0.02	2.61	2.69	**2.65** ±0.06
59	3.70	3.60	**3.65** ±0.07	3.38	3.38	**3.38** ±0.00
60	2.16	2.21	**2.19** ±0.04	2.35	2.38	**2.37** ±0.02
63	3.29	3.17	**3.23** ±0.08	2.33	2.27	**2.30** ±0.04
67	2.50	2.42	**2.46** ±0.06	2.10	2.07	**2.09** ±0.02
69	2.38	2.30	**2.34** ±0.06	2.31	2.28	**2.30** ±0.02
70	1.87	1.81	**1.84** ±0.04	2.17	2.10	**2.14** ±0.05
71	1.24	1.30	**1.27** ±0.04	2.96	2.80	**2.88** ±0.11
73	2.55	2.62	**2.59** ±0.05	2.62	2.67	**2.65** ±0.04
74	3.00	3.17	**3.09** ±0.12	3.95	3.80	**3.88** ±0.11
78	2.69	2.49	**2.59** ±0.14	2.62	2.75	**2.69** ±0.09
88	3.65	3.66	**3.66** ±0.01	3.11	3.21	**3.16** ±0.07

Table 6.14 Mean and standard deviation measurements (mm), for the full veneer gold shell crown preparation for i) total occlusal convergence (i.e. Bucco-palatal convergence) and ii) occlusal reduction from mesial view using ImageJ software

Mesial view									
Models Number	Total occlusal convergence (i.e. Bucco-palatal convergence)			Occlusal reduction					
				Buccal side			Palatal side		
	First occasion	Second occasion	Mean ± Standard deviation	First occasion	Second occasion	Mean ± Standard deviation	First occasion	Second occasion	Mean ± Standard deviation
1	19.80	18.40	**19.10** ±0.99	2.30	2.27	**2.29** ±0.02	0.83	0.85	**0.84** ±0.01
3	21.10	20.14	**20.62** ±0.68	1.82	1.91	**1.87** ±0.06	2.10	2.14	**2.12** ±0.03
4	19.58	19.68	**19.63** ±0.07	1.21	1.28	**1.25** ±0.05	1.25	1.28	**1.27** ±0.02
5	16.66	17.34	**17.00** ±0.48	1.84	1.92	**1.88** ±0.06	2.60	2.62	**2.61** ±0.01
7	22.10	21.10	**21.60** ±0.71	1.88	1.78	**1.83** ±0.07	0.95	0.90	**0.93** ±0.04
13	31.20	32.39	**31.80** ±0.84	2.03	2.00	**2.02** ±0.02	1.54	1.65	**1.60** ±0.08
14	23.10	22.20	**22.65** ±0.64	2.10	2.12	**2.11** ±0.01	1.55	1.45	**1.50** ±0.07
18	16.10	16.80	**16.45** ±0.49	1.10	1.00	**1.05** ±0.07	1.06	0.95	**1.01** ±0.08
20	22.82	23.40	**23.11** ±0.41	3.01	3.07	**3.04** ±0.04	1.62	1.65	**1.64** ±0.02
21	21.30	22.28	**21.79** ±0.69	2.00	1.95	**1.98** ±0.04	2.38	2.40	**2.39** ±0.01
25	27.90	26.77	**27.34** ±0.80	1.65	1.75	**1.70** ±0.07	1.32	1.40	**1.36** ±0.06
26	21.20	22.33	**21.77** ±0.80	1.38	1.40	**1.39** ±0.01	0.86	1.00	**0.93** ±0.10
29	22.71	22.87	**22.79** ±0.11	1.40	1.39	**1.40** ±0.01	1.65	1.70	**1.68** ±0.04
31	20.27	20.38	**20.33** ±0.08	2.80	2.71	**2.76** ±0.06	0.90	0.85	**0.88** ±0.04
51	20.30	19.50	**19.90** ±0.57	1.15	1.08	**1.12** ±0.05	0.97	0.95	**0.96** ±0.01
52	11.00	12.20	**11.60** ±0.85	1.85	1.80	**1.83** ±0.04	1.75	1.84	**1.80** ±0.06
54	20.57	21.28	**20.93** ±0.50	2.14	2.13	**2.14** ±0.01	1.46	1.43	**1.45** ±0.02
57	4.50	5.30	**4.90** ±0.57	1.43	1.40	**1.42** ±0.02	1.30	1.33	**1.32** ±0.02
58	20.15	19.68	**19.92** ±0.33	1.74	1.78	**1.76** ±0.03	0.60	0.55	**0.58** ±0.04
59	21.20	20.00	**20.60** ±0.85	3.10	3.05	**3.08** ±0.04	2.20	2.30	**2.25** ±0.07
60	20.60	19.30	**19.95** ±0.92	0.65	0.62	**0.64** ±0.02	0.44	0.50	**0.47** ±0.04
63	26.63	27.60	**27.12** ±0.69	3.30	3.38	**3.34** ±0.06	2.35	2.27	**2.31** ±0.06
67	26.00	25.20	**25.60** ±0.57	2.00	2.01	**2.01** ±0.01	2.05	2.01	**2.03** ±0.03
69	23.30	22.60	**22.95** ±0.49	1.86	1.91	**1.89** ±0.04	0.96	1.05	**1.01** ±0.06
70	16.10	16.76	**16.43** ±0.47	0.81	0.75	**0.78** ±0.04	0.48	0.45	**0.47** ±0.02
71	16.20	16.00	**16.10** ±0.14	1.40	1.42	**1.41** ±0.01	1.00	0.98	**0.99** ±0.01
73	19.10	18.90	**19.00** ±0.14	1.95	1.90	**1.93** ±0.04	2.52	2.61	**2.57** ±0.06
74	22.79	22.10	**22.45** ±0.49	1.40	1.38	**1.39** ±0.01	1.22	1.29	**1.26** ±0.05
78	19.54	19.50	**19.52** ±0.03	1.80	1.87	**1.84** ±0.05	0.82	0.76	**0.79** ±0.04
88	36.70	38.10	**37.40** ±0.99	2.95	3.00	**2.98** ±0.04	3.10	3.04	**3.07** ±0.04

Table 6.15 Mean and standard deviation measurements (mm), for the full veneer gold shell crown preparation for axial reduction from mesial view using ImageJ software

Models Number	Mesial view					
	Axial reduction					
	Mesial side			Distal side		
	First occasion	Second occasion	Mean ± Standard deviation	First occasion	Second occasion	Mean ± Standard deviation
1	0.41	0.35	**0.38** ±0.04	1.11	1.17	**1.14** ±0.04
3	1.73	1.76	**1.75** ±0.02	1.66	1.65	**1.66** ±0.01
4	0.42	0.30	**0.36** ±0.08	1.37	1.25	**1.31** ±0.08
5	0.90	0.96	**0.93** ±0.04	1.04	1.17	**1.11** ±0.09
7	1.73	1.58	**1.66** ±0.11	1.23	1.20	**1.22** ±0.02
13	1.81	1.66	**1.74** ±0.11	1.47	1.31	**1.39** ±0.11
14	1.70	1.68	**1.69** ±0.01	1.73	1.70	**1.72** ±0.02
18	0.23	0.21	**0.22** ±0.01	1.00	0.97	**0.99** ±0.02
20	1.52	1.46	**1.49** ±0.04	2.00	2.00	**2.00** ±0.00
21	0.97	0.97	**0.97** ±0.00	0.76	0.71	**0.74** ±0.04
25	0.31	0.28	**0.30** ±0.02	2.07	1.85	**1.96** ±0.16
26	1.46	1.30	**1.38** ±0.11	1.38	1.33	**1.36** ±0.04
29	0.69	0.65	**0.67** ±0.03	0.62	0.60	**0.61** ±0.01
31	1.75	1.84	**1.80** ±0.06	1.10	1.11	**1.11** ±0.01
51	0.22	0.28	**0.25** ±0.04	0.96	0.97	**0.97** ±0.01
52	0.90	0.94	**0.92** ±0.03	1.54	1.44	**1.49** ±0.07
54	1.46	1.56	**1.51** ±0.07	1.79	1.81	**1.80** ±0.01
57	0.30	0.25	**0.28** ±0.04	1.27	1.21	**1.24** ±0.04
58	0.42	0.50	**0.46** ±0.06	1.15	1.06	**1.11** ±0.06
59	2.21	2.32	**2.27** ±0.08	1.68	1.73	**1.71** ±0.04
60	1.25	1.18	**1.22** ±0.05	0.94	0.98	**0.96** ±0.03
63	1.41	1.46	**1.44** ±0.04	2.08	2.07	**2.08** ±0.01
67	1.00	0.98	**0.99** ±0.01	0.70	0.70	**0.70** ±0.00
69	1.10	1.05	**1.08** ±0.04	0.72	0.62	**0.67** ±0.07
70	0.80	0.88	**0.84** ±0.06	0.25	0.40	**0.33** ±0.11
71	0.25	0.31	**0.28** ±0.04	1.36	1.25	**1.31** ±0.08
73	1.42	1.39	**1.41** ±0.02	1.38	1.40	**1.39** ±0.01
74	2.75	2.80	**2.78** ±0.04	1.34	1.31	**1.33** ±0.02
78	1.61	1.53	**1.57** ±0.06	1.73	1.75	**1.74** ±0.01
88	2.84	2.81	**2.83** ±0.02	2.38	2.30	**2.34** ±0.06

From previous tables of SAFMs for class II amalgam cavity and full veneer gold shell crown preparations; there were acceptable random differences (errors) according to standard deviations for the most of the measurements from digital calliper, MeshLab and ImageJ measuring methods.

To determine systemic difference and agreement for SAFMs, all of the SAFMs for class II amalgam cavity and full veneer gold shell crown preparations were analysed by different statistical analysis tests in order to determine reliable measurements (see result section 6.4.3).

6.4.3 Identification of reliable measurements for class II amalgam cavity and full veneer gold shell crown preparations

To identify reliable measuring method for class II amalgam cavity and full veneer gold shell crown preparation, the steps outlined in section (6.3.3) were applied:

Step 1: *comparing mean difference between measurements of occasion one and occasion two for each specific anatomical feature by using one or two different methods*

Paired sample t-test was used for each anatomical feature measurements of class II amalgam cavity and full veneer gold shell crown preparation.

<u>Class II amalgam cavity preparation:</u>

Table 6.16 summarises mean and standard deviation (±SD) for each occasion, the mean differences between first and second occasion, the (t) value, the degree of freedom and the statistical difference, using a paired sample t-test, for each specific anatomical feature measured using the digital calliper for all of the class II amalgam cavities.

Table 6.16 Paired sample t-test for class II amalgam cavity features measured using a digital calliper

	Occasion	**Mean**	**Standard deviation**	**Means difference**	**t**	**Degree of freedom**	**Significance (p≤0.05)**
Depth of box gingivally	1	2.848	±0.605	0.017	1.034	25	0.311
	2	2.831	±0.616				
Bucco-palatal width of the box floor	1	2.939	±0.733	-0.006	-0.312	25	0.757
	2	2.944	±0.702				
Bucco-palatal width occlusally	1	2.970	±0.631	0.002	0.069	25	0.945
	2	2.969	±0.583				
Box floor (mesio-distal) depth	1	1.377	±0.270	0.019	1.179	25	0.250
	2	1.358	±0.254				
Pulpal axial wall length	1	1.532	±0.583	0.004	0.181	25	0.858
	2	1.529	±0.624				
Isthmus width occlusally	1	1.467	±0.386	-0.010	-0.577	25	0.569
	2	1.478	±0.382				
Isthmus floor width	1	1.298	±0.197	-0.009	-0.559	25	0.581
	2	1.307	±0.226				
Occlusal cavity width in the middle	1	1.185	±0.109	0.005	0.233	25	0.817
	2	1.181	±0.118				
Marginal ridge thickness	1	1.390	±0.300	-0.019	-0.527	25	0.603
	2	1.409	±0.347				
Occlusal cavity depth at palatal side in the middle	1	1.646	±0.451	-0.084	-2.023.	25	0.054
	2	1.730	±0.463				
Occlusal cavity depth at buccal side in the middle	1	1.610	±0.353	-0.063	-2.040	25	0.052
	2	1.672	±0.369				
Occlusal cavity depth at distal side	1	1.192	±0.369	0.026	0.806	25	0.428
	2	1.166	±0.332				

A paired sample t-test was conducted to compare SAFMs of the class II amalgam cavity by digital calliper on the first and second occasions. According to mean difference of measurements, there were no significant differences ($p \leq 0.05$) between the measurements made on the first occasion and the second occasion of class II amalgam cavity preparation using a digital calliper.

Table 6.17 summarises the mean and standard deviation (SD) data for each occasion and each SAFM as well as the difference of the means between the first and second occasion evaluations. In addition, (t) value, the degree of freedom and significant difference for each SAFM of class II amalgam cavity by using MeshLab software are also presented.

Table 6.17 Paired sample t-test for class II amalgam cavity features measured using MeshLab software

	Occasion	**Mean**	**Standard Deviation**	**Means difference**	**t**	**Degree of freedom**	**Significance (p≤0.05)**
Depth of box gingivally	1	2.821	±0.581	-0.014	-0.982	25	0.335
	2	2.835	±0.580				
Bucco-palatal width of the box floor	1	2.995	±0.687	0.007	0.628	25	0.535
	2	2.988	±0.689				
Bucco-palatal width occlusally	1	2.995	±0.687	0.007	0.628	25	0.535
	2	2.988	±0.689				
Box floor (mesio-distal) depth	1	1.348	±0.292	0.010	0.873	25	0.391
	2	1.338	±0.275				
Pulpal axial wall length	1	1.506	±0.603	0.008	0.583	25	0.565
	2	1.498	±0.599				
Isthmus width occlusally	1	1.581	±0.396	-0.022	-1.607	25	0.121
	2	1.603	±0.398				
Isthmus floor width	1	1.287	±0.228	-0.002	-0.165	25	0.870
	2	1.289	±0.238				
Occlusal cavity width in the middle	1	1.115	±0.149	0.003	0.153	25	0.880
	2	1.112	±0.124				
Marginal ridge thickness	1	1.550	±0.367	0.046	1.670	25	0.107
	2	1.504	±0.327				
Occlusal cavity depth at palatal side in the middle	1	1.700	±0.432	-0.021	-0.970	25	0.341
	2	1.721	±0.437				
Occlusal cavity depth at buccal side in the middle	1	1.701	±0.331	0.049	2.640	25	0.014
	2	1.652	±0.302				
Occlusal cavity depth at distal side	1	1.112	±0.372	0.029	0.914	25	0.370
	2	1.083	±0.333				

The highlighted values represent statistically significant differences.

A paired sample t-test was conducted to measure the reliability of SAFMs for class II amalgam cavity by using MeshLab software in first and second occasions. According to the mean difference between measurements, there were no significant differences in the measurements for almost all of class II amalgam cavity features except the occlusal cavity depth at the buccal side in the middle. There was a significant difference between two occasions ($p \leq 0.05$) in the measurements for occlusal cavity depth at buccal side in the middle was on the first occasion (Mean= 1.701, SD = ±0.331) and on the second occasion (Mean = 1.652, SD = ±0.302); t (25) = 2.640, p = 0.014. However, this difference is extremely small (0.049 mm) and represents a difference less than the diameter of a human hair [i.e. from 0.03 to 0.11 mm, according to De Lacharrière et al., (2001)].

The results of a paired sample t-test (SPSS) for each method showed that there were no significant differences between SAFMs of class II amalgam cavity on two different occasions for each method.

From Table 6.18, a paired sample t-test was conducted to compare mean of measurements for each specific anatomical feature of class II amalgam cavity by using the digital calliper and MeshLab software.

Table 6.18 Paired sample t-test of digital calliper and MeshLab software measurements for class II amalgam cavity preparations

		Mean	Standard Deviation	Means difference	t	Degree of freedom	Significance (p≤0.05)
Depth of box gingivally	Calliper	2.841	0.609	0.010	0.490	25	0.628
	MeshLab	2.831	0.579				
Bucco-palatal width of the box floor	Calliper	2.944	0.716	-0.011	-0.208	25	0.837
	MeshLab	2.955	0.681				
Bucco-palatal width of the box occlusally	Calliper	2.944	0.622	-0.213	-6.644	25	0.000
	MeshLab	3.157	0.603				
Box floor (mesio-distal) depth	Calliper	1.370	0.259	0.024	1.075	25	0.293
	MeshLab	1.346	0.282				
Pulpal axial wall length	Calliper	1.532	0.603	0.028	1.718	25	0.098
	MeshLab	1.504	0.600				
Isthmus width occlusally	Calliper	1.475	0.382	-0.119	-4.799	25	0.000
	MeshLab	1.594	0.396				
Isthmus floor width	Calliper	1.305	0.208	-0.015	-1.163	25	0.256
	MeshLab	1.290	0.231				
Occlusal cavity width in the middle	Calliper	1.187	0.102	0.071	3.761	25	0.001
	MeshLab	1.154	0.130				
Marginal ridge thickness	Calliper	1.402	0.311	-0.127	-2.769	25	0.010
	MeshLab	1.529	0.341				
Occlusal cavity depth at palatal side in the middle	Calliper	1.691	0.445	-0.022	-1.189	25	0.245
	MeshLab	1.713	0.430				
Occlusal cavity depth at buccal side in the middle	Calliper	1.645	0.352	-0.034	-1.181	25	0.249
	MeshLab	1.678	0.313				
Occlusal cavity depth at distal side	Calliper	1.182	0.342	0.080	2.746	25	0.011
	MeshLab	1.101	0.344				

The highlighted values represent statistically significant differences between the methods.

Some of SAFMs had significant differences ($p \leq 0.05$) between the two measurement methods. These features were the;

1. bucco-palatal width of the box occlusally,
2. isthmus width occlusally,
3. occlusal cavity width at the middle,
4. occlusal cavity depth at distal side, and
5. marginal ridge thickness.

According to examiners' opinions, means difference of these feature was also extremely small. The correlation coefficients between occasions for digital calliper and MeshLab software and between the averaged (first occasion, and second occasion) measurements from class II amalgam cavities were applied in step 2 in order to determine the impact of differences on measurement correlation.

Full veneer gold shell crown preparation:

Table 6.19 summarises the mean, standard deviation (±SD) for each occasion, means difference between first and second occasion, (t) value, the degree of freedom and significant difference for each specific anatomical feature of all of the full veneer gold shell crown preparation as measured by using ImageJ software.

Table 6.19 Paired sample t-tests for the full veneer gold shell crown preparation features measured using ImageJ software

	Occasion	**Mean**	**Standard Deviation**	**Means difference**	**t**	**Degree of freedom**	**Significance (p≤0.05)**
Buccal view							
Total occlusal convergence (TOC)	1	19.255	6.427	0.039	0.219	29	0.828
	2	19.216	6.580				
Occlusal reduction from mesial side	1	1.846	0.663	0.001	0.083	29	0.935
	2	1.845	0.660				
Occlusal reduction from distal side	1	1.544	0.530	-0.014	-1.162	29	0.255
	2	1.558	0.534				
Axial reduction (Mesial side)	1	2.386	0.604	0.101	1.116	29	0.274
	2	2.366	0.587				
Axial reduction (Distal side)	1	2.513	0.595	0.025	0.086	29	0.121
	2	2.488	0.590				
Mesial view							
Total occlusal convergence (TOC)	1	21.017	5.720	0.014	0.087	29	0.931
	2	21.003	5.744				
Occlusal reduction from buccal side	1	1.868	0.659	0.001	0.061	29	0.952
	2	1.868	0.673				
Occlusal reduction from palatal side	1	1.461	0.688	-0.012	-1.002	29	0.325
	2	1.473	0.694				
Axial reduction (Buccal side)	1	1.186	0.720	0.011	0.783	29	0.440
	2	1.175	0.723				
Axial reduction (Palatal side)	1	1.327	0.487	0.026	1.854	29	0.074
	2	1.307	0.466				

According to means difference of measurements, there were no statistically significant differences ($p \leq 0.05$) in the measurements between first occasion and second occasion for each specific anatomical feature of full veneer gold shell crown preparation by using ImageJ software. Although there was no difference between two series measurements for each specific anatomical feature, the correlation coefficients for ImageJ software (first occasion, and second occasion) measurements from full veneer gold shell crown preparation was applied in Step 2 to determine the correlation between two series of measurements for each specific anatomical feature.

Step 2: *calculating intra-class correlation to analyse the intra-examiner reliability of methods.*

Intra-examiner reliability for Class II amalgam cavity:

Intra-examiner reliability is the degree of stability observed when a measurement is repeated under identical conditions by the same examiner. It gives a value of how much homogeneity or consensus of the measurements for the same examiner. In addition, it provides reliability of measurement points for each specific anatomical feature.

Table 6.20 summarises the intra-examiner reliability (intra-class correlation coefficient - SPSS) for each specific anatomical feature measured of class II amalgam cavity using the digital calliper device.

Table 6.20 Intra-class correlation for each specific anatomical feature for the class II amalgam cavity preparation measured using a digital calliper

Feature	**Intra-class correlation**		**Significance ($p \leq 0.05$)**
	Single measures	Average measures	
Proximal part			
Depth of the box gingivally	0.991	0.996	0.000
Bucco-palatal width of the box floor	0.991	0.996	0.000
Bucco-Palatal width of the box occlusally	0.973	0.986	0.000
Box floor depth (mesio-distal)	0.950	0.974	0.000
Pulpal axial wall length	0.987	0.993	0.000
Occlusal part			
Isthmus width occlusally	0.971	0.986	0.000
Isthmus floor width	0.934	0.966	0.000
Occlusal cavity width in the middle	0.606	0.755	0.000
Occlusal cavity depth (Palatal side)	0.893	0.944	0.000
Occlusal cavity depth (buccal side)	0.905	0.950	0.000
Occlusal cavity depth (at distal side)	0.889	0.941	0.000
Marginal ridge thickness	0.842	0.914	0.000

The highlighted value represents the lowest correlation.

From Table 6.20, intra-examiner reliability using digital calliper had a strong positive correlation between first and second measurements for almost all of the specific anatomical features except occlusal cavity width in the middle. The highest significant correlation was the depth of the box and bucco-palatal width of the box gingivally whereas the lowest was the occlusal width in the middle of the cavity. This result support that reliability of examiner's measurements was excellent by using the digital calliper.

For MeshLab software measuring method, Table 6.21 summarises the intra-examiner reliability (i.e. intra-class correlation) for each anatomical feature measured of class II amalgam cavity.

Table 6.21 Intra-class correlation for the class II amalgam cavity preparation for each specific anatomical feature measured using MeshLab software

Feature	Intra-class correlation		Significance ($p \leq 0.05$)
	Single measures	Average measures	
Proximal part			
Depth of the box gingivally	0.992	0.996	0.000
Bucco-palatal width of the box floor	0.996	0.998	0.000
Bucco-Palatal width of the box occlusally	0.996	0.998	0.000
Box floor depth (mesio-distal)	0.977	0.989	0.000
Pulpal axial wall length	0.994	0.997	0.000
Occlusal part			
Isthmus width occlusally	0.983	0.992	0.000
Isthmus floor width	0.955	0.977	0.000
Occlusal cavity width in the middle	0.793	0.885	0.000
Occlusal cavity depth (palatal side)	0.967	0.983	0.000
Occlusal cavity depth (buccal side)	0.967	0.983	0.000
Occlusal cavity depth (at distal side)	0.945	0.972	0.000
Marginal ridge thickness	0.914	0.955	0.000

The highlighted value represents the lowest correlation.

From Table 6.21, intra-class correlation demonstrated that intra-examiner reliability was excellent for almost all specific anatomical features except occlusal width in the middle of the cavity. The similar result was reported in digital calliper measuring method. Therefore, measurement points for all specific anatomical features were reliable by using MeshLab measuring method. The highest significant correlation was pulpal axial wall length while the lowest was the occlusal width in the middle of the cavity.

Intra-examiner reliability of the measuring method and measurement points for almost all specific anatomical features of the class II cavity had a strong positive correlation between measurements in the first and second occasion by using the digital calliper and MeshLab software measuring methods. From these results, a very important question arose. Is the MeshLab measuring method (indirect) reliable in comparison with the digital calliper measuring method (direct)?

The answer to this question is demonstrated in the next table. Table 6.22 summarises the intra-examiner reliability (i.e. intra-class correlation) for each specific anatomical feature measured from the class II amalgam cavity by using two different methods. The mean of the first and second measurements of specific anatomical features for each class II amalgam cavity made using the digital calliper, compared with mean of first and second measurements of specific anatomical feature for same cavities made using the MeshLab software.

Table 6.22 Intra-class correlation between the mean (first occasion and second occasion) measurements for each specific anatomical feature measured for the class II amalgam cavity preparation by using digital calliper and MeshLab software

Feature	Intra-class correlation		Significance ($p \leq 0.05$)
	Single measures	Average measures	
Proximal part			
Depth of the box gingivally	0.984	0.992	0.000
Bucco-palatal width of the box floor	0.931	0.964	0.000
Bucco-Palatal width of the box occlusally	0.911	0.953	0.000
Box floor depth (mesio-distal)	0.912	0.954	0.000
Pulpal axial wall length	0.990	0.995	0.000
Occlusal part			
Isthmus width occlusally	0.907	0.951	0.000
Isthmus floor width	0.957	0.978	0.000
Occlusal cavity width in the middle	0.563	0.720	0.000
Occlusal cavity depth (palatal side)	0.976	0.988	0.000
Occlusal cavity depth (buccal side)	0.905	0.950	0.000
Occlusal cavity depth (at distal side)	0.884	0.938	0.000
Marginal ridge thickness	0.696	0.821	0.000

The highlighted data represent those features with the lowest correlation.

From Table 6.22, intra-examiner reliability was high for almost all of the specific anatomical features. The strongest positive correlation was the depth of the box and the bucco-palatal width of the box gingivally whereas the lowest correlation was for the occlusal cavity width in the middle. This means that MeshLab software was reliable as an indirect method but some of the measurement points for specific anatomical feature were not reliable, for example, occlusal cavity width in the middle.

Intra-examiner reliability for full veneer gold shell crown preparation:

Table 6.23 summarises the intra-examiner reliability of measurement points using intra-class correlation for each specific anatomical feature of full veneer gold shell crown preparation by using ImageJ software. According to Kerner et al., (2007), using ImageJ software for linear measurement was a reliable tool. Therefore, this measuring method was only used to determine the intra-examiner reliability of measurement points for each specific feature of full veneer gold shell crown preparation.

Table 6.23 Intra-class correlation for the full veneer gold shell crown preparation for each anatomical feature measured using ImageJ software

Feature	**Intra-class correlation**		**Significance ($p \leq 0.05$)**
	Single measures	Average measures	
Buccal view			
Total occlusal convergence	0.989	0.994	0.000
Occlusal reduction in the mesial side	0.995	0.998	0.000
Occlusal reduction in the distal side	0.992	0.996	0.000
Axial reduction in the mesial side	0.985	0.993	0.000
Axial reduction in the distal side	0.989	0.994	0.000
Mesial view			
Total occlusal convergence	0.989	0.994	0.000
Occlusal reduction in the buccal side	0.996	0.998	0.000
Occlusal reduction in the palatal side	0.995	0.998	0.000
Axial reduction in the buccal side	0.994	0.997	0.000
Axial reduction in the palatal	0.986	0.993	0.000

From Table 6.23, intra-class correlation of full veneer gold shell crown preparation measurements was high for all specific anatomical features. There was a strong positive correlation between two measurements. Therefore, intra-examiner reliability of measurement points was excellent and reliable.

In summary, Tables from Table 6.20 to 6.23 showed that the reliability (ICC) of digital calliper and MeshLab software measuring methods for class II amalgam cavity preparation and the ImageJ measuring method for the full veneer gold shell crown preparation was high. Moreover, the results demonstrated that measurement points by using direct and indirect methods for some specific anatomical features of class II amalgam cavity and for all full veneer gold shell crown preparation were reliable. According to Atkinson and Nevill (1998), Rankin and Stokes (1998), Bland and Altman (1999), Hopkins (2000) and Lexell and Downham, (2005), the analysis of reliability is not sufficient using only intra-class correlation. Intra-class correlation (ICC) can produce misleading results, for instance, the value of ICC may be low, if the sample is homogeneous. Therefore, in addition to intra-class correlation, Bland and Altman plots were used to identify agreement and systemic differences for each and between these methods (Rankin and Stokes, 1998, Bland and Altman, 1999).

Step 3*: Creating Bland-Altman plots for each feature of the tooth preparations using one or two different methods over two occasions.*

Bland and Altman plots were used to describe the agreement between two quantitative measurements using the same method or two different methods. The difference between the two measurements was plotted as the Y-co-ordinate against the mean of the same two measurements as the X-co-ordinate.

The next stage was to calculate a further value, known as either, 'the mean difference' or, 'bias'. This was the mean of all the previously described Y-co-ordinates. This value was displayed as a blue line on the Bland and Altman plot. This line estimated bias.

The final stage was to determine the standard deviation (± 1.96) around the mean difference (bias). This value is called 'the limit of agreement', and is displayed as two orange lines on the Bland and Altman plot and describes the upper and lower limits of agreement. Bland and Altman recommended that 95% of the data points should lie within ± 1.96 Standard deviation of the mean difference (Bland and Altman, 1999, Hanneman, 2008).

Bland and Altman plots were created using MedCalc software (version 13.0.0.0), They illustrated bias and limits of agreement for each specific anatomical feature measurement (SAFM) for the class II amalgam cavity and the full veneer gold shell crown preparation.

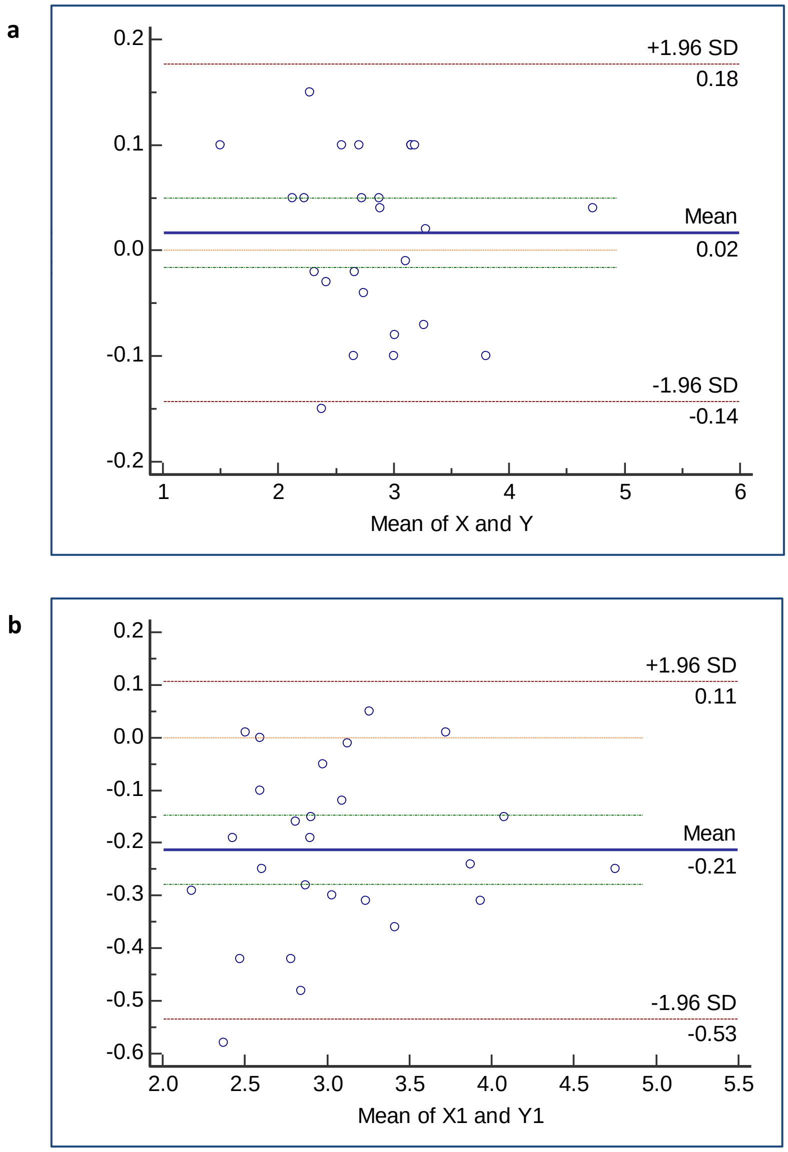
According to MedCalc software, the Bland and Altman plots illustrate several features. These are described below and refer to Figure 6.10.

1. A horizontal line of equality appears on the graph as a dotted amber line (zero value). This line is useful for detecting any systematic difference.
2. A horizontal line of mean differences appears on the graph as a solid blue line (Figure 6.10). This line represents any bias between the two different measurements or methods. In addition, this line is useful to detect which measuring method or occasion was generally higher or lower than the other. For example, measurements on the first occasion (X) from the Method One in Figure 6.10a tended to be higher than on the second occasion (Y). Thus the blue solid horizontal line is above the amber dotted line. Conversely, in Figure 6.10b, the line of mean differences (solid blue line) was below the zero value (amber dotted line), indicating that the measurements from the first occasion (X1) using

Method Two tended to be lower than the measurements made on the second occasion (Y1).

3. Two horizontal dotted green lines represent the 95% Confidence Interval of the mean difference. In other word, these lines indicate the range within which 95% of the differences from the bias are expected to be. This range is called confidence limit. These lines also illustrate the magnitude of the systematic difference (Figure 6.10). If the line of equality (dotted amber line = Zero) is not within this interval, there is a significant systematic difference between the two occasions/methods (X1, Y1) (Figure 6.10b).
4. Two horizontal brown lines on the Bland and Altman plot indicate the Limits of Agreement (LOA) (Figure 6.10). This is also known as ±1.96 SD of the mean difference. Upper limit of agreement is computed as 'mean difference (bias) + 1.96SD'. The lower limit of agreement is computed as 'mean difference (bias) − 1.96SD'. Upper LOA − lower LOA = confidence limit. The data should lie within ±1.96 SD of the mean difference (LOA). The width between the upper and lower limits of agreement is useful to identify the clinical relevance of what constitutes a significant difference between a pair of measurements. This might also depend on the examiners' opinion or previous studies. For example, In Figure 6.10a, the width of the limits of agreement is from -0.14mm to +0.18mm. In total this is a width of 0.32mm. Thus, it was not possible to detect a difference between the two evaluations when evaluation 1 was ≤ 0.32mm. Note the same value for Figure 6.10b would be ≤ 0.64mm. However, these data show a significant systematic difference between occasion/method (X1) and occasion/method (Y1) which is indicated by the fact that the dotted amber line lies outwith the two dotted green lines (see point 3 above). This means that

Figure 6.10b demonstrates that measurements on occasion/method (X1) were lower than measurements on occasion/method (Y1).



a. X measurements higher than Y measurements with no systemic difference and acceptable limit of agreement (0.18 + 0.14 = 0.32).
b. X1 measurements lower than Y1 measurements with systematic difference and unacceptable limit of agreement (0.11 + 0.53 = 0.64).

Figure 6.10 Examples of Bland and Altman plots to illustrate the differences and mean measurements from two different occasions and methods (Xn and Yn)

Class II amalgam cavity preparation:

All the Bland and Altman plots for class II amalgam cavities are illustrated in Appendix 5. These figures show Bland and Altman plots for each specific anatomical features of 26 class II amalgam cavity preparation using digital calliper and MeshLab software methods over two occasions in order to determine the agreement (repeatability) between measurements. According to the opinion of one of the senior examiners, an error of 0.50 mm between readings for the class II cavity would be acceptable as this would not be recognised clinically. Therefore, this value was acceptable, if the total number of the upper and lower limit of agreement was within this value.

Table 6.24 summarises the results of the Bland and Altman plots for SAFMs of the class II amalgam cavity preparations. Most of the values created by digital calliper and MeshLab software were within the range of ±1.96 SD which is within the limit of agreement as show in Appendix 5.

Most of the measured values from MeshLab software tended to be higher than those from the digital calliper device except for measurements of the occlusal cavity depth at the distal side. Furthermore, seven specific anatomical feature measurements (SAFMs) of the class II amalgam cavity demonstrated a systematic difference and non-acceptable range of the limits of agreement comparing digital calliper and MeshLab software (Table 6.24).

Table 6.24 Descriptive summary of SAFMs for comparisons of digital calliper and Meshlab software for the class II amalgam cavity taken from data recorded as Bland-Altman plots in Appendix 5

Specific anatomical features which were measured (SAFMs) for class II amalgam cavity preparation	**Result of comparing SAFMs using digital calliper and MeshLab software for class II amalgam cavity preparation**
1. Depth of the box gingivally	Acceptable difference, upper and lower limits of agreement
2. Bucco-palatal width of the box at gingival floor	Non-acceptable limits of agreement.
3. Bucco-palatal width of the box occlusally	Systematic difference, Non-acceptable wide limits of agreement
4. Box floor (mesio-distal) depth	Acceptable difference, upper and lower limits of agreement
5. Pulpal axial wall length	Acceptable difference, upper and lower limits of agreement
6. Isthmus width occlusally	Systematic difference.
7. Isthmus floor width	Acceptable difference, upper and lower limits of agreement
8. Occlusal cavity width in the middle	Systematic difference.
9. Occlusal cavity depth at palatal side	Acceptable difference, upper and lower limits of agreement
10. Occlusal cavity depth at buccal side	Non-acceptable limits of agreement
11. Occlusal cavity depth at distal side	Systematic difference, Non-acceptable limits of agreement
12. Marginal ridge thickness	Systematic difference, Non-acceptable limits of agreement

The high-lighted specific anatomical features according to (SAFMs) of the class II amalgam cavity have systematic difference and/or non-acceptable range of the limits of agreement

Although these features had systematic differences and non-acceptable range of limits of agreement, some of these features can be estimated, changed or deleted in order to collect reliable SAFMs for class II amalgam cavity preparation.

The accuracy and reliable of SAFMs can be used to evaluate the accuracy of some subjective decisions made by examiners when they evaluate tooth preparations. These data will be used in a later section to this effect (see section 6.4.4).

<u>Full veneer gold shell crown preparation</u>

All the Bland and Altman plots for full veneer gold shell crown preparations are illustrated in Appendix 6. These figures show Bland and Altman plots for each specific anatomical features of 30 full veneer gold shell crown preparations from *buccal and* *mesial view* on two occasions using ImageJ software over two occasions in order to determine the agreement (repeatability) between measurements..

Bland and Altman plots illustrated no or acceptable systemic differences in the ImageJ software. In addition, mean differences were very small and acceptable for all specific anatomical features according to (SAFMs) of full veneer gold shell crown preparation.

6.4.4 Comparison of a) the objective class II amalgam cavity or full veneer gold veneer shell crown preparation dimensions recorded by the researcher (AM) and those subjective tooth preparation evaluations by the senior examiners exhibiting the best agreement with b) dimensions presented in the dental literature and subsequent calibration with c) the grades of the best senior examiner

Specific anatomical feature measurements (SAFMs) and evaluation of the specific anatomical features (SAF) according the most agreed binary decision between senior academic staff examiners compared with values and features recorded in the literature followed by subsequent comparison with the grades determined by the best senior academic examiner.

The SAFMs evaluated previously, using paired sample t-tests, intra-class correlation coefficients (ICCs) and Bland and Altman plots, give a good indication which measurements are accurate and reliable. There are also data for such measurements reported in the literature. However, data from the literature is often reported as a data range or in terms of acceptable features. The researcher (AM) has been unable to identify actual tables of measurements for class II cavity preparations in premolar teeth made by other researchers although there is more data for gold shell crown preparations. Thus, it is sensible to make comparisons between measurements of tooth preparations which have been recorded by the researcher (AM) and acceptable measurements reported in the literature to determine passing and failed teeth objectively. If all SAFMs for each tooth preparation have been accepted according to acceptable measurements reported in the literature, the tooth preparation was marked as a passing tooth. If only one of SAFMs was not acceptable, this tooth was marked as a failed tooth. Such a tooth preparation was deemed not to have met the standard.

On the other hand, SAF which were unable to be measured can be evaluated according to the most agreed binary decision between three senior academic staff examiners from feedback sheets. The result of SAFM comparisons and SAF evaluations produced passing and failed tooth preparations. The final step was to compare the passing and failed tooth preparations with grades awarded by the best senior academic staff examiner. These comparisons gave weight to determine the 'best' senior academic staff examiner who was most accurate in identifying passing and failed grades as well as establishing this examiner as a gold standard examiner for this study or not.

The information within this section is complex and has the facility to become lost within the body of the text. Section 6.4.4 is set out as described in the following Figure 6.11.

1. **Step 1:** determining reliable average measurements (means) and standard deviations (±SD) for each objective measurement of SAF and subjective evaluation of SAF for each type of tooth preparation
    a. Class II amalgam cavity
        i. determining reliable average measurements (means) and standard deviations (±SD) for each objective measurement of SAF
        ii. determining subjective evaluation of SAF from other objective SAFMs or from the most agreed binary decision between three senior academic staff examiners for the class II amalgam cavity preparation
    b. Full veneer gold shell crown preparation
        i. determining reliable average measurements (means) and standard deviations (±SD) for each objective SAFMs

ii. determining subjective evaluation of SAF from the most agreed binary decision between three senior academic staff examiners for the full veneer gold shell crown preparation

2. **Step 2:** comparing the SAFM and SAF of tooth preparations which have been recorded by the researcher and acceptable measurements reported in the literature

    a. Class II amalgam cavity

        i. comparing the SAFM of the class II amalgam cavity preparations which have been recorded by the researcher and acceptable measurements reported in the literature

        ii. comparing the SAFs for the class II amalgam cavity preparation of the most agreed binary decision between three senior academic staff examiners with acceptable class II amalgam cavity features which were reported in the literature

    b. Full veneer gold shell crown preparation

        i. comparing the SAFM of the full veneer gold shell crown preparations which have been recorded by the researcher and acceptable measurements reported in the literature

        ii. comparing the SAFs for the full veneer gold shell crown preparation of the most agreed binary decision between three senior academic staff examiners with acceptable class full veneer gold shell crown preparation features which were reported in the literature

3. **Step 3:** comparing the tooth preparations with grades awarded by the best senior academic staff examiner.

    a. Class II amalgam cavity

    b. Full veneer gold shell crown preparation

Figure 6.11 Outline of a comparison of the SAFMs with values recorded in the literature and evaluation of the SAF according the most agreed binary decision between senior academic staff examiners followed by subsequent comparison with the grades determined by the best senior academic examiner

To investigate the relationship of class II amalgam cavity and full veneer gold shell crown preparation feature measurements and grades awarded by the best senior academic staff examiner, the steps outlined previously were applied.

Step 1: *determining reliable average measurements (means) and standard deviations (±SD) for each objective measurement of SAF and subjective evaluation of SAF for each type of tooth preparation*

Class II amalgam cavity preparation

I. determining reliable average measurements (means) and standard deviations (±SD) for each objective measurement of SAF

According to the results in the sections 6.4.2 and 6.4.3, the mean and standard deviation (±SD) for each reliable SAFMs, measured using the digital calliper device or MeshLab software, for the class II amalgam cavity were selected (Table 6.25). Data which exhibited systematic differences and a wide range of limits of agreement were excluded as such differences were often due to difficulty in identifying features involving subjective interpretation of the tooth preparation or its image, for example, landmarks along a curved surface (Table 6.24).

Table 6.25 demonstrates reliable mean and standard deviation (±SD) data for SAFMs made using MeshLab software for the class II amalgam cavity preparation.

Table 6.25 Means and standard deviations (±SD) for reliable SAFMs, using MeshLab software, for class II amalgam cavity preparation

Model number	Depth of the box gingivally (mm)	Box floor (mesio-distal) depth (mm)	Pulpal axial wall length (mm)	Isthmus floor width (mm)
	Mean ± standard deviation	Mean ± standard deviation	Mean ± standard deviation	Mean ± standard deviation
5	**3.14** ± 0.05	**1.83** ± 0.04	**1.69** ± 0.13	**1.25** ± 0.03
8	**2.16** ± 0.06	**1.42** ± 0.05	**1.41** ± 0.01	**1.22** ± 0.05
15	**3.35** ± 0.01	**1.65** ± 0.06	**2.02** ± 0.02	**1.03** ± 0.04
16	**2.98** ± 0.11	**1.58** ± 0.06	**2.02** ± 0.03	**1.25** ± 0.06
36	**2.66** ± 0.08	**1.28** ± 0.04	**1.27** ± 0.04	**1.36** ± 0.06
39	**3.25** ± 0.10	**1.08** ± 0.04	**2.08** ± 0.04	**1.21** ± 0.01
40	**2.43** ± 0.03	**1.14** ± 0.02	**1.08** ± 0.04	**1.44** ± 0.06
41	**2.53** ± 0.18	**1.52** ± 0.03	**1.19** ± 0.06	**1.93** ± 0.11
43	**2.40** ± 0.09	**1.44** ± 0.04	**1.35** ± 0.06	**1.28** ± 0.04
46	**3.10** ± 0.00	**1.33** ± 0.04	**1.26** ± 0.07	**1.19** ± 0.01
53	**2.73** ± 0.01	**1.05** ± 0.05	**1.13** ± 0.01	**1.19** ± 0.06
54	**2.70** ± 0.04	**0.93** ± 0.02	**1.18** ± 0.04	**1.14** ± 0.04
57	**3.25** ± 0.01	**0.90** ± 0.07	**1.12** ± 0.02	**1.00** ± 0.07
62	**2.25** ± 0.00	**1.14** ± 0.06	**1.16** ± 0.06	**1.44** ± 0.05
73	**4.53** ± 0.04	**1.73** ± 0.04	**3.78** ± 0.04	**1.20** ± 0.06
78	**2.85** ± 0.07	**1.34** ± 0.01	**1.66** ± 0.06	**0.90** ± 0.07
80	**3.14** ± 0.04	**1.18** ± 0.02	**1.80** ± 0.01	**1.48** ± 0.04
83	**2.20** ± 0.13	**1.29** ± 0.02	**1.04** ± 0.01	**1.12** ± 0.02
85	**2.48** ± 0.04	**1.21** ± 0.06	**1.13** ± 0.03	**1.45** ± 0.04
87	**2.83** ± 0.04	**1.29** ± 0.06	**1.32** ± 0.01	**1.89** ± 0.02
88	**3.30** ± 0.00	**1.42** ± 0.01	**2.00** ± 0.01	**1.40** ± 0.06
94	**2.66** ± 0.01	**0.92** ± 0.02	**0.86** ± 0.01	**1.31** ± 0.01
109	**2.73** ± 0.04	**1.31** ± 0.08	**1.72** ± 0.03	**1.32** ± 0.02
111	**1.56** ± 0.06	**1.96** ± 0.06	**0.91** ± 0.07	**1.27** ± 0.06
120	**2.49** ± 0.02	**1.29** ± 0.01	**1.00** ± 0.07	**1.14** ± 0.06
138	**3.72** ± 0.03	**1.76** ± 0.01	**1.93** ± 0.04	**1.14** ± 0.01

According to the data from section 6.4.3, the remaining SAFMs for the class II amalgam cavity preparation measured using MeshLab software were excluded due to systematic differences and unacceptable limits of agreement in comparison with measurements made using the digital calliper method. These excluded SAFMs were, bucco-palatal width of the box occlusally, bucco-palatal width of the box at gingival floor, isthmus width occlusally, occlusal width at the middle, occlusal cavity depth at buccal and distal sides and marginal ridge thickness.

There are seven SAFMs that have been found to be un-reliable and these data have been excluded. Excluded data cannot be used to evaluate the categories of the ‘Gray feedback sheet’. However, there are two other ways in which two of these previously excluded measurements could be reliably alternatively defined. The inclusion of these alternatively defined SAFMs will lead to a more comprehensive and more objective comparison with categories from the ‘Gray feedback sheet’.

The ways in which two of these previously excluded measurements could be reliably alternatively defined were <u>*substitution*</u> and <u>*calculation*</u> from reliable SAFMs and such methods were used to evaluate 26 Class II amalgam cavities more objectively.

Substitution was used for the reliable determination of occlusal cavity width. This feature was measured at both the isthmus and also at a line passing from the buccal to the palatal cusp tips. Measurement at the isthmus was more reliable than measurement along the buccal-palatal-cusp-tip-line. Therefore, the isthmus cavity width was substituted for the buccal-palatal-cusp-tip-line width (Table 6.25).

Calculation was used for the reliable determination of occlusal cavity depth. The subtraction of the reliable measurement of pulpal axial wall length from the reliable measurement of depth of the box gingivally can define the occlusal cavity depth at the isthmus (Table 6.26).

Table 6.26 Determination of calculated occlusal cavity depth at isthmus area for the class II amalgam cavity preparations

Model number	MeshLab depth of the box gingivally (mm)		MeshLab pulpal axial wall length (mm)		Difference between first occasion of depth of the box gingivally measurements and pulpal wall length measure-ment of first occasion	Difference between second occasion of depth of the box gingivally measurements and pulpal axial wall length measure-ment of second occasion	MeshLab calculated occlusal cavity depth at isthmus area (Mean ± standard deviation)
	First occasion	Second occasion	First occasion	Second occasion			
5	3.10	3.17	1.78	1.60	1.32	1.57	**1.45** ±0.18
8	2.20	2.12	1.40	1.42	0.80	0.70	**0.75** ±0.07
15	3.35	3.34	2.03	2.00	1.32	1.34	**1.33** ±0.01
16	2.90	3.05	2.04	2.00	0.86	1.05	**0.96** ±0.13
36	2.85	2.78	1.24	1.30	1.61	1.48	**1.55** ±0.09
39	3.18	3.32	2.10	2.05	1.08	1.27	**1.18** ±0.13
40	2.41	2.45	1.10	1.05	1.31	1.40	**1.36** ±0.06
41	2.65	2.55	1.13	1.15	1.52	1.40	**1.46** ±0.08
43	2.33	2.46	1.30	1.39	1.03	1.07	**1.05** ±0.03
46	3.10	3.10	1.11	1.21	1.99	1.89	**1.94** ±0.07
53	2.72	2.73	1.03	1.02	1.69	1.71	**1.70** ±0.01
54	2.67	2.73	1.15	1.11	1.52	1.62	**1.57** ±0.07
57	3.25	3.24	1.10	1.13	2.15	2.11	**2.13** ±0.03
62	2.25	2.25	1.20	1.12	1.05	1.13	**1.09** ±0.06
73	4.55	4.50	3.75	3.80	0.80	0.70	**0.75** ±0.07
78	2.80	2.90	1.62	1.70	1.18	1.20	**1.19** ±0.01
80	3.17	3.11	1.80	1.79	1.37	1.32	**1.35** ±0.04
83	2.19	2.10	1.00	1.03	1.19	1.07	**1.13** ±0.08
85	2.50	2.45	1.05	1.01	1.45	1.44	**1.45** ±0.01
87	2.80	2.86	1.12	1.11	1.68	1.75	**1.72** ±0.05
88	3.30	3.30	2.00	1.99	1.30	1.31	**1.31** ±0.01
94	2.65	2.66	0.85	0.86	1.80	1.80	**1.80** ±0.00
109	2.75	2.70	1.74	1.70	1.01	1.00	**1.01** ±0.01
111	1.51	1.60	0.86	0.96	0.65	0.64	**0.65** ±0.01
120	2.47	2.50	0.95	1.05	1.52	1.45	**1.49** ±0.05
138	3.70	3.74	1.95	1.90	1.75	1.84	**1.80** ±0.06

Reliable SAFMs, determined using MeshLab software for class II amalgam cavity preparation, were used to confirm the subjective evaluation of the most widely agreed SAF between three senior academic staff examiners. This was the presence or absence of retention form of the box preparation. For this subjective SAF, the reliable SAFMS of bucco-palatal width of the box at both the gingival floor and occlusal aspect of the box were used to determine if the box had retention form (i.e. was undercut) (Table 6.27).

II. Comparing the SAFs for the class II amalgam cavity preparation of the most agreed binary decision between three senior academic staff examiners with acceptable class II amalgam cavity features which were reported in the literature

Some of the subjective evaluations of class II amalgam cavity preparations which were part of the 'Gray feedback sheet' can be confirmed using reliable, objective SAFM data. To confirm 'retention form' of the box, reliable SAFMs (i.e. bucco-palatal box width at the gingival floor and at the occlusal aspect of the box using MeshLab software) were used to objectively determine if the walls of the box preparation were acceptable (undercut or parallel walls = YES) or not (divergent walls = NO) (Table 6.27). According to Hilton et al., (2013), buccal and palatal walls should be slightly converged or parallel walls.

In addition, Table 6.27 illustrates subjective specific features (SAF) for class II amalgam cavity which could not be confirmed by objective measurements. These remaining SAFs were deduced using those binary decisions from three senior academic staff examiners which exhibited the highest agreement.

Table 6.27 Subjective specific anatomical features (SAFs) for class II amalgam cavity preparation and calculation of retention form of the proximal box by comparing two reliable SAFMs (mm) by using MeshLab software

Model number	Is outline of the class II cavity acceptable?	Is position of the proximal box acceptable?	Is unsupported enamel existed?	Estimate retention form (converge, parallel or diverge walls) by comparing measurements of bucco-palatal width at gingival floor (mm) with bucco-palatal with of the box at occlusal (mm).		Is retention form acceptable?		Is there Occlusal lock?	Is there damage to adjacent tooth?
				Bucco-palatal width at gingival floor (mm) using MeshLab software	Bucco-palatal width of the box at occlusal (mm) using MeshLab software	According to the measurements	According to most agreed decision between examiners		
5	Yes	Yes	No	**2.45** ± 0.01	**3.01** ± 0.01	No	Yes	Yes	No
8	Yes	Yes	No	**2.43** ± 0.04	**2.68** ± 0.03	No	Yes (P)*	Yes	No
15	Yes	Yes	No	**2.38** ± 0.04	**2.73** ± 0.01	No	Yes	Yes	No
16	Yes	Yes	Yes	**2.80** ± 0.05	**2.66** ± 0.05	Yes	Yes	Yes	No
36	Yes	Yes	No	**3.13** ± 0.04	**2.98** ± 0.05	Yes	Yes	Yes	No
39	Yes	Yes	No	**2.67** ± 0.04	**2.59** ± 0.01	Yes	Yes	Yes	No
40	Yes	Yes	No	**3.48** ± 0.04	**3.23** ± 0.02	Yes	Yes	Yes	No
41	Yes	Yes	No	**3.44** ± 0.01	**3.99** ± 0.02	No	Yes (P)*	Yes	No
43	Yes	Yes	No	**2.72** ± 0.03	**2.99** ± 0.02	No	Yes	Yes	No
46	Yes	Yes	No	**2.73** ± 0.04	**3.15** ± 0.04	No	Yes	Yes	No
53	Yes	Yes	No	**2.68** ± 0.04	**3.39** ± 0.01	No	No	Yes	No
54	Yes	Yes	No	**2.27** ± 0.02	**2.64** ± 0.01	No	Yes	Yes	No
57	Yes	Yes	No	**3.59** ± 0.02	**3.13** ± 0.04	Yes	Yes	Yes	No
62	Yes	Yes	No	**2.96** ± 0.01	**3.18** ± 0.01	No	Yes	Yes	No
73	Yes	Yes	No	**2.06** ± 0.01	**2.32** ± 0.04	No	Yes	Yes	Yes
78	Yes	Yes	No	**2.55** ± 0.04	**2.50** ± 0.00	Yes	Yes	Yes	No
80	Yes	Yes	No	**2.22** ± 0.03	**2.52** ± 0.06	No	Yes	No	No
83	Yes	Yes	No	**2.85** ± 0.07	**2.99** ± 0.02	No	Yes	Yes	No
85	Yes	Yes	No	**3.68** ± 0.04	**3.72** ± 0.03	No	Yes (P)*	Yes	No
87	Yes	Yes	No	**3.14** ± 0.03	**3.59** ± 0.01	No	Yes	Yes	No
88	Yes	Yes	No	**4.68** ± 0.07	**4.88** ± 0.04	No	Yes (P)*	No	No
94	Yes	Yes	No	**3.05** ± 0.08	**3.00** ± 0.00	Yes	Yes	Yes	Yes
109	Yes	Yes	No	**2.69** ± 0.01	**3.08** ± 0.04	No	Yes	Yes	No
111	Yes	Yes	No	**4.84** ± 0.06	**4.15** ± 0.05	Yes	Yes (P)*	Yes	No
120	Yes	Yes	No	**2.65** ± 0.07	**2.89** ± 0.06	No	Yes	Yes	No
138	Yes	Yes	No	**3.68** ± 0.03	**4.09** ± 0.05	No	Yes	Yes	Yes

Highlighted yellow data indicate where a feature of the cavity preparation was not acceptable. The red rows indicate the models with one or more unacceptable feature. (*P = Parallel walls).

An interesting feature of Table 6.27 was the variation between the retention form objectively evaluated using measurements and the retention form most agreed from the subjective evaluation among senior examiners. A sub-set of the data from Table 6.27 is further compared in Table 6.28. This sub-set is for all cavities where the measured retention form of the box disagreed with the subjective evaluation by the three senior examiners. Table 6.28 demonstrates that examiners did not recognise a measurement difference of up to **0.56** mm clinically between bucco-palatal width at gingival floor and bucco-palatal width of the box at the occlusal level as acceptable retention form.

Table 6.28 Disagreement between 'Retention form' decisions according to the measurements using MeshLab software and 'Retention form' decisions according to the most agreed decision between three senior examiners for each cavity

Model number	Retention form decisions according to the measurements	Retention form decisions according to the most agreed decision between examiners*	Difference between bucco-palatal width at gingival floor and bucco-palatal with of the box at occlusal measurements (mm)
5	No	Yes	0.56
8	No	Yes (P)*	0.25
15	No	Yes	0.36
41	No	Yes (P)*	0.55
43	No	Yes	0.27
46	No	Yes	0.43
53	No	Yes	0.38
54	No	Yes	0.22
62	No	Yes	0.26
73	No	Yes	0.30
80	No	Yes	0.14
83	No	Yes	0.05
85	No	Yes (P)*	0.45
87	No	Yes	0.20
88	No	Yes (P)*	0.39
120	No	Yes	0.23
138	No	Yes	0.41
		Minimum measurement	**0.05**
		Maximum measurement	**0.56**

* P = Parallel walls

Full veneer gold shell crown preparation

I. determining the reliable average measurements (means) and standard deviations (±SD) for each objective measurement of SAF

According to the results in sections 6.4.2 and 6.4.3, all mean and standard deviation date for SAFMs were reliable when measured using ImageJ software for the full veneer gold shell crown preparation. Thus, all data were selected (Table 6.29).

Table 6.29 Means and standard deviations (±SD), using ImageJ software, of reliable SAFMs for the full veneer gold shell crown preparation

Buccal view										
Model Number	Total occlusal convergence		Occlusal reduction (mm)				Axial reduction (mm)			
			Mesial side		Distal side		Mesial side		Distal side	
	Mean ± Standard Deviation		**Mean ± Standard Deviation**		**Mean ± Standard Deviation**		**Mean ± Standard Deviation**		**Mean ± Standard Deviation**	
1	**18.35**	±1.68	**2.28**	±0.11	**2.04**	±0.05	**1.62**	±0.06	**1.54**	±0.02
3	**16.78**	±0.04	**2.67**	±0.07	**1.56**	±0.06	**2.80**	±0.15	**2.35**	±0.00
4	**16.47**	±0.08	**2.16**	±0.08	**1.83**	±0.04	**2.16**	±0.06	**2.54**	±0.02
5	**8.60**	±0.14	**1.71**	±0.08	**0.51**	±0.01	**2.21**	±0.08	**2.38**	±0.04
7	**17.80**	±0.28	**2.92**	±0.13	**0.58**	±0.10	**2.48**	±0.00	**2.40**	±0.01
13	**25.14**	±0.93	**1.82**	±0.13	**2.08**	±0.03	**2.42**	±0.01	**2.48**	±0.00
14	**27.05**	±0.35	**1.98**	±0.03	**1.76**	±0.06	**3.47**	±0.03	**3.73**	±0.04
18	**14.20**	±0.14	**1.05**	±0.07	**1.41**	±0.13	**1.90**	±0.05	**2.24**	±0.06
20	**17.15**	±0.07	**2.80**	±0.00	**1.45**	±0.14	**2.51**	±0.08	**2.06**	±0.16
21	**14.00**	±0.14	**1.35**	±0.01	**1.13**	±0.01	**1.95**	±0.01	**2.09**	±0.01
25	**17.90**	±0.28	**2.68**	±0.11	**2.80**	±0.04	**1.69**	±0.04	**2.82**	±0.02
26	**20.65**	±1.48	**1.79**	±0.10	**0.86**	±0.06	**1.93**	±0.11	**1.88**	±0.11
29	**19.58**	±0.32	**1.26**	±0.06	**0.97**	±0.09	**2.45**	±0.13	**1.45**	±0.00
31	**17.89**	±0.58	**2.06**	±0.08	**1.53**	±0.11	**2.36**	±0.01	**1.44**	±0.01
51	**16.87**	±0.52	**1.08**	±0.04	**1.86**	±0.10	**1.80**	±0.01	**2.47**	±0.04
52	**21.41**	±1.14	**2.41**	±0.08	**1.29**	±0.04	**2.56**	±0.08	**2.95**	±0.07
54	**32.35**	±1.21	**2.21**	±0.01	**1.24**	±0.09	**2.41**	±0.03	**3.27**	±0.06
57	**17.77**	±0.33	**1.26**	±0.08	**1.82**	±0.04	**1.80**	±0.09	**2.50**	±0.11
58	**16.03**	±0.32	**1.87**	±0.09	**2.15**	±0.07	**1.92**	±0.02	**2.65**	±0.06
59	**21.01**	±0.41	**3.21**	±0.01	**2.36**	±0.08	**3.65**	±0.07	**3.38**	±0.00
60	**7.40**	±0.85	**0.67**	±0.10	**0.62**	±0.02	**2.19**	±0.04	**2.37**	±0.02
63	**16.64**	±0.05	**3.56**	±0.08	**1.79**	±0.01	**3.23**	±0.08	**2.30**	±0.04
67	**19.01**	±0.58	**2.38**	±0.06	**0.47**	±0.02	**2.46**	±0.06	**2.09**	±0.02
69	**18.26**	±0.50	**1.48**	±0.04	**0.95**	±0.07	**2.34**	±0.06	**2.30**	±0.02
70	**15.84**	±0.37	**2.04**	±0.13	**2.09**	±0.02	**1.84**	±0.04	**2.14**	±0.05
71	**21.80**	±0.99	**1.95**	±0.07	**1.98**	±0.04	**1.27**	±0.04	**2.88**	±0.11
73	**18.05**	±0.35	**2.84**	±0.01	**0.92**	±0.05	**2.59**	±0.05	**2.65**	±0.04
74	**30.92**	±1.16	**3.17**	±0.10	**1.93**	±0.10	**3.09**	±0.12	**3.88**	±0.11
78	**13.15**	±0.07	**2.02**	±0.11	**0.83**	±0.11	**2.59**	±0.14	**2.69**	±0.09
88	**39.05**	±0.21	**2.92**	±0.09	**2.75**	±0.08	**3.66**	±0.01	**3.16**	±0.07

Mesial view										
Model Number	Total occlusal convergence		Occlusal reduction (mm)				Axial reduction (mm)			
			Mesial side		Distal side		Mesial side		Distal side	
	Mean ± Standard Deviation		**Mean ± Standard Deviation**		**Mean ± Standard Deviation**		**Mean ± Standard Deviation**		**Mean ± Standard Deviation**	
1	**19.10**	±0.99	**2.33**	±0.04	**0.84**	±0.01	**0.38**	±0.04	**1.14**	±0.04
3	**20.62**	±0.68	**2.92**	±0.05	**2.32**	±0.03	**1.75**	±0.02	**1.66**	±0.01
4	**19.63**	±0.07	**1.10**	±0.00	**1.76**	±0.06	**0.36**	±0.08	**1.31**	±0.08
5	**17.00**	±0.48	**2.25**	±0.07	**2.69**	±0.09	**0.93**	±0.04	**1.11**	±0.09
7	**21.60**	±0.71	**1.97**	±0.05	**1.95**	±0.07	**1.66**	±0.11	**1.22**	±0.02
13	**31.80**	±0.84	**2.54**	±0.08	**1.83**	±0.04	**1.74**	±0.11	**1.39**	±0.11
14	**22.65**	±0.64	**2.25**	±0.07	**1.49**	±0.05	**1.69**	±0.01	**1.72**	±0.02
18	**16.45**	±0.49	**1.04**	±0.08	**1.00**	±0.06	**0.22**	±0.01	**0.99**	±0.02
20	**23.11**	±0.41	**3.25**	±0.07	**1.64**	±0.02	**1.49**	±0.04	**2.00**	±0.00
21	**21.79**	±0.69	**2.21**	±0.06	**2.53**	±0.07	**0.97**	±0.00	**0.74**	±0.04
25	**27.34**	±0.80	**2.31**	±0.06	**1.67**	±0.01	**0.30**	±0.02	**1.96**	±0.16
26	**21.77**	±0.80	**1.51**	±0.08	**1.14**	±0.05	**1.38**	±0.11	**1.36**	±0.04
29	**22.79**	±0.11	**1.40**	±0.01	**2.31**	±0.05	**0.67**	±0.03	**0.61**	±0.01
31	**20.33**	±0.08	**2.85**	±0.07	**1.82**	±0.03	**1.80**	±0.06	**1.11**	±0.01
51	**19.90**	±0.57	**1.48**	±0.04	**0.96**	±0.01	**0.25**	±0.04	**0.97**	±0.01
52	**11.60**	±0.85	**3.34**	±0.05	**2.73**	±0.04	**0.92**	±0.03	**1.49**	±0.07
54	**20.93**	±0.50	**2.21**	±0.07	**1.45**	±0.01	**1.51**	±0.07	**1.80**	±0.01
57	**4.90**	±0.57	**1.68**	±0.08	**1.32**	±0.02	**0.28**	±0.04	**1.24**	±0.04
58	**19.92**	±0.33	**1.99**	±0.01	**0.58**	±0.04	**0.46**	±0.06	**1.11**	±0.06
59	**20.60**	±0.85	**2.65**	±0.07	**2.45**	±0.07	**2.27**	±0.08	**1.71**	±0.04
60	**19.95**	±0.92	**0.64**	±0.02	**0.94**	±0.08	**1.22**	±0.05	**0.96**	±0.03
63	**27.12**	±0.69	**3.54**	±0.06	**2.38**	±0.04	**1.44**	±0.04	**2.08**	±0.01
67	**25.60**	±0.57	**2.24**	±0.04	**2.25**	±0.05	**0.99**	±0.01	**0.70**	±0.00
69	**22.95**	±0.49	**2.08**	±0.03	**1.71**	±0.06	**1.08**	±0.04	**0.67**	±0.07
70	**16.43**	±0.47	**0.88**	±0.04	**0.46**	±0.01	**0.84**	±0.06	**0.33**	±0.11
71	**16.10**	±0.14	**1.64**	±0.08	**1.59**	±0.01	**0.28**	±0.04	**1.31**	±0.08
73	**19.00**	±0.14	**2.83**	±0.04	**2.65**	±0.08	**1.41**	±0.02	**1.39**	±0.01
74	**22.45**	±0.49	**1.39**	±0.01	**1.36**	±0.05	**2.78**	±0.04	**1.33**	±0.02
78	**19.52**	±0.03	**1.98**	±0.01	**0.74**	±0.06	**1.57**	±0.06	**1.74**	±0.01
88	**37.40**	±0.99	**2.96**	±0.06	**3.07**	±0.04	**2.83**	±0.02	**2.34**	±0.06

II. determining subjective evaluation of SAF from the most agreed binary decision between three senior academic staff examiners for the full veneer gold shell crown preparation

Table 6.30 demonstrates SAF for the full veneer gold shell crown preparation from the 'Mhanni feedback sheet' which cannot be measured by ImageJ software. Subjective evaluation features were established from the binary decisions of three senior academic staff examiners which exhibited the highest agreement.

Table 6.30 Subjective specific anatomical features (SAFs) for the full veneer gold shell crown preparation

Model number	Is the contour of occlusal preparation satisfied?	Is there any axial surface undercuts?	Is contour of axial surface(s) preparation follow tooth surface contour?	Is the contact area with adjacent teeth cleared?	Is location of functional bevel on the functional cusps adequate?
1	Yes	No	No	Yes	Yes
3	Yes	No	Yes	Yes	Yes
4	Yes	No	Yes	No	No
5	No	Yes	No	Yes	Yes
7	Yes	No	Yes	Yes	Yes
13	Yes	No	Yes	Yes	Yes
14	Yes	No	Yes	Yes	Yes
18	Yes	Yes	No	Yes	Yes
20	Yes	No	Yes	Yes	Yes
21	Yes	No	Yes	Yes	Yes
25	Yes	No	Yes	Yes	Yes
26	No	No	Yes	No	Yes
29	Yes	Yes	Yes	Yes	Yes
31	Yes	No	Yes	Yes	No
51	Yes	No	Yes	Yes	Yes
52	Yes	No	Yes	Yes	Yes
54	Yes	No	Yes	Yes	Yes
57	Yes	No	Yes	Yes	Yes
58	Yes	No	Yes	No	Yes
59	Yes	No	Yes	Yes	Yes
60	Yes	No	Yes	Yes	Yes
63	Yes	No	Yes	Yes	No
67	Yes	No	Yes	Yes	Yes
69	Yes	No	No	Yes	No
70	Yes	No	Yes	Yes	Yes
71	Yes	No	Yes	Yes	Yes
73	Yes	No	Yes	Yes	Yes
74	Yes	No	No	Yes	Yes
78	Yes	No	Yes	Yes	Yes
88	No	No	No	Yes	No

Model number	Is finish line of tooth preparation chamfer shape?	Is level of the finish line to gingival margin adequate?	Is texture of finish line margin adequate?	Is depth of finish line acceptable?	Is texture of final preparation except finish line margin adequate?	Is there any damage to mesial adjacent tooth?	Is there any damage to distal adjacent tooth?
1	Yes	Yes	Yes	No	No	No	No
3	No	Yes	Yes	No	Yes	No	No
4	No	No	Yes	No	No	No	Yes
5	No	Yes	Yes	No	Yes	No	No
7	Yes	Yes	Yes	Yes	Yes	No	No
13	Yes	No	No	No	Yes	No	Yes
14	Yes	Yes	Yes	Yes	Yes	No	No
18	No	Yes	Yes	No	No	No	No
20	Yes	Yes	Yes	Yes	Yes	No	No
21	Yes	Yes	Yes	No	Yes	No	No
25	Yes	Yes	Yes	Yes	Yes	No "M"*	No
26	Yes	Yes	Yes	No	Yes	Yes	Yes
29	No	Yes	Yes	No	Yes	Yes	Yes
31	Yes	Yes	Yes	Yes	Yes	No	Yes
51	Yes	Yes	Yes	Yes	Yes	No	No
52	Yes	Yes	Yes	Yes	Yes	No	No "M"*
54	No	No	Yes	No	Yes	No	No
57	Yes	Yes	Yes	Yes	Yes	No	No
58	No	No	Yes	No	Yes	No	No "M"*
59	No	Yes	Yes	No	Yes	No "M"*	Yes
60	No	No	Yes	No	Yes	No	Yes
63	No	Yes	Yes	No	Yes	No "M"*	No
67	Yes	Yes	Yes	Yes	Yes	No	No
69	Yes	Yes	Yes	Yes	Yes	No	No
70	Yes	Yes	Yes	No	Yes	No	No
71	Yes	Yes	Yes	Yes	Yes	No	No
73	No	Yes	Yes	No	Yes	No	No "M"*
74	Yes	Yes	Yes	Yes	Yes	No	No
78	Yes	Yes	Yes	Yes	Yes	No	No
88	No	Yes	Yes	No	Yes	No	No "M"*

Highlighted yellow data indicate where a feature of the tooth preparation was not acceptable. The red rows indicate the models with one or more unacceptable feature. (*"M" = Minor damage).

Step 2, *Compare i), the reliable mean of objective SAFMs from the most reliable method (using graphical representation) with ideal or acceptable measurements which were suggested in literature. For subjective evaluations, the dental literature stated whether or not a feature should be present and the subjective evaluation of the examiners reported for each tooth preparation whether or not the senior examiners agreed the feature was present.*

Class II amalgam cavity preparation

I. comparing the SAFMs of the class II amalgam cavity preparations which have been recorded by the researcher (AM) and acceptable measurements reported in the literature

Graphical representation the mean of SAFMs from the most reliable measurement was used to compare SAFMs of the class II amalgam cavity preparation which was recorded by the researcher (AM) and acceptable measurements reported in the literature.

Table 6.31 demonstrates the range of ideal and acceptable measurements for each class II amalgam cavity feature (SAFMs) which were collected from a literature review (Table 1.5).

Table 6.31 Ideal and accepted measurement range of objective features (SAFMs) of premolar teeth for class II amalgam cavity preparation

Category	Measurement range
Depth of the box gingivally*	2.50 – 4.00 mm
Proximal box depth (mesio-distal width of the box)	0.80 – 1.50 mm
Proximal extension of the box** (Bucco-palatal width of the box)	Clears adjacent tooth 0.50 to < 1.00 mm (3.00 - 4.00 mm)
Pulpal axial wall (Distance from floor of the occlusal cavity to the floor of the box 'gingival floor')	1.00 – 1.5 mm
Occlusal depth / Isthmus depth	1.50 - 2.00 mm
Occlusal width / Isthmus width***	0.8 – 1.50 mm

* The acceptable measurement of the depth of the box gingivally was calculated by adding isthmus depth measurement and pulpal axial wall measurement. ** Proximal extension of the box was calculated by the same method of isthmus width measuring. *** ¼ and ⅓ between intercuspal distance was measured from the picture and 3D unprepared premolar tooth by using ImageJ and MeshLab software respectively to determine the ideal and acceptable isthmus width measurement of premolar tooth.

All the bar charts for class II amalgam cavity preparations are illustrated in Appendix 7. Appendix 7 shows the mean of SAFMs from the reliable measurements recorded by the researcher (AM) compared with acceptable measurements reported in the literature. They also show the numbers of the models which lie between the acceptable upper and lower measurements for each objective feature.

Table 6.32 summarises the results of the bar charts for SAFMs of the class II amalgam cavity preparations. It demonstrates the percentage of the models which lie between the acceptable upper and lower measurements for each objective feature of class II amalgam cavity according to acceptable measurements reported in the literature.

Table 6.32 Percentages of the models which lie between the acceptable range of upper and lower measurements, defined in the literature search, based on objective features (SAFMs) of class II amalgam cavity preparations taken from bar charts in Appendix 7

Objective features (SAFMs) of class II amalgam cavity preparation	Percentage of the numbers of the models lie between the acceptable upper and lower measurements
Box depth in the gingival direction	69%
Box (mesio-distal) depth	73%
Bucco-palatal width of the box at gingival floor	31%
Pulpal axial wall length	54%
Occlusal cavity depth at isthmus area	27%
Occlusal cavity floor width at isthmus area	92%

Table 6.33 summarises the pass and fail scores and the number of errors for each class II amalgam cavity model according to reliable SAFMs. According to Knight (1997), any model with one SAFM error was excluded and scored overall as a fail model.

Table 6.33 Scoring the models according to reliable SAFMs for each class II amalgam cavity preparation

Model number	Box depth gingivally	Box floor (mesio-distal) depth	Bucco-palatal of the box at gingival floor	Pulpal axial wall length	Occlusal cavity depth at isthmus area	Occlusal cavity width at isthmus area	Score
5	✓	×	×	×	×	✓	Fail
8	×	✓	×	✓	×	✓	Fail
15	✓	×	×	×	×	✓	Fail
16	✓	×	×	×	×	✓	Fail
36	✓	✓	✓	✓	✓	✓	Pass
39	✓	✓	×	×	×	✓	Fail
40	×	✓	✓	✓	×	✓	Fail
41	✓	×	✓	✓	×	×	Fail
43	×	✓	×	✓	×	✓	Fail
46	✓	✓	×	✓	✓	✓	Fail
53	✓	✓	×	✓	✓	✓	Fail
54	✓	✓	×	✓	✓	✓	Fail
57	✓	✓	✓	✓	×	✓	Fail
62	×	✓	×	✓	×	✓	Fail
73	×	×	×	×	×	✓	Fail
78	✓	✓	×	×	×	✓	Fail
80	✓	✓	×	×	×	✓	Fail
83	×	✓	×	✓	×	✓	Fail
85	✓	×	✓	✓	×	✓	Fail
87	✓	✓	✓	✓	✓	×	Fail
88	✓	✓	×	×	×	✓	Fail
94	✓	✓	✓	×	✓	✓	Fail
109	×	✓	×	×	×	✓	Fail
111	×	×	×	×	×	✓	Fail
120	✓	✓	×	✓	×	✓	Fail
138	✓	×	✓	×	✓	✓	Fail

(✓ = measurement was ideal or acceptable, × = measurement was not acceptable)

The total number of models which have acceptable measurements according to reliable SAFMs was only one. This model was number 36 (Figure 6.12).

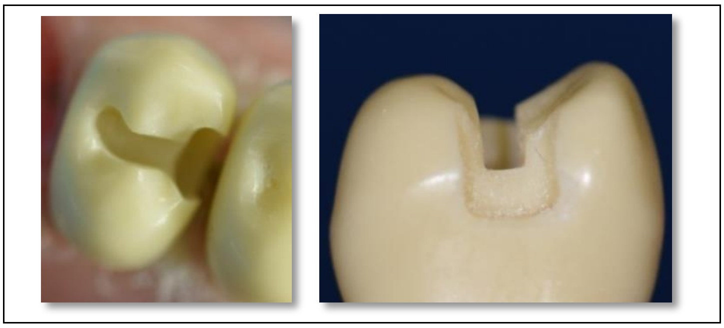


Figure 6.12 Picture of class II amalgam cavity number 36

II. Comparing the SAFs for the class II amalgam cavity preparation of the most agreed binary decision between three senior academic staff examiners with acceptable class II amalgam cavity features which were reported in the literature

Table 6.27 identified the class II amalgam cavities which had acceptable features of SAF evaluation determined by a majority decision of the three senior academic. Highlighted yellow data indicate where a feature of the tooth preparation was not acceptable in comparison with acceptable feature design in the literature (Table 6.27). Table 6.34 demonstrates the acceptable SAFs of the class II amalgam cavity preparation features according to the available literature (Baum et al., 1995, Roberson et al., 2002, Haj-Ali and Feil, 2006, Akpata et al., 2013, Hilton et al., 2013, Ahmed et al., 2016). According to Knight (1997), the number of the models which have no error (acceptable) ranked as a pass score and for this set of models there were only six tooth preparations.

Table 6.34 Table of responses to questions about SAF which represent an 'ideal' or 'acceptable' Class II amalgam cavity preparation.

SAF	Acceptable feature
Is the outline of class II amalgam cavity acceptable?	Yes
Is the position of the proximal box acceptable?	Yes
Does unsupported enamel exist?	No
Is the cavity wall retention form acceptable?	Converging or Parallel
Is there an occlusal lock?	Yes
Is there any damage to the adjacent tooth?	No or Minor

The final stage in this step for the class II amalgam restoration was to determine more objective grades demonstrated in the Table 6.35. Table 6.35 shows the comparison between scores of SAFM (Pass or fail) with subjective SAF evaluations scores to determine a more objective score for 26 tooth cavities and the number of errors.

Table 6.35 Score of SAFMs and SAFs for each class II amalgam cavity preparation to determine a more objective total score and the number of errors for these teeth

Model number	Scores according to SAFMs	Scores according to SAFs	Total score for class II amalgam cavity	Number of errors
5	Fail	Fail	Fail	5
8	Fail	Fail	Fail	4
15	Fail	Fail	Fail	5
16	Fail	Fail	Fail	5
36	Pass	Pass	Pass	0
39	Fail	Pass	Fail	3
40	Fail	Pass	Fail	2
41	Fail	Fail	Fail	4
43	Fail	Fail	Fail	4
46	Fail	Fail	Fail	2
53	Fail	Fail	Fail	2
54	Fail	Fail	Fail	2
57	Fail	Pass	Fail	1
62	Fail	Fail	Fail	4
73	Fail	Fail	Fail	7
78	Fail	Pass	Fail	3
80	Fail	Fail	Fail	5
83	Fail	Fail	Fail	4
85	Fail	Fail	Fail	3
87	Fail	Fail	Fail	2
88	Fail	Fail	Fail	5
94	Fail	Fail	Fail	2
109	Fail	Fail	Fail	5
111	Fail	Pass	Fail	5
120	Fail	Fail	Fail	3
138	Fail	Fail	Fail	4

According to Knight (1997), any model which has no errors will pass the evaluation overall. In relation to these data this was only for model number 36.

Full veneer gold shell crown preparation

I. Comparison of the SAFM of the full veneer gold shell crown preparations which have been recorded by the researcher (AM) and acceptable measurements reported in the literature

Table 6.36 demonstrates the range of acceptable specific anatomical feature measurements (SAFMs) for full veneer gold shell crown preparation features according to the available literature (Table 1.6).

Table 6.36 Ideal or acceptable ranges for objective SAFM for an upper molar tooth prepared for a full veneer gold shell crown preparation

Category	Measurement range (mm)
Total occlusal convergence (Mesio-distal and bucco-palatal convergence angle)	3° - 20°
Occlusal reduction (Functional cusps)	1.50 - < 2.00
Occlusal reduction (Non-functional cusps)	1.00 - < 1.50
Axial reduction	0.50 – < 1.50

All the bar charts in Appendix 8 are illustrated for full veneer gold shell crown preparations. Graphical representation of the mean of SAFMs from the most reliable measurement was used to compare SAFMs of the full veneer gold shell crown preparation which were recorded by the researcher (AM) and acceptable measurements reported in the literature. These figures also show the numbers of the models which lie between the acceptable upper and lower measurements for each objective feature (Appendix 8).

Table 6.37 summarises the results of the bar charts for SAFMs of the full veneer gold shell crown preparations. It demonstrates the percentage of the models which lie between the acceptable upper and lower measurements for each objective feature of full veneer gold shell crown preparation according to acceptable measurements reported in the literature.

Table 6.37 Percentage of models which lie between the acceptable range of upper and lower measurements defined in the literature search, based on objective features (SAFMs) of full veneer gold shell crown preparations taken from bar charts in Appendix 8

Objective features (SAFMs) of full veneer gold shell crown preparations	**Percentage of the numbers of the models lie between the acceptable upper and lower measurements**
Occlusal reduction at the mesial side from buccal view	27%
Occlusal reduction at the distal side from buccal view	33%
Occlusal reduction at the buccal side from mesial view	27%
Occlusal reduction at the palatal side from mesial view	17%
Axial reduction at the mesial side from buccal view	3%
Axial reduction at the distal side from buccal view	7%
Axial reduction at the buccal side from mesial view	43%
Axial reduction at the palatal side from mesial view	67%
Total occlusal convergence angle from buccal view and	70%
Total occlusal convergence angle from mesial view	43%

From previous graphs, Table 6.38 summarises the pass and fail models for the full veneer gold shell crown preparation using all SAFMs in a more objective way. Total evaluation for each of the occlusal convergence angles, occlusal reduction and axial reduction features were also reported according to Knight's recommendation. For example, if there was any one wall of the four axial wall reductions which did not meet the acceptable measurement in the literature, the total axial reduction assessment was fail (Knight, 1997).

Table 6.38 Scoring the models according to reliable SAFMs for each full veneer gold shell crown preparation

Model number	Occlusal convergence angle		Total	Occlusal reduction				Total	Axial reduction				Total	Score
	Buccal side	Mesial side		Mesio-buccal	Disto-buccal	Bucco-mesial	Palato-mesial		Mesio-buccal	Disto-buccal	Bucco-mesial	Palato-mesial		
1	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	Fail
3	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	Fail
4	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	Fail
5	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	Fail
7	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	Fail
13	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	Fail
14	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓	✗	Fail
18	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	Fail
20	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✓	✓	✗	Fail
21	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	Fail
25	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	Fail
26	✗	✗	✗	✓	✗	✓	✗	✗	✗	✗	✓	✗	✗	Fail
29	✓	✗	✗	✓	✓	✓	✓	✓	✗	✓	✓	✗	✗	Fail
31	✓	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	Fail
51	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	Fail
52	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗	Fail
54	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✓	✓	✗	Fail
57	✓	✓	✓	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗	Fail
58	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	Fail
59	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	Fail
60	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	Fail
63	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	Fail
67	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗	✓	✗	✗	Fail
69	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗	✓	✗	✗	Fail
70	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	Fail
71	✗	✓	✗	✓	✓	✓	✗	✗	✓	✗	✗	✗	✗	Fail
73	✓	✓	✓	✗	✓	✗	✗	✗	✗	✗	✓	✗	✗	Fail
74	✗	✗	✗	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	Fail
78	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	Fail
88	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	Fail

(✓ = measurement was ideal or acceptable, ✗ = measurement was not acceptable)

The total number of full veneer gold shell crown preparation models which had acceptable measurements according to reliable SAFMs was <u>zero</u>.

II. Comparison of the SAFs for the full veneer gold shell crown preparation (defined as the most agreed binary decision between three senior academic staff examiners) with acceptable full veneer gold shell crown preparation features reported in the literature

Table 6.30 shows the full veneer gold shell crown preparations which have acceptable features according to the SAF evaluation defined by the most agreed decisions between three senior academic staff examiners. Highlighted yellow data indicate where a feature of the tooth preparation was not acceptable in comparison with acceptable feature design in the literature. Table 6.39 demonstrates the acceptable SAFs of the full veneer gold shell crown preparation according to the available literature (Goodacre et al., 2001, Blair et al., 2002, O'Sullivan, 2005, Rosenstiel et al., 2006, Ricketts and Bartlett, 2011, Shillingburg et al., 2012).

Table 6.39 Table of responses to questions about SAF which represent an 'ideal' or 'acceptable' full veneer gold shell crown preparation

SAF	Acceptable feature
Is the contour of the occlusal preparation stratified?	Yes
Are there any axial surface undercuts?	No
Does the contour of the axial surface follow the unprepared tooth surface contour?	Yes
Has the contact area with the adjacent teeth been cleared?	Yes
Is the location of the functional bevel adequate?	Yes
Is the finish line of the tooth preparation a chamfer?	Yes
Is the level of the finish line to gingival margin adequate?	Yes
Is the texture of the finish line acceptable?	Yes
Is the depth of the finish line acceptable?	Yes
Is the texture of the final preparation (except finish line margin) adequate?	Yes
Is there any damage to mesial adjacent tooth?	No or Minor
Is there any damage to distal adjacent tooth?	No or Minor

According to Knight (1997), the number of the models which have no error were defined as a pass score. For this cohort that was 13 tooth preparations.

The final stage in this step to determine more objective grades is shown in Table 6.40. Table 6.40 shows the comparison between score of SAFMs (Pass or fail) with subjective SAF evaluations scores to determine more objective score for 30 tooth preparations. In addition, the numbers of error (negative points) were reported from scores of SAFMs and SAFs for each full veneer gold shell crown preparation.

Table 6.40 Score of SAFMs and SAFs for each full veneer gold shell crown preparation to determine a more objective total score and the number of errors for these teeth

Model number	Scores according to SAFMs	Scores according to SAFs	Total score for FVGSC preparation	Number of errors
1	Fail	Fail	Fail	11
3	Fail	Fail	Fail	10
4	Fail	Fail	Fail	12
5	Fail	Fail	Fail	12
7	Fail	Pass	Fail	9
13	Fail	Fail	Fail	13
14	Fail	Pass	Fail	8
18	Fail	Fail	Fail	10
20	Fail	Pass	Fail	6
21	Fail	Fail	Fail	9
25	Fail	Pass	Fail	8
26	Fail	Fail	Fail	12
29	Fail	Fail	Fail	8
31	Fail	Fail	Fail	10
51	Fail	Pass	Fail	5
52	Fail	Pass	Fail	7
54	Fail	Fail	Fail	10
57	Fail	Pass	Fail	6
58	Fail	Fail	Fail	12
59	Fail	Fail	Fail	12
60	Fail	Fail	Fail	11
63	Fail	Fail	Fail	10
67	Fail	Pass	Fail	7
69	Fail	Fail	Fail	9
70	Fail	Fail	Fail	7
71	Fail	Pass	Fail	5
73	Fail	Fail	Fail	8
74	Fail	Fail	Fail	8
78	Fail	Pass	Fail	7
88	Fail	Fail	Fail	14

According to Knight (1997), any model which has no errors should be classified as a pass. The total number of the models which have a pass score was zero tooth preparations.

Step 3: *comparing the tooth preparations with grades awarded by the best senior academic staff examiner.*

Class II amalgam cavity

The following Table 6.41 shows objective scores for 26 class II amalgam cavity preparations. These scores were compared with grades awarded from the best senior examiner on the first occasion. Grades of the best examiner were converted into fail and pass scores. To accept or reject the null hypothesis, converted scores (pass and fail) from grades of three senior academic staff examiners on the first and second occasion and average values of the both occasions' were selected. All these grades were changed into pass and fail in order to compare them with more objective scores for the class II amalgam cavity preparation.

Table 6.41 Comparison of A) the Developed Standard scores for the class II amalgam cavities with the scores (derived from grades) of B) the best examiner, C) three senior examiners on two separate occasions and D) three senior examiners for two combined occasions

Model number	A) Developed Standard score for class II amalgam cavity	B) Grades and scores of the best examiner on the first occasion		C) Average grades and scores of three senior examiners for each of two occasions				D) Average grades and scores of three senior examiners for two combined occasions	
		Grade	*Score*	*Grades of occasion one*	*Score 1*	*Grades of occasion two*	*Score 2*	*Grade*	*Score*
5	Fail	3	Fail	3	Fail	2	Fail	3	Fail
8	Fail	2	Fail	3	Fail	3	Fail	3	Fail
15	Fail	4	Pass	4	Pass	4	Pass	4	Pass
16	Fail	2	Fail	4	Pass	4	Pass	4	Pass
36	Pass	5	Pass	5	Pass	5	Pass	5	Pass
39	Fail	5	Pass	5	Pass	5	Pass	5	Pass
40	Fail	4	Pass	4	Pass	4	Pass	4	Pass
41	Fail	3	Fail	4	Pass	4	Pass	4	Pass
43	Fail	3	Fail	4	Pass	3	Fail	3	Fail
46	Fail	5	Pass	5	Pass	5	Pass	5	Pass
53	Fail	2	Fail	3	Fail	3	Fail	3	Fail
54	Fail	4	Pass	3	Fail	4	Pass	4	Pass
57	Fail	2	Fail	3	Fail	3	Fail	3	Fail
62	Fail	3	Fail	3	Fail	3	Fail	3	Fail
73	Fail	2	Fail	2	Fail	2	Fail	2	Fail
78	Fail	3	Fail	3	Fail	3	Fail	3	Fail
80	Fail	4	Pass	4	Pass	4	Pass	4	Pass
83	Fail	5	Pass	4	Pass	5	Pass	5	Pass
85	Fail	2	Fail	3	Fail	3	Fail	3	Fail
87	Fail	4	Pass	4	Pass	5	Pass	5	Pass
88	Fail	2	Fail	2	Fail	2	Fail	2	Fail
94	Fail	2	Fail	2	Fail	2	Fail	2	Fail
109	Fail	5	Pass	4	Pass	4	Pass	4	Pass
111	Fail	2	Fail	2	Fail	2	Fail	2	Fail
120	Fail	4	Pass	5	Pass	4	Pass	5	Pass
138	Fail	2	Fail	2	Fail	2	Fail	2	Fail
Agreement percentage with objective score			**62%**		**54%**		**54%**		**54%**

Table 6.41 shows that agreement percentage of the best examiner grades (scores) on the first occasion with objective score for 26 class II amalgam cavity preparations was 62%. It was the highest agreement percentage between objective scores and other grades. From this result, grades awarded from the best senior examiner on the first occasion were the best grades for the 26 class II amalgam cavities.

In order to determine the level of agreement between objective scores from SAFMs and SAFs with:

- the best examiner scores,
- the scores on the first and second occasion the three senior examiners and,
- the average scores of the three senior examiners,

Cohen's kappa coefficient test (SPSS) was used (Table 6.42).

Table 6.42 Cohen's kappa coefficient test between objective scores from SAFMs and SAFs with i) the best examiner scores and ii) scores of the examiners on the first and second occasions and iii) the average scores of the three examiners

<table>
<tr><th colspan="2"></th><th>Objective scores from SAFMs and SAFs (Developed Standard scores)</th><th>Significance ($p \leq 0.05$)</th></tr>
<tr><td colspan="2">The best senior examiner scores</td><td>0.103</td><td>0.103</td></tr>
<tr><td rowspan="3">scores of the three senior examiners</td><td>Occasion one scores</td><td>0.077</td><td>0.308</td></tr>
<tr><td>Occasion two scores</td><td>0.077</td><td>0.308</td></tr>
<tr><td>Average scores</td><td>0.077</td><td>0.308</td></tr>
</table>

Table 6.42 showed that the agreements were not significant. According to Landis and Koch (1977), agreement was slight for the best examiner scores whereas fair for the three senior examiners.

Full veneer gold shell crown preparation

From Table 6.43, more objective score for the full veneer gold shell crown preparation compared with grades awarded from the best senior examiner on the first occasion. The grades of the best senior examiner were converted into fail and pass. To accept or reject the null hypothesis, grades awarded from three senior examiners on the first and second occasion and averaged values of both occasions' grades were selected. All these grades were changed into pass and fail scores in order to compare them with more objective scores for the full veneer gold shell crown preparation.

Table 6.43 Comparison of A) the Developed Standard scores for the full veneer gold shell crown preparations with the scores (derived from grades) of B) the best examiner, C) three senior examiners on two separate occasions and D) three senior examiners for two combined occasions

Model number	**A) Developed Standard score for FVGSC preparation**	**B) Grades and scores of the best examiner on the first occasion**		**C) Average grades and scores of three senior examiners for each of two occasions**				**D) Average grades and scores of three senior examiners for two combined occasions**	
		Grade	*Score*	*Grades of occasion one*	*Score*	*Grades of occasion two*	*Score*	*Grade*	*Score*
1	Fail	3	Fail	3	Fail	3	Fail	3	Fail
3	Fail	4	Pass	3	Fail	4	Pass	4	Pass
4	Fail	3	Fail	3	Fail	3	Fail	3	Fail
5	Fail	3	Fail	3	Fail	4	Pass	4	Pass
7	Fail	4	Pass	4	Pass	4	Pass	4	Pass
13	Fail	3	Fail	3	Fail	3	Fail	3	Fail
14	Fail	5	Pass	4	Pass	4	Pass	4	Pass
18	Fail	3	Fail	3	Fail	3	Fail	3	Fail
20	Fail	5	Pass	4	Pass	4	Pass	4	Pass
21	Fail	5	Pass	4	Pass	4	Pass	4	Pass
25	Fail	3	Fail	3	Fail	4	Pass	4	Pass
26	Fail	2	Fail	2	Fail	2	Fail	2	Fail
29	Fail	3	Fail	3	Fail	3	Fail	3	Fail
31	Fail	5	Pass	4	Pass	4	Pass	4	Pass
51	Fail	5	Pass	4	Pass	4	Pass	4	Pass
52	Fail	4	Pass	3	Fail	4	Pass	4	Pass
54	Fail	3	Fail	3	Fail	3	Fail	3	Fail
57	Fail	5	Pass	4	Pass	4	Pass	4	Pass
58	Fail	3	Fail	3	Fail	3	Fail	3	Fail
59	Fail	2	Fail	2	Fail	2	Fail	2	Pass
60	Fail	3	Fail	3	Fail	3	Fail	3	Fail
63	Fail	3	Fail	3	Fail	4	Pass	3	Fail
67	Fail	4	Pass	4	Pass	4	Pass	4	Pass
69	Fail	4	Pass	4	Pass	4	Pass	4	Pass
70	Fail	3	Fail	3	Fail	3	Fail	3	Fail
71	Fail	5	Pass	4	Pass	4	Pass	4	Pass
73	Fail	4	Pass	4	Pass	3	Fail	4	Pass
74	Fail	3	Fail	3	Fail	4	Pass	4	Pass
78	Fail	5	Pass	4	Pass	4	Pass	4	Pass
88	Fail	2	Fail	2	Fail	2	Fail	2	Fail
Agreement percentage with objective score			**53%**		**60%**		**43%**		**40%**

Table 6.43 demonstrated that agreement percentage of the best examiner grades or scores with more objective scores for 30 full veneer gold shell crown preparations was 53%. The highest agreement percentage between objective scores and average scores of three senior academic staff examiners on the first occasion was 60%. From this result, average grades awarded from three senior academic staff examiners on the first occasion have the best agreement with objective scores. Although averaged values of three senior examiner staff grades were more representative than the best senior examiner grades, the agreement percentage was still low.

In order to determine the level of agreement between objective scores from SAFMs and SAFs with:

- the best examiner scores,
- the scores on the first, second occasion for the three senior examiners and,
- the average scores of the three senior examiners,

Cohen's kappa coefficient test (SPSS) was used (Table 6.44).

Table 6.44 Cohen's kappa coefficient test between objective scores from SAFMs and SAFs with i) the best examiner scores, ii) scores of the three senior examiners on the first and second occasions and iii) the average scores of the three senior examiners

		Objective scores from SAFMs and SAFs (Developed Standard scores)
The best examiner scores		0.000
Scores of the three senior examiners	Occasion one scores	0.000
	Occasion two scores	0.000
	Average scores	0.000

Table 6.44 showed that no statistics were computed by using SPSS because objective scores from SAFMs and SAFs were constant.

6.5 Discussion

Accurate evaluation of dental student work in a clinical skills laboratory is a most critical component of the dental education process. Assessment should provide consistent and accurate feedback for students in order to assist them to achieve a high level of competency before working on real patients (Renne et al., 2013).

The traditional method, at most dental schools, is still assessment of students' pre-clinical dental work using visual inspection by examiners who are often experienced clinical specialists (Cardoso et al., 2006). According to several studies (Sherwood and Douglas, 2014, Kateeb et al., 2016, and Zou et al., 2016), it is difficult to guarantee reliable grades and consistent feedback for the student using visual assessment alone and this method does not generally agree with objective evaluations of tooth preparations. According to several authors (Natkin and Guild, 1967, Houpt and Kress, 1973, Helft et al., 1987, Sharaf et al., 2007, Kateeb et al., 2016) the most common reasons for this lack of agreement are:

- Misuse of a grading scale,
- Lack of calibration of examiners
- Insufficient training of examiners,
- The need for a combination of objective and subjective evaluations, and
- Misinterpretation of the component of any assessment tool (e.g. checklist criteria).

These matters have been also investigated by other researchers in order to find a better way to analyse tooth preparations completed by dental students and thus provide more accurate and consistent feedback and assessment (Cardoso et al., 2006, Haj-Ali and Feil, 2006, Sharaf et al., 2007, Kateeb et al., 2016). These researchers used measurements and features which were recommended in the literature as a Developed Standard to analyse the tooth preparations. Figure 6.13 summarises the principle findings of Chapter 6.

Step 1: Identify reliable and repeatable measuring method for class II amalgam cavity and full veneer gold shell crown preparations:

Un-reliable measurements for class II amalgam cavity:
- Bucco-palatal width of the box occlusally,
- Isthmus width occlusally,
- Occlusal cavity width in the middle,
- Marginal ridge thickness,
- Occlusal cavity depth at buccal side in the middle, and
- Occlusal cavity depth at distal side.

Reliable measurements for class II amalgam cavity:
- Bucco-palatal width of the box floor
- Depth of the box gingivally,
- Box floor (mesio-distal) depth
- Isthmus floor width
- Pulpal axial wall length
- Occlusal cavity depth at palatal side

For full veneer gold shell crown preparation, all measurements of full veneer gold shell crown preparation were reliable.

Step 2: Comparison of the SAFMs with values recorded in the literature and evaluation of the SAF according to the most agreed binary decision between senior academic staff examiners:

Class II amalgam cavity preparation:

For SAFMs comparing with recommended measurements for the class II amalgam cavity preparation:
- Most of the contact points between the box cavity and adjacent tooth were not cleared bucco-palatally.
- Most of the cavity floor was not flat. At the isthmus area, the depth was shallow for almost cavities.
- According to SAFMs, only one cavity passed the entire feature of the cavity objectively (Model No. = 36).

For SAFs of class II amalgam cavity preparation:
- There was no retention form on the proximal box according to two objective measurements. This SAF was the most prominent SAFs which reduced the number of passed models.
- According to SAFs, there were six cavities passed the entire feature of the cavity (No. models = 36, 39, 40, 57, 78, 111).

The result of comparison between scores of SAFM (Pass or fail) with subjective SAF evaluations scores to determine a more objective score for 26 tooth cavities was only one tooth passed objectively (model No. = 36).

Full veneer gold shell crown preparation:

For SAFMs comparing with recommended measurements for the full veneer gold shell crown preparation:
- Most of the finish line depth was uneven.
- Ideal or acceptable depth on the buccal and the palatal sides was easier to achieve than on the mesial and the distal sides, especially on the buccal side, was the easiest. On the other hand, finish line distally was the hardest to achieve.
- The most of the occlusal reduction was non-symmetrically.
- Occlusal reduction on the palatal cusps was more than on the buccal cusps.
- Most of the axial reduction was non-symmetrically.
- The worst reduction was the axial reduction, especially on the mesial and distal sides.

According to SAFMs, no preparation passed the entire feature of full veneer gold shell crown preparation objectively.

For SAFs of full veneer gold shell crown preparation:
- The most prominent feature was given fail was finish line of tooth preparation chamfer shape or not.
- According to SAFs, there were 13 tooth preparations passed the entire feature of the cavity subjectively (model number 7, 14, 20, 21, 25, 51, 52, 57, 67, 69, 70, 71 and 78).

The result of comparison between scores of SAFM (Pass or fail) with subjective SAF evaluations scores to determine a more objective score for 30 tooth preparations was no model passed objectively.

Step 3: Comparing the tooth preparations scores (Pass/Fail) from the step 2 with grades (converted to Pass/Fail scores) awarded by the best senior academic staff examiner.

Class II amalgam cavity preparation:

Agreement percentage (62%) was low between objective scores and grades awarded by the best senior academic staff examiner, although it was the highest agreement in comparison with average grades awarded from three senior examiners.

Full veneer gold shell crown preparation:

Agreement percentage (53%) was low between objective scores and grades awarded by the best senior academic staff examiner. Additionally, the highest agreement percentage (60%) was for average grades awarded from three senior staff academic examiners.

Figure 6.13 Outline of the principle findings of Chapter 6

From Chapter 5, examiner agreement was improved by using a feedback sheet. In addition, the most reliable examiners for both class II amalgam cavity and full veneer gold shell crown preparation were established. In this chapter, examiner 3 and examiner 1 were confirmed as the best examiner for the class II amalgam cavity and the full veneer gold shell crown preparation, respectively, by correlating their grades and negative points. From these results a question arose. Can the grades with good agreement and reliability from these examiners be used as gold standard grades?

To identify the answer to this question, grades awarded from the best examiners should be compared with gold standard grades. This method is called calibration. Most of the studies in this area have used measurements of tooth preparations recorded by the researchers or examiners and compared them with ideal, or acceptable, measurements reported in the literature (Jokstad and Mjor, 1987, Haj-Ali and Feil, 2006, Sharaf et al., 2007, Yoon et al., 2014, Tiu et al., 2015, Kateeb et al., 2016). Evaluation of the methodology used in these studies, failed to demonstrate a clear and consistent method by which objective measurements of tooth preparation, recorded by researchers, were compared with ideal, or acceptable, measurements reported in the literature.

Therefore, a plan was devised to compare the best examiner evaluations with developed standard evaluations from acceptable (which includes ideal) measurements reported in the literature in order to determine the amount of agreement between the two parameters. Thus, do the best examiner grades offer a true reflection of acceptable (which includes ideal) measurements reported in the literature?

The gold standard alluded to in the previous paragraphs is itself not clearly defined. In recent years, calibration has become the main aim for almost all studies in this area. Most of the assessment tools developed were calibrated in order to provide accurate and consistent evaluation and feedback for the student (Mays and Branch-Mays, 2016). To calibrate assessment tools with examiner's grades, gold standard grades are essential.

Therefore, how can gold standard grades, which represent the tooth preparations truly, be determined?

In order to determine gold standard grades, several studies selected the examiner who had the highest specialty or greatest experience, as a gold standard against which to calibrate other examiners (Curtis et al., 2008, Cho et al., 2010, Mays and Levine, 2014, Tuncer et al., 2015, Alhumaid et al., 2016). Few studies used an averaged value from all examiners (Cho et al., 2010, Callen et al., 2015). Authors of these studies selected gold standard examiner and grades without determining whether or not these grades reflected the tooth preparations truly.

In addition, the researcher (AM) in this thesis faced with the problem of developing such a standard based on the published literature which described acceptable tooth preparations. This cannot be called a 'Gold Standard'. Therefore, the phrase used in this thesis will be 'Developed Standard'. Ultimately, this Developed Standard may acquire 'gold' status but this will be for others to judge. Furthermore, the results from this part of thesis have shown that it is not a simple matter of selecting experienced examiners or an average value of examiners as both methods may have a profound impact on the gold standard grades. The way to identify a standard examiner is to increase the use of objective measurements of tooth preparations and decrease the use of subjective evaluations and also encourage the use of binary (yes/no) responses to these subjective evaluations.

The objective measurements of tooth preparations to provide consistent and accurate feedback and grades for students can be achieved by using technology (Kateeb et al., 2016). Most of these tools were calibrated using specific anatomical feature measurements (SAFMs). Examples of these devices include, E4D compare software (Renne et al., 2013), DentSim by Denx (Rose et al., 1999, Jasinevicius et al., 2004, Esser, et al., 2006, Welk et al., 2008), Virtual Reality Dental Training System by Novint

(Buchanan, 2001, Jasinevicius et al., 2004), PreAssistant by Kavo (Arnetzl and Dornhofer, 2004, Kournetas et al., 2004, Cardoso et al., 2006) and the Cavity Preparation Skill Evaluation System (CPSES) (Zou et al., 2016).

These tools, and their associated software, provided 70% of feedback and grades for dental students when using specific anatomical feature measurements (SAFMs) for full crown preparations. These systems cannot make subjective evaluations, for example, assessment of clearance and damage to adjacent teeth (Arnetzl and Dornhofer 2004, Cardoso et al., 2006).

Therefore, in this thesis scanning methodology, image processing and direct measurements were the methods used to record specific anatomical feature measurements (SAFMs) for both the class II amalgam cavity and the full veneer gold shell crown preparations. The results demonstrated good reliability and repeatability of these measurements.

Previous studies (Buchanan, 2001, Jasinevicius et al., 2002, Arnetzl and Dornhofer, 2004, Cardoso et al., 2006, Haj-Ali and Feil, 2006, Kateeb et al., 2016 and Zou et al., 2016) have suggested that both objective and subjective evaluations should be used to complement each other in order to provide a more accurate assessment. The combined use of both forms of evaluation can often exploit the advantages of each. For example, retention form and damage to the adjacent tooth for the class II amalgam cavity can only be evaluated subjectively. Because of that specific anatomical features of tooth preparations were divided into objective (SAFMs) and subjective (SAFs) evaluations. Furthermore, subjective anatomical features (SAFs) were designed to be binary (yes/no) decisions.

There were also several studies which utilised the SAFMs to compare these objective evaluations with values recommended in the literature (Sato et al., 1998, Patel et al., 2005, Ayad et al., 2005, Yoon et al., 2014, Tiu et al., 2014, Tiu et al., 2015). The values recommended in the literature constitute what passes for a gold standard. These values often fall within an acceptable range rather than one specific value. According to Knight (1997), if there are no errors in a preparation then the preparation should be defined as ideal or acceptable and therefore be designated a 'pass' grade. Errors present, defined as SAFMs which fall outside the recommended range within the literature, in preparations should be defined as non-ideal or not-acceptable and therefore be designated a 'fail' grade. Therefore, Developed Standard scores (pass / fail) can be determined by using SAFMs.

The difficulty of incorporating subjective evaluations into tooth preparation assessment is real. There is scant literature reporting clearly-defined values or ranges for subjective evaluations of tooth preparations. Often, phrases are used such as, *"Maintaining the bur parallel to the long axis of the tooth crown creates facial, lingual, and distal walls with a slight occlusal convergence, which provides favourable amalgam angles at the margins…. The occlusal convergence of the facial and lingual walls and the dovetail design (if needed) provide sufficient retention form to the occlusal portion of the tooth preparation"*, (Roberson et al., 2002). From the scant literature, a list of SAFs was drawn up for each type of tooth preparation and, defining when a specific subjective evaluation (SAF) was present relied on agreement between two or more examiners. Once again, Knight (1997) was used to identify ideal or acceptable tooth preparations using SAF criteria defined in this way.

In this part of the study, the Developed Standard scores for 26 Class II amalgam cavities and 30 full veneer gold shell crown preparations were determined from SAFMs made by the researcher (AM) and compared with SAFMs recommended in the literature and from the agreed, subjective, SAFs. Thus, defined pass / fail scores could be determined. The next stage was to determine the level of agreement with the grades awarded by the senior academic staff examiners.

Reporting range of SAFMs for each feature from literature is essential to identify the Developed Standard score for each tooth preparation. For example, the ideal occlusal convergence (TOC) angle varied in different studies. Smith and Howe (2007) suggested that when the TOC angle for the full veneer gold shell crown preparation exceeds 30°, loss of retention becomes common. They recommended a TOC angle of 7° as the best angle to produce maximum retention with minimum cement film thickness. This angle is very difficult to achieve without producing some undercuts and damage to adjacent teeth. Indeed, the human eye cannot detect the difference between a parallel preparation and 10° angle. Therefore, the taper of posterior teeth preparation that have been successful is approximately 20° (Smith and Howe, 2007). Goodacre et al., (2001) reviewed several papers and concluded that a 10° – 20° angle was as an ideal total occlusal convergence angle. In addition, other studies recommended TOC angle values which, based on in vitro testing, have ranged from as low as 3° to 14° for optimal retention and resistance form (Jørgensen, 1955, Gilboe and Teteruck, 1974, Ohm and Silness, 1978, Johnston et al., 1986, Wilson and Chan, 1994, Shillingburg et al., 2012). Therefore, the widest range 3° to 20° was selected in this study as an acceptable range for total occlusal convergence (TOC) angle.

After reporting acceptable range measurement for each feature from the literature, one or two different methods were used to measure Specific Anatomical Features (SAFMs) for the class II amalgam cavity and full veneer gold shell crown preparation. Three different statistical analyses were used to determine reliable SAFMs for both type of tooth preparations.

Several studies have been described and reported paired sample t-tests as an initial analysis of presented data in order to detect the systematic error of two series of measurement. Houston, (1983) suggested 25 models as a minimum sample number to determine the reliable method or measurement. Therefore, all samples of both type of tooth preparation were used to identify reliable SAFMs. In addition, the standard deviation was also reported to quantify the amount of variation (Random difference/error) of a set of two or more measurements (Bland and Altman, 1996). The most common reason for this variation (systematic and random errors) is that many landmarks of tooth preparation were difficult to identify, and author's opinion about the exact location of the point may vary at random (Taylor et al., 2013). The main disadvantage of paired sample t-test is that comparing two means of two groups. The means of two set of measurements sometimes can be equal while the (random) differences between measurements can be huge (Chhapola et al., 2015). Therefore, intra-class correlation was used to determine the reliable measurement for each feature.

Using intra-class correlation was better suited to determine the direction of any differences between datasets may take. However, a high correlation, by using Intra-class Correlation Coefficient (ICC), does not necessarily imply that there is a high agreement between two measurements for one or more different measuring methods. In addition, the Intra-class Correlation Coefficient (ICC) fails to provide information on the type of association between the measurements. It may show excellent correlation despite the presence of significant systematic difference. The correlation cannot distinguish between the random or systemic differences in two measurements (Van Stralen et al., 2008). In this part of study, most of the data had a high positive correlation. According to Rankin and Strokes, (1998), "*The intra-class correlation and Bland and Altman tests are appropriate for analysis of reliability studies of similar design to that described, but neither test alone provides sufficient information and it is recommended that both are used*". Therefore, Bland and Altman plots were constructed to evaluate the agreement and systemic differences between two measurements for the same method or for two different methods (Bland and Altman, 1999). From result of these statistical analyses, the reliable measuring method and measurement points for the class II amalgam cavity and measurement points for the full veneer gold shell crown preparation were reported. Almost all of SAFMs were reliable for the class II amalgam cavity preparation by using MeshLab software whereas all SAFMs of full veneer gold shell crown preparation were reliable by using ImageJ software. On the other hand, there were some specific anatomical features (SAFs) which could not be measured using previous methods. For example, measuring of the finish line depth for full crown tooth preparation was very difficult because of the inability to select specific measuring points on a rounded tooth surface at the margins (Beschnidt and Strub, 1999) (Figure 6.14).

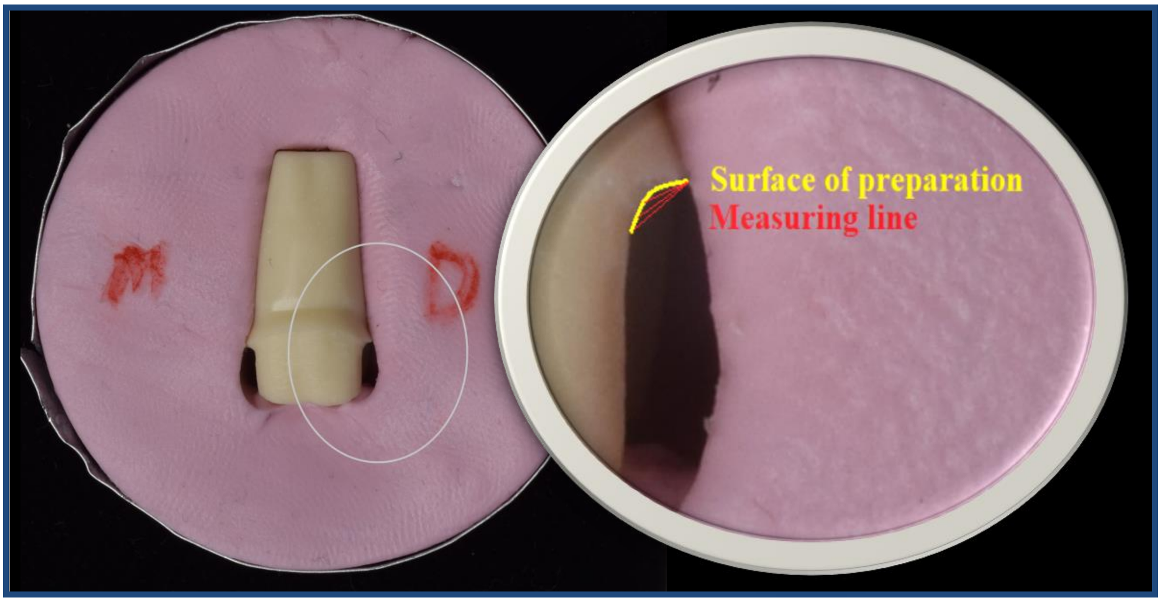


Figure 6.14 Photograph to show the difficulty of measuring chamfer finish line depth

The most agreed decision of a dichotomous scale for each SAF evaluation among three senior academic staff examiners was selected as the agreed SAF evaluation. A dichotomous scale was chosen, because two point rating scales produced scores which were more valid and reliable (Houpt and Kress, 1973). Houpt and Kress (1973) compared two point rating scales and five-point rating scales. The result of their study showed that using two point rating scale had greater inter-examiner agreement than the five-point rating scale. In addition, a two point rating scale increased the reliability and validity (Houpt and Kress, 1973). There were other studies which concluded similar results (Hinkelman and Long, 1973, Deranleau et al., 1983). If an individual had an extremely positive or negative attitude toward an object, a dichotomous scale (e.g. ‘‘Yes’’ or ‘‘No’’) easily permitted reporting that attitude (Krosnick and Presser, 2010). Therefore, binary decisions (Yes/No) were indicated for SAFs in this part of the study.

Reliable SAFMs were compared with the acceptable range of measurements which was recommended in the literature. This comparison was presented graphically. The purpose of these graphs were to compare clinically achieved tooth preparations for class II amalgam cavities and full veneer gold shell crown preparations by dental students with recommended values in the literature. There were few studies which used a similar way to determine this comparison for specific tooth preparation features (Seo et al., 2014, Tiu et al., 2015). According to Tiu et al., (2015), this way used to compare specific anatomical features (i.e. total occlusal convergence angle and finish line) of full ceramic crown preparations which were prepared by experienced examiners using the measurements recommended in the literature. For class II amalgam cavity preparation, there was much less literature available for this type of preparation but the same methodology was used to SAFMs of class II amalgam cavity preparations with the few recommended measurements in the literature. Comparison of measurements using a graph was the simplest way to determine the Developed Standard score for each feature. These scores were used to calibrate the grades of the best senior examiner.

Graphically, if every mean for each reliable SAFM from MeshLab software (indirect measuring) method was between acceptable ranges of measurement in the literature, a passing score was given. According to Knight (1997), a passing score requires that every criterion or measurement of the criterion be clinically acceptable and within acceptable measurements in the literature. Even if there was only one criterion measurement which did not meet the standard, the preparation did not receive a pass score overall (Knight, 1997). This step was used to identify the level of pass-fail scores for each SAFM.

For class II amalgam cavity preparation, the findings showed that students struggled to prepare teeth with acceptable measurements and features. The comparison of SAFMs for class II amalgam cavity preparation demonstrated that most of the contact points between the box cavity and adjacent tooth were not cleared bucco-palatally, and the cavity depth was too shallow at the isthmus area (Appendix 7).

The comparison of SAFMs for full veneer gold shell crown preparations demonstrated most of the occlusal reduction was non-symmetric. An acceptable occlusal reduction on the buccal cusps was easier to achieve than on the palatal cusps. The axial reduction was also non-symmetric with the worst reduction being over-reduction on the mesial and distal sides. Most of the full veneer gold shell crown preparations have an acceptable proximal occlusal convergence angle (mesio-distal direction). In addition, the mean of the bucco-lingual convergence angle for full veneer gold shell crown preparations were higher values compared to means of mesio-distal convergence angle (Appendix 8). Several studies had reported that the bucco-lingual convergence angle of tooth preparation was higher than the mesio-distal convergence angle (Ohm and Silness, 1978, Al-Omari et al., 2004, Ayad et al., 2005). Some of these results supported the Chapter 2 findings.

For specific anatomical features (SAFs) which cannot be measured, the Developed Standard evaluation for each non-measured SAFs of tooth preparation was determined by selecting the highest criterion agreement of SAF from three senior examiners. The answers for these decisions were Yes or No for each SAF. If the most-agreed decision for SAF for each tooth preparation was consistent with recommended acceptable features described in the literature, the particular SAF was designated as a Pass score. If it was not consistent with the literature then the feature was designated as a Fail score. This ranking form was based on the recommendation by Knight (1997).

Some of the SAF evaluations of class II amalgam cavity preparations which were part of the ‘Gray feedback sheet’ could be confirmed using reliable, objective SAFM data, for example, confirming ‘Retention form’ of the box by two reliable SAFMs. These two measurements were used to objectively determine if the box preparation was acceptable (undercut or parallel walls = YES) or not (divergent walls = NO) (Table 6.27). There was no retention form on the proximal box for the majority of the class II amalgam cavities. This SAF evaluation was the most commonly-occurring SAF to reduce the overall number of models with passing scores.

The comparison was made between the retention forms from i) objective evaluation using two reliable SAFMs and ii) the retention form from the binary decisions reported from the three senior academic staff examiners (Table 6.27). A sub-set from this table demonstrated that examiners did not recognise a measurement difference of up to 0.56 mm clinically between the two SAFMs as an acceptable retention form (Table 6.28). This supports the argument that, if the variance between the two measurements at the base of the box and the marginal ridge was less than 0.50 mm, it was very difficult to recognise this clinically. While this measured difference of 0.56 mm was difficult for the senior examiners in this thesis to detect, this might also account for the widely reported range of +/- 0.50 mm often reported in the literature for various aspects of cavity preparation. For example, the ideal depth of the occlusal cavity is between 1.50 to 2.00 mm (Roberson et al., 2002, Haj-Ali and Feil, 2006, Hilton et al., 2013).

For the class II amalgam cavity preparations, there was only one cavity that satisfied all the criteria based on SAFM evaluations while six cavities satisfied all the criteria for all SAF evaluations. For the full veneer gold shell crown preparation, there were no preparations that were acceptable based on SAFM evaluations while seven preparations satisfied all the criteria for all SAF evaluations according to Knight's recommendation (Knight. 1997).

The scores of SAFMs and SAFs for the class II amalgam cavity preparation and full veneer gold shell crown preparation were reported. From these scores, the total score for each tooth preparation was also determined according to the Knight (1997) ranking system. According to Knight (1997), the passing of tooth preparation requires that every feature is clinically acceptable (Pass score) according to recommendations in the literature. Thus any unacceptable feature constitutes an overall fail for the tooth preparation. Pass or Fail scores for each tooth preparation were called Developed Standard scores (Tables 6.41 and 6.43).

Thus, taking into account both the objective measurements (SAFM) and the subjective features (SAF), only one class II amalgam cavity preparation passed and no full veneer gold shell crown preparations passed.

Developed Standard scores of the tooth preparations were reported. The purpose of reporting Developed Standard scores was to calibrate the grades awarded from the best senior examiner with these Developed Standard scores to determine whether the best senior examiner grades reflected tooth preparations truly or not. In other words, using these criteria, there was an exercise to confirm the best senior examiner.

In this part of study, the grades of the best senior examiner were also converted into Pass and Fail scores before comparing with Developed Standard scores. The reason for converting the best senior examiner grades to Pass and Fail scores was to determine the agreement percentage between the two sets of data. For class II amalgam cavity preparations, the agreement percentage was low (62%) between objective pass/fail scores and pass/fail scores awarded by the best senior academic staff examiner on the first occasion. This was better than the percentage agreement using pass/fail scores derived from average grades awarded by three senior examiners (54%). For the full veneer gold shell crown preparation, agreement percentage was also low (53%) between objective pass / fail scores and pass/fail scores awarded by the best senior academic staff examiner on the first occasion. However, for this type of preparation, a higher agreement percentage (60%) was derived from average grades awarded by three senior examiners. Thus, for these examiners and for these preparations it would appear that there was no consistency between whether the best examiner is determined or an average of several examiners was used when compared with a Developed Standard. Ideally, this observation should be tested with another group of examiners but, if this is a generalisable observation then studies reported in the literature which select a gold standard examiner based on, for example, length of previous experience, may be fundamentally flawed. The results of this thesis supported the studies of Lilley et al., (1968), Hinkleman and Long (1973), Deanleau et al., (1983) and Jenkins et al., (1998). These authors concluded that pass-fail differences seemed to be unrelated to the experience of the examiner.

The variation described in the above paragraph has been reported previously in the literature. Jenkins et al., (1998) reported that there was a great variation between examiners assessing class II amalgam cavity preparations using a, ‘glance and grade’ method. Preparations that were initially given a passing mark, were scored as a failure after a second evaluation (Jenkins et al., 1998). In this part of study, the level of pass-fail difference was reported between grades of the best senior examiner and scores from SAFMs and SAFs (Developed Standard Scores).

In this thesis, the percentage agreement between pass/fail scores derived through a combination of SAFMs and SAFs and pass/fail scores derived by combinations of examiners for both the class II amalgam cavity and full veneer gold shell crown preparation was low. The level agreement determined by Cohen’s Kappa confidence test showed that the agreements were not significant. According to Landis and Koch (1977), agreement was fair and slight for the best examiner scores only.

In this study, repeated evaluation of tooth preparations by any examiner (senior or otherwise) tended to result in an increased number of passing scores compared with the Developed Standard. Thus, over time all examiners had a tendency to become more dove-like. The researcher (AM) has not been able to find any literature to support this although there is literature to support the fact that examiner behavior does change when multiple evaluations of the same tooth preparations are compared (Renne et al., 2013). It would seem prudent therefore, to have a fixed, gold standard against which to describe differences between examiners or change in examiner status (e.g. from hawk to dove). Such problems have also been addressed by Knight (1997) who recommended that passing criteria should be defined and differentiated from criterion not met within a checklist. Therefore, the passing criteria reflect the preparation truly and will provide accurate feedback. In addition, Haj-Ali and Feil (2006) concluded that, training using a gold standard, inter-examiner reliability can be improved. So, calibrated assessment tools are an essential part to create an objective assessment.

6.6 Conclusion

Dundee Dental School did not provide for dental students sufficiently accurate tooth preparation guidelines and sufficiently acceptable measurements of the tooth preparations at the Clinical Skills Laboratory before commencing tooth preparation (see Chapter 2). Thus, Dundee dental students struggled to prepare teeth as recommended in the literature.

In addition, it was very difficult to select the best examiner grades/scores or even an average grade/scores from three examiners as a standard which reflected the tooth

preparation truly without calibration. Therefore, calibration of the assessment method is essential to provide more accurate and consistent feedback and grades for the student.
The ‘Gray and Mhanni feedback sheet’ enhanced the repeatability and reproducibility of the examiners grades but they did not truly reflect the entirety of the tooth preparations consequently. Lack of standardization (calibration) in the feedback sheets was identified as a major reason why assessments have been considered as un-reliable and un-fair for students. According to Renne et al., (2013), *“It is widely agreed faculty members should be calibrated in an attempt to overcome variability in assessment”.*
The findings of both types of tooth preparations were consistent with the null hypothesis that all grades awarded by the best senior academic staff examiner did not agree with the passing and failing grade features for tooth preparations which was reported in the literature and measured objectively by the researcher (AM). Thus, the null hypothesis of this chapter was accepted. From this chapter, Appendix 9 summarises the steps to determine the grades of the best examiner and whether or not they can be selected as standard grades at Dundee Dental School.

In the next chapter, a new measuring tool to assess student performance in tooth preparation, will be developed, in order to improve intra- and inter-examiner agreement grades and consistency. Developed Standard scores from SAFM- and SAF- evaluations for 26 class II amalgam cavities and for 30 full veneer gold shell crown preparations will be utilised to determine the ascertain whether or not these are truly representative of the key aspects of tooth preparations. These scores and their number of negative points (Tables 6.35 and 6.40) will be used in order to determine standard setting (Cut-off point) of the checklist to justify the passing score in order to maintain a valid and reliable assessment (Puryer and O'Sullivan, 2015).

Chapter 7 : Identification of reliable tools and development of a new checklist to assess class II amalgam cavities and full veneer gold shell crown preparations

	Chapters
Chapter 2	**Evaluating current assessment and feedback methods at Dundee Dental School in the Clinical Skills Laboratory** **Aims:** The aims of this Chapter were to identify, in relation to the class II amalgam cavity preparation and the full veneer gold shell crown preparation undertaken in the Clinical Skills Laboratory at Dundee Dental School: 1. the types and quality of feedback and assessment provided for students and, 2. the disadvantages associated with current feedback.
Chapter 3	**Independent evaluation of student class II amalgam cavity and full gold shell-crown preparations by senior academic staff** **Aim:** The aim of this study was to evaluate inter-examiner agreement between a convenience sample of three senior academic staff members, who have more than twenty years of clinical and teaching experience each, when evaluating operative procedures (class II amalgam cavity and full veneer gold shell crown preparation) on plastic teeth, from a sample year-cohort of dental students, at Dundee Dental School.
Chapter 4	**Evaluation of selected student class II amalgam cavity and full gold shell-crown preparations by senior academic staff and additional teaching staff** **Aims:** 1. To develop a sub-set of class II amalgam cavity, and full veneer gold shell crown, preparations. 2. To determine intra-examiner agreement for a group of senior academic staff when evaluating a sub-set of class II amalgam cavity preparations and full-veneer gold shell crown preparations. 3. To study a large group of additional teaching staff at Dundee Dental School in order to determine intra- and inter-examiner agreement when evaluating a sub-set of class II amalgam cavity preparations and full-veneer gold shell crown preparations.
Chapter 5	**Evaluation of selected student class II amalgam cavity and full gold shell crown preparations by using specific additional tools and feedback sheets** **Aims:** 1. To identify and develop specific additional tools to assist with evaluation of tooth preparations. 2. To determine examiner agreement for both class II amalgam cavity and full veneer gold shell crown preparations using the specific additional tools and feedback sheets for both grades awarded and repeatability and reproducibility of detailed feedback provided by senior academic staff.
Chapter 6	**Identification of gold standard (representative grades) for 26 class II amalgam cavities and 30 full gold shell-crown preparations** **Aims:** These objectives can be set out as a number of steps, each of which had a specific aim which is expanded within the methodology of this chapter. Thus, the two aims of this chapter were: 1. to identify the best senior academic staff examiner who had the highest correlation between grades awarded with the number of negative points identified; 2. to identify those tooth preparations which were awarded a grade by the best examiner in order to: • determine the specific feature measurements of this sub-group of grade tooth preparations and: • compare the objective class II amalgam cavity or full veneer gold shell crown preparation dimensions (SAMFs) recorded by the researcher and those subjective tooth preparation evaluations (SAFs) for similar teeth by the senior examiner with dimensions presented in the dental literature and subsequent calibration with the grades awarded by the best examiner.
Chapter 7	**Identification of reliable tools and development of a new checklist to assess class II amalgam cavities and full veneer gold shell crown preparations** **Aims:** **1. To identify a reliable specific additional measurement tool.** **2. To develop two reliable new checklists (New Checklist CII Preparation and New Checklist Gold shell Crown Preparation).** **3. To determine intra- and inter-examiner agreement using the grades and consistency of the New Checklists (Class II Preparation and Gold Shell Crown Preparation) and compare these data with that from previously used ‘Gray and Mhanni feedback sheets’ to ascertain if the agreement and consistency of the New Checklists have been improved.**
Chapter 8	**General conclusions, recommendation, and further studies**

7.1 Introduction

Evaluation of dental student performance of operative skills is essential for the students themselves to treat teeth affected by caries. According to Cowpe et al., (2009), *"restoring the diseased and damaged teeth, and management of dental caries by direct or indirect means using material and techniques that maintain pulp vitality and restore the tooth form, function and appearance acceptable to the patient in ways which prevent further diseases and damage, and help to promote the health of adjacent soft tissues"* (Cowpe et al., 2009). One of the fundamental methods a student can acquire these skills is by practising tooth preparation in the Clinical Skills Laboratory. This training helps dental students to acquire the necessary ability level and skills for them to progress to treating patients. The use of valid and reliable assessment tools will enable useful feedback and assessment of progress within the course (Brown, 1930, Houpt and Kress, 1973, Renne et al., 2013).

The assessment methods used in Clinical Skills Laboratories are widely variable. The 'glance and grade' method and 'objective checklist' are the most common assessment tools which are used (Vann et al., 1983, Manogue et al., 2001). The 'glance and grade' method utilises a subjective global assessment of the student performance without specific evaluation of each component of the skill (Vanek, 1969). Using the 'glance and grade' method with limited feedback for the students did not motivate them for deeper learning (Satterthwaite and Grey, 2008). The 'glance and grade' assessment should be supplemented with other forms of assessment, such as a checklist, to improve reliability (Mackenzie, 1973). The checklist method utilises a more analytical form of evaluation where each criterion of the entire performance is assessed separately. These methods can be used to drive students to learn and acquire new skills (Plasschart et al., 2007). Nowadays, some Institutions use computers to assess student performance. For validity, this software utilises measurement of specific features of tooth preparations to provide reliable and objective assessment. The disadvantages of this type of assessment are that the software is very expensive, takes longer to evaluate than a visual assessment, and assesses only about 70% of each tooth preparation leaving 30% to still be assessed by using subjective evaluation (Arnetzl and Dornhofer, 2004, Cardoso et al., 2006) and therefore is less efficient.

7.1.1 Self Assessment

According to Satterthwaite and Grey (2008), self-assessment is being performed in both preclinical and clinical environment but it cannot be used summative examinations because some of the students who were ‘high achievers’ tended to be overly-critical in their self-assessment, while less able students tended to over rate their work when self-assessing. These findings are supported by several other authors (Falchikov, 1986, Orsmond et al., 1997, Cho et al., 2010). Therefore, several authors focussed on methods which are expected to provide greater validity, reliability and agreement for examiners (Fuller, 1972, Gaines et al., 1974, Goepfred and Kerber, 1980, Vann et al., 1983, Chambers et al., 1997, Manogue et al., 2001).

7.1.2 Checklists

The Checklist is the most common assessment tool used within the Clinical Skills Laboratory (Sherwood and Douglas, 2014). Although the Checklist provides feedback for students to raise the level of student achievement and skills, there are also some aspects which might discourage them from learning. Examples of these related to scaling systems and criteria definitions (Brown 1930, Gaines et al., 1974, Feil, 1982, Helft et al., 1987). These issues lead the authors to conclude that there is very low inter-examiner agreement when Checklists are used. Gaines et al., (1974) compared two types of checklist for cavity preparation; one of them contained of six assessment areas each scoring 0 to 5, while the second checklist contained objective statements for each score in each area. The conclusion of Gaines' study supported the work of Brown which was low agreement among seven examiners (inter-examiner agreement) (Gaines et al., 1974). In addition, bias of the examiner and incorrect interpretation of rating scale were other common problems when marking student performance (Feil, 1982). Therefore, checklists with poorly-designed criteria are likely to result in mis-interpretation by examiners (Helft et al., 1987).

Haj-Ali and Feil (2006) concluded that when trying to evaluate student performance as a simply acceptable or unacceptable after calibration, examiners often estimated the work as acceptable when it was actually un-acceptable.

Checklist and scale designs should be carefully evaluated to reduce problems outlined in the previous paragraphs and Chapters. Therefore, construct validity is the first step to create items and scales for a new checklist. Construct validity is *"the degree to which a test measures what it claims, or purports, to be measuring"* (Cronbach and Meehl, 1955). Focus groups, clinical observation, theory, literature and expert opinion are all used to construct scales and items of a checklist. Next, content validity is essential to produce a valid checklist and scales to evaluate the object (e.g. class II amalgam cavity preparations). Content validity is when the item must be relevant to what judges need to assess. To demonstrate content validity, more than two examiners are needed (Streiner and Norman, 2008). Content validity is used to ensure that the items and scales cover all features of the tooth preparation the area and does not include irrelevant content. According to results of Chapter 6, the grades were not representative of the tooth preparations. The reason for this was a lack of calibration of the grades used. In addition, the criteria of the 'Gray and Mhanni feedback sheets' were not defined clearly enough. In addition, the standard setting of the errors number (negative points) in comparison with the cut-off score was also not clear. Therefore, the feedback sheets produce non-representative grades for the tooth preparations and unfair assessment of student performance (see Chapter 6).

According to Knight (1997), "*Without valid and reliable criteria, calibration would be an unattainable goal*". Knight introduced a three phases in order to calibrate examiners at University of Detroit Mercy. The first phase was Criteria Development. The second phase in creating calibrated examiners was development of Training Programmes for clinical and preclinical faculty in dentistry. These programmes focused on training in techniques and materials as well as calibration for faculty. This was undertaken by providing training sessions for both part-time and full-time faculty. The third phase was to confirm that calibration and accuracy of assessment was relevant. The faculty ratified a revision of a, "Rank and Tenure", document stating that one of the considerations for promotion and tenure was evidence of calibration (Knight, 1997). Thus, the concept of Knight was used in this chapter to develop checklists and therefore improve student feedback and assessment.

7.1.3 Checklist improvement / development

Building on the work from Chapter 5, checklists can be defined using three levels. These are category level, criterion level and level of performance. Taking the example of the class II amalgam cavity preparation, the category level can be the occlusal preparation, the criterion level can be the depth of this occlusal preparation and the level of performance would be how this depth is evaluated. Knight (1997) has said that the criteria and level of performance are the most important part of the checklist to calibrate examiners and should be the first to be defined. There are two ways in which the criteria level can be defined. These ways utilise categorical scales and continuous scales.

Each level of performance must assess a specific clinically-relevant criterion and category (e.g. the occlusal reduction is reduced to create a uniform 1.5 mm for full veneer gold shell crown preparation). In addition, all the categories, criteria and level of performance must be collectively valid for the tooth preparation being evaluated. For example, occlusal reduction is not valid for a class II amalgam cavity preparation. Level of performance for each criterion must be stated clearly and be clinically-acceptable to describe what should be evaluated. For example a millimetre scale can be used to describe occlusal cavity depth for a Class II amalgam cavity but that scale should not exceed the likely depth of the dental pulp. Thus a 10mm depth of occlusal cavity would not be valid. In addition, levels of performance for each criterion must be independent. For example, an overall occlusal cavity depth cannot be 'excellent', 'acceptable' and 'does not meet standard' at the same time (Knight, 1997). Construct validity might be used to identify the correlation between 'criteria which do not meet standard' and 'the total score of the checklist' (Keszei et al., 2010). After developing categories, criteria and level of performance, reliability is the next phase.

According to Knight (1997), "*the criteria must be sequenced to reflect the procedure; they must be accompanied by specifically described tests; the number of degrees of excellence must be clearly defined; and each criterion must exhibit consistency of terminology"*. Thus, to improve reliability, a valid checklist alone as an assessment tool might be not enough to improve the intra-or inter-examiner reliability. Schiff et al., (1975) designed a device called the 'pulpal floor measuring instrument' to measure the profile of the preparation, including depth, smoothness, and flatness of the pulpal floor and an intra-coronal cavity. They reported significant improvement in operator consistency using this device (Schiff et al., 1975). Therefore, the reliability of grades might be improved by using calibrated tools with the checklist.

Several authors created scaling systems and criteria according to Knight's phases but still the agreement among examiners was not high (Haj-Ali and Feil, 2006, Sherwood and Douglas, 2014). In addition, absence of clear guidelines on how the examiners should evaluate student performance by using an assessment tool was the main reason for inconsistency and lack of fair grading (Polyzois et al., 2010). The results of 'Gray and Mhanni feedback sheets' supported these conclusions (see Chapter 6).

7.1.4 Standard setting

In order to assess whether students acquired skills, a valid and reliable assessment should be developed that employs an appropriate standard setting (Taylor et al., 2013). Standard setting is better to be accompanied by other analytical methods utilising a checklist that is effective in determining whether the minimum requirement of the skill is met. Therefore, absolute standard setting was established in order to identify the maximum acceptable number of errors to pass. Jenkins et al., (1998) concluded that there was variety in the level of pass-fail between examiners during assessment of student performances (Jenkins et al., 1998). Therefore, standard setting or minimum pass level is mandatory to justify the pass score in order to maintain a valid and reliable assessment such as Objective Structured Clinical Examination (OSCE) (Rajiah et al., 2014, Puryer and O'Sullivan, 2015).

7.1.5 Development of a new checklist and tool for assessment

In this Chapter, the new checklists will be developed according to the concepts of Lynn (1986), Knight (1997), Streiner and Norman (2008) and Puryer and O'Sullivan (2015). These new checklists will be used with reliable specific additional tools.

The specific additional tool with the highest agreement was named as the reliable measurement tool used with the new checklist for each type of tooth preparation in this chapter.

New checklists with clearly defined criteria and levels of performance were created with absolute standard setting and the reliable specific additional measurement tool were used to re-evaluate class II amalgam cavity and full veneer gold shell crown preparations in an attempt to improve intra- and inter-examiner agreement and consistency.

7.2 Aims and Null hypothesis:

Aims of this chapter are:

To identify a reliable specific additional measurement tool by:

- determining the highest agreement between the Developed Standard scores of 26 cavities (see Chapter 6) and scores awarded when the condenser and the bur instrument were used to develop grades for the class II amalgam cavity preparation, and
- determining the highest agreement between the Developed Standard scores of 30 preparations (see Chapter 6) and scores awarded when the bur instrument and impression index were used to develop grades for the full veneer gold shell crown preparation.

To develop two reliable new checklists (New Checklist CII Preparation and New Checklist Gold shell Crown Preparation) by:

- determining absolute standard setting of the number of negative points (errors) for the Developed Standard scores for 26 class II amalgam cavities and 30 full veneer gold shell crown preparations for each grade.

- developing and defining new categories, criteria and levels of performance for the class II amalgam cavities and full veneer gold shell crown preparations.

To determine intra- and inter-examiner agreement using the grades and consistency of the New Checklists (Class II Preparation and Gold Shell Crown Preparation) and compare these data with that from previously used 'Gray and Mhanni feedback sheets' to ascertain if the agreement and consistency of the New Checklists have been improved.

Null hypothesis:

The new CII preparation (nCIIpc) and Gold Shell Crown preparation checklists (nGSCpc), when each used with a reliable specific measurement tool(s), does not improve intra- and inter-examiner agreement and consistency of grades awarded by all senior examiners in comparison with previous intra- and inter-examiner agreement and consistency of the 'Gray and Mhanni feedback sheet' grades from Chapter 5.

7.3 Material and methods

7.3.1 Identification of a reliable specific measurement tool to evaluate the class II amalgam cavity and full veneer gold shell crown preparation

From Chapter 5, different grades were awarded when different specific additional tools were used to evaluate 26 class II amalgam cavities (amalgam condenser and fissure bur) and 30 full veneer gold shell crown preparations (tapered high-speed diamond with a rounded tip 'Chamfer' bur and impression index) . The grades were converted to Pass/fail scores and compared with a Developed Standard scores which itself was determined from SAFM and SAF evaluations outlined in Chapter 6 (see Appendix 10). Thus, data from Chapters 5 and 6 were used to identify the most reliable specific additional tool to be used with the new checklist. By selecting the highest agreement between,

1. the scores for the Developed Standard and,
2. the scores from grades awarded using each specific additional tool for each type of tooth preparation.

Cohen's Kappa agreement test (SPSS) and agreement percentage were used in order to identify reliable tool for each type of tooth preparation.

7.3.2 New class II amalgam cavity (nCIIpc) and full veneer gold shell crown preparation (nGSCpc) checklist development.

a. Absolute standard setting

No study reports steps to calculate an absolute standard where a checklist with only pass and fail categories is used. Thus, a new checklist with more than two categories is required for absolute standard setting. A ranking system with more than two categories provides more information for the student for feedback and, with absolute standard setting, can be used to ascertain the number of errors acceptable to pass.

In this part of study, the Developed Standard scores (pass/fail) derived by SAFM and SAF evaluations were used. Knight (1997), said that the presence of any negative point about a preparation would constitute a fail. Conversely, the passing preparation would have no negative features. Using these criteria, for the class II amalgam cavity preparation there was only one such passing preparation and there were no such passes for the full veneer gold shell crown preparation (see Chapter 6).

Using the grades awarded subjectively by three senior examiners on two occasions using the feedback sheets, grades one, two and three constituted a fail grade whereas grades four and five constituted a pass grade. Thus, there were six grades available for each tooth preparation (i.e. three senior examiners * two occasions = 6 grades for each tooth preparation) (see Chapter 5).

It has been established that these subjective grades are not always repeatable or reproducible for all examiners (see Chapter 5) but it has not been established how they may reflect the objective evaluation of the tooth preparations. Thus, this chapter attempts to further compare the two sets of data.

This comparison is made by using the Developed Standard scores (pass/fail) for each tooth preparation and then searching through the subjective senior examiner grades to find the most frequently occurring (mode) grade that agrees with the Developed Standard score (pass/fail).

Having established the mode subjective evaluation grade that agrees with the Developed Standard scores (pass/fail), the next step was to determine the number of objective errors for SAFM and SAF evaluations (see Chapter 6) within a tooth preparation which defined a 'fail' preparation or would be accepted for a 'pass' preparation. In order to do this a borderline linear regression analysis was required (Schoonheim-Klein et al., 2009, Puryer and O'Sullivan, 2015) using:

1. the most frequently occurring (mode) grade agreeing with the Developed Standard score (pass/fail) and,
2. the number of objective errors from SAFM and SAF evaluations within a tooth preparation.

This exercise was performed for 26 class II amalgam cavity preparations and 30 full - veneer gold shell crown preparations. The number of acceptable negative points (errors) to pass was determined graphically using Microsoft Excel 2013.

Having established the number of acceptable errors for a passing preparation (grade 4 or 5), the next step was to establish which errors were acceptable for these grades. This was undertaken to determine the face validity of further evaluation. In order to establish which errors were acceptable, the feedback sheets from the senior examiners were collected and the errors identified for preparations awarded a grade 4 or a grade 5 mark were evaluated. Any negative points for these preparations could be identified on between one and six occasions. A list of these identified negative points was created for both the class II amalgam cavity preparation and the full veneer gold shell crown preparation. These lists of negative points were provided to the three senior academic staff examiners in order for them to accept, modify or reject these criteria.

b. New checklist development:

Almost all categories and criteria of a new checklist were created based on the 'Gray feedback sheet' for the class II amalgam cavity and the 'Mhanni feedback sheet' for the full veneer gold shell crown preparation. In addition, levels of performance for each criterion of the new checklists were created and developed according to the literature, textbooks and protocols used in some of clinical laboratory courses. Thus, acceptable and un-acceptable SAFMs and SAF evaluations of class II amalgam cavities and full veneer gold shell crown preparations were used to create levels of performance for the new checklists. In addition, a scale rating system was designed and developed by the author in order to provide:

1. general feedback for student (Figures 7.1 and 7.2), and
2. specific feedback for each feature of tooth preparation (Figures 7.7 and 7.11).

In this case, examiners can be able to better convey to a student the reasons why a preparation has been accepted (pass) or rejected (fail). The final grade or score from new checklists was developed based on recommendations by Knight (1997).

Grade	Description
Grade 1	The student prepared wrong tooth, unprepared tooth, or prepared different cavity design
Grade 2	The class II amalgam cavity has not met the standard (not acceptable)
Grade 3	The class II amalgam cavity needs modification (not acceptable)
Grade 4	The class II amalgam cavity is generally acceptable (acceptable)
Grade 5	The class II amalgam cavity is ideal and meets the standard

Figure 7.1 Grades and their descriptions for the class II amalgam cavity preparation

Grade	Description
Grade 1	The student prepared wrong tooth, unprepared tooth, or prepared different full crown preparation design
Grade 2	The full veneer gold shell crown preparation has not met the standard (not acceptable)
Grade 3	The full veneer gold shell crown preparation needs modification (not acceptable)
Grade 4	The full veneer gold shell crown preparation is generally acceptable (acceptable)
Grade 5	The full veneer gold shell crown preparation is ideal and meets the standard

Figure 7.2 Grades and their descriptions for the full veneer gold shell crown preparation

<u>For the class II amalgam cavity preparation:</u>

To create new checklist for class II amalgam preparation, **<u>ten</u>** criteria were selected based on a widely accepted literature, textbooks, protocols used in the clinical skills laboratory courses, and the 'Gray feedback sheet'. The criteria divided into **<u>three</u>** main categories: occlusal, proximal box and adjacent tooth. Some of criteria were assessed objectively by using reliable tool and others assessed subjectively. For each criterion, **<u>up to three</u>** levels of performance were specifically described. Most of the criteria have two levels; 'acceptable' (including the ideal range or feature) and 'not acceptable'. Sometimes there are **<u>three</u>** levels for each criterion, 'ideal', 'acceptable' and 'not acceptable'. This type of checklist can be utilised to provide pass or fail score for student. In this study, this checklist was 'Stage 1' (Figure 7.6).

To determine the final form of "nCIIpc" and provide grades and feedback for the student (Figure 7.7) a "nCIIpc - Stage 2" checklist was developed which was essentially and expanded for of the "nCIIpc - stage 1". However, before starting the "nCIIpc - Stage 2", absolute standard setting was determined by using a linear regression graph

(Figure 7.5). From the linear regression graph, the maximum number of acceptable errors was determined in order to give the student minimum passing grades (Grade 4).

After determining the number of acceptable negative points for passing grades, definition for each negative point for passing grade (grade 4) was determined according to the highest negative points in the grade 4 which collected from three senior academic staff examiners assessment using 'Gray feedback sheet'. To confirm the acceptable negative points for grade 4, these negative points were provided to all senior academic staff' examiners in order to accept, modify or reject. This process is called face validity for acceptable errors.

According to the previous paragraphs, 'new class II preparation checklist (nCIIpc) of stage 2' was commenced. **Up to four** levels of performance for each criterion were created. Each level represented grade (i.e. grade 5 the best and grade 2 the worst). Grade 1 represented a tooth that could not be evaluated (e.g. unprepared tooth and wrong tooth prepared) (Figure 7.1). Some criteria had four levels of performance 'ideal and acceptable', 'acceptable only', 'needs modification', 'not acceptable' and other had three (Figure 7.7). Descriptors of these levels were created according to acceptable and not acceptable of SAFMs and SAF evaluation in the literature and protocols of other institutions. This information was used to provide clearly defined level of performance for each criterion of the new checklist.

For the full veneer gold shell crown preparation:

Ten criteria of the full veneer gold shell crown preparation were selected based on widely accepted literature, textbooks, protocols used in the clinical laboratory courses and 'Mhanni feedback sheet'. The criteria were divided into **four** main categories: occlusal surface, axial surfaces, finish line and adjacent teeth. Some of the criteria were assessed objectively by using one reliable tool and other criteria were assessed subjectively. This process was the same as that used for the “nCIIpc” and was named the, ‘new Gold Shell Crown preparation checklist’, (nGSCpc).

For each criterion of the “nGSCpc stage 1”, **up to three** levels of performance were specifically described: most of the criteria have two levels, ‘acceptable’ (which includes the ideal range or feature) and ‘not acceptable’. This type of checklist is only to identify the pass or fail score for the student (Figure 7.10).

To provide more details for student performance, absolute standard setting was also determined by using a borderline linear regression graph (Figure 7.9). From the linear regression graph, the maximum number of acceptable errors was determined in order to give the student minimum passing grades (i.e. grade 4).

After determining the number of acceptable negative points for passing grade (grade 4), definition for each negative point for passing grade (grade 4) was determined according to the highest negative points in the grade 4 which collected from three senior academic staff examiners assessment using ‘Mhanni feedback sheet’.

This was called “nGSCpc stage 2” and contained **up to four** levels of performance to provide more accurate feedback for students. Some criteria have four levels while others have three levels of performance, ‘ideal and acceptable’ and/or ‘acceptable only’, ‘needs modification’, ‘not acceptable’ (Figure 7.11). The “nGSCpc stage 2” was also used to provide grades ranked from 2 to 5 for students.

All of the levels of performance were created according to recommended SAFMs and SAF evaluations or, for example when not acceptable, lay outwith recommended values in the literature or protocols of several institutions. SAFMs and SAF evaluations provided clearer definitions for each criterion of the “nGSCpc”.

7.3.3 Determination of intra- and inter-examiner agreement and consistency for new checklists

For purpose of this part of the study, the same samples of class II amalgam cavity and full veneer gold shell crown preparations were used as in the previous chapters. The preparations were evaluated using the "nCIIpc" for the class II amalgam cavity preparations with the selected reliable tool and using the "nGSCpc" for the full veneer gold shell crown preparation, again, with the selected reliable tool. In addition, the researcher (AM) provided instructions on how to use the new checklist with selected reliable tool(s) for both type of preparations (Appendices 10 and 11).

Twenty-six class II amalgam cavities and 30 full-veneer gold shell crown preparations were by dental students in 2014 and 2015 and evaluated by three senior academic staff examiners. The examiners evaluated the Typodont models held in their hands. One week after the first evaluation, the preparations were again evaluated by the same senior academic staff examiners for the second occasion using same method of scoring and the same selected reliable tool. Intra- and inter-examiner agreement and consistency were calculated for each occasion. For intra-examiner agreement and consistency, Cohen's Kappa agreement test (SPSS) was used. For inter-examiner agreement and consistency, intra-class correlation $_{(3,1)}$ (SPSS) was used.

7.3.4 Comparison of intra- and inter-examiner agreement and consistency for the new checklists ("nCIIpc" and "nGSCpc") with intra- and inter-examiner agreement and consistency for the 'Gray and Mhanni Feedback sheets'

The agreement and consistency of the new checklists ("nCIIpc" and "nGSCpc") both within (intra-examiner) and between (inter-examiner) senior academic staff examiners was determined. This was then compared with data from feedback sheets detailed in Chapter 5 to determine whether or not the new checklists ("nCIIpc" and "nGSCpc") with selected reliable tools improved intra- and inter-examiner agreement and consistency. Values of Cohen's Kappa and intra-class correlation (ICC) tests of the new checklists ("nCIIpc" and "nGSCpc") were compared with values of Cohen's Kappa and ICC tests of the 'Gray and Mhanni Feedback sheets'.

7.4 Results:

7.4.1 Identification of a reliable tool for the class II amalgam cavity and full veneer gold shell crown preparation

Table 7.1 shows Cohen's Kappa agreement values (SPSS) between the scores for specific additional tools which were used to assess 26 class II amalgam cavities in Chapter 5, with the 'Developed Standard' scores from SAFMs and SAFs evaluations for the same class II amalgam cavities in Chapter 6. Although the agreement between the 'specific additional tools' scores and the 'Developed Standard' scores was not statistically significant, the **condenser,** when used for **occasion 1** provided the highest level of Kappa agreement compared with the bur. To confirm this result, the agreement percentage was also calculated (Table 7.1) and this was also highest for the use of the condenser. Thus, the **condenser** became the selected reliable tool for the new checklist "nCIIpc" for the class II amalgam cavity preparation assessment in this part of the study (Figure 7.3).

Table 7.1 Summary table of Cohen's Kappa agreement values and agreement percentages between the scores of specific additional tools and the Developed Standard scores from SAFMs and SAFs evaluations (Appendix 10) for the class II amalgam cavities for each of the three examiners

Examiner One				
Specific addition tools	**Condenser**		**Bur**	
Occasions	*Occasion 1*	*Occasion 2*	*Occasion 1*	*Occasion 2*
Kappa value	0.165	0.077	0.140	0.089
Significance ($p \leq 0.05$)	0.126	0.308	0.161	0.271
Agreement percentage %	73	54	69	58
Examiner Two				
Specific addition tools	**Condenser**		**Bur**	
Occasions	*Occasion 1*	*Occasion 2*	*Occasion 1*	*Occasion 2*
Kappa value	0.196	0.140	0.140	0.120
Significance ($p \leq 0.05$)	0.093	0.161	0.161	0.197
Agreement percentage %	77	69	73	65
Examiner Three				
Specific addition tools	**Condenser**		**Bur**	
Occasions	*Occasion 1*	*Occasion 2*	*Occasion 1*	*Occasion 2*
Kappa value	0.196	0.089	0.140	0.089
Significance ($p \leq 0.05$)	0.093	0.271	0.161	0.271
Agreement percentage %	73	54	69	58

The highlighted value indicates the highest agreement

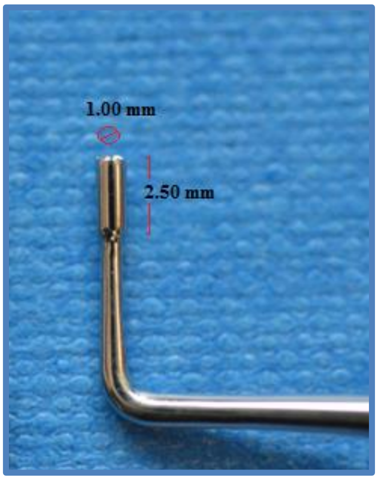


Figure 7.3 Picture of amalgam condenser with dimensions (mm)

Table 7.2 shows Cohen's Kappa agreement values (SPSS) and agreement percentages between the scores for 'specific additional tools' which were awarded after assessing 30 full veneer gold shell crown preparations in Chapter 5, with the 'Developed Standard' scores from SAFMs and SAF evaluations for same full veneer gold shell crown preparations in Chapter 6.

Table 7.2 Summary table of Cohen's Kappa agreement values and agreement percentages between the scores of specific additional tools and the Developed Standard scores from SAFMs and SAFs evaluations (from Appendix 13) for the full veneer gold shell crown preparations for each of the three examiners

Examiner One				
Specific addition tools	**Bur**		**Impression index**	
Occasions	*Occasion 1*	*Occasion 2*	*Occasion 1*	*Occasion 2*
Kappa value	-	-	-	-
Agreement percentage	37	47	40	43
Examiner Two				
Specific addition tools	**Bur**		**Impression index**	
Occasions	*Occasion 1*	*Occasion 2*	*Occasion 1*	*Occasion 2*
Kappa value	-	-	-	-
Agreement percentage	33	33	30	33
Examiner Three				
Specific addition tools	**Bur**		**Impression index**	
Occasions	*Occasion 1*	*Occasion 2*	*Occasion 1*	*Occasion 2*
Kappa value	-	-	-	-
Agreement percentage %	73	60	97	70

The highlighted value indicates the highest agreement percentage.

From Table 7.2, **no** measures of association were computed for Cohen's Kappa agreement test for all senior academic staff examiners with Developed Standard scores because all Developed Standard scores for 30 full veneer gold shell crown preparations were constant (see Chapter 6). Therefore, agreement percentage was calculated.

According to the result of the agreement percentage, there was no substantial difference between using the bur or an impression index as an assessment tools with new checklist. Thus, it was very difficult to select one reliable tool from the previous table. Therefore, both the bur and impression index were used to assess 30 full veneer gold shell crown preparations. In addition, assessment using a periodontal probe was suggested by one senior examiner instead of using bur to evaluate some features. For example, depth of finish line and TOC angle (Appendix 15). It was very difficult to measure TOC angle (20°) using any of the assessment tool(s) (i.e. bur, periodontal probe or impression index). Therefore, a proximal TOC angle of 20° was evaluated by measuring the distance between the mesial and distal marginal ridges of each prepared tooth using a periodontal probe. For a bucco-palatal TOC angle of 20°, the buccal-to-palatal width of the occlusal table of the prepared tooth was also evaluated using the same instrument. This distance for both a proximal and a buccal-palatal convergence of 20^{o} was 7 mm. Thus, the periodontal probe was considered suitable to assess total occlusal convergence (TOC) angles.

After the evaluation, one of the senior examiners admitted they were not familiar with the evaluation of the student full veneer gold shell crown preparation and this could, and did, have a substantial bearing on the results. Thus, a calculated decision was taken to remove the data from this examiner (i.e. examiner 3) and so the data from examiners 1 and 2 using the either the bur or periodontal probe with the impression index were analysed (Figure 7.4). For example, axial and occlusal reduction can be assessed by impression index and periodontal probe or tapered high-speed diamond with a rounded tip (Chamfer) bur (Appendix 15).

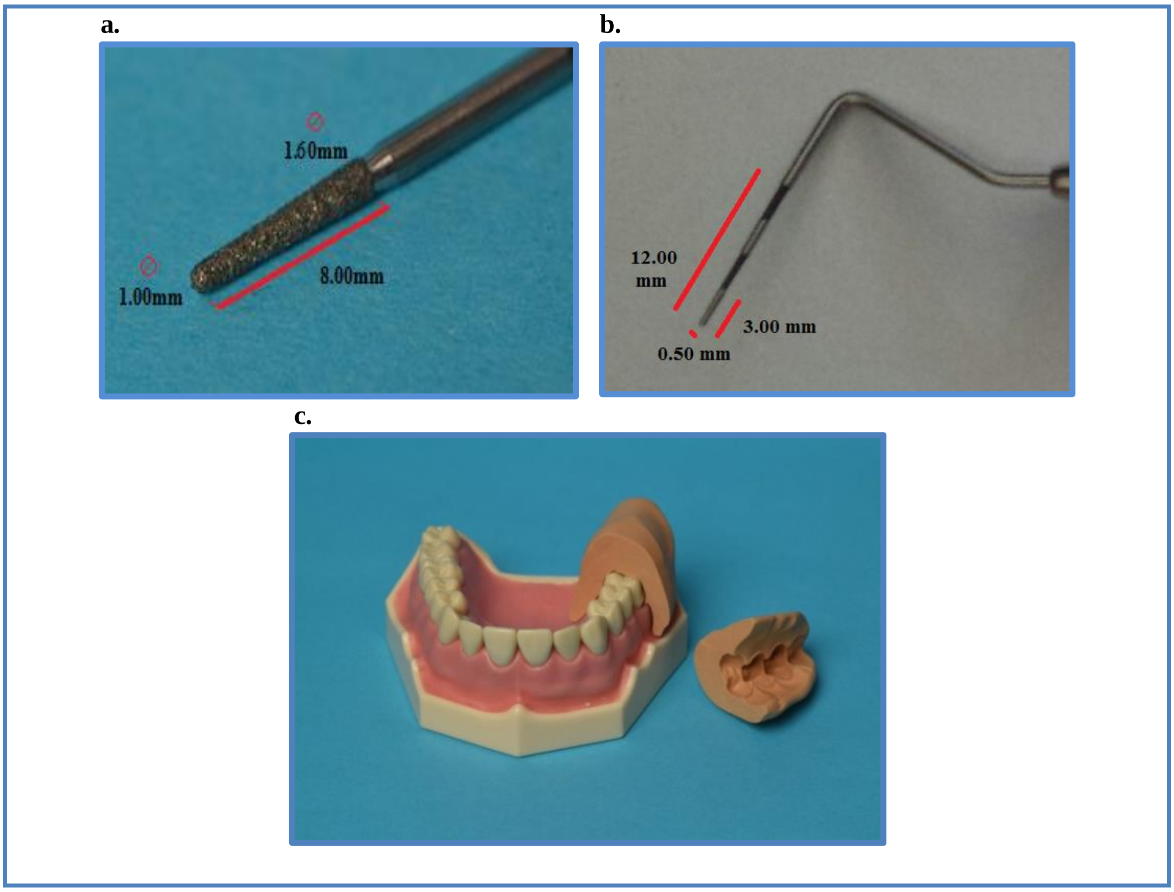


Figure 7.4 Pictures of a) tapered high-speed diamond bur with a rounded tip (Chamfer) and its dimensions (mm), b) CP12 periodontal probe with its dimensions and c) impression index

7.4.2 New checklists development for the class II amalgam cavity "nCIIpc" and full veneer gold shell crown preparation "nGSCpc"

a. For the class II amalgam cavity preparation

Initially, absolute standard setting (cut-off grades with number of negative points) was determined graphically by using borderline linear regression (Microsoft Excel 2013) in order to create valid and reliable a new checklist "nCIIpc".

In order to determine the maximum number of negative points acceptable to pass, the Developed Standard score (pass or fail) for each model was converted to a Developed Standard Grade (grade 1 to 5) (Table 7.3). According to type of the Developed Standard score (pass or fail) for each model (column 2), the most frequently (mode) occurring grade by three senior examiners over two occasions using ‘Gray feedback sheet’ was selected as a Developed Standard grade (column 3). For example, the Developed Standard score for the model number 5 (i.e. fail score) was converted to the fail grade (i.e. grade 2). Grade 2 for this model was selected because it was the most frequently fail grade (mode) awarded by three senior examiners over two occasions (see Appendix 12).

For the class II amalgam cavity preparation, Table 7.3 shows:

- the model number (column 1),
- Developed Standard score (pass/fail) awarded from SAFMs and SAFs evaluations (column 2),
- the most frequently occurring (mode) grade awarded subjectively by the three senior examiners over two occasions using the ‘Gray feedback sheet’ (column 3) (see Appendix 12) and,
- the number of objective negative points (errors) from SAFMs and SAF evaluations (column 4).

Table 7.3 The Developed Standard scores were converted to grades which the most frequently occurring (mode) grade awarded subjectively by three senior examiners on two occasions using the 'Gray feedback sheet' and the number of objective negative points (errors) from SAFMs and SAF evaluations within a class II amalgam cavity preparation

Model number	Developed Standard scores	Grades (modes) awarded by three senior examiners on two occasions using the feedback sheets	the number of objective negative points (errors) from SAFM and SAF evaluations
5	Fail	2	5
8	Fail	3	4
15	Fail	3	5
16	Fail	3	5
36	Pass	5	0
39	Fail	Pass	-
40	Fail	Pass	-
41	Fail	3	4
43	Fail	3	4
46	Fail	Pass	-
53	Fail	3	2
54	Fail	3	2
57	Fail	2	1
62	Fail	3	4
73	Fail	2	7
78	Fail	3	3
80	Fail	3	5
83	Fail	3	4
85	Fail	3	3
87	Fail	Pass	-
88	Fail	2	5
94	Fail	2	2
109	Fail	3	5
111	Fail	2	5
120	Fail	Pass	-
138	Fail	2	4

Red rows indicate that the models with no frequently occurring grade by the three senior academic examiners over two occasions using 'Gray feedback sheet' in comparison to Developed Standard scores.

Grades (modes) awarded by three senior examiners on two occasions using the feedback sheets and the number of objective negative points (errors) from SAFM and SAF evaluations were used to determine the maximum number of negative points acceptable to pass. From Table 7.3 and Appendix 12, five class II amalgam cavities were excluded because the grades awarded from the three senior examiners did not agree with the Developed Standard scores. To estimate the maximum number of negative points (errors) for cut-off grade (i.e. grade 3), 21 cavities were selected instead of 26 cavities.

Figure 7.5 demonstrates the maximum number of negative points (errors) which were acceptable to pass for 21 class II amalgam cavity preparations. The number of negative points below the linear regression of grade 3 (cut-off grade) was the acceptable number of errors to pass. In this case, the maximum number of negative points (errors) that was acceptable to pass was **three** out of ten (30%) according to SAFMs and SAFs.

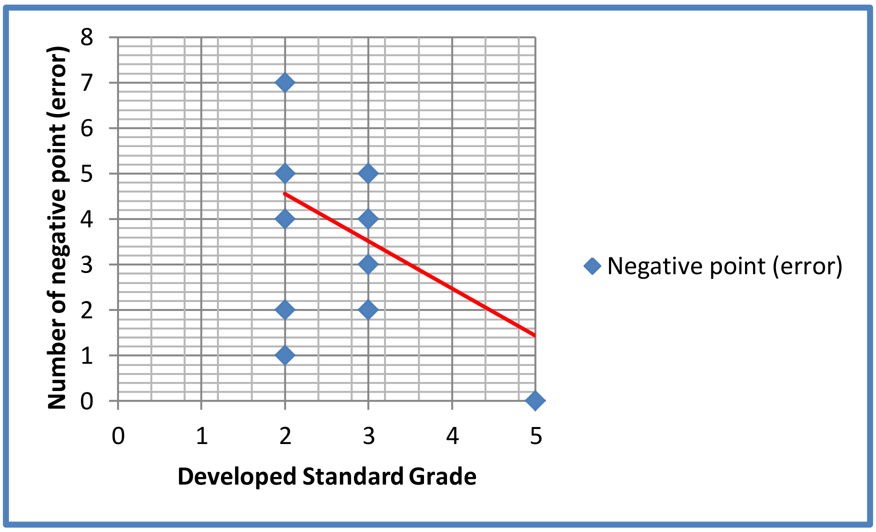


Figure 7.5 Linear regression between Developed Standard Grades and the number of negative point (error) of the 'Gray feedback sheet' for 26 class II amalgam cavity preparations

From Figure 7.5, the maximum number of errors which acceptable to pass was three errors out of ten. To define these acceptable errors for grade 4, the three most prevalent descriptors of level of performance in relation to the negative points for grades 4 which were assessed by three senior academic staff examiners using the 'Gray feedback sheet' were selected as acceptable levels of performance (see Chapter 5). These descriptors according to the level of performance were:

1. one or two **parallel walls** of the occlusal cavity,
2. one or two **parallel walls** of the proximal box and,
3. **minor damage** to adjacent tooth (Table 7.4).

The third criterion according to the 'Gray feedback sheet' was an occlusal lock. This feature is usually, but not always, a type of retention form for the occlusal aspect of this tooth preparation. Thus, buccal and palatal/lingual **parallel walls** were a feature of the occlusal cavity in this new checklist which was noted widely in the literature as an acceptable error. This feature as a retention form was used instead of a rough or irregular occlusal key from the 'Gray feedback sheet'.

Table 7.4 demonstrates the passing grades (5 and 4 grades) with number of negative points (errors) for each criterion according to the 'Gray feedback sheet'.

Table 7.4 Demonstration of the passing grades (5 and 4 grades) with number of negative points (errors) for each criterion of class II amalgam cavity preparation according to the 'Gray feedback sheet'

Criteria and their *level of performance*	**Grade 5**	**Grade 4**
Outline		
Rough/irregular	*0*	*0*
Position of the box		
Too far B/P	*0*	*3*
Depth gingivally		
Deep	*0*	*3*
Shallow	*0*	*5*
B-P width of the box		
Too wide	*0*	*4*
Too narrow	*0*	*5*
M-D depth of the box		
Too deep	*0*	*1*
Too shallow	*0*	*3*
Unsupported enamel *(yes)*	*1*	*2*
Retention		
Parallel walls	***3***	***7***
Divergent walls	*0*	*0*
Key		
Rough/irregular	***2***	***8***
Not follow fissure	*1*	*3*
Depth of occlusal cavity		
Too deep	*0*	*0*
Too shallow	*0*	*6*
Width of occlusal cavity		
Too wide	*0*	*0*
Too narrow	*0*	*0*
Damage to adjacent tooth		
Minor	***2***	***16***
Moderate or severe	*0*	*0*

The highlighted figures are the highest number of negative point which represented the three acceptable errors for passing grades.

These descriptors of the levels of performance for three criteria were provided for all senior academic staff examiners. They accepted these descriptors as acceptable negative points (errors) for minimum passing grades (grade 4). This was a type of face validity.

For the next step, the author created a new checklist "nCIIpc" as Stage 1 to provide pass and fail scores for the students. This is shown in Figure 7.6. Most of the 'Gray feedback sheet' criteria were selected to create "nCIIpc – Stage 1" with the exception of the occlusal key/ lock criterion. In addition to the feature of occlusal cavity walls as a retention form instead of occlusal key, the marginal ridge criterion was also chosen for "nCIIpc – Stage 1" instead of the occlusal key criterion. There were three reasons for this. First, the occlusal key lock has different shapes (e.g. dovetails or occlusal locks) described in the literature (Roberson et al., 2002, Akpata et al., 2013) and other literature demonstrated it may not even be needed at all (Roberson et al., 2002). The second reason was the consistency of the occlusal key was very difficult to evaluate (see Chapter 5). The third reason was a reduction in the number of criteria evaluated subjectively provided greater accuracy and reliability of grades and feedback for the student. Taking these three reasons into account, marginal ridge thickness, which helps to preserve the fracture resistance of the tooth (Shahrbaf et al., 2007), was selected and evaluated objectively (see section 7.5).

"nCIIpc - Stage 1"				
Category	Criteria	**Level of performance (Descriptors)**		
		Ideal	*Acceptable*	*Not acceptable*
Occlusal cavity	*Depth*	1.50 – 2.00 mm		Above or below the previous range
	Width	0.80 – 1.50 mm		Above or below the previous range
	Retention	Two **converge** walls	One **converge** and other **parallel** or two **parallel** walls	One or two **diverge** walls
	Marginal ridge thickness	> 1.00 mm		below the previous range
Proximal box	*Depth (occlusal-gingival direction)*	2.50 – 4.00 mm At or below contact area		Above or below the previous range
	Depth (mesial-distal direction) at gingival floor	0.80 – 1.50 mm		Above or below the previous range
	Width (buccal-lingual/palatal direction) at gingival floor	3.00 – 4.00 mm		Above or below the previous range
	Retention	Two **converge** walls	One **converge** and other **parallel** or two **parallel** walls	One or two **diverge** walls
	Position and unsupported enamel	In the **middle** between tip of two cusps without unsupported enamel		In the **middle** with unsupported enamel or far buccally, lingually/palatally with or without unsupported enamel
Adjacent tooth	*Damage*	**None**	**Minor**	**Moderate or severe**
Scores		***If select all levels of performance from this side, the score is Pass***		***If select one from this side, the score is Fail***

Figure 7.6 Schematic representation of the "nCIIpc - Stage 1" checklist for the class II amalgam cavity preparation to determine pass/fail score for each student

To provide more information for the student, "nCIIpc - Stage 2" was created. Figure 7.7 demonstrates the "nCIIpc - Stage 2" and provides grades ranked from Grade 2 to Grade 5. Grade 1 represents the wrong tooth being prepared or no tooth being prepared. Grade 2 was the worst while the Grade 5 was the best grade. Grades 1, 2, and 3 were fail grades whereas Grades 4 and 5 were pass grades (Figure 7.1). The "nCIIpc - Stage 2" produced grade as well as feedback for the student to determine the weak and strong points of student's performance. Therefore, the Stage 2 of "nCIIpc" was created to improve student feedback.

"nCIIpc stage 2"					
Category	Criteria	**Level of performance (Descriptors)**			
		Ideal/ Acceptable	*Acceptable only*	*Needs modification*	*No modification can apply*
Occlusal cavity	*Depth*	1.50 – 2.00 mm		Below the previous range	Above the previous range
	Width	0.80 – 1.50 mm		Below the previous range	Above the previous range
	Retention form	Two **converge** walls	One **converge** and other **parallel** or two **parallel** walls	One **converge** or **parallel** and other **diverge**	Two **diverge** walls
	Marginal ridge thickness	>1.00 mm			Below the previous range
Proximal box	*Depth (occlusal-gingival direction)*	2.50 – 4.00 mm At or below contact area		Below the previous range	Above the previous range
	Depth (mesial-distal direction) at gingival floor	0.80 – 1.50 mm		Below the previous range	Above the previous range
	Width (buccal-lingual/palatal direction) at gingival floor	3.00 – 4.00 mm		Below the previous range	Above the previous range
	Retention form	Two **converge** walls	One **converge** and other **parallel** or two **parallel** walls	One **converge** or **parallel** and other **diverge**	Two **diverge** walls
	Position and unsupported enamel	In the **middle** between tip of two cusps without unsupported enamel		In the **middle** with unsupported enamel	Far buccally, lingually or palatally
Adjacent tooth	*Damage*	**None**	**Minor**		**Moderate or severe**
Grades		***Grade 5***	***Grades 4***	***Grade 3***	***Grades 2***
Marks*		**10**	***9 to 7***	***6 to -24***	***-25 to-340***

Figure 7.7 Schematic representation of the "nCIIpc - Stage 2" checklist for the class II amalgam cavity preparation to provide grade and feedback for each student

* Marks for the grades were used to develop electronic version of new checklist and to determine examiner agreement according to the level of performance for each criterion. These marks were calculated as follows:

- **Grade 5:** All 10 criteria must be ideal. For each ideal criterion a +1 mark is awarded. Thus, 10 criteria results in a mark of 10 for a grade 5 preparation.
- **Grade 4:** if one or two parallel walls of occlusal cavity and/or one or two parallel walls of proximal box and/or minor damage to the adjacent tooth are selected, the overall grade will be 4. The mark for each criterion for grade 4 is (0).
- **Grade 3:** if one or more of the descriptors of level of performance are selected which are listed as a Grade 3, the overall grade will be 3. The mark for each criterion for grade 3 is (-3).
- **Grade 2:** if one or more of the descriptors of level of performance are selected which are listed as a Grade 2, the overall grade will be 2. The mark for each criterion for grade 2 is (-34).
- **Grade 1:** if the tooth was unprepared or wrong tooth prepared, the overall grade will be 1 immediately.

For subjective criterion evaluation in the "nCIIpc", the definition of the criterion must be clearly defined. For example, buccal and lingual or palatal walls of the occlusal cavity and proximal box must converge toward the occlusal aspect of the tooth preparation to be classified as 'ideal'. Parallel walls of the occlusal cavity and/or proximal box provide acceptable retention while divergent walls are not acceptable. Thus retention form is clearly defined (Figure 7.8).

Thus, the presence of any 'not-acceptable' level of performance will result in the award of a Grade 2. In the absence of any 'not-acceptable' level of performance, the presence of any 'needs-modification' level will result in a Grade 3. In the absence of either of these levels, the presence of any 'acceptable-only' level of performance will result in a Grade 4 while the presence of only 'ideal/acceptable' level of performance for all criteria will result in a grade 5.

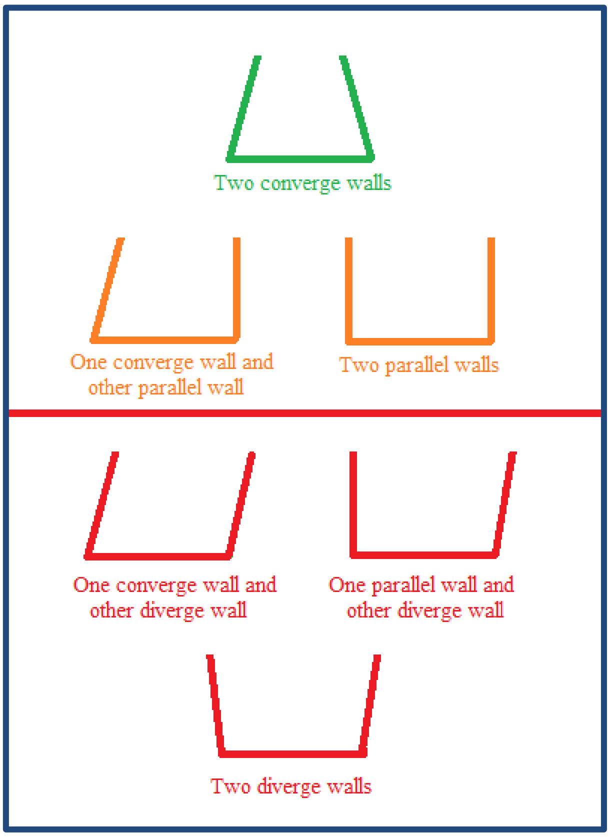


Figure 7.8 Diagram to show combinations of retention form criteria for class II amalgam cavity preparation

Iatrogenic damage to the adjacent tooth is due to the fact that the teeth are in close contact. Damage to the adjacent tooth surface has been classified by both degree of damage and pattern of damage.

In relation to degree of damage, Moopnar and Faulkner, (1991) adapted the following scale:

- No damage visible to the naked eye or under magnifying glass. **(None)**
- Slight damage visible to the naked eye and identifiable with a magnifying glass. **(Minor)**
- Obvious damage. **(Moderate or severe)**

Although this classification gives the examiner a reasonable definition, assessment of damage to the adjacent tooth surface is still difficult. Therefore, the pattern of damage according to Medeiros and Seddon, (2000) was also considered:

- Undamaged: a sound surface with regular curved proximal surface without loss of contour. **(None)**
- Scratches: narrow, shallow score-lines, usually multiple with a consistent orientation. **(Minor)**
- Indentation: a regular defect without an orientation, roughly circular or irregular in shape. **(Minor)**
- Groove: a deeper defect, length greater than width with a vertical or horizontal orientation. **(Moderate)**
- Extensive damage: damage involving a large area of the proximal surface. **(Severe)**

In order to capture elements from both these scales, a definition of minor damage was "slight damage (i.e. scratches or indentations) visible to the naked eye".

b. For the full veneer gold shell crown preparation

Initially, absolute standard setting (based on cut-off grades and the number of negative points) was determined graphically using a borderline linear regression method (Microsoft Excel 2013) in order to create a valid and reliable new checklist "nGSCpc". For the full veneer gold shell crown preparation, Table 7.5 shows:

- the model number (column 1),
- the agreement of column 3 with the Developed Standard score (pass/fail) (column 2)
- the most frequently occurring (mode) grade awarded subjectively by the three senior examiners over two occasions using the 'Mhanni feedback sheet' (column 3) (see Appendix 13) and,
- the number of objective negative points (errors) from SAFMs and SAF evaluations (column 4).

Table 7.5 The most frequently occurring (mode) grade by using the grades awarded subjectively by three senior examiners on two occasions using the 'Mhanni feedback sheet', the Developed Standard score and, the number of objective negative points (errors) from SAFMs and SAF evaluations within a full veneer gold shell crown preparation

Model number	Developed Standard scores	Grades (modes) awarded by three senior examiners on two occasions using the feedback sheets	the number of objective negative points (errors) from SAFM and SAF evaluations
1	Fail	3	11
3	Fail	3	10
4	Fail	3	12
5	Fail	3	12
7	Fail	3	9
13	Fail	3	13
14	Fail	3	8
18	Fail	3	10
20	Fail	3	6
21	Fail	3	9
25	Fail	3	8
26	Fail	2	12
29	Fail	3	8
31	Fail	3	10
51	Fail	3	5
52	Fail	3	7
54	Fail	3	10
57	Fail	3	6
58	Fail	3	12
59	Fail	2	12
60	Fail	3	11
63	Fail	3	10
67	Fail	3	7
69	Fail	3	9
70	Fail	3	7
71	Fail	3	5
73	Fail	3	8
74	Fail	3	8
78	Fail	3	7
88	Fail	2	14

Grades (modes) awarded by three senior examiners on two occasions using the feedback sheets and the number of objective negative points (errors) from SAFM and SAF evaluations were used to determine the maximum number of negative points acceptable to pass. Figure 7.9 demonstrates the maximum number of negative points (errors) which were acceptable to pass for 30 full veneer gold shell crown preparations. The number of negative points below the linear regression of grade 3 (cut-off grade) was the acceptable number of errors to pass. In this case, the maximum number of negative points (errors) that was acceptable to pass was **eight** out of twenty two (42%) according to SAFMs and SAFs.

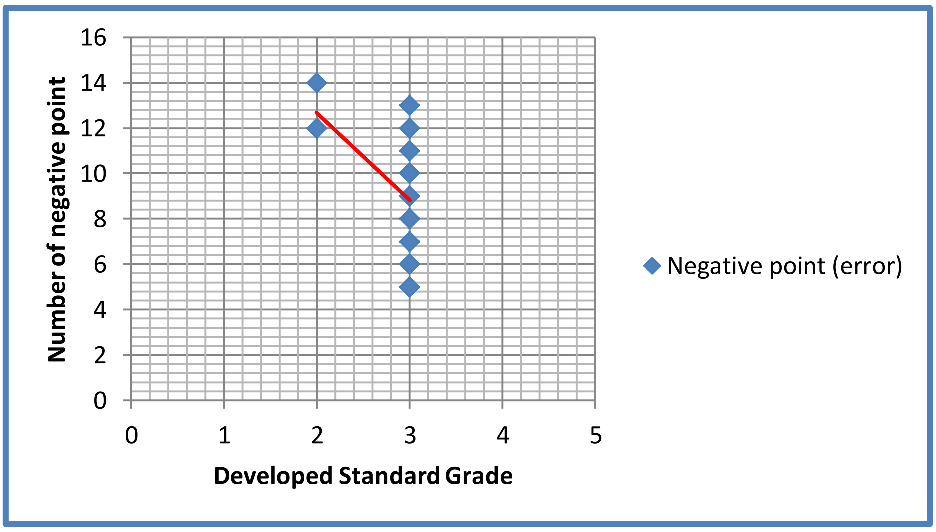


Figure 7.9 Linear regression between Developed Standard Grades and the number of negative point (errors) of the 'Mhanni feedback sheet' for 30 full veneer gold shell crown preparations

- The maximum number of errors which were acceptable to pass was eight errors out of twenty criteria of 'Mhanni feedback sheet'. The new checklist contained only ten criteria. Therefore, the total number of acceptable negative points (errors) for the new checklist should be not more than **4** (10*42% = 4) acceptable negative points (errors). To define these acceptable negative points (errors) for grade 4, the four most frequently observed negative points for grades four and/or five assessed by three senior academic staff examiners using 'Mhanni feedback sheet' were selected as an acceptable negative point for passing grades. All of these acceptable negative points were from subjective criteria. These subjective criteria were:
  - **functional cusp bevel** presence and location (it should be noted that no senior examiner considered this necessary to be acceptable),
  - **type of finish line**,
  - **level of finish line in relation to gingival margin,** and
  - **damage to one or more adjacent teeth** (Table 7.6).

The remaining criteria in the new checklist were objective; for example, occlusal reduction, axial reduction, and total occlusal convergence (TOC), and were defined as ideal and/or acceptable based on data from the literature.

Table 7.6 demonstrates the passing grades (4 and 5 grades) with number of negative points (errors) for each criterion according to the ‘Mhanni feedback sheet’.

Table 7.6 Demonstration of the passing grades (5 and 4 grades) with number of negative points (errors) for each criterion for the full veneer gold shell crown preparation according to the 'Mhanni feedback sheet'

Criterion and their *level of performance*	**Grade 5**	**Grade 4**
Occlusal reduction		
Under-prepared	*0*	*4*
Over-prepared	*2*	*10*
Contour of occlusal preparation		
No (does not follow the contour tooth surfaces)	*0*	*3*
Buccal reduction		
Under-prepared	*0*	*12*
Over-prepared	*0*	*3*
Lingual reduction		
Under-prepared	*2*	*11*
Over-prepared	*0*	*5*
Mesial reduction		
Under-prepared	*0*	*1*
Over-prepared	*1*	*11*
Distal reduction		
Under-prepared	*0*	*1*
Over-prepared	*1*	*10*
Undercuts		
Yes	*0*	*5*
Bucco-lingual convergence		
Improper convergence	*0*	*7*
No (destructive shape)	*0*	*0*
Proximal convergence		
Improper convergence	*2*	*11*
No (destructive shape)	*0*	*0*
Contour of axial surfaces preparation		
One of axial surface not follow the contour	*0*	*6*
More than one of axial surface not follow the contour	*0*	*1*
Contact area with adjacent teeth		
Cleared – only one side	*0*	*0*
Not clear	*0*	*0*
Functional cusp bevel reduction		
Not - symmetrical	*0*	***13***
There is NO functional cusp bevel	*0*	*5*
Location of functional cusp bevel		
Location of functional cusp bevel is NOT good	*2*	***14***
Chamfer finish line (Type of finish line)		
Type of finish line is not chamfer (i.e. Knife edge)	*0*	***8***
Level of finish line to gingival margin		
Supra-gingival (more than 0.5mm)	*2*	***6***
Subgingival and/or Supra-gingival	*0*	*0*
Depth of finish line all around		
Uneven	*1*	*11*
Deep	*0*	*0*
Texture of final preparation except margin		
Rough (irregular) Sharp edges	*0*	*8*
Texture of margin		
Rough (irregular)	*0*	*2*
Mesial tooth		
Minor damage	*1*	***6***
Moderate/severe damage	*0*	*0*
Distal tooth		
Minor damage	*1*	***9***
Moderate/severe damage	*0*	*0*

The highlighted figures are the highest number of negative point which represented the acceptable errors for passing grades

These descriptors of level of performance for subjective criteria were provided to all senior academic staff examiners. They reviewed and accepted these descriptors as the acceptable negative points (errors) for the minimum passing grade (grade 4).

For the next step, the researcher (AM) created a new checklist "nGSCpc" as Stage 1 to provide pass and fail scores for the students. This is shown in Figure 7.10. Most of the 'Mhanni feedback sheet' criteria were selected to create "nGSCpc" especially the objectively measured criteria. The researcher (AM) selected more objective than subjective criteria for the new checklist to provide more accurate, valid and reliable feedback and grades for students.

"nGSCpc - Stage 1"				
Category	Criteria	**Level of performance (Descriptors)**		
		Ideal	*Acceptable*	*Not acceptable*
Occlusal surface	*Functional cusps reduction*	1.50 – 2.00 mm with or with non-symmetrical bevel		Above or below the previous range with or without bevel
	Non-functional cusps reduction	1.00 – 1.50 mm		Above or below the previous range
	Bucco-lingual/palatal occlusal convergence	3° - 20° or between 7.00 to 9.00 mm		Above or below the previous range angles or <7.00 mm or >9.00 mm
	Proximal occlusal convergence	3° - 20° or between 7.00 to 9.00 mm		Above or below the previous range angles or <7.00 mm or >9.00 mm
Axial surface(s)	*Axial reduction*	0.50 – 1.50 mm (for each axial surface)		Above or below the previous range even on one surface
Finish line	*Type*	Chamfer with or without small lip		Another type
	Depth	0.50 - <1.00 mm On all sides		Above or below the previous range even on one side
	Level	At or above gingival line (maximum 1.00mm from gingival line) On all sides		Above or below the previous range even on one side
Adjacent teeth	*Contact area with adjacent teeth*	Cleared on both sides		Not clear on one or both sides or > 0.5mm clearance on one or both sides
	Damage	**None for one or both**	**Minor for one or both**	**Moderate or severe for one or both**
Scores		***If select all levels of performance from this side, the score is Pass***		***If select one from this side, the score is Fail***

Figure 7.10 Schematic representation of the "nGSCpc - Stage 1" checklist for the full veneer gold shell crown preparation to determine pass/fail score for each student

To provide more information for the student, “nGSCpc - Stage 2” was created. Figure 7.11 demonstrates the “nGSCpc - Stage 2” and provided grades ranked from Grade 2 to Grade 5. Grade 1 represented the wrong tooth being prepared or no tooth being prepared. Grade 2 was the worst-, while the Grade 5 was the best-grade. Grades 1, 2, and 3 were fail grades whereas Grades 4 and 5 were pass grades (Figure 7.2). The “nGSCpc Stage 2” produced feedback for the student to determine the weak and strong points of student’s performance. Thus, the “nGSCpc - Stage 2” was created to improve student feedback.

“nGSCpc - Stage 2”					
Category	Criteria	**Level of performance (Descriptors)**			
		Ideal/ acceptable	*Acceptable only*	*Needs modification*	*Not acceptable*
Occlusal surface	*Functional cusps reduction*	1.50 – 2.00 mm with symmetrical bevel	1.50 – 2.00mm With non-symmetrical bevel	Below the previous range with or without bevel	Above the previous range with or without bevel
	Non-functional cusps reduction	1.00 – 1.50 mm		Below the previous range	Above the previous range
	Bucco-lingual/palatal occlusal convergence	3° - 20° or between 7.00 to 9.00 mm		Below the previous range angle or >9.00 mm	Above the previous range angle or <7.00 mm
	Proximal occlusal convergence	3° - 20° or between 7.00 to 9.00 mm		Below the previous range angle or >9.00 mm	Above the previous range or <7.00 mm
Axial surface(s)	*Axial reduction*	0.50 – 1.50 mm on all surfaces		Below the previous range even on one surface	Above the previous range even on one surface
Finish line	*Type*	Chamfer	Chamfer with small lip	Knife edge	Shoulder
	Depth	0.50 - 1.00 mm on all sides		Below the previous range even on one side	Above the previous range even on one side
	Level	0 to 0.50 mm supragingival on all sides	>0.50 to 1.00mm supra-gingival even on one side	>1.00 mm supragingival even on one side	<0.00mm Subgingival even on one side
Adjacent teeth	*Contact area with adjacent teeth*	Cleared on both sides		Not clear on one or both sides	> 0.50mm clearance on one or both sides
	*Damage**	**None**	**Minor for one or both teeth**		**Moderate or severe for one or both teeth**
Grades		***Grade 5***	***Grade 4***	***Grade 3***	***Grade 2***
Marks*		10	9-6	5 to -36	-37 to - 460

Figure 7.11 Schematic representation of the “nGSCpc - Stage 2” checklist for the full veneer gold shell crown preparation to provide grade and feedback for each student (Damage* defined as in “nCIIpc – Stage 2” checklist – see pages 321 and 322)

* Marks for the grades were used to develop an electronic version of new checklist and to determine examiner agreement according to the level of performance for each criterion. These marks were calculated as follows:

- **Grade 5:** All 10 criteria must be ideal. For each ideal criterion a +1 mark is awarded. Thus, 10 criteria results in a mark of 10 for a grade 5 preparation.

- **Grade 4:** if functional cusp reduction with non-symmetrical bevel is within the acceptable range and/or chamfer finish line with small lip and/or level of the finish line above gingival line from >0.50 to 1.00 mm and/or minor damage to the adjacent tooth is selected, the overall grade will be 4. The mark for each criterion for grade 4 is (0).
- **Grade 3:** if one or more of the descriptors of level of performance are selected which are listed as a Grade 3, the overall grade will be 3. The mark for each criterion for grade 3 is (-4).
- **Grade 2:** if one or more of the descriptors of level of performance are selected which are listed as a Grade 2, the overall grade will be 2. The mark for each criterion for grade 2 is (-46).
- **Grade 1:** if the tooth was unprepared or wrong tooth prepared, the overall grade will be 1 immediately.

For subjective criteria evaluation using the “nGSCpc”, each subjective criterion and levels of performance must be clearly defined. For example, the ‘contact area with the adjacent teeth’ should be cleared. This means that there is small space (<0.5mm) between the full veneer gold shell crown tooth preparation with the adjacent teeth. This descriptor is classified as ‘ideal’. For ‘damage to adjacent teeth’, same definition should be used as in the “nCIIpc”. Thus, the presence of any ‘not acceptable’ level will result in the award of a Grade 2. In the absence of any ‘not acceptable’ level of performance, the presence of any ‘needs modification’ level will result in a Grade 3. In the absence of either of these levels, the presence of any ‘acceptable-only’ level of performance will result in a Grade 4 while the presence of ‘only ideal/acceptable’ level will result in a grade 5.

The “nCllpc - Stage 2” and “nGSCpc - Stage 2” were provided for three senior examiners to assess tooth preparations according to instructions which were also provided (Appendices 14 and 15).

7.4.3 Determination of intra- and inter-examiner agreement and consistency for the new checklists

a. For the class II amalgam cavity preparation

Intra-examiner agreement for class II amalgam cavity preparation according to grades by using the new Class II preparation checklist “nCIIpc”

Table 7.7 shows, for examiners 1, 2, and 3, the grades awarded for each cavity on each occasion of grading. These are colour-coded to indicate agreement and disagreement (Figure 5.9). Table 7.7 also shows the number of negative points awarded from the “nCIIpc” on each occasion and by the same colour coding convention indicates agreement and disagreement. The number of times from the first and second grading occasions, where the same final conclusion was reached, is also summarised for each examiner in the next tables.

Table 7.7 Three senior academic staff examiners' grades awarded for each class II amalgam cavity preparation on each occasion of grading and the number of negative points on each occasion with their agreement percentages using "nCIIpc" and Developed Standard scores

Model number	**Examiner 1**				**Examiner 2**				**Examiner 3**				Developed Standard scores
	Grades awarded		Negative point number		**Grades awarded**		Negative point number		**Grades awarded**		Negative point number		
	First occasion	Second occasion	First occasion	Second occasion	First occasion	Second occasion	First occasion	Second occasion	First occasion	Second occasion	First occasion	Second occasion	
5	**2**	**2**	3	4	**2**	**2**	4	4	**2**	**2**	4	4	Fail
8	**3**	**3**	5	4	**3**	**3**	4	6	**3**	**2**	3	6	Fail
15	**2**	**2**	7	5	**2**	**2**	7	8	**3**	**3**	5	4	Fail
16	**3**	**2**	4	5	**3**	**2**	6	8	**3**	**3**	5	3	Fail
36	**2**	**3**	4	3	**2**	**3**	3	2	**3**	**3**	2	3	Pass
39	**3**	**3**	3	4	**3**	**2**	5	4	**3**	**3**	3	5	Fail
40	**3**	**3**	3	4	**3**	**2**	4	3	**2**	**3**	2	2	Fail
41	**3**	**2**	4	4	**3**	**3**	6	6	**2**	**2**	4	2	Fail
43	**2**	**2**	5	5	**2**	**2**	6	7	**2**	**2**	4	5	Fail
46	**3**	**2**	4	6	**3**	**2**	4	6	**3**	**2**	3	5	Fail
53	**3**	**3**	4	4	**3**	**3**	3	4	**2**	**2**	4	5	Fail
54	**3**	**3**	5	4	**3**	**3**	5	5	**3**	**3**	4	4	Fail
57	**3**	**2**	5	5	**3**	**2**	4	6	**2**	**3**	4	4	Fail
62	**2**	**3**	4	4	**2**	**2**	5	4	**2**	**2**	4	4	Fail
73	**2**	**2**	7	9	**2**	**2**	8	8	**2**	**2**	7	8	Fail
78	**2**	**3**	6	5	**2**	**2**	4	4	**2**	**2**	5	6	Fail
80	**3**	**3**	5	5	**3**	**3**	5	4	**2**	**3**	5	5	Fail
83	**3**	**3**	4	5	**3**	**3**	6	4	**3**	**3**	4	4	Fail
85	**2**	**2**	5	5	**2**	**3**	8	5	**2**	**2**	5	5	Fail
87	**2**	**2**	4	5	**2**	**2**	3	3	**2**	**2**	5	5	Fail
88	**2**	**2**	5	5	**2**	**2**	6	6	**2**	**2**	6	7	Fail
94	**2**	**2**	4	5	**2**	**2**	5	7	**2**	**2**	6	6	Fail
109	**3**	**3**	2	3	**3**	**3**	5	4	**3**	**3**	4	3	Fail
111	**2**	**2**	7	8	**2**	**2**	7	7	**2**	**2**	6	5	Fail
120	**2**	**3**	5	4	**2**	**3**	5	5	**3**	**3**	4	4	Fail
138	**2**	**2**	6	7	**2**	**2**	8	6	**2**	**2**	7	6	Fail
Dis-agreement	**8** (31%)		18 (69%)		**8** (31%)		17 (65%)		**5** (19%)		15 (58%)		
Agreement	**18** (69%)		8 (31%)		**18** (69%)		9 (35%)		**21** (81%)		11 (42%)		

Table 7.7 shows that the best of agreement percentage according to grade and number of negative point was for examiner 3. By comparing the grades awarded by three senior academic staff examiners on each occasion with Developed Standard scores according to SAFMs and SAF evaluations (Table 7.7), the percentage agreement of these grades with Developed Standard scores (i.e. grade 5 and 4 = Pass, and 1, 2, and 3 = Fail) was 96% for each occasion. The percentage shows that most of the grades awarded from three senior examiners reflected 26 class II amalgam cavities truly.

Table 7.8 shows time taken to assess class II amalgam cavity preparations using “nCIIpc” for each senior examiner.

Table 7.8 Time spent (seconds) to assess class II amalgam cavity preparations by using the new Class II preparation checklist “nCIIpc”

Method	**Occasion**	**Examiner 1**	**Examiner 2**	**Examiner 3**	**Average time spent per stage** (seconds)	**Average time spent per model** (seconds)
		Time spent for evaluation (seconds)				
The new Class II preparation checklist “nCIIpc”	1	3240	3350	3254	3277	126
The new Class II preparation checklist “nCIIpc”	2	2697	2658	2472	2.609	100
						Average 113

Table 7.9 shows for each examiner of grading the Cohen’s Kappa statistic as calculated in SPSS to assess agreement. To calculate the strength of agreement, Landis and Koch (1977) have proposed values as standards for the strength of agreement (see Chapter 5).

Table 7.9 Measure of Kappa Agreement for each examiner according to grades by using the new Class II preparation checklist "nCIIpc"

	Examiner 1		Examiner 2		Examiner 3	
Method	*Kappa Value*	*Significance (p≤0.05)*	*Kappa Value*	*Significance (p≤0.05)*	*Kappa Value*	*Significance (p≤0.05)*
The new Class II preparation checklist "nCIIpc"	0.380	0.052	0.373	0.054	0.601	0.002

The high-lighted values represent statistically significant agreement differences.

From Table 7.9 it is apparent that, using the new Class II preparation checklist "nCIIpc", there is moderate intra-examiner agreement for only examiner 3. In addition, it shows that examiners 1 and 2 do not have good intra-examiner agreement according to grades.

Inter-examiner agreement for class II amalgam cavity preparation according to grades by using the new class II amalgam cavity preparation checklist "nCIIpc" for each occasion

Table 7.10 Inter-examiner agreement of class II amalgam cavity preparation for each occasion according to grades by using the new class II preparation checklist "nCIIpc"

Inter-examiner agreement for Class II amalgam cavity preparation					
Method	*Occasion*	*Number of examiners*	*Single measurement ICC*	*95 % of CI*	*Best single measurement if examiner deleted*
The new Class II preparation checklist "nCIIpc"	1	3	0.593	0.376 - 0.772	1.000 if examiner 3 is excluded
The new Class II preparation checklist "nCIIpc"	2	3	0.433	0.190 – 0.662	0.540 if examiner 3 is excluded

The highlighted values represent the highest inter-examiner agreement.

Table 7.10 demonstrated the agreement between senior academic staff by using the new Class II preparation checklist "nCIIpc" for evaluation of 26 class II amalgam cavity preparations, indicating the occasion 1 assessment was better than the occasion 2.

Table 7.10 also shows that the new Class II preparation checklist "nCIIpc" produced the best agreement among some senior examiners, if one of senior academic staff (examiner 3) was excluded.

In general, the "nCIIpc" with amalgam condenser produced similar inter-examiner agreement according to awarded grades compared with when the same examiners used 'Gray feedback sheet' (see Chapter 5).

Intra-examiner agreement for class II amalgam cavity preparation according to negative points by using the new class II amalgam cavity preparation checklist "nCIIpc" for each examiner

Intra-examiner agreement for the number negative points awarded by each senior academic staff examiner on each of two occasions is shown in Table 7.11. The highest value was for examiner 3 while the lowest value was for examiner 1.

Table 7.11 Cohen's Kappa of the number of negative points in the first and second occasion for each examiner by using the new class II preparation checklist "nCIIpc"

Examiners	Kappa value	Significance ($p \leq 0.05$)
Examiner 1	0.056	0.593
Examiner 2	0.205	0.019
Examiner 3	0.266	0.005

The high-lighted value represents statistically significant agreement differences.

Inter-examiner agreement for class II amalgam cavity preparation according to negative points by using the new class II amalgam cavity preparation checklist "nCIIpc" for each occasion

The inter-examiner agreement was evaluated using intra-class correlation (ICC). As displayed in Table 7.12, there was moderate agreement among three senior academic staff, (0.543) and (0.503), for occasion 1 and 2 respectively. By process of elimination, this table also shows that the best internal consistency was for examiner 3 who was the only examiner not excluded when the best single measurement was determined.

Table 7.12 Inter-examiner agreement among senior academic staff who assessed class II amalgam cavity preparations according to negative points for each occasion by using "nCIIpc"

Occasion	**Intra-class correlation** single measure	**95% confidence interval**		**Significance ($p \leq 0.05$)**	**Best single measurement if examiner deleted**
		Lower bound	Upper bound		
1	0.559	0.325	0.751	0.000	0.570 if examiner 1 is excluded
2	0.495	0.265	0.704	0.000	0.614 if examiner 2 is excluded

Intra- and inter-examiner agreement according to the number of negative points by using "nCIIpc" was lower than the result of intra- and inter-examiner agreement of 'Gray feedback sheet' (see Chapter 5).

Intra-examiner repeatability according to criteria (consistency) of the new checklist for class II amalgam cavity preparation "nCIIpc"

Table 7.13 summarises the Cohen's Kappa scores for intra-examiner agreement at the level of the criteria for the new checklist for class II amalgam cavity preparation "nCIIpc" for each senior examiner.

Table 7.13 Intra-examiner agreement and percentage agreement between three senior examiners according to criteria of the new checklist for class II amalgam cavity preparations "nCIIpc" on two occasions

	Examiner 1			Examiner 2			Examiner 3		
	Kappa agreement	Significance ($p \leq 0.05$)	%	Kappa agreement	Significance ($p \leq 0.05$)	%	Kappa agreement	Significance ($p \leq 0.05$)	%
Occlusal cavity									
Depth of the occlusal cavity at the isthmus area	0.615	0.000	81	0.500	0.000	69	0.677	0.000	85
Width of the occlusal cavity at isthmus area	0.435	0.007	84	0.435	0.007	73	0.345	0.020	73
Retention form	1.000	0.000	96	0.581	0.001	81	0.480	0.000	92
Marginal ridge thickness	1.000	0.000	100	0.661	0.000	89	1.000	0.000	100
Proximal box									
Depth of the box gingivally (occlusal-gingival direction)	0.668	0.000	89	0.534	0.000	77	0.639	0.000	96
Depth (mesio-distal direction) at gingival floor	0.736	0.000	89	0.603	0.000	85	0.765	0.000	92
Width (buccal-lingual/palatal direction) at gingival floor	0.636	0.000	85	0.740	0.000	85	0.577	0.000	77
Retention form	-0.062	0.586	73	0.388	0.004	69	0.180	0.143	62
Position of the box and unsupported enamel	0.304	0.036	77	0.541	0.005	77	0.538	0.002	77
Damage to adjacent tooth	0.872	0.000	92	0.315	0.016	58	0.728	0.000	85

The high-lighted values represent statistically significant agreement differences.

According to Landis and Koch (1977), examiner 1 demonstrated almost perfect intra-examiner agreement for 3/10 criteria of the "nCIIpc", substantial intra-examiner agreement for a further 4/10 criteria, moderate intra-examiner agreement for 1/10 criteria, poor intra-examiner agreement for 1/10 and fair intra-examiner agreement for 1/10 criteria. Examiner 1 had problem to assess 'retention form', 'position of the box and unsupported enamel' by using "nCIIpc" with amalgam condenser. Examiner 2 demonstrated substantial intra-examiner agreement for 2/10 criteria of the "nCIIpc", moderate agreement for a further 6/10 criteria, and fair intra-examiner agreement for

2/10 criteria. Examiner 2 had problem to assess ‘retention form’ and ‘damage to the adjacent tooth’ by using “nCIIpc” with amalgam condenser. Examiner 3 demonstrated almost perfect intra-examiner agreement for 1/10 criteria of the, “nCIIpc”, substantial agreement for a further 4/10 criteria, moderate intra-examiner agreement for 3/10 criteria and fair intra-examiner agreement for 2/10 criteria. Examiner 3 had problem to assess ‘retention form’ and ‘width of the occlusal cavity at isthmus’ by using “nCIIpc” with amalgam condenser.

Furthermore, there was no almost perfect intra-examiner agreement by all three senior examiners for any criterion of the, “nCIIpc”, while there was substantial intra-examiner agreement by all examiners for ‘Depth (mesio-distal direction) at gingival floor’ criterion of the, “nCIIpc”. For examiner 1 and 3, almost perfect intra-examiner agreement was observed for ‘Marginal ridge thickness’ criterion of the “nCIIpc”.

All other intra-examiner agreement varied between fair and moderate. Indeed, the lowest level of intra-examiner agreement and was not significant for criteria of the, “nCIIpc”, was only for, ‘Retention form of the proximal box’ for examiners 1 and 3. According to the agreement percentage, the percentages were acceptable for all criteria of all examiners. Again, the lowest percentage agreement was for ‘Retention form of the proximal box’.

By comparison, these results are better than the results of intra-examiner agreement using the ‘Gray feedback sheet’ according to the levels of criteria [see Chapter 5 (Table 5.23)].

Inter-examiner reproducibility according to criteria (consistency) of the new checklist for class II amalgam cavity preparation “nCIIpc”

Table 7.14 Inter-examiner agreements (single measures) and confidence intervals for criteria for the new checklist for class II amalgam cavity preparations “nCIIpc” among three senior examiners for each occasion

	Occasion 1				Occasion 2			
	Intra-class correlation *(Single measures)*	95% confident interval *lower*	95% confident interval *upper*	Significance ($p \leq 0.05$)	Intra-class correlation *(Single measures)*	95% confident interval *lower*	95% confident interval *upper*	Significance ($p \leq 0.05$)
Occlusal cavity								
Depth of the occlusal cavity at the isthmus area	0.523	0.292	0.725	0.000	0.576	0.353	0.762	0.000
Width of the occlusal cavity at isthmus area	0.245	0.016	0.505	0.019	0.251	0.029	0.506	0.013
Retention form	-0.044	-0.204	0.196	0.650	0.030	-0.097	0.231	0.336
Marginal ridge thickness	0.498	0.262	0.708	0.000	0.731	0.559	0.857	0.000
Proximal box								
Depth (occlusal-gingival direction)	0.524	0.291	0.726	0.000	0.580	0.362	0.763	0.000
Depth (mesio-distal direction) at gingival floor	0.604	0.391	0.779	0.000	0.667	0.471	0.818	0.000
Width (buccal-lingual/palatal direction) at gingival floor	0.768	0.612	0.878	0.000	0.703	0.518	0.841	0.000
Retention form	0.022	-0.167	0.284	0.407	0.119	-0.090	0.384	0.145
Position of the box and unsupported enamel	-0.003	-0177	0.248	0.495	0.028	-0.151	0.280	0.384
Damage to adjacent tooth	0.685	0.470	0.835	0.000	0.765	0.605	0.877	0.000

The highlighted values represent statistically significant agreement differences.

Table 7.14 summarises the intra-class correlation measurements (single measures) to determine inter-examiner agreement at the level of the criteria for the “nCIIpc” among senior examiners.

There was moderate to substantial inter-examiner agreement for the ‘depth of the occlusal cavity at the isthmus area’, ‘the marginal ridge thickness’, ‘the depth of the proximal box in occlusal-gingival direction’, ‘the depth of the proximal box (mesio-distal direction) at gingival floor’, ‘the width of the proximal box (buccal-lingual/palatal direction) at gingival floor’ as well as ‘damage to the adjacent tooth’. Each of these features was evaluated using the “nCIIpc” with the amalgam condenser. On the other hand, there was also poor inter-examiner agreement for ‘retention form of occlusal cavity and proximal box’, ‘position of the proximal box and unsupported enamel’ features.

In general, inter-examiner agreement according to criteria for occasion 2 of using “nCIIpc” was better than occasion 1. There was also better inter-examiner agreement according to criteria using “nCIIpc” than ‘Gray feedback sheet’ [see Chapter 5 (Table 5.24)].

b. For the full veneer gold shell crown preparation

<u>Intra-examiner agreement for full veneer gold shell crown preparation according to grades by using the new full veneer gold shell crown preparation checklist “nGSCpc”</u>

Table 7.15 demonstrates three senior academic staff examiners’ grades awarded for each full veneer gold shell crown preparation on each occasion of grading and the number of negative point on each occasion with their agreement percentages using “nGSCpc” and Developed Standard scores.

Table 7.15 Three senior academic staff examiners' grades awarded for each full veneer gold shell crown preparation on each occasion of grading and the number of negative points on each occasion with their agreement percentages using "nGSCpc" and Developed Standard scores

Model number	Examiner 1				Examiner 2				Examiner 3				Developed Standard scores
	Grades awarded		Negative point number		Grades awarded		Negative point number		Grades awarded		Negative point number		
	First occasion	Second occasion	First occasion	Second occasion	First occasion	Second occasion	First occasion	Second occasion	First occasion	Second occasion	First occasion	Second occasion	
1	**3**	**3**	4	4	**3**	**2**	2	4	**3**	**2**	3	4	Fail
3	**2**	**2**	3	3	**2**	**2**	7	4	**2**	**2**	2	4	Fail
4	**2**	**2**	5	4	**2**	**2**	7	3	**2**	**3**	3	1	Fail
5	**3**	**3**	1	2	**3**	**3**	1	1	**3**	**3**	5	1	Fail
7	**3**	**2**	1	2	**3**	**3**	1	2	**3**	**3**	1	1	Fail
13	**2**	**2**	5	6	**2**	**2**	3	4	**3**	**2**	1	3	Fail
14	**2**	**2**	3	3	**2**	**2**	4	4	**2**	**2**	3	5	Fail
18	**2**	**2**	5	6	**2**	**2**	7	7	**2**	**2**	6	6	Fail
20	**2**	**2**	1	1	**2**	**2**	2	1	**3**	**3**	2	1	Fail
21	**2**	**2**	2	2	**2**	**2**	3	2	**3**	**3**	2	6	Fail
25	**3**	**3**	1	2	**3**	**2**	1	3	**3**	**3**	1	3	Fail
26	**2**	**2**	3	4	**2**	**2**	4	4	**2**	**2**	3	3	Fail
29	**2**	**3**	3	3	**3**	**3**	1	2	**3**	**3**	4	3	Fail
31	**2**	**2**	2	3	**2**	**2**	2	2	**3**	**3**	1	1	Fail
51	**3**	**3**	3	2	**3**	**2**	3	3	**3**	**3**	1	2	Fail
52	**2**	**2**	1	3	**2**	**2**	2	3	**3**	**2**	1	3	Fail
54	**2**	**2**	5	4	**2**	**2**	6	5	**2**	**2**	4	5	Fail
57	**3**	**3**	2	2	**2**	**2**	5	4	**3**	**3**	1	2	Fail
58	**2**	**2**	5	5	**2**	**2**	6	7	**2**	**2**	5	5	Fail
59	**2**	**2**	6	7	**2**	**2**	5	6	**2**	**2**	6	6	Fail
60	**3**	**3**	2	2	**2**	**2**	6	3	**2**	**3**	2	1	Fail
63	**2**	**2**	4	3	**2**	**2**	5	4	**2**	**2**	3	3	Fail
67	**3**	**2**	1	2	**2**	**3**	4	2	**3**	**3**	4	4	Fail
69	**2**	**2**	3	3	**2**	**2**	2	4	**2**	**2**	2	3	Fail
70	**3**	**3**	2	2	**2**	**2**	5	3	**2**	**3**	3	1	Fail
71	**3**	**3**	3	2	**2**	**2**	5	2	**3**	**3**	1	2	Fail
73	**2**	**2**	3	2	**2**	**2**	5	5	**2**	**2**	1	2	Fail
74	**2**	**2**	3	3	**2**	**2**	5	4	**3**	**2**	1	2	Fail
78	**2**	**2**	2	2	**2**	**2**	4	3	**3**	**3**	3	2	Fail
88	**2**	**2**	6	7	**2**	**2**	8	8	**2**	**2**	5	5	Fail
Dis-agreement	**3** (10%)		**17** (57%)		**4** (13%)		**22** (73%)		**7** (23%)		**21** (70%)		
Agreement	**27** (90%)		**13** (43%)		**26** (87%)		**8** (27%)		**23** (77%)		**9** (30%)		

Table 7.15 showed that the best agreement percentage according to grade and number of negative point was for examiner 1. By comparing the grades awarded by three senior academic staff examiners on each occasion with Developed Standard scores according to SAFMs and SAF evaluations, the agreement percentage of these grades with Developed Standard scores (i.e. grade 5 and 4 = Pass, and 1, 2, and 3 = Fail) was 100% for each occasion. The percentage shows that most of the grades awarded from senior examiners reflected 30 full veneer gold shell crown preparations truly.

Table 7.16 shows time taking to assess full veneer gold shell crown preparation using “nGSCpc” for each senior examiner.

Table 7.16 Time spent (second) to assess full veneer gold shell crown preparations by using the new gold shell crown preparation checklist “nGSCpc”

Method	**Occasion number**	**Examiner 1**	**Examiner 2**	**Examiner 3**	**Average time spent per stage** (seconds)	**Average time spent per model** (seconds)
		Time spent for evaluation (seconds)				
the new gold shell crown preparation checklist “nGSCpc”	1	3518	3793	3933	3748	125
the new gold shell crown preparation checklist “nGSCpc”	2	2985	2849	2632	2822	94
						Average 110

Table 7.17 shows for each examiner of grading the Cohen’s Kappa statistic as calculated in SPSS to assess agreement.

Table 7.17 Measure of Kappa Agreement for each examiner according to grades by using the new gold shell crown preparation checklist "nGSCpc"

	Examiner 1		**Examiner 2**		**Examiner 3**	
Method	*Kappa Value*	*Significance ($p \le 0.05$)*	*Kappa Value*	*Significance ($p \le 0.05$)*	*Kappa Value*	*Significance ($p \le 0.05$)*
The new gold shell crown preparation checklist "nGSCpc"	0.769	0.000	0.524	0.000	0.533	0.003

Table 7.17 is apparent that using the new gold shell crown preparation checklist "nGSCpc" shows tremendous intra-examiner agreement for each of three examiners. In addition, it shows that examiner 1 ultimately demonstrate substantial agreement whereas examiner 2 and 3 display moderate agreement. All examiners have demonstrated improvement in agreement according to grades awarded.

Inter-examiner agreement for full veneer gold shell crown preparation according to grades by using the new gold shell crown preparation checklist "nGSCpc"

Table 7.18 Inter-examiner agreement of full veneer gold shell crown preparations for each occasion according to grades by using the new gold shell crown preparation checklist "nGSCpc"

Occasion	**Intra-class correlation** single measure	**95% confidence interval**		**Significance ($p \le 0.05$)**	**Best single measurement if examiner deleted**
		Lower bound	Upper bound		
1	0.437	0.216	0.647	0.000	0.508 if examiner 3 is excluded
2	0.318	0.103	0.546	0.000	0.475 if examiner 2 is excluded

Table 7.18 shows the agreement between senior academic staff by using the new gold shell crown preparation checklist "nGSCpc" for evaluation of 30 full veneer gold shell crown preparations,. Table 7.18 indicates occasion 1 assessment was better than occasion 2. Furthermore, the table shows that the new gold shell crown preparation

checklist “nGSCpc” produced better agreement among some senior examiners, if one of senior academic staff (i.e. examiner 3 for occasion 1, and examiner 2 for occasion 2) was excluded.

The new gold shell crown preparation checklist “nGSCpc” with assessment tools produced the better inter-examiner agreement according to grades than when the same examiners used ‘Mhanni feedback sheet’ [see Chapter 5 (Table 5.20)].

Intra-examiner agreement for full veneer gold shell crown preparation according to negative points by using the new gold shell crown preparation checklist “nGSCpc” for each examiner

Intra-examiner agreement for the number of negative points awarded by each senior academic staff examiner on each of two occasions is shown in Table 7.19. The highest value was for examiner 1 while the lowest value was for examiner 3.

Table 7.19 Cohen’s Kappa of the number of negative points on the first and second occasions for each examiner by using the new gold shell crown preparation checklist “nGSCpc”

Examiners	Kappa value	Significance ($p \leq 0.05$)
Examiner 1	0.304	0.000
Examiner 2	0.153	0.023
Examiner 3	0.130	0.134

The high-lighted value represents statistically significant agreement differences.

Inter-examiner agreement for full veneer gold shell crown preparation according to negative points by using the new gold shell crown preparation checklist “nGSCpc” for each occasion

Table 7.20 shows that there was moderate and substantial inter-examiner agreement among three senior academic staff, (0.417) and (0.660), for occasions 1 and 2, respectively. By process of elimination, Table 7.20 also shows that the best inter-examiner agreement according to the number of negative points was for examiner 1 who was the only examiner not excluded when the best single measurement was determined.

Table 7.20 Inter-examiner agreement among senior academic staff who assessed full veneer gold shell crown preparations according to negative points for each occasion by using “nGSCpc”

Occasion	**Intra-class correlation** single measure	**95% confidence interval**		**Significance ($p \leq 0.05$)**	**Best single measurement if examiner deleted**
		Lower bound	Upper bound		
1	0.417	0.186	0.635	0.000	0.520 if examiner 3 is excluded
2	0.660	0.475	0.805	0.000	0.767 if examiner 3 is excluded

Intra-examiner repeatability according to criteria (consistency) of the new checklist of full veneer gold shell crown preparation “nGSCpc”

Table 7.21 summarises the Cohen’s Kappa scores for intra-examiner agreement at the level of the criteria for the new checklist of full veneer gold shell crown preparation “nGSCpc” for each senior examiner.

Table 7.21 Intra-examiner agreement and percentage agreement between three senior examiners according to criteria of new full veneer gold shell crown preparation checklist "nGSCpc" on two occasions

	Examiner 1			**Examiner 2**			**Examiner 3**		
	Kappa agreement	Significance ($p \leq 0.05$)	%	Kappa agreement	Significance ($p \leq 0.05$)	%	Kappa agreement	Significance ($p \leq 0.05$)	%
Occlusal surface									
Functional cusps reduction	0.702	0.000	83	0.572	0.000	77	0.051	0.519	57
Non-functional cusps reduction	0.571	0.000	77	0.529	0.000	73	0.096	0.082	30
Bucco-lingual/palatal occlusal convergence	1.000	0.000	100	1.000	0.000	100	0.186	0.022	77
Proximal occlusal convergence	0.630	0.000	87	0.467	0.008	70	0.526	0.001	90
Axial surface(s)									
Axial reduction	0.526	0.001	90	0.399	0.001	63	0.224	0.060	60
Finish line									
Type of finish line	0.809	0.000	90	0.713	0.000	87	0.294	0.019	53
Depth of finish line all around	0.268	0.042	70	0.250	0.059	60	0.430	0.001	70
Level of finish line to gingival margin	0.263	0.017	63	0.769	0.000	90	-0.013	0.649	3
Adjacent teeth									
Contact area with the adjacent teeth	0.889	0.000	97	0.789	0.000	90	0.037	0.023	13
Damage to adjacent teeth	0.574	0.000	77	0.563	0.001	77	0.091	0.073	60

The high-lighted values represent statistically significant agreement differences.

According to Landis and Koch (1977), examiner 1 demonstrated almost perfect intra-examiner agreement for 3/10 criteria of the "nGSCpc", and substantial intra-examiner agreement for a further 2/10 criteria, moderate intra-examiner agreement for 3/10 criteria and fair intra-examiner agreement for 2/10 criteria. Examiner 2 demonstrated almost perfect intra-examiner agreement for 1/10 criteria of the "nGSCpc", substantial intra-examiner agreement for a further 3/10 criteria, moderate intra-examiner agreement for 4/10 criteria, fair intra-examiner agreement for 1/10 criteria and one criterion (i.e. depth of finish line all around) had significant agreement difference. Examiner 3 demonstrated moderate intra-examiner agreement for 2/10 criteria, fair intra-examiner agreement for 1/10 criteria, and slight intra-examiner agreement for 2/10 criteria while 5/10 criteria had significant agreement difference.

On no occasion was there the same level of agreement for all three examiners for any criteria of the full veneer gold shell crown preparation using the, "nGSCpc". For examiners 1 and 2, almost perfect intra-examiner agreement was observed for the, 'bucco-lingual/palatal occlusal convergence' criterion of the, "nGSCpc". In addition, substantial intra-examiner agreement was produced for the, 'type of finish line' and 'contact area with the adjacent teeth' criteria by examiner 1 and 2. Moderate intra-examiner agreement was observed for the, 'non-functional cusps reduction' criterion of "nGSCpc". All other intra examiner agreement varied between substantial and slight.

According to the result of the examiner 2 and 3, the lowest level of agreement was not significant for criteria of the, "nGSCpc", but was significant for 'functional cusps reduction', 'non-functional cusps reduction', 'axial reduction', 'depth of finish line all around', 'level of finish line to gingival margin', and 'damage to adjacent teeth'.

These results are better than results of intra-examiner agreement of the 'Mhanni feedback sheet' according to criteria [see Chapter 5 (Table 5.25)].

Inter-examiner reproducibility according to criteria (consistency) of the new checklist for full veneer gold shell crown preparation "nGSCpc"

Table 7.22 summarises the intra-class correlation measurements (single measures) to determine inter-examiner agreement at the level of the criteria for the "nGSCpc" among senior examiners. These data were the same data generated to determine intra-examiner agreement which had to take place over two occasions of evaluation. Thus, inter-examiner agreement for each occasion could be determined and there are some interesting comparisons between the two occasions.

Table 7.22 Inter-examiner agreements (single measures) and confidence intervals for criteria of "nGSCpc" among three senior examiners for each occasion

	Occasion 1				Occasion 2			
	Intra-class correlation *(Single measures)*	95% confident interval *lower*	*upper*	Significance (p≤0.05)	Intra-class correlation *(Single measures)*	95% confident interval *lower*	*upper*	Significance (p≤0.05)
Occlusal surface								
Functional cusps reduction	0.136	-0.056	0.378	0.091	0.285	0.060	0.524	0.000
Non-functional cusps reduction	0.313	0.097	0.543	0.001	0.239	0.039	0.471	0.004
Bucco-lingual/palatal occlusal convergence	0.230	0.014	0.474	0.018	0.326	0.104	0.558	0.002
Proximal occlusal convergence	0.481	0.264	0.681	0.000	0.327	0.081	0.569	0.000
Axial surface(s)								
Axial reduction	0.347	0.114	0.580	0.000	0.356	0.139	0.579	0.000
Finish line								
Type of finish line	0.587	0.386	0.756	0.000	0.463	0.241	0.668	0.000
Depth of finish line all around	0.045	-0.144	0.292	0.328	0.414	0.193	0.630	0.000
Level of finish line to gingival margin	0.237	0.034	0.470	0.002	0.068	-0.033	0.228	0.063
Adjacent teeth								
Contact area with the adjacent teeth	0.434	0.205	0.647	0.000	0.294	0.080	0.527	0.001
Damage to adjacent teeth	0.378	0.154	0.601	0.000	0.751	0.599	0.862	0.000

The high-lighted values represent statistically significant agreement differences.

Table 7.22 showed that there was fair to substantial inter-examiner agreement for two occasions. All of these features were evaluated using the "nGSCpc" with assessment tools (impression index, bur, and/or periodontal probe). On the other hand, there was no significant inter-examiner agreement for 'functional cusps reduction', 'depth of finish line all around' and 'level of finish line to gingival margin' criteria of the, "nGSCpc".

In general, there was better inter-examiner agreement according to criteria by using "nGSCpc" than 'Mhanni feedback sheet' [see Chapter 5 (Table 5.26)].

7.4.4 Comparison of intra- and inter-examiner agreement and consistency for the new checklists ("nCIIpc" and "nGSCpc") with intra- and inter-examiner agreement and consistency for the 'Gray and Mhanni Feedback sheets'

a. Intra- and inter-examiner agreement according to grades

Table 7.23 demonstrates comparison of Cohen's Kappa test (intra-examiner agreement) for the new checklist ("nCIIpc" and "nGSCpc") with Cohen's Kappa test (intra-examiner agreement) for the 'Gray and Mhanni Feedback sheets' according to the grades.

Table 7.23 Intra-examiner agreement according to grades awarded from three examiners using different methods

	Examiner 1		**Examiner 2**		**Examiner 3**	
Method	*Kappa Value*	*Significance ($p \leq 0.05$)*	*Kappa Value*	*Significance ($p \leq 0.05$)*	*Kappa Value*	*Significance ($p \leq 0.05$)*
The new Class II preparation checklist "nCIIpc"	0.380	0.052	0.373	0.054	0.601	0.002
'Gray feedback sheet' (see Chapter 5)	0.583	0.000	0.637	0.000	0.739	0.000
Examiners in different sequence	**Examiner 1**		**Examiner 2**		**Examiner 3**	
Method	*Kappa Value*	*Significance ($p \leq 0.05$)*	*Kappa Value*	*Significance ($p \leq 0.05$)*	*Kappa Value*	*Significance ($p \leq 0.05$)*
The new full veneer gold shell crown preparation checklist "nGSCpc"	0.769	0.000	0.524	0.000	0.533	0.003
'Mhanni feedback sheet' (see Chapter 5)	0.573	0.000	0.409	0.001	0.268	0.032

The high-lighted values represent statistically significant agreement differences.

According to the Chapter 6, the grades awarded from the examiners using feedback sheets were not representative for class II amalgam cavity and full veneer gold shell crown preparation. Although intra-examiner agreement by using new checklists was lower than using feedback sheets, the grades reflected the tooth preparations truly (i.e. valid and reliable grades).

Table 7.24 demonstrates comparison of inter-examiner agreement for the new checklist ("nCIIpc" and "nGSCpc") with inter-examiner agreement for the 'Gray and Mhanni Feedback sheets' according to the grades.

Table 7.24 Inter-examiner agreement according to grades awarded from three examiners using different methods

Inter-examiner agreement according to grades for class II cavities			
Method	Occasion	*Single measurement ICC*	*Significance ($p\leq0.05$)*
The new Class II preparation checklist "nCIIpc"	1	0.593	0.000
	2	0.433	0.000
'Gray feedback sheet' (see Chapter 5)	1	0.540	0.000
	2	0.692	0.000
Inter-examiner agreement according to grades for full gold crown preparations			
Method	Occasion	*Single measurement ICC*	*Significance ($p\leq0.05$)*
The new full veneer gold shell crown preparation checklist "nGSCpc"	1	0.437	0.000
	2	0.318	0.000
'Mhanni feedback sheet' (see Chapter 5)	1	0.342	0.000
	2	0.375	0.000

For class II amalgam cavity preparation, Table 7.24 shows that inter-examiner agreement according to grades by using "nCIIpc" tended to be higher than inter-examiner agreement by using 'Gray feedback sheet' on occasion 1. On occasion 2, inter-examiner agreement of grades by using new checklist was lower than using 'Gray feedback sheet'. Although occasion 2 was low, the grades reflected 26 class II amalgam cavities truly (i.e. valid and reliable grades). In addition, there were **eleven** class II amalgam cavities which had an agreed grade among three examiners on two occasions by using "nCIIpc" and amalgam condenser (Table 7.7), while 'Gray feedback sheet' there were only **four** class II amalgam cavities which had agreed grade among three examiners on two occasions (see Chapter 5).

For full veneer gold shell crown preparation, Table 7.24 shows also that inter-examiner agreement according to grades by using “nGSCpc” was higher than inter-examiner agreement by using ‘Mhanni feedback sheet’ on occasion 1. On occasion 2, inter-examiner agreement using new checklist was lower than using ‘Mhanni feedback sheet’. Although occasion 2 for “nGSCpc” was low, the grades reflected 30 full veneer gold Shell crown preparation truly according to SAFMs and SAF evaluations. From Table 7.15, there were **twelve** full veneer gold shell crown preparations that had agreed grades among three examiners on two occasions by using “nGSCpc” with selected reliable tools, whereas ‘Mhanni feedback sheet’ was only **four** full veneer gold shell crown preparations which had agreed grades among three examiners on two occasions (see Chapter 5).

In general, inter-examiner agreement for class II amalgam cavities and full veneer gold shell crown preparations was improved by using new checklists (i.e. “nCIIpc” and “nGSCpc”) with selected reliable tools.

b. Intra- and inter-examiner agreement according to criteria (consistency)

Intra- and inter-examiner agreements according to criteria (consistency) for the new checklists (i.e. “nCIIpc” and “nGSCpc”) with selected reliable specific tools were sometimes improved. Tables 5.23, 5.24, 5.25, 5.26, 7.13, 7.14, 7.21 and 7.22, demonstrate these improvements in comparison with intra- and inter-examiner agreement according to criteria (consistency) for the ‘Gray and Mhanni Feedback sheets’.

For intra-examiner agreement according to criteria (consistency) for class II amalgam cavity preparation, it is clear that there was high level of intra-examiner agreement according to criteria (consistency) for almost of the “nCIIpc” criteria. There was only one criterion of “nCIIpc” which had low intra-examiner agreement. This criterion was ‘retention form of the proximal box’ (Table 7.13). On the other hand, the ‘Gray feedback sheet’ had more than one criterion which had low intra-examiner agreement according to criteria (Table 5.23).

Furthermore, there was high level of inter-examiner agreement according to criteria (consistency) for almost of the “nCIIpc” criteria. There were three criteria of the “nCIIpc” which had low intra-examiner agreement according to criteria (i.e. not significant agreement). These criteria were ‘retention form of the occlusal cavity’, ‘retention form of the proximal box’ and ‘position of the box and unsupported enamel’ (Table 7.14). On the other hand, the ‘Grey feedback sheet’ had more than one criterion which had low intra-examiner agreement according to criteria (Table 5.24).

For intra-examiner agreement according to criteria (consistency) for full veneer gold shell crown preparation, it is also clear that there was high level of intra-examiner agreement according to criteria (consistency) for almost all of the “nGSCpc” criteria. All the criteria of “nGSCpc” for examiner 1 improved. On the other hand, there were up to four criteria of “nGSCpc” which had low intra-examiner agreement according to criteria (i.e. not significant agreement) for examiner 2 and examiner 3 (Table 7.21). On the other hand, ‘Mhanni feedback sheet’ had more than one criterion which had low intra-examiner agreement according to criteria (Table 5.25).

According to inter-examiner agreement according to criteria (consistency) for full veneer gold shell crown preparation, there was also a higher level of inter-examiner agreement according to criteria (consistency) for almost all of the “nGSCpc” criteria than ‘Mhanni feedback sheet’. There were two criteria on occasion 1 and one criterion of “nGSCpc” on occasion 2 that had low inter-examiner agreement according to criteria (i.e. not significant agreement) (Table 7.22). On the other hand, there were four criteria on occasion 1 and seven criteria of ‘Mhanni feedback sheet’ on occasion 2 had low inter-examiner agreement according to criteria (Table 5.26).

7.5 Discussion

Accurate evaluation of dental students’ ability to prepare teeth in the clinical skills laboratory is a most critical component of the dental education process. For evaluation to be effective it should provide consistent and accurate feedback and grades for students to help them to achieve a high level of competency before proceeding to patient care (Renne et al., 2013). Designing a system of assessment is not easy. Several studies in the dental literature have described assessment tools but they all have advantages and disadvantages (see Chapter 1).

According to the course guide for Conservative and Fixed Prosthodontic at Dundee Dental School (see Chapter 2), the Conservation Course uses a system assessment form ranked from grade 1 to grade 5, while the Fixed Prosthodontic Course uses a system assessment form according to a pass/fail decision. Therefore, this researcher (AM) developed the “nCIIpc” and “nGSCp” forms to be used in assessment of class II amalgam cavity and full veneer gold shell crown preparation respectively. In formative feedback, the pass/fail decision does not provide sufficient constructive information to the students to further develop their performance. For example, if the student fails, they do not know what the part of tooth preparation needs work to pass at the next attempt. Therefore, formative feedback along with a ranking system is helpful and useful. On the other hand, summative assessment can be used to determine pass/fail scores.

Criteria and levels of performance without objective statements are the most common problem associated with the feedback sheet(s) or checklist(s). Brown (1930) did not find any inter-examiner agreement utilising checklists without objective definition for each criterion. Gaines et al., (1974) compared two checklists for crown preparation. The first checklist consisted of six assessment items each scoring 0 to 5. The second checklist contained objective statements for each score in each item. Inter-examiner agreement using the first checklist was 0.26, this increased to 0.56 with the second checklist (Gaines et al., 1974). Therefore, creating a reliable checklist with objective descriptions of the performance (descriptors) for each criterion will increase the agreement. The researcher (AM) created criteria of the new “nCIIpc” and “nGSCpc” according to criteria of feedback sheet(s) which were the most repeatable and reproducible according to the grades awarded. In addition, “nCIIpc” and “nGSCpc” with up to 70% objective criteria were also created. Thus, setting up an assessment system (new checklist) was done.

Additionally, the researcher (AM) determined the most reliable specific additional tool(s) which had already been used in Chapter 5 to be also used with the new checklists. There were few studies which mentioned the advantages of using tools to assess tooth preparations (Schiff et al., 1975, Haj-Ali and Feil, 2006, Ahmed et al., 2016). Schiff et al., (1975) developed a tool called the ‘pulpal floor measuring instrument’ to measure the profile of preparations including depth, smoothness, and flatness of the pulpal floor. Haj-Ali and Feil (2006), and Ahmed et al., (2016) selected the periodontal probe as a measuring tool. According to Ahmed et al., (2016), the study concluded that there was wide intra-examiner variation was noted while inter-examiner reliability was improved after calibration by using the instrument. After 6 months intra- and inter-examiner reliability decreased. The authors demonstrated a decrease in inter- and intra-examiner reliability because there was not enough frequent calibration sessions for the examiners to maintain an optimum level of calibration (Ahmed et al., 2016).

Most of the full crown preparations were assessed using computer assisted devices and software. Although the authors reported there was significant improvement in the reliability and consistency, this researcher (AM) preferred to select a familiar instrument to be used as a tool from the tooth preparation kit. Tool selection for assessment from the kit or procedure also helps the students because they had already used or were familiar with these tools during tooth preparation procedures.

Therefore, items such as the amalgam condenser were selected as a reliable assessment tool for class II amalgam cavity preparation according to agreement with the Developed Standard scores which were awarded from SAFMs and SAF evaluations, while impression index and chamfer bur or periodontal probe (according to opinion of one senior examiner) were selected as reliable assessment tools for full veneer gold shell crown preparation. The researcher (AM) was unable to find a study which explained the way to select a reliable tool in dentistry. Thus, the way which was mentioned in this chapter cannot compare with other studies.

Standard setting is the cut-off score/grade that identifies the consequences of the assessment and determines who passes and who fails. Therefore, it is very important to develop a new checklist with a cut-off point. Borderline regression was used to determine the number of negative points (errors) is acceptable to still pass the assessment. This is most commonly used for OSCEs but can be also used to other forms of assessments (Puryer and O’Sullivan, 2015). Pass and fail scores (Developed Standard scores) from SAFMs and SAF evaluations were converted to grades (five scale point) and the number of negative points (errors) were used to calculate an absolute standard setting. The absolute standard setting was determined by using a borderline linear regression analysis between grades and the number of negative points (errors) (Schoonheim-Klein et al., 2009, Puryer and O’Sullivan, 2015). Scores and errors which were awarded from SAFMs and SAF evaluations might reduce the subjective nature of the standard setting. Puryer and O'Sullivan (2015) stated that *"All of the absolute standard methods require judgment… examiners are required to observe and rate a student's performance… subjective judgement is used and this may be criticised".* In addition, Zeiky et al., (2006) reported that there was no purely objective method for determining the cut-off score. In this chapter, the researcher (AM) reduced the subjectivity by selecting the scores from SAFMs and SAF evaluations according to Knight's recommendations as well as the number of negative points (errors) for each tooth preparation.

After determining the reliable tools and standard setting, new checklists were created according to 'Gray and Mhanni feedback sheets' and from the literature. New checklists with specific criteria allowed a more analytical form of assessment where each criterion contributing to the entire performance is evaluated separately according to a written checklist. These checklists defined specific levels of performance of each criterion. This method was used in the study of Haj-Ali and Feil (2006) and Ahmed et al., (2016). To confirm that the new checklists covered all features of the tooth preparation; face validity was also used as sub-type of content validity. It was used because the other types of validity were already used for 'Gray and Mhanni feedback sheets' in the Chapter 5.

Some features of tooth preparation were excluded because the measurement for some features was very difficult to estimate such as occlusal key feature. From the text books, occlusal extension has different shapes, occlusal lock (Akpata et al., 2013), and Dovetail (Roberson et al., 2002). This subjective feature may not even be needed at all (Roberson et al., 2002). According to Hilton et al., (2013), the preparation for class II amalgam preparations should not be extended further into a sound occlusal surface to provide retention of the proximal restoration, because this will weaken the tooth resistance to fracture. Although opening of the occlusal fissure has been advised in dental school, which is in line with Black's famous concept of 'extension for prevention' (Black, 1955), Almquist et al., (1973) proposed the omission of the opening of the occlusal fissure except to treat a carious lesion. The old axiom of 'extension for prevention' was thus discarded.

Reliable tools and new checklists were provided for three senior academic staff examiners to assess 26 class II amalgam cavities and 30 full veneer gold shell crown preparations in order to determine examiner reliability and consistency to reduce subjectivity. These new checklists were provided up to 70% objective assessment according to SAFMs, while up to 40% subjective assessment according to SAF evaluations. Even with a greater guarantee of objective assessment, the new checklists were not able to assess all the criteria of tooth preparation purely objective. SAFMs were used to develop new checklists which were also utilised in assessment using software such as Kavo Prepassistant machine to evaluate full crown preparations in clinical skills laboratories. Similarly, 70% of assessment of full crown preparation was evaluated objectively by using Kavo Prepassistant (Cardoso et al., 2006).

By using new checklists with assessment tools, the time spent to assess each tooth preparation was about two minutes. If the examiners spent more than 30 minutes assessing the tooth preparations, they might become tired and impacted on the result of the assessment. According to Caro et al., (1979), long period of observation can be tiring and vigilance will decrease over time. They also suggested no significant differences between records made in the first and last 10 minutes of a 30 minute session (Caro et al., 1979).

The best agreement percentage of grades and the negative point number of "nCIIpc" were for examiner 3. This means that the examiner 3 was reliable to assess the students' performances by using "nCIIpc". Furthermore, the percentage of agreement for examiner 3 scores (pass/fail) with Developed Standard scores (pass/fail) was 96% for each occasion. The grades or scores of the examiner 3 truly reflected 25 class II amalgam cavities. Thus, examiner 3 can be selected as a standard examiner.

For full veneer gold shell crown preparation, grades of examiner 1 were reliable. In addition, the percentage of agreement for examiner 1 scores (pass/fail) with Developed Standard scores was 100% for each occasion. Thus, examiner 1 can be selected as a standard examiner.

Most other studies used the gold standard examiner or grades according to the level of qualification or the average of the examiners. This selection might be un-reliable because the results might not be a true reflection of the tooth preparations.

From the results of new checklists in this chapter, it is apparent that examiners must be calibrated and trained in using the new method until they become familiar with the new method (Knight, 1997). Consequently, the examiner will be more likely to provide reliable and fair feedback to students.

According to the consistency of class II amalgam cavity evaluation, the results of intra- and inter-examiner agreement according to criteria (consistency) were better than results of intra- and inter-examiner agreement of the 'Gray feedback sheet'. All of the objective criteria and their levels of performance using "nCIIpc" were repeatable and reproducible.

According to the results of full veneer gold shell crown preparation, intra- and inter-examiner agreement according to criteria (consistency) were improved for some criteria in comparison with intra- and inter-examiner agreement of the 'Mhanni feedback sheet'. Although intra- and inter-examiner agreement according to criteria by using new checklists ("nCIIpc" and "nGSCpc") were sometimes lower than using feedback sheets ('Gray and Mhanni Feedback sheets'), the assessments still produced valid and reliable grades the tooth preparations. According to Helft et al., (1986), clearly defined criteria and their levels of performance improve the consistency of the assessment method.

7.6 Conclusion

The amalgam condenser was selected as an assessment tool to assess class II amalgam preparations with new checklist “nCIIpc”. An impression index and bur (or similar such as a periodontal probe with millimetre graduations) tools with new checklist “nGSCpc” were used to assess full veneer gold shell crown preparations. These tools were used to determine intra- and inter-examiner agreement according to the grades and consistency of criteria for class II amalgam preparation and full veneer gold shell crown preparations and there data compared with data from the ‘Gray and Mhanni feedback sheets’ to ascertain if the agreement and consistency of the new Checklists had been improved.

By using new checklists with assessment tools, the time spent assessing each tooth preparation was about two minutes. The researcher (AM) concluded that the examiners could spend about 20 to 30 minutes assessing the tooth preparations without becoming tired and bored. If the total evaluation time was more than 30 minutes, this might have impacted on the result of the assessment.

According to Cohen’s Kappa and intra-class correlation tests, the new class II amalgam preparation “nCIIpc” and full veneer gold shell crown preparation checklists “nGSCpc”, sometimes improved intra- and inter-examiner agreement and consistency of grades awarded by all senior examiners in comparison with previous intra- and inter-examiner agreement and consistency of the ‘Gray and Mhanni feedback sheet’ grades from Chapter 5. This can be explained as one of the senior examiners was not familiar with the new checklists. Thus, the results of this examiner impacted on intra- and inter-examiner agreement. According to agreement percentages for the new checklists, there is significant enhancement. In addition, the new checklists improved the intra- and inter-examiner agreement according to the number of negative points in comparison with the feedback sheets. The consistency of the criteria for both checklists was improved.

In this study, the grades/scores of the examiner 3 for class II amalgam preparations and the grades/scores of the examiner 1 for full veneer gold shell crown preparations using new checklists can be selected as valid grades for these preparations. This is because these grades, awarded from the new checklists, are more representative of the tooth preparations than grades obtained using the other feedback sheets. On the other hand, using the new checklist(s) for the sample has not provided both categorical scores (Pass/Fail). Most of the scores were fails due to the sample of tooth preparations was created only by students. Thus, these cavities were prepared by inexperienced operators and therefore were generally of lower quality which narrowed the range of cavities available for assessment. Therefore, this was the main drawback of this study. Sample selection according to experience may confirm validity and reliability of the new checklist(s) and scores (Pass/Fail).

In addition, some of the examiner(s) demonstrated reduced intra- and inter-examiner agreement and consistency. The reasons are possibly: i) the examiner was not familiar with the new checklists and/or assessment tool(s), and ii) spending too long time (more than 30 minutes) assessing might have had a significant impact on the results.

Therefore, it is essential to provide training sessions to familiarise and calibrate examiners who should not assess for longer than 30 minutes without a break in order to achieve better agreement.

In addition, the result of this part of study emphasises the difficulty of assessing tooth preparations using a feedback sheet in comparison of checklist.

Eventually, the null hypothesis is rejected because the new checklists generally improved the agreement and consistency.

Chapter 8：General conclusions, recommendations and further studies

Chapters	
Chapter 2	**Evaluating current assessment and feedback methods at Dundee Dental School in the Clinical Skills Laboratory** **Aims:** The aims of this Chapter were to identify, in relation to the class II amalgam cavity preparation and the full veneer gold shell crown preparation undertaken in the Clinical Skills Laboratory at Dundee Dental School: 1. the types and quality of feedback and assessment provided for students and, 2. the disadvantages associated with current feedback.
Chapter 3	**Independent evaluation of student class II amalgam cavity and full gold shell-crown preparations by senior academic staff** **Aim:** The aim of this study was to evaluate inter-examiner agreement between a convenience sample of three senior academic staff members, who have more than twenty years of clinical and teaching experience each, when evaluating operative procedures (class II amalgam cavity and full veneer gold shell crown preparation) on plastic teeth, from a sample year-cohort of dental students, at Dundee Dental School.
Chapter 4	**Evaluation of selected student class II amalgam cavity and full gold shell-crown preparations by senior academic staff and additional teaching staff** **Aims:** 1. To develop a sub-set of class II amalgam cavity, and full veneer gold shell crown, preparations. 2. To determine intra-examiner agreement for a group of senior academic staff when evaluating a sub-set of class II amalgam cavity preparations and full-veneer gold shell crown preparations. 3. To study a large group of additional teaching staff at Dundee Dental School in order to determine intra- and inter-examiner agreement when evaluating a sub-set of class II amalgam cavity preparations and full-veneer gold shell crown preparations.
Chapter 5	**Evaluation of selected student class II amalgam cavity and full gold shell crown preparations by using specific additional tools and feedback sheets** **Aims:** 1. To identify and develop specific additional tools to assist with evaluation of tooth preparations. 2. To determine examiner agreement for both class II amalgam cavity and full veneer gold shell crown preparations using the specific additional tools and feedback sheets for both grades awarded and repeatability and reproducibility of detailed feedback provided by senior academic staff.
Chapter 6	**Identification of gold standard (representative grades) for 26 class II amalgam cavities and 30 full gold shell-crown preparations** **Aims:** These objectives can be set out as a number of steps, each of which had a specific aim which is expanded within the methodology of this chapter. Thus, the two aims of this chapter were: 1. to identify the best senior academic staff examiner who had the highest correlation between grades awarded with the number of negative points identified; 2. to identify those tooth preparations which were awarded a grade by the best examiner in order to: • determine the specific feature measurements of this sub-group of grade tooth preparations and: • compare the objective class II amalgam cavity or full veneer gold shell crown preparation dimensions (SAMFs) recorded by the researcher and those subjective tooth preparation evaluations (SAFs) for similar teeth by the senior examiner with dimensions presented in the dental literature and subsequent calibration with the grades awarded by the best examiner.
Chapter 7	**Identification of reliable tools and development of a new checklist to assess class II amalgam cavities and full veneer gold shell crown preparations** **Aims:** 1. To identify a reliable specific additional measurement tool. 2. To develop two reliable new checklists (New Checklist CII Preparation and New Checklist Gold shell Crown Preparation). 3. To determine intra- and inter-examiner agreement using the grades and consistency of the New Checklists (Class II Preparation and Gold Shell Crown Preparation) and compare these data with that from previously used 'Gray and Mhanni feedback sheets' to ascertain if the agreement and consistency of the New Checklists have been improved.
Chapter 8	**General conclusions, recommendations, and further studies**

8.1 General conclusion

The principle findings of this PhD thesis are:

8.1.1 Chapter 2

The Researcher (AM) attended Clinical Skills Laboratory sessions at Dundee Dental School, for the class II amalgam cavity preparation and full veneer gold shell crown preparation to determine whether or not the feedback and assessment satisfied the undergraduate students.

The observations of the researcher (AM) together with students' comments, feedback concluded that assessment by tutors was preferred to reading books or peer evaluation from other students. Although feedback from the tutor was preferred, the students complained that the feedback was not prompt, not clear, and did not contain objective evaluations, such as acceptable dimensions or features, to guide them. To ensure these comments from the students were addressed, Chapters 3, 4, 5 and 6 provided information about the validity and reliability of the quality of the assessment and consistency of feedback at Dundee Dental School for work undertaken in the Clinical Skills Laboratory.

8.1.2 Chapter 3

The ability of the senior academic staff examiners to evaluate operative procedures on plastic teeth for students in relation to the class II amalgam cavity and full veneer gold shell crown preparation was evaluated. The levels of inter-examiner agreement of three senior academic examiners for the class II amalgam cavity preparation and the full gold shell crown preparation were low and slight, respectively. This was most likely because senior examiners came from a varied background and used their own scaling systems and criteria.

8.1.3 Chapter 4

The additional teaching staff examiners demonstrated poor to moderate intra- and inter-examiner agreement. A minority of these teachers had moderate agreement. The level of intra-examiner agreement for senior examiners was low.

In general, the levels of intra- and inter-examiner agreement were disappointing for both senior academic staff and the large group of additional teaching staff at Dundee Dental School. These results indicate the challenge of providing reliable assessment at Dundee Dental School. These results support the comments of the dental students in the Chapter 2.

8.1.4 Chapter 5

In this Chapter, for the three senior examiners, a five-stage process using five different assessment methods was used to determine the best method(s) to improve intra- and inter-examiner agreement. The results for intra-examiner agreement showed the grades awarded for the class II amalgam cavity preparation improved through the cumulative stages of grading whereas the full veneer gold shell crown preparation did not always improve. The results for inter-examiner agreement of the grades awarded for the class II amalgam cavity and the full veneer gold shell crown preparation did not improve.

According to the number of negative points awarded from the feedback sheet (Fifth stage), intra-examiner agreement of senior academic staff did not improve, while inter-examiner agreement did improve. To confirm these results were reliable, intra- and inter-examiner agreement according to consistency for each criterion of feedback sheets was calculated.

The repeatability and reproducibility of detailed comments from the feedback sheets was also analysed. The repeatability for each senior examiner (intra-examiner agreement) was better than reproducibility among examiners (inter-examiner agreement). In addition, the criteria which had objective levels (SAFMs) provided better agreement than the subjective criteria (SAFs). Although the use of specific additional tools and the feedback sheets to assess tooth preparations sometimes improved intra- and inter-examiner agreement according to criteria (consistency), the overall repeatability and reproducibility of feedback for the students was low.

These results indicated that the senior academic staff examiners had sometimes good grade agreement with themselves but feedback to the student was not consistent. In addition, specific additional tools and feedback sheets which were used did not always provide repeatable comments for the student. The reason is that definition of levels for each criterion was not clear and/or subjective. Therefore, non-consistent evaluation using the feedback sheet could be attributed to assessor bias and misinterpretation. In addition, the first occasion assessment using feedback sheet was better than the second occasion. This supports the argument that the first occasion feedback sheet assessment was not used as a 'practice session' for the examiners.

Ultimately, valid, clear descriptions and reliable feedback sheet(s) with tool(s) are essential to provide grade and fair comments for the student performance in the Clinical Skills Laboratory. From the results of Chapter 5, there is a very important question raised from this feedback sheet assessment stage. This question is, "How can researchers know whether their repeatability or reproducibility relates to valid observations without gold standard data?"

8.1.5 Chapter 6

This chapter further investigates the validity of examiner evaluations. While examiner repeatability is important, it is equally important that examiner grading reflects what is truly known about the tooth preparation (Construct validity). The selection of grades from the examiner who had the highest specialty or greatest experience as a gold standard is one of the most commonly-recommended methods. Therefore, the senior academic staff who had the best grade agreements and had the best correlation with the number of negative points, was selected, in order to determine whether or not the grades awarded by the best senior examiner can be selected as gold standard grades which reflect the tooth preparation truly.

The 'Gray and Mhanni feedback sheet' sometimes enhanced agreement of the examiners' grades but they did not reflect the tooth preparations truly. Because of that the highest reliability of grades did not always mean that these grades reflected the tooth preparation truly. Lack of standardisation (calibration) in the feedback sheets was a major reason why assessments have been considered as un-reliable and un-fair for students.

In addition, this chapter demonstrated that the examiners were unable to recognise a measurement difference of less than 0.60 mm (actually 0.56 mm), clinically. Therefore, the difference of acceptable range measurement (include ideal measurement) for each feature should be at least 0.60 mm to provide wide range for student to prepare ideal or acceptable class II amalgam cavity or full veneer gold shell crown.

In summary, all grades awarded from the best senior academic staff examiner did not agree with the pass and fail score features for tooth preparations reported in the literature and measured objectively by the researcher (AM). Therefore, the grades were not representative the tooth preparations. Additionally to lack of calibration of the grades, the criteria of the 'Gray and Mhanni feedback sheets' were not defined clearly enough and standard setting of the errors number (negative points) in comparison with the cut-off score was also not clear. Thus, the feedback sheets produced non-representative grades for the tooth preparations and unfair assessment and feedback for the students on their performances. This was disappointing. Perhaps, the main drawback of this study was that almost all scores using SAFMs and SAF evaluations were 'fail' due to the preparations being completed by students. The incorporation of some preparations completed by experienced dentists might have provided a wider range of scores.

Consequently, Chapter 7 was focused on i), the selection of reliable specific additional tools and ii), the development of new checklists (nCIIpc and nGSCpc) with cut-off points, in order to solve these problems.

8.1.6 Chapter 7

The reliable specific additional measurement tool were used to re-evaluate class II amalgam cavity and full veneer gold shell crown preparations in order to improve intra- and inter-examiner agreement and consistency. In addition, the new checklists (nCIIpc and nGSCpc) were developed according to the concepts of Lynn (1986), Knight (1997), Streiner and Normann (2008) and Puryer and O'Sullivan (2015).

The checklist(s) improved intra- and inter-examiner agreement, according to grades awarded and consistency of feedback, for the three senior academic staff examiners in comparison with previous intra- and inter-examiner agreement and consistency of the 'Gray and Mhanni feedback sheet' grades.

In general, the checklist(s) with reliable tool(s) produced valid and reliable comments and grades for the student, if the examiner was familiar with this checklist and tool. However it is recommended that the maximum time spent to assess tooth preparations using the checklist with tools should range from 20 to 30 minutes for 10 to 15 tooth preparations only. More time than this, means that examiners become tired and their results are less valid. Thus it is recommended time for assessment should be not more than two minutes per tooth preparation and only a maximum 10 to 15 teeth should be assessed before resting.

To improve reliability and agreement of grades and feedback,

- the duration of assessment, the number of categories and their criteria should be limited to two minutes,
- proper description for each criterion should be considered,
- assessment tool(s) with checklists should be used and,
- training for examiners on an assessment methods is essential to ensure familiarity.

Figure 8.1 shows the *Mhanni protocol* which can be used to evaluate and improve assessment method for other departments or institutions.

<table>
<tr><th>Stage</th><th>Sub-stage description</th><th>Step(s)</th></tr>
<tr><td rowspan="3">Evaluation stage</td><td>Evaluation stage for specific department at Dental School</td><td>- Assess student(s) performance by examiner(s).
- Determining intra- and inter-examiner agreement according to the grades awarded.</td></tr>
<tr><td>Evaluation stage for ability of the examiner(s) to assess</td><td>- Examiner(s) assesses student(s) performance by using three or more stages (e.g. eyeball, specific tool(s) and feedback sheet).
- The tool(s) must be selected according to what the student(s) used in the session.
- Determining intra- and inter-examiner agreement according to grades for each stage.
- Determining intra- and inter-examiner agreement according to negative points from the feedback sheet.
- Determining correlation between grades and negative points awarded from the examiner(s) using feedback sheet.
- Determining intra- and inter-examiner consistency to ensure that comment(s) from feedback sheet for the student(s) performance is/are consistent.</td></tr>
<tr><td>Evaluation stage of feedback sheet and its grade awarded whether or not these grades represent the object truly</td><td>- Divide the categories of feedback sheet into objective evaluation which can be measured by specific tool(s) and subjective evaluation which can be evaluated by examiner(s).
- Measuring object feature(s) to determine objective evaluation(s) by using reliable measurement tool(s) or software and then compare these measurement(s) with the measurement(s) which were recommended in the literature for this object.
- Determining the subjective evaluation of some features by limiting this to a binary response of the examiner(s) such as yes/no. If there is more than one examiner the most frequent response is selected as the response representing the object feature truly.
- Determining the grades/scores by using previous points for each object according to Knight (1997) recommendation.
- Comparing the grades awarded from examiner(s) using feedback sheet with grades/scores awarded from previous point by using agreement test (e.g. Kappa test or agreement percentage).</td></tr>
<tr><td rowspan="2">Developmental stage</td><td>Determine reliable tool(s) for assessment</td><td>- Selecting the grades awarded from tool(s) which was used in ‘Evaluation stage for ability of the examiner(s) to assess’.
- Determining the most reliable tool(s) by comparing grades awarded from the tool(s) with the grades/scores awarded from objective and subjective evaluations in the ‘Evaluation stage of feedback sheet and its grade awarded whether or not these grades represent the object truly’.</td></tr>
<tr><td>Develop or improve new checklist(s) to assess the object</td><td>- Develop the new checklist according to previous feedback sheet, other protocol(s), literature and/or other concepts and recommendations.
- New checklist with the most reliable tool(s) is used to assess student(s) performance by examiner(s).
- Determining intra- and inter-examiner agreement according to grades awarded from new checklist.
- Determining intra- and inter-examiner agreement according to negative points from new checklist.
- Determining correlation between grades and negative points awarded from the examiner(s) using new checklist.
- Determining intra- and inter-examiner consistency to ensure that comment(s) from the new checklist for the student(s) performance is/are consistent.</td></tr>
</table>

Figure 8.1 Outline of Mhanni protocol to evaluate and improve feedback/assessment for the students in the Clinical Skills Laboratory

8.2 Recommendations

A larger group of teachers with different level of experience and/or students as a sample should be selected to confirm the findings of this study.

The students should know the ideal and acceptable feature measurements [(SAFMs) and (SAFs)] before commencing each clinical session. In addition, SAFMs should be calibrated according to reliable tools which will be used to assess the student's performance.

If the plastic teeth which are used in the clinical skills laboratory have different sizes, a ratio is used instead of SAFMs to evaluate features which can be measured.

Practice on the checklist is essential part in order to familiarise and calibrate the examiners as well as to identify the most reliable examiner(s).

Assessors should demonstrate repeatable or reproducible results using the new checklist, before acting as an examiner.

The new checklist can be used as a formative assessment for the students in order to provide them with feedback on their preparations during the course.

The new checklist can also be used as a summative assessment. If the new checklist provides a grade 4 or 5 for student performance, the student can pass the exam immediately. If the checklist provides grade 3 for student performance, the student has chance to correct the errors. After that, this tooth preparation should be assessed once more using the new checklist.

The examiners can change the descriptions of level of performance (SAFMs or SAFs) for criteria of the checklist according to their recommendation. Before changing the measurements, the examiners should inform the student what are the ideal and acceptable features and measurements according to the reliable assessment tool.

8.3 Further studies

Further work could be carried out to:

- support the outcomes of this thesis by using the new checklist(s) on a larger group of teachers and/or students as a sample.
- try the new checklist as a formative assessment for student for another year or more as a training session for the examiners and then used it as a summative assessment in the clinical laboratory skills.
- use the new checklist(s) as a self-assessment tool or peer assessment for the students in order to determine the impact of the checklist(s) on performance of students.
- confirm that the optimum time for examiners to spend assessing the tooth preparations to 20 – 30 minutes before they become so fatigued as to be unreliable..
- create other checklists which relate to other departments and skills to improve feedback and assessment for dental students at Dundee Dental School by using similar methods described in this thesis (Mhanni Protocol).

References

Ahmed, S.N., Sturdevant, J., Wilder, R., Kowlowitz, V. and Boushell, L. 2016. Development and assessment of discrimination exercises for faculty calibration in preclinical operative dentistry. *Journal of Dental Education*, *80*(8), 994-1003.

Akpata, E.S., Alomari, Q.D., and AlShamary, A.R. 2013. *Principles and Practice of Operative Dentistry: A Modern Approach*. Quintessence Publishing Company Incorporated.

Albino, J.E., Young, S.K., Neumann, L.M., Kramer, G.A., Andrieu, S.C., Henson, L., Horn, B. and Hendricson, W.D. 2008. Assessing dental students' competence: best practice recommendations in the performance assessment literature and investigation of current practices in predoctoral dental education. *Journal of Dental Education*, *72*(12), 1405-1435.

AlHumaid, J., El Tantawi, M., Al-Ansari, A.A. and Al-Harbi, F.A. 2016. Agreement in Scoring Preclinical Dental Procedures: Impact on Grades and Instructor-Related Determinants. *Journal of Dental Education*, *80*(5), 553-562.

Almquist, T.C., Cowan, R.D. and Lambert, R.L. 1973. Conservative amalgam restorations. *The Journal of Prosthetic Dentistry*, *29*(5), 524-528.

Al-Omari, W.M. and Al-Wahadni, A.M. 2004. Convergence angle, occlusal reduction, and finish line depth of full-crown preparations made by dental students. *Quintessence International*, *35*(4).

Arnetzl, G. and Dornhofer, R. 2004. PREPassistant: a system for evaluating tooth preparations. *International Journal of Computerized Dentistry*, *7*(2), 187-197.

Assessment of Clinical Skills Protocol. 2016. [online] Available at: https://ndeb-bned.ca/sites/ndeb/files/pdf/Protocols/2016/protocol_acs_2016_0.pdf

Atkinson, G. and Nevill, A.M. 1998. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Medicine*, *26*(4), 217-238.

Ayad, M.F., Maghrabi, A.A. and Rosenstiel, S.F. 2005. Assessment of convergence angles of tooth preparations for complete crowns among dental students. *Journal of Dentistry*, *33*(8), 633-638.

Barr, R.B. and Tagg, J. 1995. From teaching to learning—A new paradigm for undergraduate education. *Change: The Magazine of Higher Learning*, *27*(6), 12-26.

Bastien, C.H., Vallières, A. and Morin, C.M., 2001. Validation of the Insomnia Severity Index as an outcome measure for insomnia research. *Sleep Medicine*, *2*(4), 297-307.

Baum, L., Phillips, R.W. and Lund, M.R., 1995. 3rd ed. *Textbook of Operative Dentistry*. WB Saunders Company.

Beanland, C., Schneider, Z., LoBiondo-Wood, G. and Haber, J. 1999. Nursing research: Methods, Critical Appraisal and Utilization. Sydney: Mosby.

Beard, J.D. 2005. Setting standards for the assessment of operative competence. *European Journal of Vascular and Endovascular Surgery*, *30*(2), 215-218.

Beck, C.T. and Gable, R.K., 2001. Ensuring content validity: An illustration of the process. *Journal of Nursing Measurement*, *9*(2), 201-215.

Ben-David, M.F. 2000. AMEE Guide No. 18: Standard setting in student assessment. *Medical Teacher*, *22*(2), 120-130.

Berk, R.A. 1990. Importance of expert judgment in content-related validity evidence. *Western Journal of Nursing Research*, *12*(5), 659-671.

Beschnidt, S.M. and Strub, J.R. 1999. Evaluation of the marginal accuracy of different all-ceramic crown systems after simulation in the artificial mouth. *Journal of Oral Rehabilitation*, *26*(7),.582-593.

Biggs, J. and Tang, C. 2011. *Teaching for quality learning at university,* Maidenhead, Berkshire, UK, McGraw-Hill Education Companies, Incorporated.

Black, G.V. 1914. A work on operative dentistry. *Volume II. The technical procedures in filling teeth. A work on operative dentistry in two volumes. Chicago: Medico-dental Publishing Company.*

Black, G.V. 1955. *Operative Dentistry* (Vol. 2). Medico-Dental Publishing Company.

Black, P. and Wiliam, D. 1998. Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7-74.

Blair, F.M., Wassell, R.W. and Steele, J.G. 2002. Crowns and other extra-coronal restorations: preparations for full veneer crowns. *British Dental Journal*, *192*(10), 561-571.

Bland, J.M. and Altman, D.G. 1996. Statistics notes: measurement error. *British Medical Journal*, *313*(7059), 744.

Bland, J.M. and Altman, D.G. 1999. Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, *8*(2), 135-160.

Bland, J.M. and Altman, D.G., 1997. Cronbach's alpha. *BMJ: British Medical Journal*, *314*(7080), 572.

Bloxham, S. 2009. Marking and moderation in the UK: false assumptions and wasted resources. *Assessment & Evaluation in Higher Education*, *34*(2), 209-220.

Bolarinwa, O.A. 2015. Principles and methods of validity and reliability testing of questionnaires used in social and health science researches. *Nigerian Postgraduate Medical Journal*, *22*(4), 195.

Bond, L.A. 1996. Norm-and criterion-referenced testing. *Practical Assessment, Research & Evaluation*, *5*(2), 120-125.

Boud, D. and Falchikov, N. 2007. *Rethinking assessment in higher education: Learning for the longer term*. Oxon, UK, Taylor and Francis.

Boud, D. and Molloy, E. 2013. *Feedback in higher and professional education: understanding it and doing it well*, Oxon, UK, Routledge.

Boursicot, K., Etheridge, L., Setna, Z., Sturrock, A., Ker, J., Smee, S. and Sambandam, E. 2011. Performance in assessment: consensus statement and recommendations from the Ottawa conference. *Medical Teacher*, *33*(5), 370-383.

Branch Jr, W.T. and Paranjape, A. 2002. Feedback and reflection: teaching methods for clinical settings. *Academic Medicine*, *77*(12, Part 1), 1185-1188.

Brink, P.J. and Wood, M.J. 1998. *Advanced Design in Nursing Research (2nd ed.)*. Thousand Oaks, CA: Sage.

Brown, G., Manogue, M. and Martin, M. 1999. The validity and reliability of an OSCE in dentistry. *European Journal of Dental Education*, *3*(3), 117-125.

Brown, N. and Cooke, L. 2009. Giving effective feedback to psychiatric trainees. *Advances in Psychiatric Treatment*, *15*(2), 123-128.

Brown, R.K. 1930. Research in the use of a rating scale as a means of evaluating the personalities of senior dental students. *Journal of Dental Research*, *10*(3), 271-279.

Buchanan, J.A. 2001. Overview of three years experience with virtual reality based technology in dental education. *Journal of Dental Education*, *65*(1), 58.

Callan, R.S., Haywood, V.B., Cooper, J.R., Furness, A.R. and Looney, S.W. 2015. The Validity of Using E4D Compare's "% Comparison" to Assess Crown Preparations in Preclinical Dental Education. *Journal of Dental Education*, *79*(12),1445-1451.

Campbell, D.T. and Fiske, D.W. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81.

Cardoso, J.A., Barbosa, C., Fernandes, S., Silva, C.L. and Pinho, A., 2006. Reducing subjectivity in the evaluation of pre-clinical dental preparations for fixed prosthodontics using the Kavo PrepAssistant®. *European Journal of Dental Education*, *10*(3), pp.149-156.

Carless, D. 2006. Differing perceptions in the feedback process. *Studies in Higher Education*, *31*(2), 219-233.

Carless, D., Salter, D., Yang, M. and Lam, J. 2011. Developing sustainable feedback practices. *Studies in Higher Education*, *36*(4), 395-407.

Caro, T.M., Roper, R., Young, M. and Dank, G.R. 1979. Inter-observer reliability. *Behaviour*, *69*(3), 303-315.

Case, S.M. and Swanson, D.B. 1998. *Constructing written test questions for the basic and clinical sciences* (2nd ed., pp. 22-25). Philadelphia, PA: National Board of Medical Examiners.

Chadwick, S.M. and Holsgrove, G.J. 2009. New developments in assessment in orthodontics. *Journal of Orthodontics*, *36*(2), 122-129.

Chambers, D.W., Leknius, C. and Woodson, R. 1997. Predictive validity of instructor judgment in preclinical technique courses. *Journal of Dental Education*, *61*(9), 736-740.

Chhapola, V., Kanwal, S.K. and Brar, R. 2015. Reporting standards for Bland–Altman agreement analysis in laboratory research: a cross-sectional survey of current practice. *Annals of Clinical Biochemistry*, *52*(3), 382-386.

Chmura Kraemer, H., Periyakoil, V.S. and Noda, A., 2002. Kappa coefficients in medical research. *Statistics in Medicine*, *21*(14), 2109-2129.

Cho, G.C., Chee, W.W. and Tan, D.T. 2010. Dental students' ability to evaluate themselves in fixed prosthodontics. *Journal of Dental Education*, *74*(11), 1237-1242.

Cicchetti, D.V. 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284.

Cicchetti, D.V. and Sparrow, S.A., 1981. Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. *American Journal of Mental Deficiency*, *86*, 127-137.

Cizek, G.J. and Bunch, M.B. 2007. Standard Setting: A guide to establishing and evaluating performance standards on tests. SAGE Publications Ltd.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37-46.

Cohen, J., 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213.

Cowpe, J., Plasschaert, A., Harzer, W., Vinkka-Puhakka, H. and Walmsley, A.D. 2010. Profile and competences for the graduating European dentist–update 2009. *European Journal of Dental Education*, *14*(4), 193-202.

Cronbach, L.J. and Meehl, P.E. 1955. Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281.

Crooks, T., 2001, September. The validity of formative assessments. In *British Educational Research Association Annual Conference, University of Leeds,* 13-15.

Crossley, J., Humphris, G. and Jolly, B., 2002. Assessing health professionals. *Medical Education, 36*(9), 800-804.

Curtis, D.A., Lind, S.L., Dellinges, M., Setia, G. and Finzen, F.C. 2008. Dental students' self-assessment of preclinical examinations. *Journal of Dental Education*, *72*(3), 265-277.

Davies, H., Archer, J., Southgate, L. and Norcini, J. 2009. Initial evaluation of the first year of the Foundation Assessment Programme. *Medical Education*, *43*(1), 74-81.

Davis, L.L. 1992. Instrument review: Getting the most from a panel of experts. *Applied Nursing Research*, *5*(4), 194-197.

De Lacharrière, O., Deloche, C., Misciali, C., Piraccini, B.M., Vincenzi, C., Bastien, P., Tardy, I., Bernard, B.A. and Tosti, A. 2001. Hair diameter diversity: a clinical sign reflecting the follicle miniaturization. *Archives of Dermatology*, *137*(5), 641-646.

de Vet, H.C., Terwee, C.B., Knol, D.L. and Bouter, L.M. 2006. When to use agreement versus reliability measures. *Journal of Clinical Epidemiology*, *59*(10), 1033-1039.

Deranleau, N.J., Feiker, J.H. and Beck, M., 1983. Effect of percentage cut-off scores and scale point variation on preclinical project evaluation. *Journal of Dental Education*, *47*(10), 650-655.

DeVellis RF. 2013. *Scale development: theory and applications*. Thousand Oaks, Calif: SAGE Publications.

DeVon, H.A., Block, M.E., Moyle-Wright, P., Ernst, D.M., Hayden, S.J., Lazzara, D.J., Savoy, S.M. and Kostas-Polston, E. 2007. A psychometric toolbox for testing validity and reliability. *Journal of Nursing Scholarship*, *39*(2), 155-164.

Dhuru, V.B., Rypel, T.S. and Johnston, W.M. 1978. Criterion-oriented grading system for preclinical operative dentistry laboratory course. *Journal of Dental Education*, *42*(9), 528-531.

Dolmans, D.H. and Tigelaar, D. 2012. Building bridges between theory and practice in medical education using a design-based research approach: AMEE Guide No. 60. *Medical Teacher*, *34*(1), 1-10.

Downing, S.M. and Haladyna, T.M. 2004. Validity threats: overcoming interference with proposed interpretations of assessment data. *Medical Education*, *38*(3), 327-333.

Duncan, W.K. 1979. The Effects of Criterion Based Evaluation on Reliability and Self-Evaluation in a Preclinical Dental Laboratory.

Earl, L.M. 2012. *Assessment as learning: Using classroom assessment to maximize student learning*. Corwin Press.

Epstein, R.M. 2007. Assessment in medical education. *New England Journal of Medicine*, *356*(4), 387-396.

Epstein, R.M. and Hundert, E.M. 2002. Defining and assessing professional competence. *JAMA*, *287*(2), 226-235.

Esser, C., Kerschbaum, T., Winkelmann, V., Krage, T. and Faber, F.J. 2006. A comparison of the visual and technical assessment of preparations made by dental students. *European Journal of Dental Education*, *10*(3), 157-161.

Eva, K.W., Armson, H., Holmboe, E., Lockyer, J., Loney, E., Mann, K. and Sargeant, J. 2012. Factors influencing responsiveness to feedback: on the interplay between fear, confidence, and reasoning processes. *Advances in Health Sciences Education*, *17*(1), 15-26.

Evans, C. 2013. Making sense of assessment feedback in higher education. *Review of Educational Research*, *83*(1), 70-120.

Falchikov, N. 1986. Product comparisons and process benefits of collaborative peer group and self assessments. *Assessment and Evaluation in Higher Education*, *11*(2), 146-166.

Feil, P.H. 1982. An analysis of the reliability of a laboratory evaluation system. *Journal of Dental Education*, *46*(8), 489-494.

Fernandez, J.J. 1967. *Evaluation of student clinical performance in dental schools: construction and validation of a scale for the evaluation of cavity preparations and silver amalgam restorations in the primary dentition* (Doctoral dissertation, University of North Carolina at Chapel Hill).

Field, A., 2013. *Discovering statistics using IBM SPSS statistics*. Sage Publications.

Fish, D. and Coles, C. 2005. *Medical Education: Developing A Curriculum For Practice: Developing a Curriculum for Practice*. McGraw-Hill Education (UK).

Fleiss, J.L., 1986. Reliability of measurement. *The Design and Analysis of Clinical Experiments*, 1-32.

Fowell, S.L., Fewtrell, R. and McLaughlin, P.J. 2008. Estimating the minimum number of judges required for test-centred standard setting on written assessments. Do discussion and iteration have an influence?. *Advances in Health Sciences Education*, *13*(1), 11-24.

Fuller, J.L., 1972. The effects of training and criterion models on interjudge reliability. *Journal of Dental Education*, *36*(4), 19-22.

Gabel, A.B. 1951. Mechanical principles of operative dentistry. *The Journal of the American Dental Association*, *43*(2), 153-160.

Gaines, W.G., Bruggers, H. and Rasmussen, R.H. 1974. Reliability of ratings in preclinical fixed prosthodontics: effect of objective scaling. *Journal of Dental Education*, *38*(12), 672-675.

Garland, K.V. and Newell, K.J. 2009. Dental hygiene faculty calibration in the evaluation of calculus detection. *Journal of Dental Education*, *73*(3), 383-389.

Gerrow, J.D., Murphy, H.J., Boyd, M.A. and Scott, D.A., 2003. Concurrent validity of written and OSCE components of the Canadian dental certification examinations. *Journal of Dental Education*, *67*(8), pp.896-901.

Gibbons, J.D. and Chakraborti, S. 2011. Nonparametric statistical inference. In *International Encyclopedia of Statistical Science*. Springer Berlin Heidelberg.

Gilboe, D.B. and Teteruck, W.R. 1974. Fundamentals of extracoronal tooth preparation. Part I. Retention and resistance form. *The Journal of Prosthetic Dentistry*, *32*(6), 651-656.

Gipps, C.V. 1994. *Beyond testing: Towards a theory of educational assessment*. Psychology Press.

Gisev, N., Bell, J.S. and Chen, T.F., 2013. Interrater agreement and interrater reliability: key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, *9*(3), 330-338.

Glaser, R. 1963. Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist*, 18, 519–521.

Goepferd, S.J. and Kerber, P.E. 1980. A comparison of two methods for evaluating primary class II cavity preparations. *Journal of Dental Education*, *44*(9), 537-542.

Goodacre, C.J., Campagni, W.V. and Aquilino, S.A. 2001. Tooth preparations for complete crowns: an art form based on scientific principles. *The Journal of Prosthetic Dentistry*, *85*(4), 363-376.

Gratton, D.G., Kwon, S.R., Blanchette, D. and Aquilino, S.A. 2016. Impact of digital tooth preparation evaluation technology on preclinical dental students' technical and self-evaluation skills. *Journal of Dental Education*, *80*(1), 91-99.

Greenhalgh, T. 2014. *How to read a paper: the basics of evidence-based medicine*. John Wiley & Sons.

Grigg, P. and Stephens, C.D. 1998. Computer-assisted learning in dentistry a view from the UK. *Journal of Dentistry*, *26*(5-6), 387-395.

Guenzel, P.J., Knight, G.W., and Feil, P.H. 1995. Designing preclinical instruction for psychomotor skills (III): Instructional engineering - design phase. *Journal of Dental Education*, *59*(3), 412-418.

Gwet, K.L. 2014. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

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

Hanneman, S.K. 2008. Design, analysis and interpretation of method-comparison studies. *AACN Advanced Critical Care*, *19*(2), 223.

Harden, R.M. and Laidlaw, J.M. 2012. *Essential Skills for a Medical Teacher: An Introduction to Teaching and Learning in Medicine.* Elsevier Health Sciences UK.

Harlen, W. 2007. *Assessment of learning*, London, UK, SAGE Publications.

Harlen, W. and James, M. 1997. Assessment and learning: differences and relationships between formative and summative assessment. *Assessment in Education: Principles, Policy & Practice*, *4*(3), 365-379.

Hattie, J. and Timperley, H. 2007. The power of feedback. *Review of Educational Research*, *77*(1), 81-112.

Hauser, A.M. and Bowen, D.M. 2009. Primer on preclinical instruction and evaluation. *Journal of Dental Education*, *73*(3), 390-398.

Hays, R.B., Davies, H.A., Beard, J.D., Caldon, L.J.M., Farmer, E.A., Finucane, P.M., McCrorie, P., Newble, D.I., Schuwirth, L.W.T. and Sibbald, G.R. 2002. Selecting performance assessment methods for experienced physicians. *Medical Education*, *36*(10), 910-917.

Helft, M., Pilo, R., Cardash, H.S. and Baharav, H., 1987. Reliability of ratings in evaluation of crowns. *The Journal of Prosthetic Dentistry*, *58*(5), 647-649.

Higgins, R., Hartley, P. and Skelton, A. 2002. The conscientious consumer: Reconsidering the role of assessment feedback in student learning. *Studies in Higher Education*, *27*(1), 53-64.

High Education Funding Council for England 2011. The National Survey: Finding and Trending 2006-2010. Bristol: Higher Education Funding Council for England. Page 8-14.

Hilton, T.J., Ferracane, J.L. and Broome, J.C. eds., 2013. *Summitt's Fundamentals of Operative Dentistry: A Contemporary Approach*. Quintessence Publishing Company Incorporated.

Hinkelman, K.W. and Long, N.K. 1973. Method for decreasing subjective evaluation in preclinical restorative dentistry. *Journal of Dental Education*, *37*(9), 13-18.

Hopkins, W.G. 2000. Measures of reliability in sports medicine and science. *Sports Medicine*, *30*(1), 1-15.

Houpt, M. I., and Kress, G. 1973. Accuracy of measurement of clinical performance in dentistry. *Journal of Dental Education*, 37(7), 34-46.

Houston, W.J.B. 1983. The analysis of errors in orthodontic measurements. *American Journal of Orthodontics*, *83*(5), 382-390.

Hubley, A.M. and Wagner, S., 2004. Using alternate forms of the multidimensional health locus of control scale: Caveat emptor. *Social Indicators Research*, *65*(2), pp.167-186.

Hyman, R.T. 1980. Fielding student questions. *Theory Into Practice*, *19*(1), 38-44.

Ivers, N., Jamtvedt, G., Flottorp, S., Young, J.M., Odgaard-Jensen, J., French, S.D., O'Brien, M.A., Johansen, M., Grimshaw, J. and Oxman, A.D. 2012. Audit and feedback: effects on professional practice and healthcare outcomes. *The Cochrane Library*.

Jakobsson, U. and Westergren, A., 2005. Statistical methods for assessing agreement for ordinal data. *Scandinavian Journal of Caring Sciences*, *19*(4), 427-431.

Jasinevicius, T.R., Landers, M., Nelson, S. and Urbankova, A. 2004. An evaluation of two dental simulation systems: virtual reality versus contemporary non-computer-assisted. *Journal of Dental Education*, *68*(11), 1151-1162.

Jenkins, S.M., Dummer, P.M.H., Gilmour, A.S.M., Edmunds, D.H., Hicks, R. and Ash, P., 1998. Evaluating undergraduate preclinical operative skill; use of a glance and grade marking system. *Journal of Dentistry*, *26*(8), 679-684.

Johnston, J.F., Dykema, R.W., Phillips, R.W. and Goodacre, C.J., 1986. *Johnston's Modern Practice in Fixed Prosthodontics*. WB Saunders Co, p. 36-39.

Jokstad, A. 1992. *Class 2 cavity preparations and restoration performance*. University of Oslo, Dental Faculty, Department of Anatomy and NIOM, Scandinavian Institute of Dental Materials.

Jokstad, A. and Mjör, I.A. 1987. Cavity designs for class II amalgam restorations: a literature review and a suggested system for evaluation. *Acta Odontologica Scandinavica*, *45*(4), 257-273.

Jørgensen, K.D. 1955. The relationship between retention and convergence angle in cemented veneer crowns. *Acta Odontologica Scandinavica*, 13(1), 35-40.

Kateeb, E.T., Kamal, M.S., Kadamani, A.M., Abu Hantash, R.O. and Abu Arqoub, M.M. 2016. Utilising an innovative digital software to grade pre-clinical crown preparation exercise. *European Journal of Dental Education*.

Kerner, S., Etienne, D., Malet, J., Mora, F., Monnet-Corti, V. and Bouchard, P. 2007. Root coverage assessment: validity and reproducibility of an image analysis system. *Journal of Clinical Periodontology*, *34*(11), 969-976.

Keszei, A.P., Novak, M. and Streiner, D.L., 2010. Introduction to health measurement scales. *Journal of Psychosomatic Research*, *68*(4), pp.319-323.

Kilminster, S. and Roberts, T. 2004. Standard setting for OSCEs: trial of borderline approach. *Advances in Health Sciences Education*, *9*(3), 201-209.

Kluger, A.N. and Van Dijk, D. 2010. Feedback, the various tasks of the doctor, and the feedforward alternative. *Medical Education*, *44*(12), 1166-1174.

Knight, G.W. 1997. Toward faculty calibration. *Journal of Dental Education*, *61*(12), 941-946.

Kottner, J., Audigé, L., Brorson, S., Donner, A., Gajewski, B.J., Hróbjartsson, A., Roberts, C., Shoukri, M. and Streiner, D.L. 2011. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *International Journal of Nursing Studies*, *48*(6), 661-671.

Kournetas, N., Jaeger, B., Axmann, D., Groten, M., Lachmann, S., Weber, H. and Geis-Gerstorfer, J. 2004. Assessing the reliability of a digital preparation assistant system used in dental education. *Journal of Dental Education*, *68*(12), 1228-1234.

Kramer, A., Muijtjens, A., Jansen, K., Düsman, H., Tan, L. and Van Der Vleuten, C. 2003. Comparison of a rational and an empirical standard setting procedure for an OSCE. *Medical Education*, *37*(2), 132-139.

Kramer, G.A., Albino, J.E., Andrieu, S.C., Hendricson, W.D., Henson, L., Horn, B.D., Neumann, L.M. and Young, S.K. 2009. Dental student assessment toolbox. *Journal of Dental Education*, *73*(1), 12-35.

Krosnick, J.A. and Presser, S. 2010. Question and questionnaire design. *Handbook of survey research*, *2*(3), 263-314.

Landis, J.R. and Koch, G.G., 1977. The measurement of observer agreement for categorical data. *biometrics*, 159-174.

LeBreton, J.M. and Senter, J.L., 2008. Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, *11*(4), 815-852.

Lexell, J.E. and Downham, D.Y. 2005. How to assess the reliability of measurements in rehabilitation. *American Journal of Physical Medicine & Rehabilitation*, *84*(9), 719-723.

Licari, F.W., Knight, G.W. and Guenzel, P.J. 2008. Designing evaluation forms to facilitate student learning. *Journal of Dental Education*, *72*(1), 48-58.

Light, G., Calkins, S. and Cox, R. 2009. *Learning and teaching in higher education: The Reflective Professional*, London, UK, SAGE Publications.

Lilley, J.D., ten Bruggen, C.H., Holloway, P.J., Holt, J.K. and Start, K.B., 1968. Reliability of practical tests in operative dentistry. *British Dental Journal*, *125*(5), 194-197.

Lindvall, C.M. and Nitko, A.J. 1975. *Measuring pupil achievement and aptitude*. Houghton Mifflin Harcourt P.

Livingston, S.A. and Zieky, M.J. 1982. Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests.

Lynn, M. R. 1986. Determination and quantification of content validity. *Nursing Research, 35,* 382–385.

Mackenzie, R.S. 1973. Defining clinical competence in terms of quality, quantity, and need for performance criteria. *Journal of Dental Education*, *37*(9), 37-44.

Manogue, M., Brown, G. and Foster, H. 2001. Clinical assessment of dental students: values and practices of teachers in restorative dentistry. *Medical Education*, *35*(4), 364-370.

Manogue, M., Kelly, M., Bartakova Masaryk, S., Brown, G., Catalanotto, F., Choo-Soo, T., Delap, E., Godoroja, P., Morio, I., Rotgans, J. and Saag, M. 2002. 2.1 Evolving methods of assessment. *European Journal of Dental Education*, *6*(3), 53-66.

Manogue, M., McLoughlin, J., Christersson, C., Delap, E., Lindh, C., Schoonheim-Klein, M. and Plasschaert, A., 2011. Curriculum structure, content, learning and assessment in European undergraduate dental education–update 2010. *European Journal of Dental Education*, *15*(3), 133-141.

Marra, L.M. 1970. An historical review of full coverage of the natural dentition. *The New York State Dental Journal*, *36*(3), 147.

Martuza, V.R. 1977. *Applying norm-referenced and criterion-referenced measurement in education*. Allyn & Bacon, Incorporated.

Mays, K.A. and Branch-Mays, G.L. 2016. A systematic review of the use of self-assessment in preclinical and clinical dental education. *Journal of Dental Education*, *80*(8), 902-913.

Mays, K.A. and Levine, E. 2014. Dental students' self-assessment of operative preparations using CAD/CAM: a preliminary analysis. *Journal of Dental Education*, *78*(12), 1673-1680.

McDowell, L., Wakelin, D., Montgomery, C. and King, S. 2011. Does assessment for learning make a difference? The development of a questionnaire to explore the student response. *Assessment & Evaluation in Higher Education*, *36*(7), 749-765.

McGraw, K.O. and Wong, S.P., 1996. Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30-46.

McKay, S.R. 1978, October. A review of student stress in nursing education programs. In *Nursing Forum* (Vol. 17, No. 4, 376-393). Blackwell Publishing Ltd.

Medeiros, V.A.F. and Seddon, R.P. 2000. Iatrogenic damage to approximal surfaces in contact with Class II restorations. *Journal of Dentistry*, *28*(2), 103-110.

Meek, G.E., Ozgur, C. and Dunning, K., 2007. Comparison of the t vs. Wilcoxon Signed-Rank Test for Likert scale data and small samples. *Journal of Modern Applied Statistical Methods*, *6*(1), p.10.

Messick, S. 1995. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741.

Miller, G.E., 1990. The assessment of clinical skills/competence/performance. *Academic Medicine*, *65*(9), S63-7.

Moopnar, M. and Faulkner, K.D.B. 1991. Accidental damage to teeth adjacent to crown-prepared abutment teeth. *Australian Dental Journal*, *36*(2), 136-140.

Myers, B. 1977. Beliefs of dental faculty and students about effective clinical teaching behaviors. *Journal of Dental Education*, *41*(2), 68-76.

Natkin, E. and Guild, R.E., 1967. Evaluation of preclinical laboratory performance: a systematic study. *Journal of Dental Education*, *31*(2), 152.

Neher, J.O., Gordon, K.C., Meyer, B. and Stevens, N. 1992. A five-step “microskills” model of clinical teaching. *The Journal of the American Board of Family Practice*, *5*(4), 419-424.

Netemeyer, R.G., Bearden, W.O. and Sharma, S. 2003. *Scaling procedures: Issues and applications*. Sage Publications.

Nichols, P., Twing, J., Mueller, C.D. and O'Malley, K. 2010. Standard-Setting Methods as Measurement Processes. *Educational Measurement: Issues and Practice*, *29*(1), 14-24.

Nicol, D. 2010. From monologue to dialogue: improving written feedback processes in mass higher education. *Assessment & Evaluation in Higher Education*, *35*(5), 501-517.

Nicol, D., Thomson, A. and Breslin, C. 2014. Rethinking feedback practices in higher education: a peer review perspective. *Assessment & Evaluation in Higher Education*, *39*(1), 102-122.

Nicol, D.J. and Macfarlane-Dick, D. 2006. Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, *31*(2), 199-218.

Noonan, J.E. and Goldfogel, M.H. 1991. Convergence of the axial walls of full veneer crown preparations in a dental school environment. *The Journal of Prosthetic Dentistry*, *66*(5), 706-708.

Norcini, J., Anderson, B., Bollela, V., Burch, V., Costa, M.J., Duvivier, R., Galbraith, R., Hays, R., Kent, A., Perrott, V. and Roberts, T., 2011. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. *Medical Teacher*, *33*(3), 206-214.

Norman, G.R. and Eva, K.W. 2010. Diagnostic error and clinical reasoning. *Medical Education*, *44*(1), 94-100.

Norman, G.R., Vleuten, C.P.M. and Graaff, E. 1991. Pitfalls in the pursuit of objectivity: issues of validity, efficiency and acceptability. *Medical Education*, *25*(2), 119-126.

Nunnally, J.C. and Bernstein, I.H., 1994. Psychological theory (3rd ed.) *New York, NY: MacGraw-Hill*.

O’Donnell, J.A., Oakley, M., Haney, S., O’Neill, P.N. and Taylor, D. 2011. Rubrics 101: a primer for rubric development in dental education. *Journal of Dental Education*, *75*(9), 1163-1175.

Ohm, E. and Silness, J. 1978. The convergence angle in teeth prepared for artificial crowns. *Journal of Oral Rehabilitation*, *5*(4), 371-375.

Orsmond, P., Maw, S.J., Park, J.R., Gomez, S. and Crook, A.C. 2013. Moving feedback forward: theory to practice. *Assessment & Evaluation in Higher Education*, *38*(2), 240-252.

Orsmond, P., Merry, S. and Reiling, K. 1997. A study in self-assessment: tutor and students' perceptions of performance criteria. *Assessment & Evaluation in Higher Education*, *22*(4), 357-368.

Orsmond, P., Merry, S. and Reiling, K. 2000. The use of student derived marking criteria in peer and self-assessment. *Assessment & Evaluation in Higher Education*, *25*(1), 23-38.

Orsmond, P., Merry, S. and Reiling, K. 2005. Biology students' utilization of tutors' formative feedback: a qualitative interview study. *Assessment & Evaluation in Higher Education*, *30*(4), 369-386.

O'Sullivan, M. ed., 2005. *Fixed Prosthodontics in Dental Practice*. Quintessence.

Park, R.D. Susarla, S.M., Howell, T.H. and Karimbux, N.Y., 2009. Differences in clinical grading associated with instructor status. *European Journal of Dental Education*, *13*(1), 31-38.

Paskins, Z., Kirkcaldy, J., Allen, M., Macdougall, C., Fraser, I. and Peile, E., 2010. Design, validation and dissemination of an undergraduate assessment tool using SimMan® in simulated medical emergencies. *Medical Teacher*, *32*(1), e12-e17.

Patel, P.B., Wildgoose, D.G. and Winstanley, R.B. 2005. Comparison of convergence angles achieved in posterior teeth prepared for full veneer crowns. *European Journal of Prosthodontics and Restorative Dentistry*, *13*(3), 100.

Patridge, M.I. and Mast, T.A. 1978. Dental clinical evaluation: a review of the research. *Journal of Dental Education*, *42*(6), 300-305.

Plasschaert, A.J.M., Manogue, M., Lindh, C., McLoughlin, J., Murtomaa, H., Nattestad, A. and Sanz, M. 2007. Curriculum content, structure and ECTS for European dental schools. Part II: methods of learning and teaching, assessment procedures and performance criteria. *European Journal of Dental Education*, *11*(3), 125-136.

Polit Denise, F. and Hungler Bernadette, P., 1999. Nursing research principles and methods, 79-80.

Polit, D.F. and Beck, C.T. 2004. *Nursing research: Principles and methods*. Philadelphia: Lippincott Williams & Wilkins.

Polit, D.F. and Beck, C.T. 2006. The content validity index: are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health*, *29*(5), 489-497.

Pollard, D.J. and Davenport, J.C. 1993. An evaluation of training general dental practitioners in partial denture design using a computer-assisted learning program. *British Dental Journal*, *177*(11-12), 405-409.

Polyzois, I., McLouglin, J., Kelly, A. and Claffey, N. 2010. Clinical teaching in restorative dentistry and the variation between students' and supervisors' perceptions of its effectiveness. *European Journal of Dental Education*, *14*(2), 92-98.

Pratt, J.W. 1959. Remarks on zeros and ties in the Wilcoxon signed rank procedures. *Journal of the American Statistical Association*, *54*(287), 655-667.

Price, M., Handley, K. and Millar, J. 2011. Feedback: Focusing attention on engagement. *Studies in Higher Education*, *36*(8), 879-896.

Prothero, J.H. 1923. Prosthetic dentistry. Chicago: medico-dental publishing co.

Puryer, J. and O'Sullivan, D. 2015. An introduction to standard setting methods in dentistry. *British Dental Journal*, *219*(7), 355.

Quinn, F., Keogh, P., McDonald, A. and Hussey, D., 2003. A pilot study comparing the effectiveness of conventional training and virtual reality simulation in the skills acquisition of junior dental students. *European Journal of Dental Education*, *7*(1), 13-19.

Quizlet. 2017. *Class II preparations Flashcards | Quizlet*. [online] Available at: https://quizlet.com/21177892/class-ii-preparations-flash-cards/

Rajiah, K., Veettil, S.K. and Kumar, S. 2014. Standard setting in OSCEs: a borderline approach. *The Clinical Teacher*, *11*(7), 551-556.

Ramaprasad, A. 1983. On the definition of feedback. *Systems Research and Behavioral Science*, *28*(1), 4-13.

Rankin, G. and Stokes, M., 1998. Reliability of assessment tools in rehabilitation: an illustration of appropriate statistical analyses. *Clinical Rehabilitation*, *12*(3), 187-199.

Rao, V., George, A.M., Sahu, S.K. and Krishnaswamy, N.R. 2011. Surface roughness evaluation of enamel after various stripping methods by using profilometer. *Archive of Oral Science and Research*, *1*(4), 190-7.

Renne, W.G., McGill, S.T., Mennito, A.S., Wolf, B.J., Marlow, N.M., Shaftman, S. and Holmes, J.R. 2013. E4D compare software: an alternative to faculty grading in dental education. *Journal of Dental Education*, *77*(2), 168-175.

Ricketts, D. and Bartlett, D.W. 2011. *Advanced Operative Dentistry*, Edinburgh, Churchill Livingstone.

Roberson, T.M., Heymann, H. and SWIFT JR, E., 2002. 4th ed. Sturdevant's art and science of operative dentistry, St. Louis: Mosby.

Rodda, J.C. 1972. Modern class 2 amalgam cavity preparations. *The New Zealand Dental Journal*, *68*(312), 132.

Rolfe, I. and McPherson, J. 1995. Formative assessment: how am I doing?. *The Lancet*, *345*(8953), 837-839.

Rose, J.T., Buchanan, J.A. and Sarrett, D.C. 1999. The dentsim system. *Journal of Dental Education*, *63*(5), 421-3.

Rosenberg, H., Grad, H.A. and Matear, D.W. 2003. The effectiveness of computer-aided, self-instructional programs in dental education: a systematic review of the literature. *Journal of Dental Education*, *67*(5), 524-532.

Rosenstiel, S.F., Land, M.F. and Fujimoto, J. 2006. *Contemporary Fixed Prosthodontics*. Elsevier Health Sciences.

Sadler, D.R. 1983. Evaluation and the improvement of academic learning. *The Journal of Higher Education*, *54*(1), 60-79.

Sadler, D.R. 1989. Formative assessment and the design of instructional systems. *Instructional Science*, *18*(2), 119-144.

Sadler, D.R. 2005. Interpretations of criteria-based assessment and grading in higher education. *Assessment & Evaluation in Higher Education*, *30*(2), 175-194.

Sadler, D.R. 2010. Beyond feedback: Developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education*, *35*(5), 535-550.

Safrit, M.J. 1976. *Reliability theory*. American Alliance for Health, Physical Education, and Recreation.

Salvendy, G., Root, C., Cunningham, P.R. Ferguson, G.W., Hinton, W.M., Baum, S. and Khan, L., 1973. Skills analysis of cavity preparations: class 1 in mandibular right first molar. *Journal of Dental Education*, *37*(10), 11-18.

Sato, T., Mutawa, N.A., Okada, D. and Hasegawa, S., 1998. A clinical study on abutment taper and height of full cast crown preparations. *Journal of Medical and Dental Sciences*, *45*, 205-210.

Satterthwaite, J.D. and Grey, N.J.A. 2008. Peer-group assessment of pre-clinical operative skills in Restorative Dentistry and comparison with experienced assessors. *European Journal of Dental Education*, *12*(2), 99-102.

Satwik, A., and Neelakantan, P. 2016. Quality of Class II Cavity Preparations Performed by Preclinical Students With and Without Magnifying Loupes. https://www.oralhealthgroup.com/features/quality-class-ii-cavity-preparations-performed-preclinical-students-without-magnifying-loupes/

Sawilowsky, S.S. and Blair, R.C. 1992. A more realistic look at the robustness and Type II error properties of the t test to departures from population normality. *Psychological Bulletin*, *111*(2), 352.

Schiff, A.J., Salvendy, G., Root, C.M., Ferguson, G.W. and Cunningham, P.R. 1975. Objective evaluation of quality in cavity preparations. *Journal of Dental Education*, *39*(2), 92-96.

Schoonheim-Klein, M., Muijtjens, A., Habets, L., Manogue, M., Van Der Vleuten, C. and Van der Velden, U. 2009. Who will pass the dental OSCE? Comparison of the Angoff and the borderline regression standard setting methods. *European Journal of Dental Education*, *13*(3), 162-171.

Schoonheim-Klein, M.E., Habets, L.L.M.H., Aartman, I.H.A., Van der Vleuten, C.P., Hoogstraten, J. and Van der Velden, U. 2006. Implementing an objective structured clinical examination (OSCE) in dental education: effects on students' learning strategies. *European Journal of Dental Education*, *10*(4), 226-235.

Schultz, K.S., and Whitney, D.J. 2005. Measurement theory in action. Thousand Oaks, CA: Sage

Schuwirth, L.W. and van der Vleuten, C.P. 2010. How to design a useful test: the principles of assessment. In: SWANWICK, T. (ed.) *Understanding medical education: Evidence, Theory and Practice*, West Sussex, UK: Wiley-Blackwell.

Scruggs, R.R., Daniel, S.J., Larkin, A. and Stoltz, R.F. 1989. Effects of specific criteria and calibration on examiner reliability. *Journal of Dental Hygiene: JDH*, *63*(3), 125-129.

Seo, Y.J., Kwon, T.K., Han, J.S., Lee, J.B., Kim, S.H. and Yeo, I.S. 2014. The reliability of an easy measuring method for abutment convergence angle with a computer-aided design (CAD) system. *The Journal of Advanced Prosthodontics*, *6*(3), 185-193.

Seymour, M.D. 1989. *Effects of a psychomotor taxonomy performance scale on state anxiety in student nurses during the evaluation process* (Doctoral dissertation, Lethbridge, Alta.: University of Lethbridge, Faculty of Education, 1989).

Shahrbaf, S., Mirzakouchaki, B., Oskoui, S.S. and Kahnamoui, M.A. 2007. The effect of marginal ridge thickness on the fracture resistance of endodontically-treated, composite restored maxillary premolars. *Operative Dentistry*, *32*(3), 285-290.

Sharaf, A.A., AbdelAziz, A.M. and El Meligy, O.A. 2007. Intra-and inter-examiner variability in evaluating preclinical pediatric dentistry operative procedures. *Journal of Dental Education*, *71*(4), 540-544.

Sherwood, I.A. and Douglas, G.V. 2014. A study of examiner variability in assessment of preclinical class II amalgam preparation. *Journal of Education and Ethics in Dentistry*, *4*(1), 12-17.

Shillingburg, H.T., Sather, D.A., Wilson, E.L., Cain, J.R., Mitchell, D.L., Blanco, L.J. and Kessler, J.C., 2012. *Fundamentals of Fixed Prosthodontics*. Quintessence Publishing Company.

Shrout, P.E. 1998. Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research*, *7*(3), 301-317.

Shrout, P.E. and Fleiss, J.L. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420-428.

Shuttleworth, M., 2009. Concurrent Validity. *Retrieved November*, *7*, .2014.

Sim, J. and Wright, C.C. 2005. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical Therapy*, *85*(3), 257-268.

Sirajudeen, M.S., Pillai, P., Shah, U. and Mohan, N. 2012. Content validity and inter-rater reliability of a checklist to assess the ergonomic practice of computer professionals. *International Journal of Therapy and Rehabilitation Research*, *1*, 11-18.

Smee, S.M. and Blackmore, D.E. 2001. Setting standards for an objective structured clinical examination: the borderline group method gains ground on Angoff. *Medical Education*, *35*(11), 1009-1010.

Smith, B. G. N., and Howe, L. C. 2007. *Planning and Making Crowns and Bridges*. Abingdon, Oxon, UK, Informa Healthcare.

Söderholm, K.J., Tyas, M.J. and Jokstad, A. 1998. Determinants of quality in operative dentistry. *Critical Reviews in Oral Biology & Medicine*, *9*(4), 464-479.

Streiner, D.L., and Norman, G.R. 2008. 4th ed. *Health measurement scales: a practical guide to their development and use*. Oxford University Press, UK.

Suvinen, T.I., Messer, L.B. and Franco, E. 1998. Clinical simulation in teaching preclinical dentistry. *European Journal of Dental Education*, *2*(1), 25-32.

Taras, M. 2008. Summative and formative assessment: Perceptions and realities. *Active Learning in Higher Education*, *9*(2), 172-192.

Taylor, C.L., Grey, N. and Satterthwaite, J.D. 2013. Assessing the clinical skills of dental students: a review of the literature. *Journal of Education and Learning*, *2*(1), 20-31.

THE CONSORTIUM, Overseas Registration Examination, Part 2. xa.yimg.com/kq/groups/31044306/141675020/name/dmexercisedescriptions.pdf

The General Dental Council. 2013. Annual review of education. London: GDC.

Tinsley, H.E. and Weiss, D.J. 1975. Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology*, *22*(4), 358.

Tiu, J., Al-Amleh, B., Waddell, J.N. and Duncan, W.J. 2015. Clinical tooth preparations and associated measuring methods: a systematic review. *The Journal of Prosthetic Dentistry*, *113*(3), 175-184.

Tiu, J., Waddell, J.N., Al-Amleh, B., van Vuuren, W.A.J. and Swain, M.V. 2014. Coordinate geometry method for capturing and evaluating crown preparation geometry. *The Journal of Prosthetic Dentistry*, *112*(3), 481-487.

Trochim, W.M. 2006. Research methods knowledge base.

Tuncer, D., Arhun, N., Yamanel, K., Çelik, Ç. and Dayangaç, B. 2015. Dental Students' Ability to Assess Their Performance in a Preclinical Restorative Course: Comparison of Students' and Faculty Members' Assessments. *Journal of Dental Education*, *79*(6), 658-664.

Turnbull, J., Gray, J. and MacFadyen, J. 1998. Improving In-Training Evaluation Programs. *Journal of General Internal Medicine*, *13*(5), 317-323.

Turnbull, J.M. 1989. What is… normative versus criterion-referenced assessment. *Medical Teacher*, *11*(2), 145-150.

Tylman, S.D. and Malone, W. F. P. 1970. *Theory and practice of crown and fixed partial prosthodontics*. St. Louis, the CV Mosby Co.

Van de Ridder, J.M., Stokking, K.M., McGaghie, W.C. and Ten Cate, O.T.J. 2008. What is feedback in clinical education?. *Medical Education*, *42*(2), 189-197.

van Der Vleuten, C.P. 1996. The assessment of professional competence: developments, research and practical implications. *Advances in Health Sciences Education*, *1*(1), 41-67.

van der Vleuten, C.P., Schuwirth, L.W.T., Driessen, E.W., Dijkstra, J., Tigelaar, D., Baartman, L.K.J. and van Tartwijk, J., 2012. A model for programmatic assessment fit for purpose. *Medical Teacher*, *34*(3), 205-214.

van der Vleuten, C.P., Schuwirth, L.W.T., Driessen, E.W., Dijkstra, J., Tigelaar, D., Baartman, L.K.J. and van Tartwijk, J. 2012. A model for programmatic assessment fit for purpose. *Medical Teacher*, *34*(3), 205-214.

Van Stralen, K.J., Jager, K.J., Zoccali, C. and Dekker, F.W. 2008. Agreement between methods. *Kidney International*, *74*(9), 1116-1120.

Vanek, H.G. 1969. Objective evaluation of dental student technic products. *Journal of Dental Education*, *33*(1), 140-144.

Vann, W.F., Machen, J.B. and Hounshell, P.B. 1983. Effects of criteria and checklists on reliability in preclinical evaluation. *Journal of Dental Education*, *47*(10), 671-675.

Waltz, C.F., Strickland, O.L. and Lenz, E.R. 4th ed. 2010. *Measurement in nursing and health research*. Springer Publishing Company.

Wass, V., Van der Vleuten, C., Shatzer, J. and Jones, R. 2001. Assessment of clinical competence. *The Lancet*, *357*(9260), 945-949.

Watson, M.J., Wong, D.M., Kluger, R., Chuan, A., Herrick, M.D., Ng, I., Castanelli, D.J., Lin, L., Lansdown, A.K. and Barrington, M.J. 2014. Psychometric evaluation of a direct observation of procedural skills assessment tool for ultrasound-guided regional anaesthesia. *Anaesthesia*, *69*(6), 604-612.

Weiner, J. 2007. Measurement: reliability and validity measures. *Bloomberg School of Public Health, Johns Hopkins University, mimeo (Power Point Presentation) at* http://ocw.jhsph.edu/courses/hsre/PDFs/HSRE_lect7_weiner.pdf

Welk, A., Maggio, M.P., Simon, J.F., Scarbecz, M., Harrison, J.A., Wicks, R.A. and Gilpatrick, R.O. 2008. Computer-assisted learning and simulation lab with 40 DentSim units. *International Journal of Computerized Dentistry*, *11*(1), 17-40.

Wilson, A.H. and Chan, D.C. 1994. The relationship between preparation convergence and retention of extracoronal retainers. *Journal of Prosthodontics*, *3*(2), 74-78.

Wingate, U. 2010. The impact of formative feedback on the development of academic writing. *Assessment & Evaluation in Higher Education*, *35*(5), 519-533.

Woolley, A.S. 1977. The long and tortured history of clinical evaluation. *Nursing Outlook*, *25*(5), 308-315.

Wynd, C.A., Schmidt, B. and Schaefer, M.A. 2003. Two quantitative approaches for estimating content validity. *Western Journal of Nursing Research*, *25*(5), 508-518.

Yates, J.L., Hembree Jr, J.H. and McKnight, J.P. 1975. A comparison of the effect of cavity design on the fracture strength of three dental alloy systems. *ASDC Journal of Dentistry for Children*, *43*(3),187-191.

Yen, M. and Lo, L.H. 2002. Examining test-retest reliability: an intra-class correlation approach. *Nursing Research*, *51*(1), 59-62.

Yoon, S.S., Cheong, C., Preisser, J., Jun, S., Chang, B.M. and Wright, R.F. 2014. Measurement of total occlusal convergence of 3 different tooth preparations in 4 different planes by dental students. *The Journal of Prosthetic Dentistry, 112*(2), 285-292.

Zieky, M., Perie, M. and Livingston, S. 2006. A primer on setting cut scores on tests of educational achievement. *Princeton, NJ: Educational Testing Service*, *320*.

Zou, H., Jin, S., Sun, J. and Dai, Y. 2016. A Cavity Preparation Evaluation System in the Skill Assessment of Dental Students. *Journal of Dental Education*, *80*(8), 930-937.

Appendices

APPENDIX 1: Details of test number for class II amalgam cavity preparation along with ticket number, grade awarded and number of negative points awarded for each student using 'Gray feedback sheet' in 2014/2015

Test number	Ticket number	Grade	Number of negative points
1	754	4	1
1	475	4	2
1	533	4	2
1	437	4	2
1	479	4	1
1	362	4	1
1	973	4	2
1	213	4	1
1	253	4	1
1	832	4	1
1	378	2	11
1	537	2	11
1	797	2	11
1	304	3	5
1	792	2	2
1	686	2	11
1	299	3	2
1	727	3	2
1	942	3	4
1	696	3	3
1	552	3	3
1	802	2	4
1	249	3	4
1	342	2	3
1	968	2	11
1	599	2	6
1	632	2	4
1	922	2	4
1	827	2	4
1	732	2	11
1	598	2	11
1	667	3	3
1	583	3	3
1	472	3	3
1	865	3	1
1	312	3	3
1	920	3	2
1	959	3	2
1	538	4	1
1	550	4	1
1	875	4	1
1	254	4	1
1	367	5	0
1	460	2	11
1	433	2	9
1	313	2	3
1	392	2	4
1	305	2	11
1	880	2	2
1	963	2	2
1	907	2	11
1	849	2	11
1	457	2	4
1	372	3	3
1	597	3	1
1	424	3	2
1	487	3	1
1	447	3	4
1	288	4	1
1	522	4	1

Test number	Ticket number	Grade	Number of negative points
1	807	4	2
1	798	5	0
1	637	4	1
1	547	4	1
1	857	4	1
2	398	4	4
2	399	3	3
2	894	3	3
2	628	4	0
2	504	4	2
2	965	4	2
2	634	3	3
2	738	3	2
2	513	4	2
2	577	1	11
2	483	2	11
2	964	2	11
2	345	2	11
2	397	2	4
2	893	2	2
2	933	2	11
2	854	2	11
2	949	2	11
2	744	2	3
2	278	2	4
2	294	3	1
2	677	3	2
2	259	3	2
2	948	2	6
2	905	2	11
2	944	2	4
2	319	2	8
2	908	5	0
2	344	5	0
2	780	4	2
2	505	4	2
2	915	4	4
2	363	4	2
2	733	4	1
2	682	3	2
2	358	3	2
2	745	3	1
2	455	3	2
2	557	3	2
2	914	3	2
2	895	3	2
2	265	3	3
2	510	3	3
2	454	3	3
3	899	2	4
3	652	2	3
3	509	2	11
3	753	2	3
3	944	2	3
3	768	2	6
3	818	2	4
3	494	3	5
3	379	3	3
3	877	3	3
3	653	3	4
3	717	3	4
3	708	2	2
3	800	2	8
3	434	5	0
3	273	4	2

Test number	Ticket number	Grade	Number of negative points
3	544	3	4
3	430	3	2
3	842	3	1
3	353	5	0
3	696	5	0
3	923	5	1
3	283	5	0
3	918	5	0
3	324	5	0
3	808	5	0
3	473	5	0
3	830	5	0
3	844	5	0
3	663	4	2
3	629	4	3
3	934	3	2
3	274	3	3
3	489	3	3
3	643	2	5
3	793	2	6
3	568	2	6
3	794	2	11
3	638	2	4
3	373	2	11
3	518	2	4
3	279	2	4
3	763	2	5
3	448	2	4
3	924	2	5
3	548	2	3
3	323	2	5
3	833	2	11
3	734	2	3
3	569	2	11
3	788	2	4
3	328	2	11
3	709	3	4
3	458	3	3
3	543	3	3
3	718	4	1
3	843	4	3
3	803	4	2
3	823	4	2
3	293	4	2
3	284	5	0
3	648	5	0
3	759	5	0
3	388	5	0
3	668	3	2
3	927	4	1
4	334	5	0
4	387	5	0
4	928	5	0
4	929	5	0
4	339	5	0
4	292	5	0
4	809	4	2
4	338	4	2
4	570	4	2
4	858	4	3
4	329	5	0
4	649	5	1
4	549	3	2
4	719	3	4
4	678	3	1

Test number	Ticket number	Grade	Number of negative points
4	715	2	11
4	449	2	11
4	828	2	11
4	484	2	11
4	639	2	11
4	898	2	11
4	935	2	3
4	406	5	0
4	520	5	0
4	429	5	0
4	730	5	0
4	654	5	0
4	795	5	0
4	393	5	0
4	321	4	5
4	515	4	1
4	11	4	3
4	740	4	2
4	435	3	2
4	724	3	2
4	670	3	2
4	878	3	3
4	820	2	4
4	804	2	2
4	369	2	3
4	879	2	3
4	394	2	2
4	514	2	5
4	490	2	2
4	450	1	11
4	735	5	0
4	389	5	0
4	645	5	0
4	360	5	0
4	562	5	0
4	974	5	0
4	859	5	0
4	359	5	0
4	714	4	1
4	474	4	3
4	485	4	1
4	739	3	1
4	630	3	1
4	495	3	2
4	635	3	3
4	697	3	2
4	459	3	1
4	919	2	2
4	783	2	2
4	699	2	11
4	925	2	3
4	644	2	2

APPENDIX 2: Details of test number for full veneer gold shell crown preparation, along with ticket number, grade awarded and number of negative points awarded for each student using 'Mhanni feedback sheet' in 2014/2015

Ticket number	Grade*	Number of negative points
1	3	4
3	4	6
4	3	8
5	2	10
7	4	0
13	2	20
14	4	4
15	4	2
16	4	6
18	2	11
20	4	2
21	4	0
22	2	11
25	3	9
26	2	20
29	3	11
31	3	2
37	3	8
40	2	20
52	3	2
54	3	11
57	4	3
58	3	10
59	2	20
60	3	8
63	3	10
65	4	3
67	4	5
69	4	2
70	4	7
71	4	4
73	3	7
74	3	7
78	4	1
88	2	12
89	3	4
90	4	6
91	3	8
92	2	10
93	4	0
94	2	20
95	4	4
96	4	2
97	4	6
98	2	11
99	4	2
100	4	0
101	2	11
102	3	9
103	2	20
104	3	11
105	3	2
106	3	8
107	2	20
108	3	2
109	3	11
110	4	3
111	3	10
112	2	20
113	3	8

Ticket number	Grade*	Number of negative points
114	3	10
115	4	3
116	4	5
117	4	2
118	4	7
119	4	4
120	3	7
121	3	7
122	4	1
123	2	12

*(Grades of full veneer gold shell crown preparation are 1=D, 2=C, 3=B, and 4=A)

APPENDIX 3: Summary of agreement and correlation of specific additional tools and feedback sheets for three senior academic staff examiners

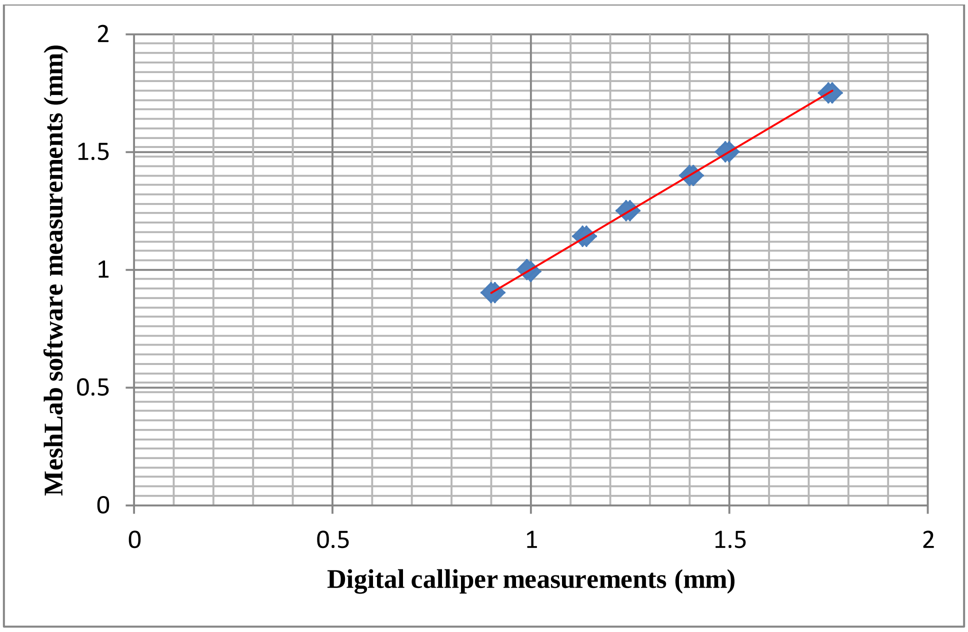
Evaluation type *Scaling system*	Evaluation with Condenser			Evaluation with Bur			Evaluation with Gray feedback sheet (GFS)			Spearman Correlation	Number of negative points from (GFS)		
	Scale = 1-5			*Scale = 1-5*			*Scale = 1-5*				*Scale = 1-11*		
	Occasion one O_1	KAPPA	Occasion two O_2	Occasion one O_1	KAPPA	Occasion two O_2	Occasion one O_1	KAPPA	Occasion two O_2		Occasion one O_1	KAPPA	Occasion two O_2
Examiner 1	1 ⇧ 26	K= 0.43	1 ⇧ 26	1 ⇧ 26	K=0.43	1 ⇧ 26	1 ⇧ 26	K= 0.58	1 ⇧ 26	O_1 O_1 r = -0.92 O_2 O_2 r = -0.86 O_{1+2} O_{1+2} r = -0.87	1 ⇧ 26	K= 0.38	1 ⇧ 26
Examiner 2	1 ⇧ 26	K= 0.32	1 ⇧ 26	1 ⇧ 26	K= 0.43	1 ⇧ 26	1 ⇧ 26	K= 0.64	1 ⇧ 26	O_1 O_1 r = -0.80 O_2 O_2 r = -0.72 O_{1+2} O_{1+2} r = -0.76	1 ⇧ 26	K= 0.21	1 ⇧ 26
Examiner 3	1 ⇧ 26	K= 0.53	1 ⇧ 26	1 ⇧ 26	K= 0.48	1 ⇧ 26	1 ⇧ 26	K= 0.74	1 ⇧ 26	O_1 O_1 r = -0.95 O_2 O_2 r = -0.93 O_{1+2} O_{1+2} r = -0.93	1 ⇧ 26	K= 0.59	1 ⇧ 26
Intra-Class Correlation	0.71		0.71	0.56		0.70	0.54		0.69		0.79		0.80

a. Class II amalgam cavity examination by three senior examiners on two occasions (Occasion one "O1" and Occasion two "O2") to determine examiner agreement and correlation for 26 cavities. The highlighted data indicates the highest agreement and correlation

Evaluation type *Scaling system*	Evaluation with Bur			Evaluation with Impression index			Evaluation with Mhanni feedback sheet (MFS)				Number of negative points from (MFS)		
	Scale = 1-5			*Scale = 1-5*			*Scale = 1-5*				*Scale = 1-20*		
	Occasion one O_1	KAPPA	Occasion two O_2	Occasion one O_1	KAPPA	Occasion two O_2	Occasion one O_1	KAPPA	Occasion two O_2	Spearman Correlation	Occasion one O_1	KAPPA	Occasion two O_2
Examiner 1	1 ⇧ 30	K= 0.35	1 ⇧ 30	1 ⇧ 30	K= 0.49	1 ⇧ 30	1 ⇧ 30	K= 0.57	1 ⇧ 30	O_1 O_1 r = -0.94 O_2 O_2 r = -0.88 O_{1+2} O_{1+2} r = -0.90	1 ⇧ 30	K= 0.20	1 ⇧ 30
Examiner 2	1 ⇧ 30	K= 0.39	1 ⇧ 30	1 ⇧ 30	K= 0.55	1 ⇧ 30	1 ⇧ 30	K= 0.41	1 ⇧ 30	O_1 O_1 r = -0.89 O_2 O_2 r = -0.88 O_{1+2} O_{1+2} r = -0.89	1 ⇧ 30	K= -0.01	1 ⇧ 30
Examiner 3	1 ⇧ 30	K=0.05	1 ⇧ 30	1 ⇧ 30	K= 0.15	1 ⇧ 30	1 ⇧ 30	K= 0.27	1 ⇧ 30	O_1 O_1 r = -0.74 O_2 O_2 r = -0.66 O_{1+2} O_{1+2} r = -0.62	1 ⇧ 30	K= 0.14	1 ⇧ 30
Intra-Class Correlation	0.49		0.38	0.25		0.35	0.34		0.38		0.56		0.65

b. Full veneer gold shell crown examination by three senior examiners on two occasions (Occasion one "O1" and Occasion two "O2") to determine examiner agreement and correlation for 30 preparations. The highlighted data indicates the highest agreement and correlation.

APPENDIX 4: Correlation between the digital calliper and MeshLab software measurements (mm) of the ParaPostXP, parallel-sided, impression plastic posts



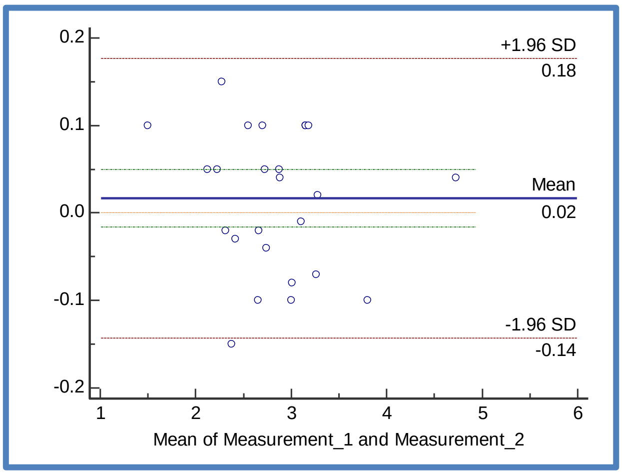
The results were ($r^2 = 0.9992$, $y = 0.9978x + 0.0028$).

($r^2 = 1$, $y=1x + 0$).

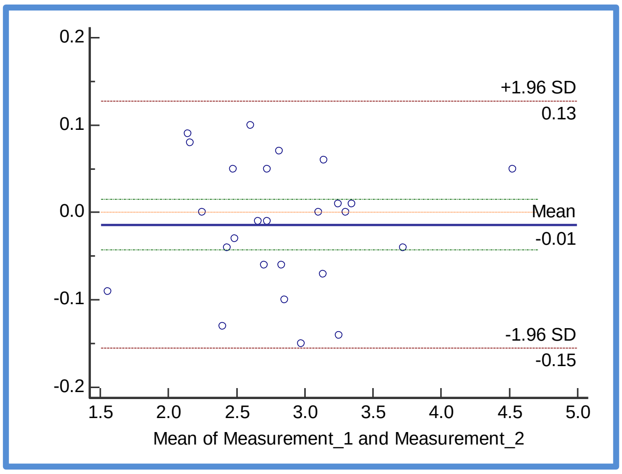
i.e. ($y = x$).

APPENDIX 5: Bland and Altman plots of differences and mean measurements for SAFMs of class II amalgam cavities

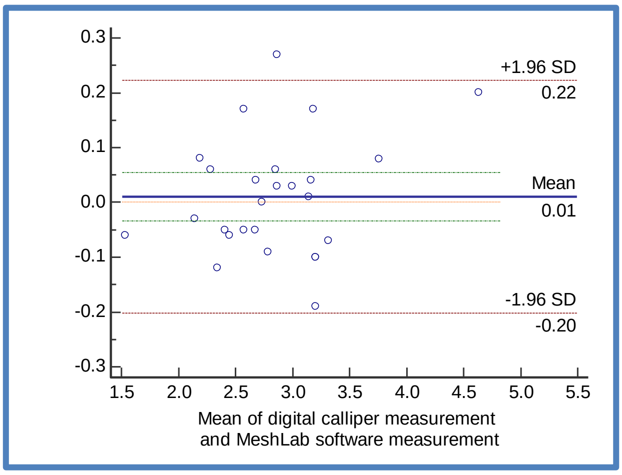
These figures illustrated that mean difference and limit of agreement for depth of the box in gingival direction for class II amalgam cavity by using two different methods on two occasions. Measurements on the first occasion from the digital calliper tended to be higher than on the second occasion while measurements on the first occasion from the MeshLab software were lower than on the second occasion [Figures (a) and (b)]. From Figure (c), the measured values from the MeshLab software were lower than those from the digital calliper. The value of mean difference between the two methods was extremely small. In addition, the widths of the limit of agreement for each plot were acceptable clinically, according to the examiners' opinion, demonstrating differences of less than 0.50 mm (actually maximum = 0.45mm) in every case. Therefore, measuring the depth of the proximal box mesio-distally by using MeshLab software (an indirect measurement) was reliable under assumption that the digital calliper method (a direct measurement) was reliable from previous steps in this section.



a. Digital calliper 1st and 2nd measurements



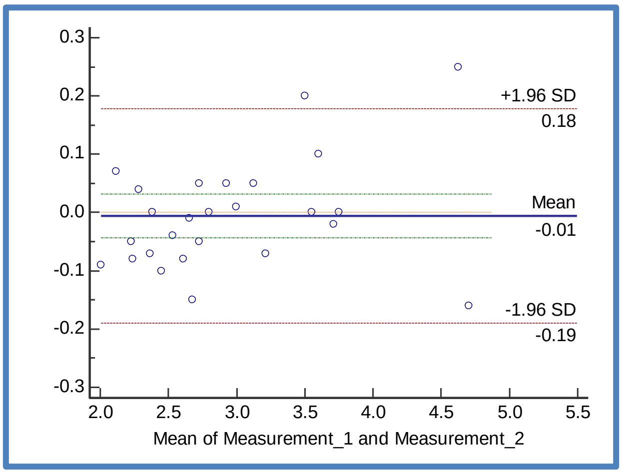
b. MeshLab software 1st and 2nd measurements



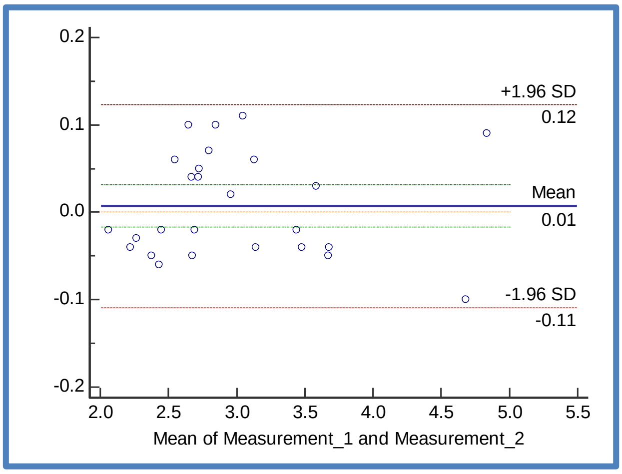
c. Digital calliper and MeshLab software measurements

Bland and Altman plots of differences and mean measurements of **depth of the box in gingival direction for class II amalgam cavity** by using two different methods on two occasions (a and b) and between methods (c)

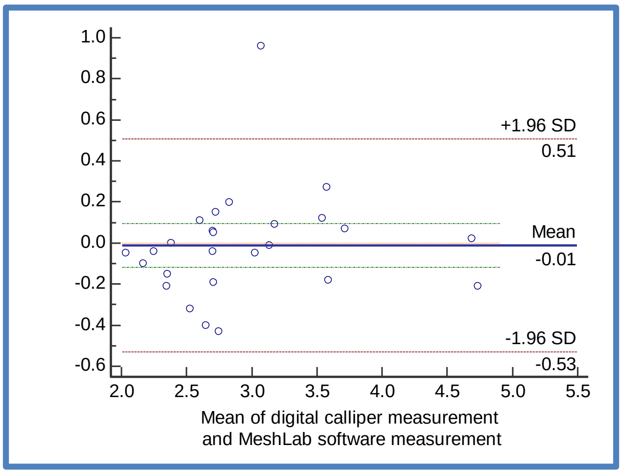
These figures demonstrated that mean difference and limit of agreement for bucco-palatal width of the box floor for class II amalgam cavity by using two different methods on two occasions. These figures also showed that the value of mean difference between the two occasions and methods was extremely small. On the other hand, Figure (c) showed that there was a wide limit of agreement between the digital calliper and MeshLab software measurements (> 0.50 mm) of the bucco-palatal width of the box floor.



a. Digital calliper 1st and 2nd measurements



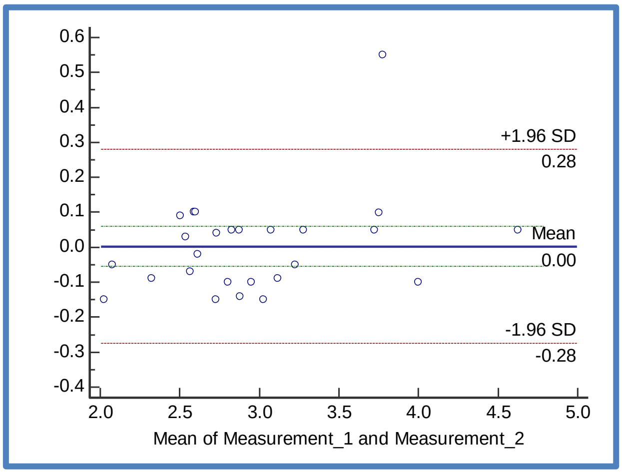
b. MeshLab software 1st and 2nd measurements



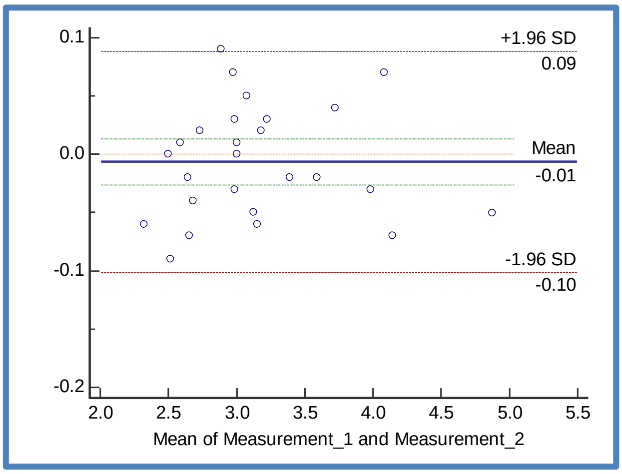
c. Digital calliper and MeshLab software measurements

Bland and Altman plot of differences and mean measurements of **bucco-palatal width of the box floor for class II amalgam cavity** by using two different methods on two occasions (a and b) and between methods (c)

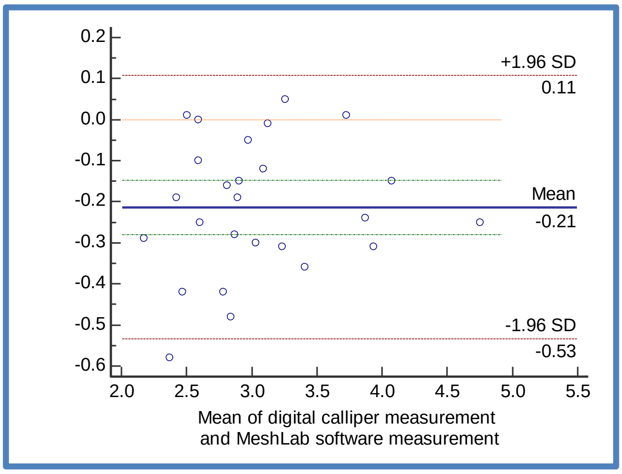
These figures illustrated that mean difference and limit of agreement for bucco-palatal width of the box occlusally for class II amalgam cavity by using two different methods on two occasions. Figure (c) demonstrated the mean measurements from the MeshLab software tend to be higher than those made using the digital calliper. When the two methods were compared, the limit of agreement width was wide (-0.53 – 0.11) and mean difference was -0.21. This indicates that there was a systematic difference between the digital calliper and the MeshLab software measurements. One reason for this was that the points between which the measurements are made for the bucco-palatal width of the box occlusally both lie on a curve and, as such, are very difficult to identify repeatability.



a. Digital calliper 1st and 2nd measurements



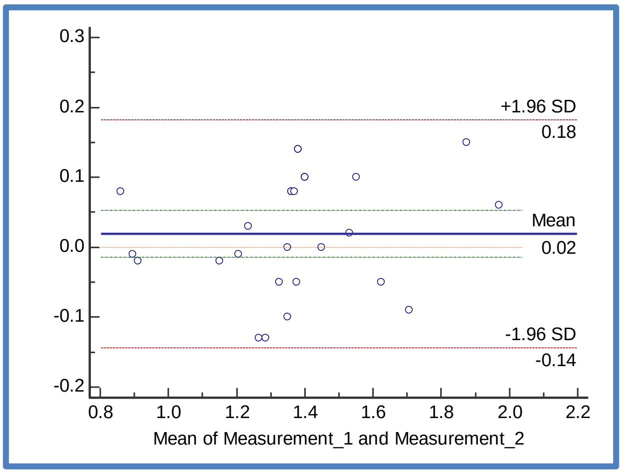
b. MeshLab software 1st and 2nd measurements



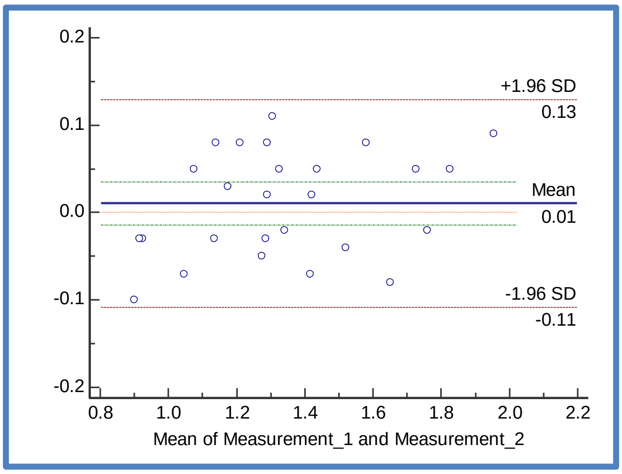
c. Digital calliper and MeshLab software measurements

Bland and Altman plot of differences and mean measurements of **bucco-palatal width of the box occlusally for class II amalgam cavity** by using two different methods on two occasions (a and b) and between methods (c)

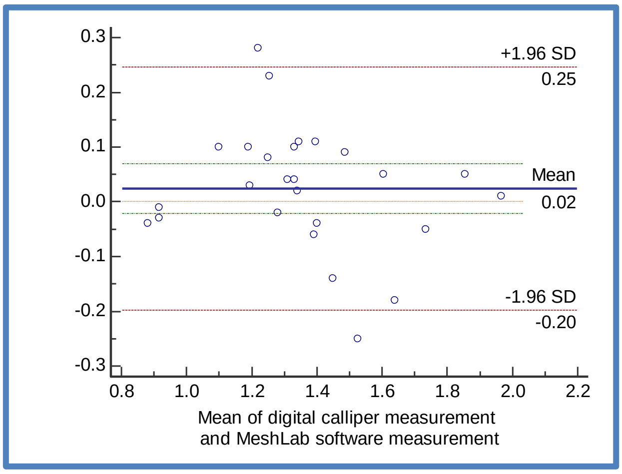
These figures illustrated that mean difference and limit of agreement for mesio-distal depth of the box floor for class II amalgam cavity by using two different methods on two occasions. Figures also demonstrated that measuring methods had acceptable mean difference and limits of agreement for this feature. This figure demonstrated that measurements from the digital calliper and MeshLab software had low mean difference and a narrow limit of agreement. From Figure (c), the measured values from the MeshLab software were lower than those from the digital calliper.



a. Digital calliper 1st and 2nd measurements



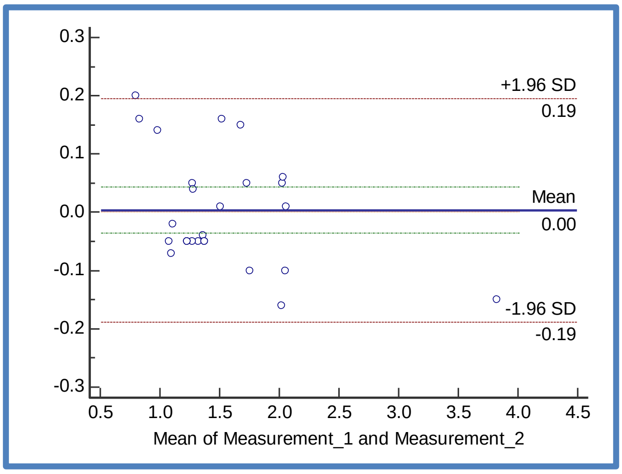
b. MeshLab software 1st and 2nd measurements



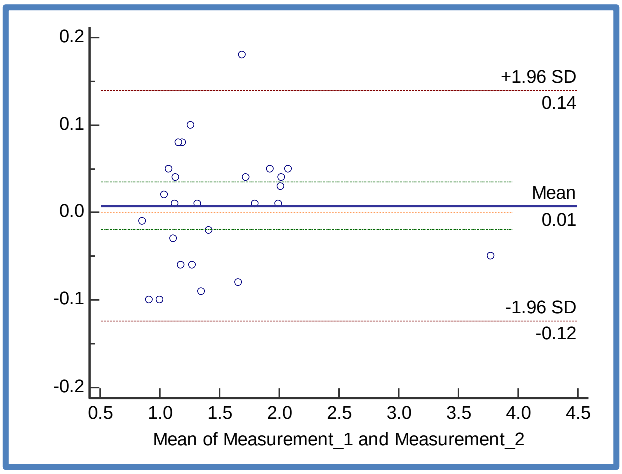
c. Digital calliper and MeshLab software measurements

Bland and Altman plot of differences and mean measurements of **mesio-distal depth of the box floor for class II amalgam cavity** by using two different methods on two occasions (a and b) and between methods (c)

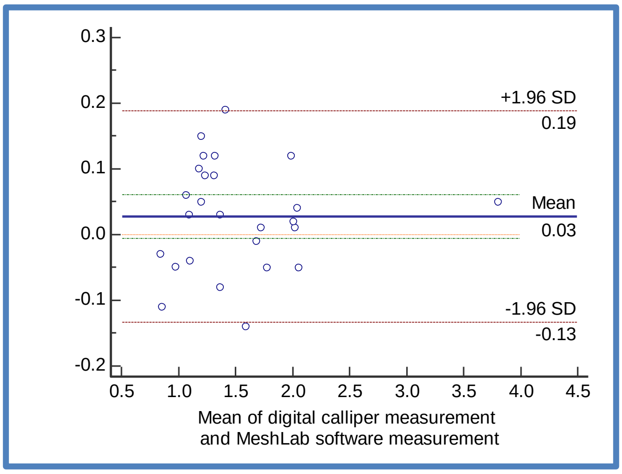
These figures illustrated that mean difference and limit of agreement for pulpal axial wall length of class II amalgam cavity by using two different methods on two occasions. Figures (a) and (b) demonstrated that both measuring methods had acceptable limit of agreement and no mean difference for this feature. Figure (c) demonstrated that averaged measurements from the digital calliper were higher than measurements made using MeshLab software but with an acceptable limit of agreement.



a. Digital calliper 1st and 2nd measurements



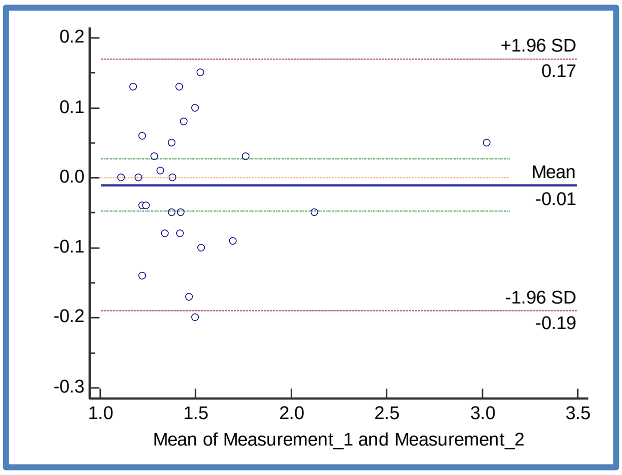
b. MeshLab software 1st and 2nd measurements



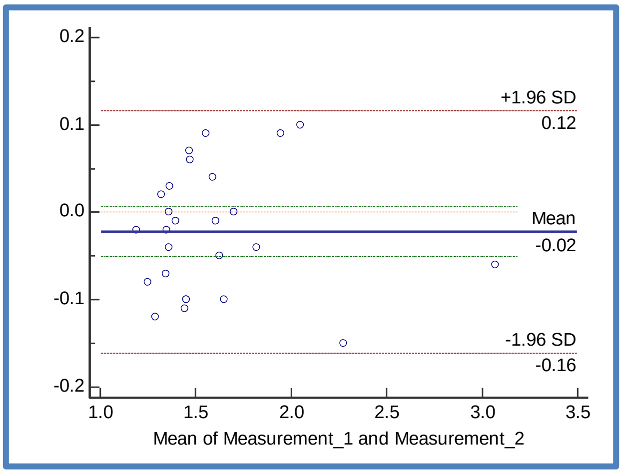
c. Digital calliper and MeshLab software measurements

Bland and Altman plot of differences and mean measurements of **pulpal axial wall length of class II amalgam cavity** by using two different methods on two occasions (a and b) and between methods (c)

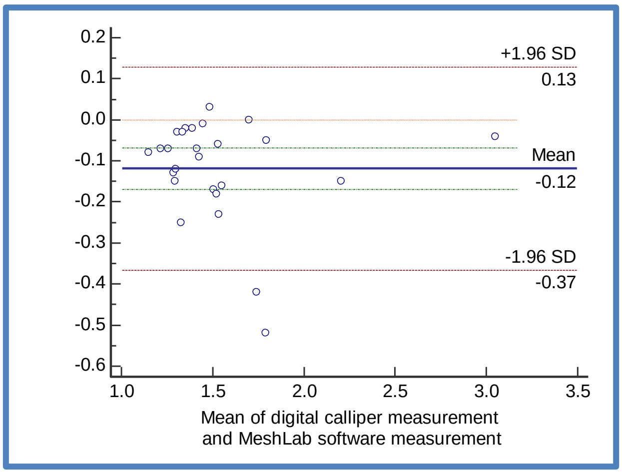
Figures illustrated that mean difference and limit of agreement for isthmus width at occlusal of class II amalgam cavity by using two different methods on two occasions. Figures (a) and (b) demonstrated that both measuring methods had acceptable mean difference and agreement for this feature. Although Figure (c) demonstrated that measurements by using the digital calliper were lower than measurements by using MeshLab software with acceptable limits of agreement, it is clear that there was a systematic difference between two methods.



a. Digital calliper 1st and 2nd measurements



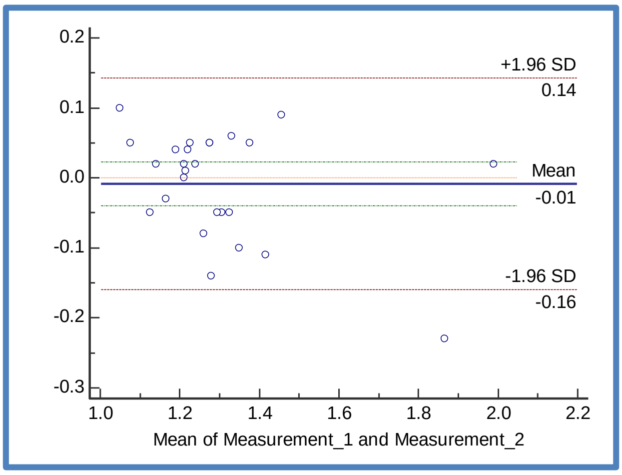
b. MeshLab software 1st and 2nd measurements



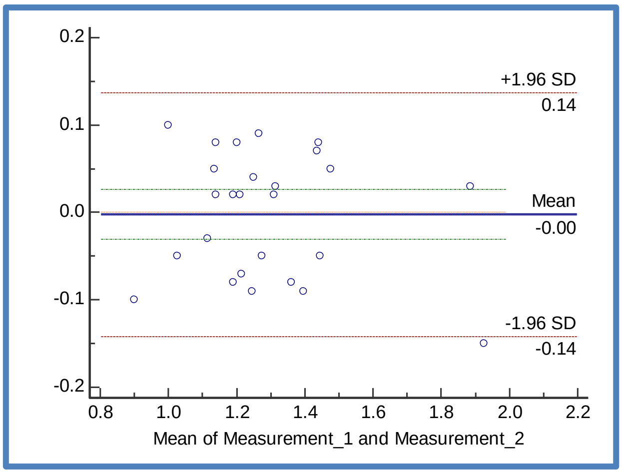
c. Digital calliper and MeshLab software measurements

Bland and Altman plot of differences and mean measurements of **isthmus width at occlusal of class II amalgam cavity** by using two different methods on two occasions (a and b) and between methods (c)

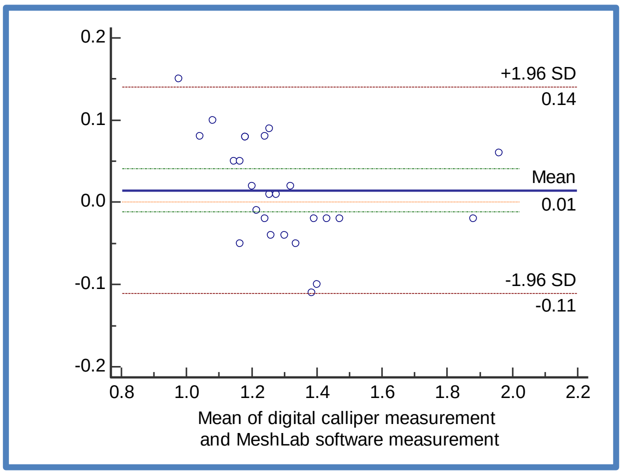
Figures illustrated that mean difference and limit of agreement for isthmus floor width of class II amalgam cavity by using two different methods on two occasions. Figures also demonstrated that both measuring methods had acceptable mean difference and limits of agreement for this feature. It also demonstrated that measurements from the digital calliper and MeshLab software had low mean difference and narrow limits of agreement. In addition, measurements from the digital calliper were slightly higher than measurements from MeshLab software.



a. Digital calliper 1st and 2nd measurements



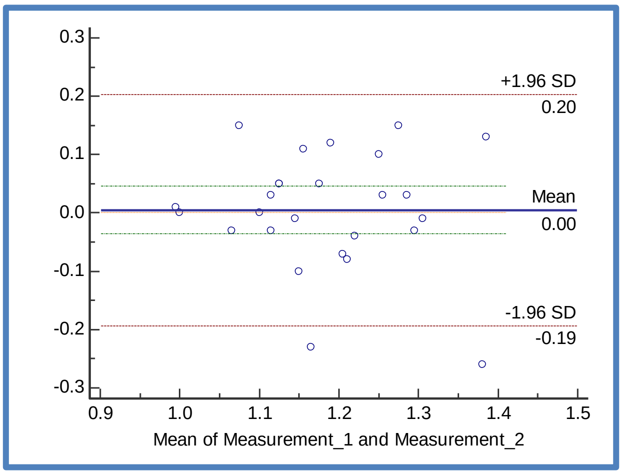
b. MeshLab software 1st and 2nd measurements



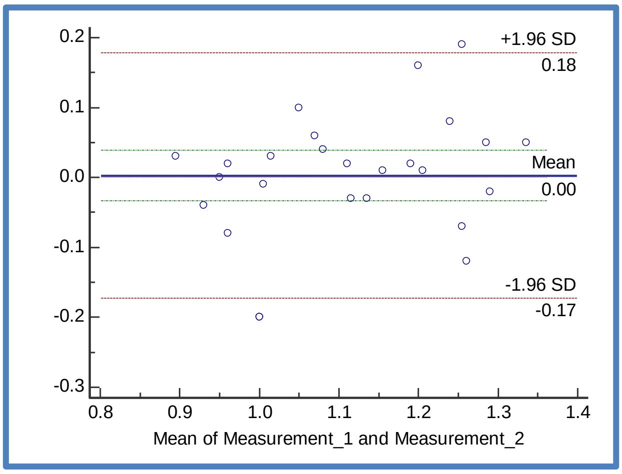
c. Digital calliper and MeshLab software measurements

Bland and Altman plot of differences and mean measurements of **isthmus floor width of class II amalgam cavity** by using two different methods on two occasions (a and b) and between methods (c)

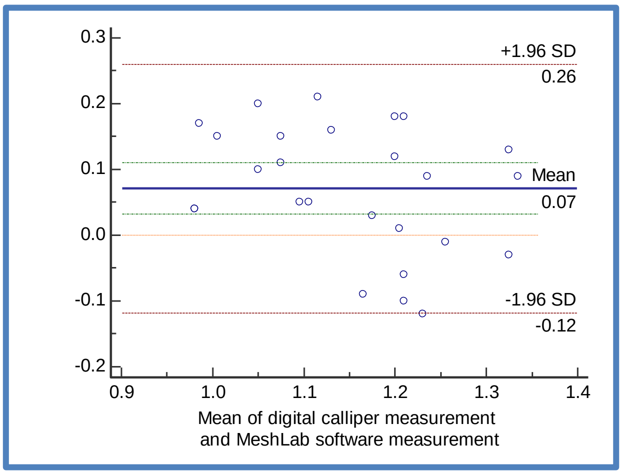
Figures (a) and (b) demonstrated that measuring methods for occlusal cavity width in the middle of class II amalgam cavity had acceptable mean difference and limit of agreement. Figure (c) demonstrated that measurements from the digital calliper were higher than measurements from MeshLab software with an acceptable limit of agreement (-0.12 – 0.26).



a. Digital calliper 1st and 2nd measurements



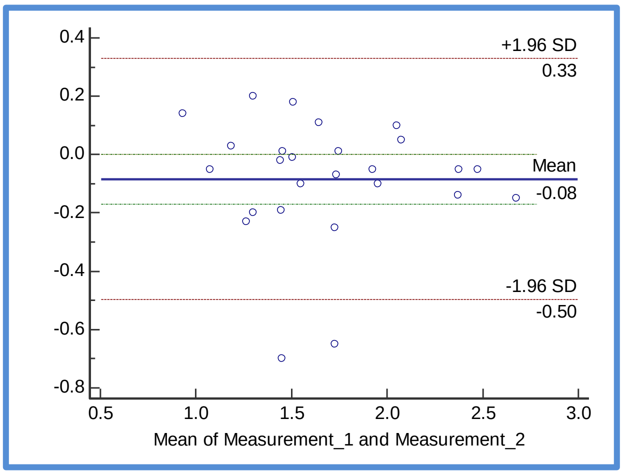
b. MeshLab software 1st and 2nd measurements



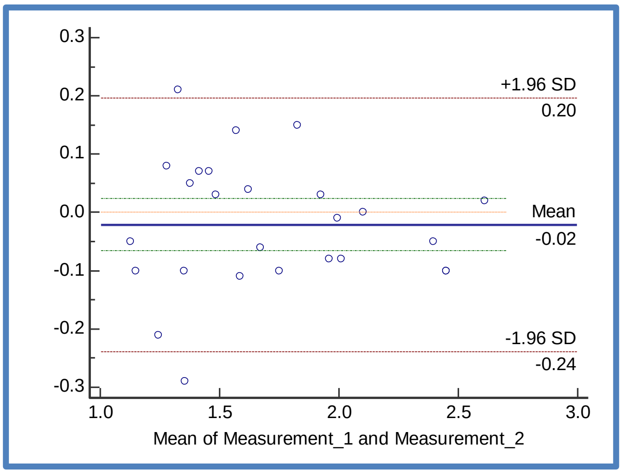
c. Digital calliper and MeshLab software measurements

Bland and Altman plot of differences and mean measurements of **occlusal cavity width in the middle of class II amalgam cavity** by using two different methods on two occasions (a and b) and between methods (c)

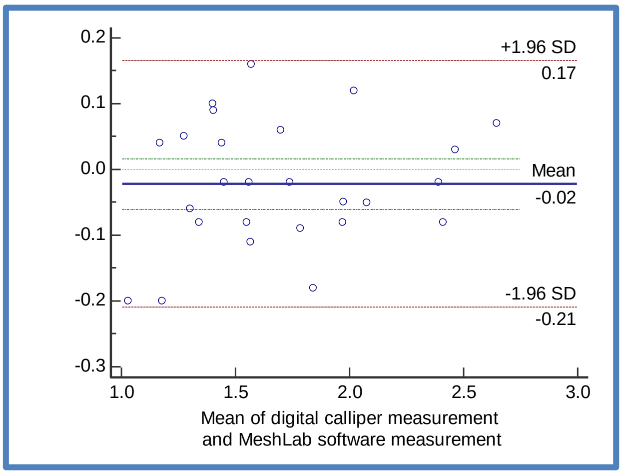
Figures illustrated that mean difference and limit of agreement for occlusal cavity depth (palatal side in the middle) of class II amalgam cavity by using two different methods on two occasions. Figure (a) demonstrated the first measurements from the digital calliper were lower than the second measurements with wide limits of agreement while Figures (b) and (c) demonstrated that measuring methods had acceptable mean difference and limits of agreement for occlusal cavity depth (palatal side in the middle) feature of class II amalgam cavity. Figure (a) also showed that there was (-0.08) mean difference between the two occasions from the digital calliper.



a. Digital calliper 1st and 2nd measurements



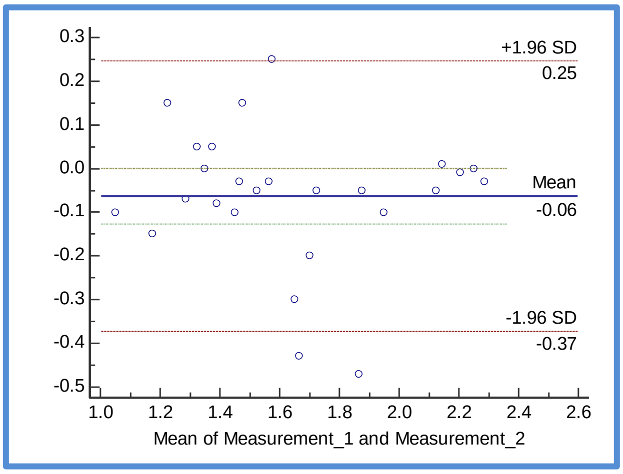
b. MeshLab software 1st and 2nd measurements



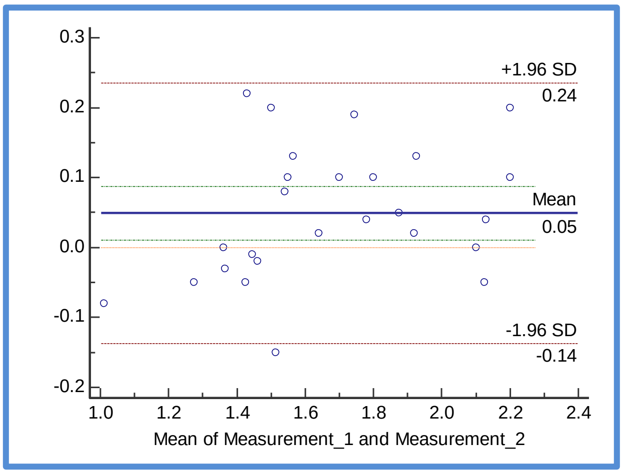
c. Digital calliper and MeshLab software measurements

Bland and Altman plot of differences and mean measurements of **occlusal cavity depth (palatal side in the middle) of class II amalgam cavity** by using two different methods on two occasions (a and b) and between methods (c)

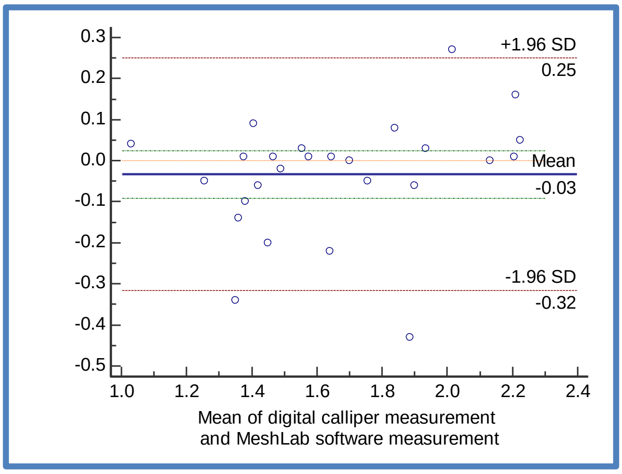
Figures illustrated that mean difference and limit of agreement for occlusal cavity depth (buccal side in the middle) of class II amalgam cavity by using two different methods on two occasions. Figure (a) demonstrated that the first measurements from the digital calliper were lower than the second measurements while Figure (b) demonstrated that the first measurements from the MeshLab were higher than the second measurements. The limit of agreement was wide for measurements from the digital calliper method but narrow for measurements from the MeshLab software method. Figure (c) illustrated that the measuring methods had (-0.03) mean difference and (-0.32 – 0.25) limit of agreement for this feature.



a. Digital calliper 1st and 2nd measurements



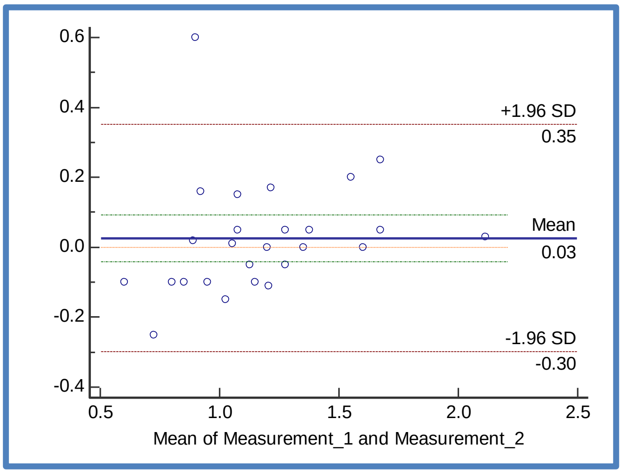
b. MeshLab software 1st and 2nd measurements



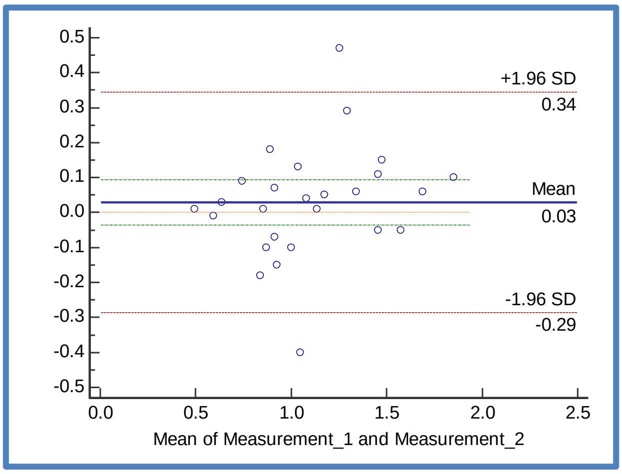
c. Digital calliper and MeshLab software measurements

Bland and Altman plot of differences and mean measurements of **occlusal cavity depth (buccal side in the middle) of class II amalgam cavity** by using two different methods on two occasions (a and b) and between methods (c)

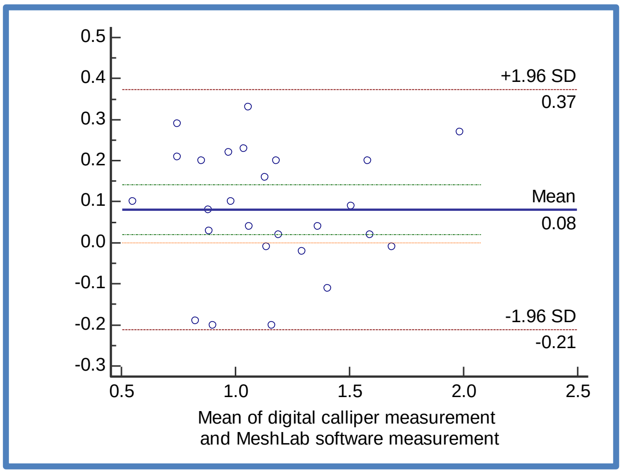
Figures illustrated that mean difference and limit of agreement for occlusal cavity depth (at distal side) of class II amalgam cavity by using two different methods on two occasions. Figures also demonstrated that the first measurement values from the digital calliper and MeshLab software were higher than the second measurement values. In addition, measurements from the digital calliper were higher than measurements from MeshLab software. The limits of agreement from the three graphs were wide. This is due to the identification of repeatable landmarks for this measurement was difficult by using two methods. Because of that, there was a systematic difference between methods.



a. Digital calliper 1st and 2nd measurements



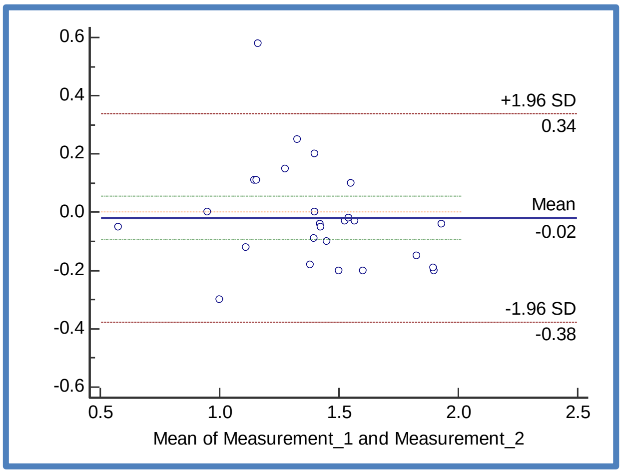
b. MeshLab software 1st and 2nd measurements



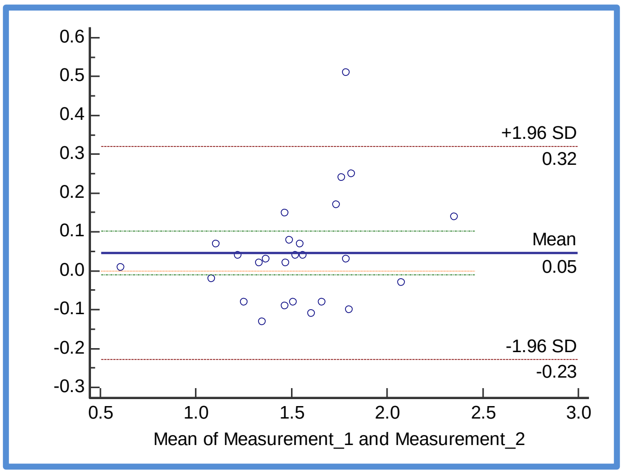
c. Digital calliper and MeshLab software measurements

Bland and Altman plot of differences and mean measurements of **occlusal cavity depth (at distal side) of class II amalgam cavity** by using two different methods on two occasions (a and b) and between methods (c)

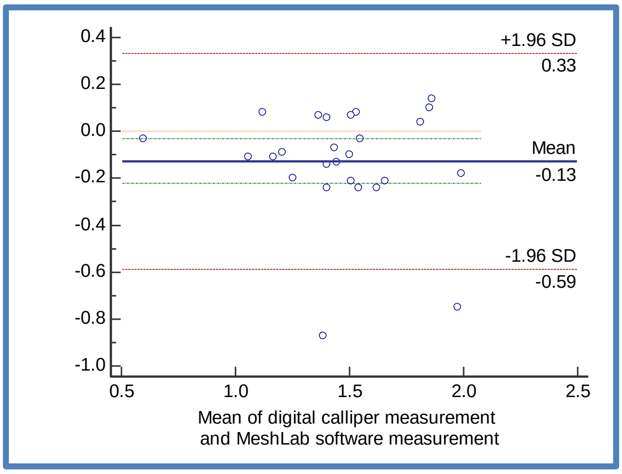
Figures illustrated that mean difference and limit of agreement for marginal ridge thickness of class II amalgam cavity by using two different methods on two occasions. Figures demonstrated that the measuring methods had wide limits of agreement for this feature. The mean difference was (-0.02, 0.05, and -0.13) for measurements using the digital calliper, MeshLab software and between two methods respectively. Because of the identification of repeatable landmarks for this measurement was also difficult by using MeshLab software, a systematic difference was appeared between methods in the Figure (c).



a. Digital calliper 1st and 2nd measurements



b. MeshLab software 1st and 2nd measurements

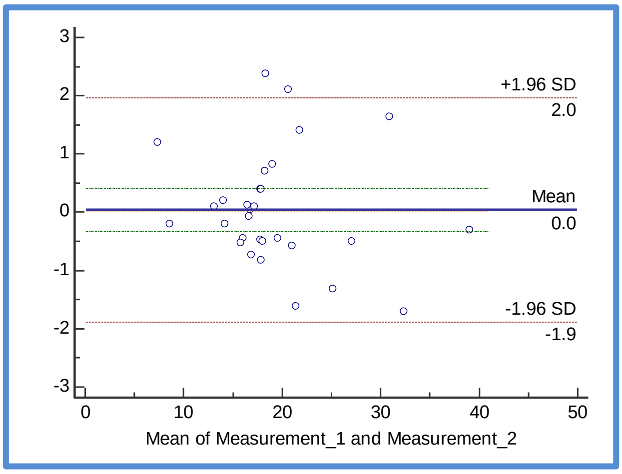


c. Digital calliper and MeshLab software measurements

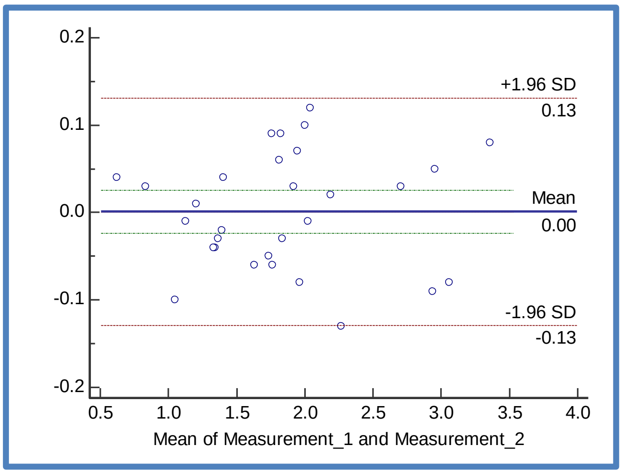
Bland and Altman plot of differences and mean measurements of **marginal ridge thickness of class II amalgam cavity** by using two different methods on two occasions (a and b) and between methods (c)

APPENDIX 6: Bland and Altman plots of differences and mean measurements for SAFMs of full veneer gold shell crown preparations

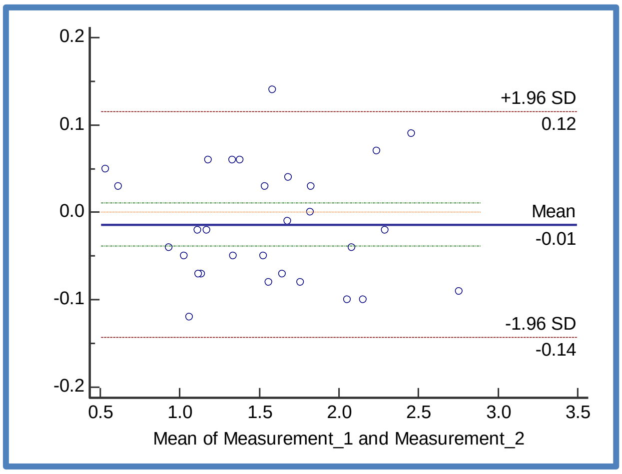
a. The following figures demonstrated Bland and Altman plots of specific anatomical feature measurements for 30 Full veneer gold shell crown preparation from the *buccal view* on two occasions by using ImageJ software:



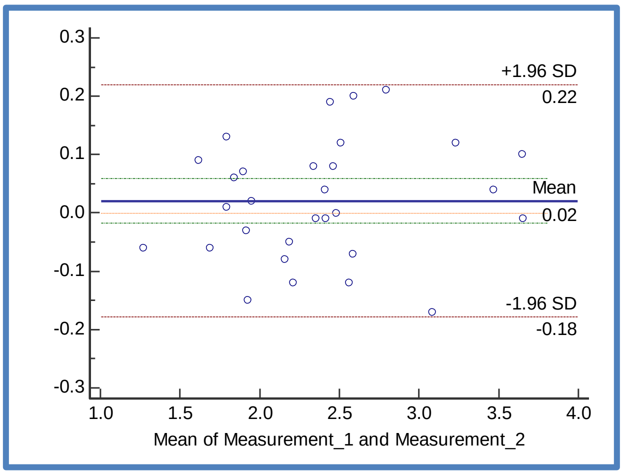
Measurements, using ImageJ software, of **angle of total occlusal convergence** (TOC)



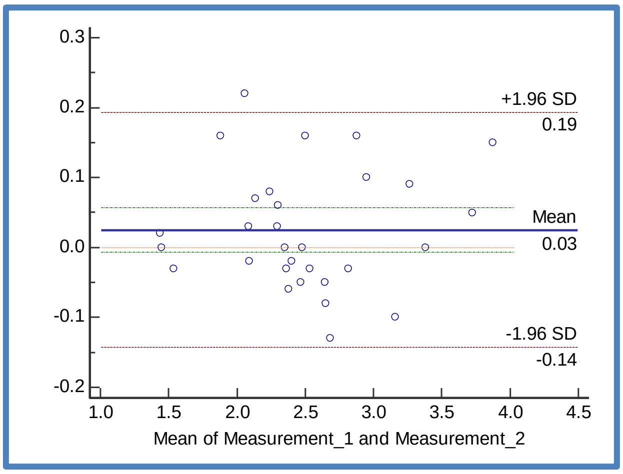
Measurements, using ImageJ software, of **occlusal reduction from mesial side**



Measurements, using ImageJ software, of **occlusal reduction from distal side**

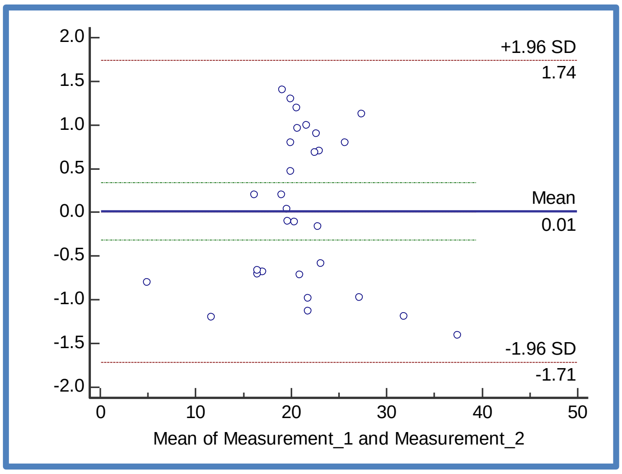


Measurements, using ImageJ software, of **axial reduction from mesial side**

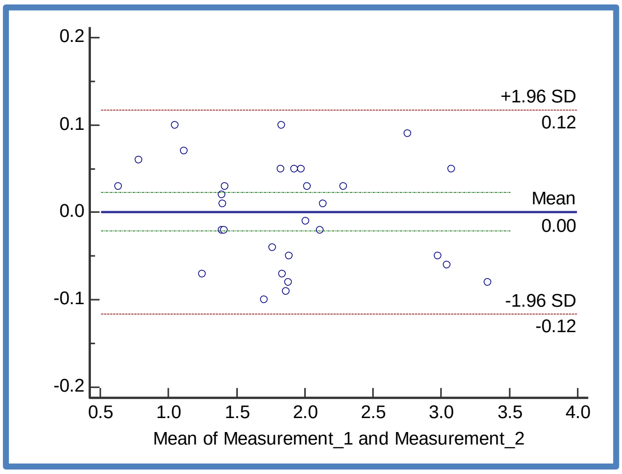


Measurements, using ImageJ software, of **axial reduction from distal side**

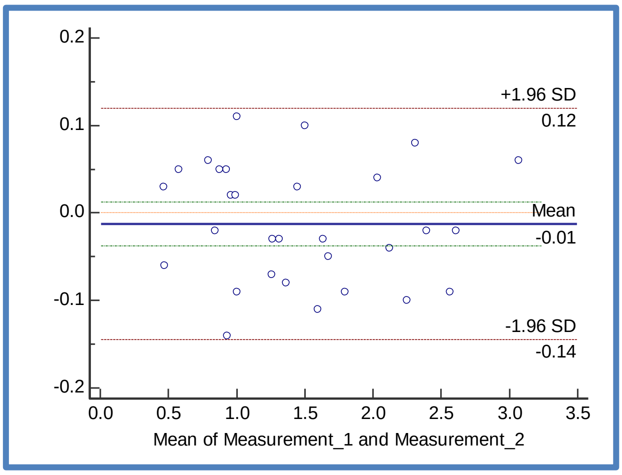
b. The following figures showed Bland and Altman plots of specific anatomical feature measurements for 30 Full veneer gold shell crown preparation from the *mesial view* on two occasions by using ImageJ software:



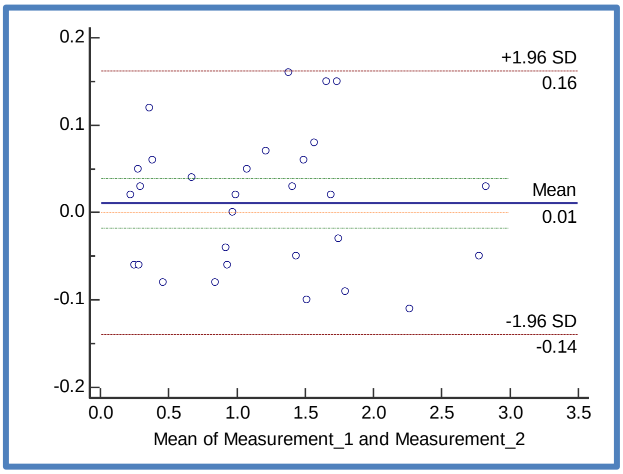
Measurements, using ImageJ software, of **the angle of total occlusal convergence** (TOC)



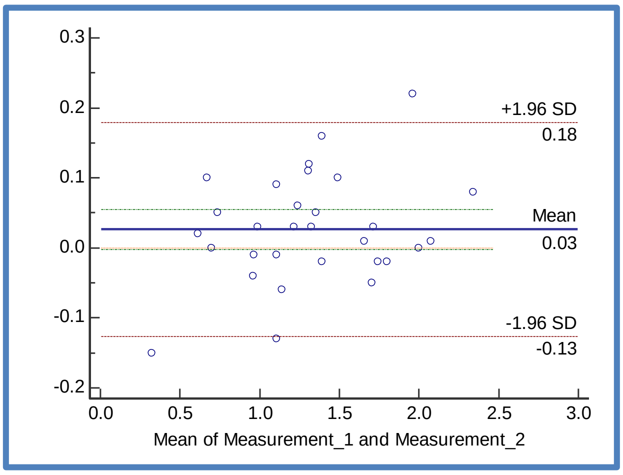
Measurements, using ImageJ software, of **occlusal reduction from buccal side**



Measurements, using ImageJ software, of **occlusal reduction from palatal side**

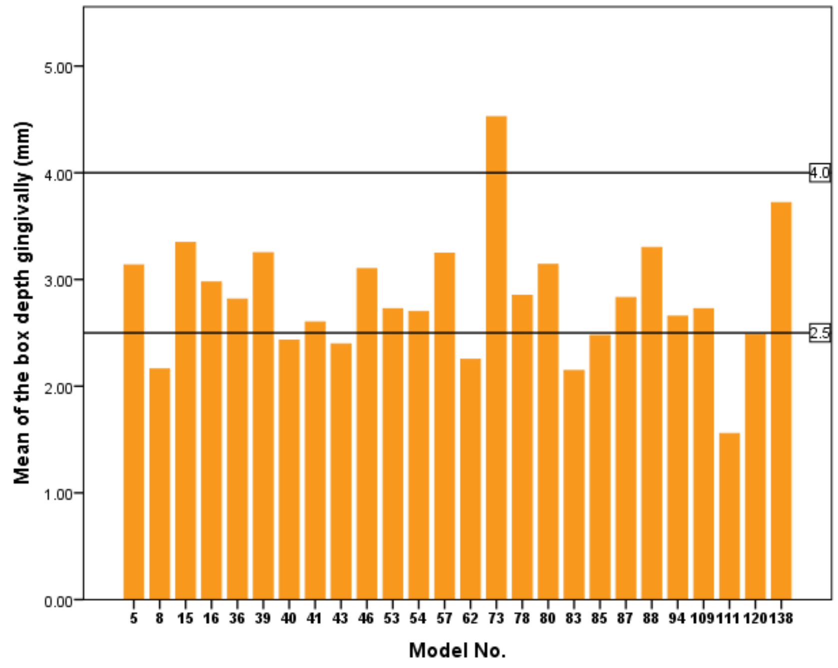


Measurements, using ImageJ software, of **axial reduction from buccal side**

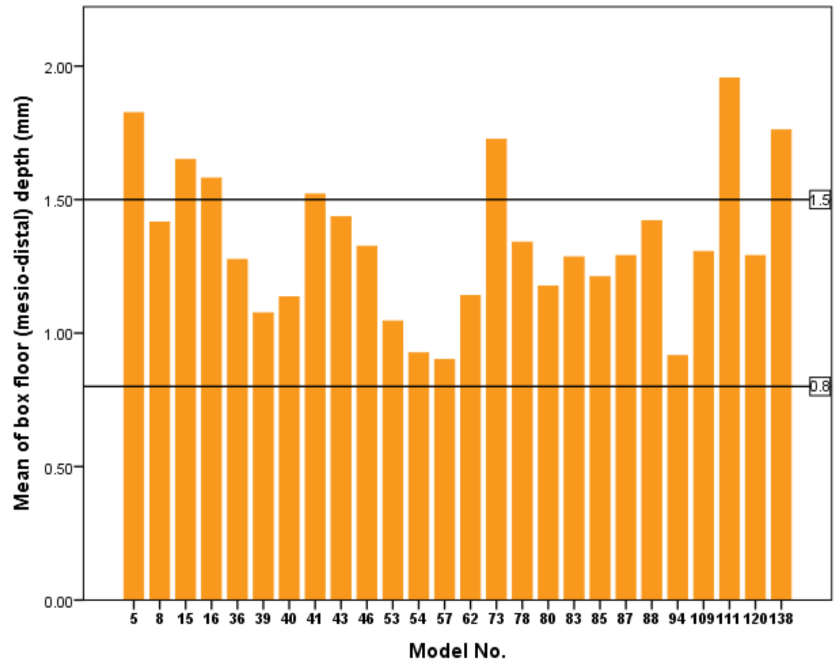


Measurements, using ImageJ software, of **axial reduction from palatal side**

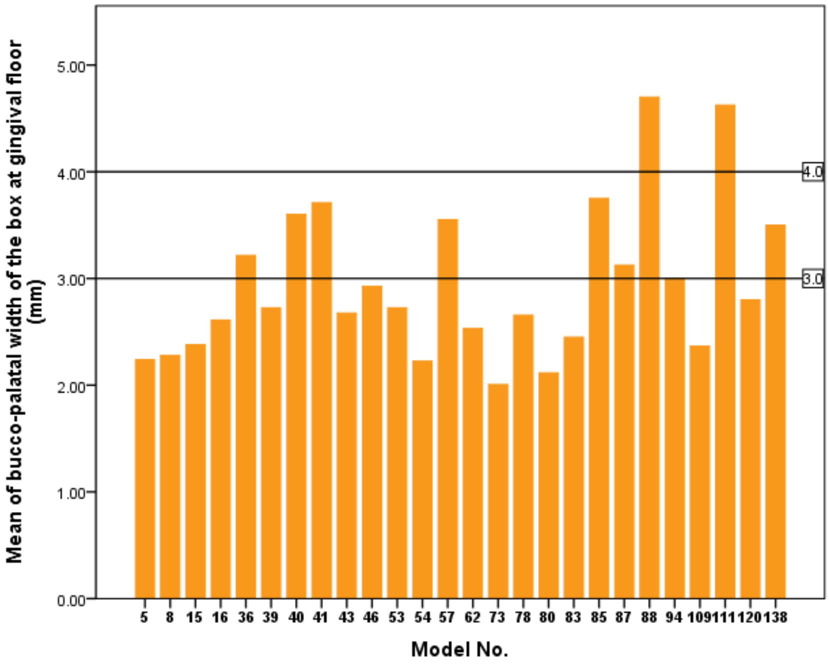
APPENDIX 7: Mean for measurements (mm) of SAFMs for each class II amalgam cavity preparation compared with recommended measurements in literature



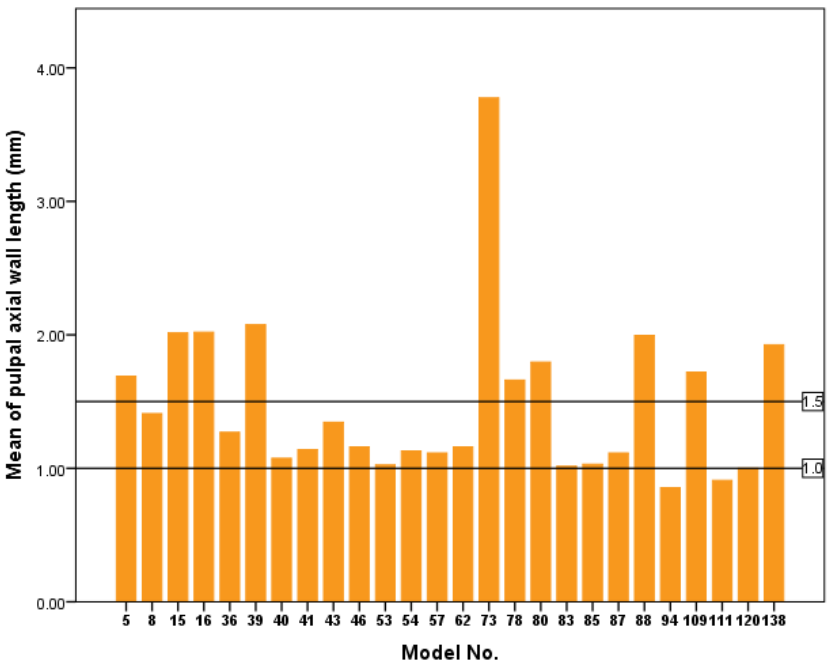
Mean for measurements (mm) of **the box depth gingivally for each class II amalgam cavity preparation** and compared to recommended measurements in literature. The horizontal lines at 2.5 and 4.0 mm represent the range of acceptable box depths reported in the literature



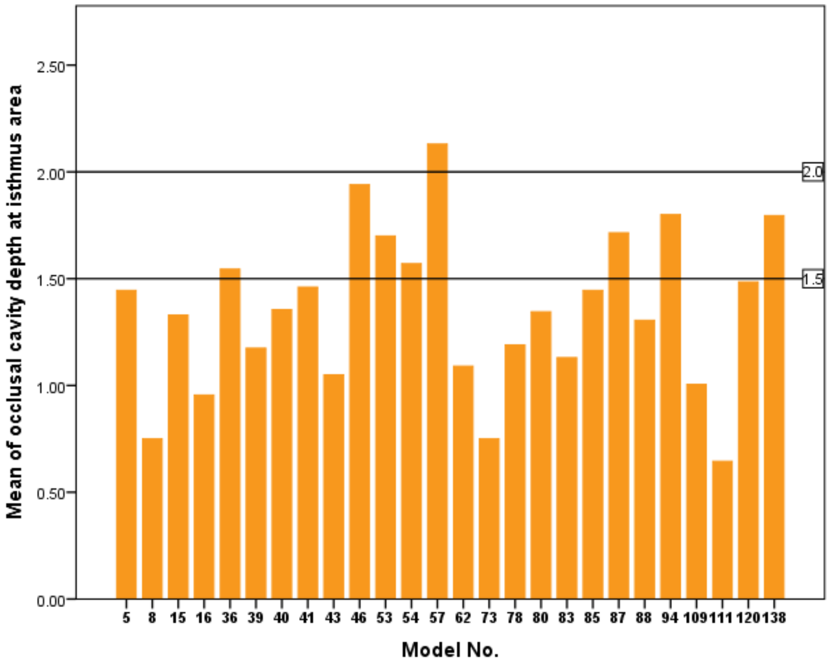
Mean for measurements (mm) of **the box floor (mesio-distal) depth for each class II amalgam cavity preparation** and compared to recommended measurements in literature. The horizontal lines at 0.8 and 1.5 mm represent the range of acceptable box floor (mesio-distal) depth reported in the literature



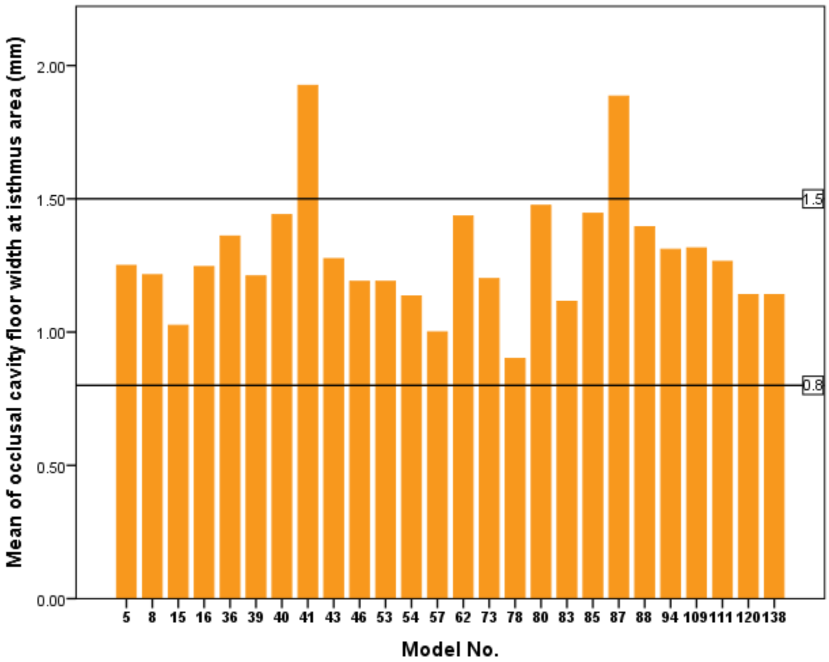
Mean for measurements (mm) of **bucco-palatal width of the box at gingival floor for each class II amalgam cavity preparation** and compared to recommended measurements in literature. The horizontal lines at 3.0 and 4.0 mm represent the range of acceptable bucco-palatal width of the box at gingival floor reported in the literature



Mean for measurements (mm) of **the pulpal axial wall length for each class II amalgam cavity preparation** and compared to recommended measurements in literature. The horizontal lines at 1.0 and 1.5 mm represent the range of acceptable box depths reported in the literature



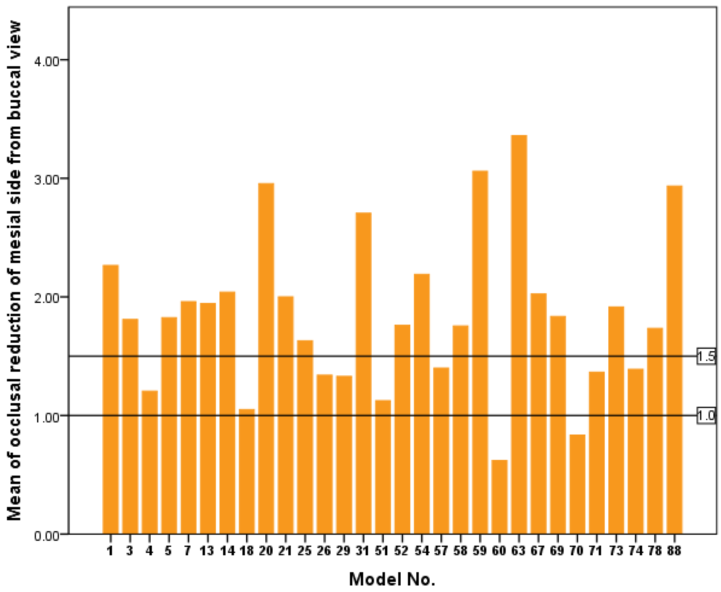
Mean for measurements (mm) of **the occlusal cavity depth at isthmus area for each class II amalgam cavity preparation** and compared to recommended measurements in literature. The horizontal lines at 1.5 and 2.0 mm represent the range of acceptable occlusal cavity depth at isthmus area reported in the literature



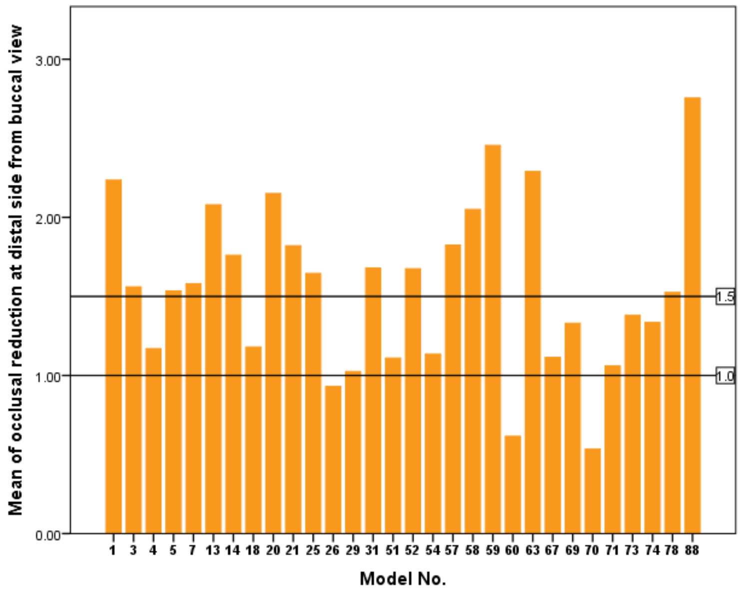
Mean for measurements (mm) of **the occlusal cavity floor width at isthmus area for each class II amalgam cavity preparation** and compared to recommended measurements in literature. The horizontal lines at 0.8 and 1.5 mm represent the range of acceptable occlusal floor width at isthmus area reported in the literature

APPENDIX 8: Mean for measurements (mm) of SAFMs for each full veneer gold shell crown preparation compared with recommended measurements in the literature.

a.



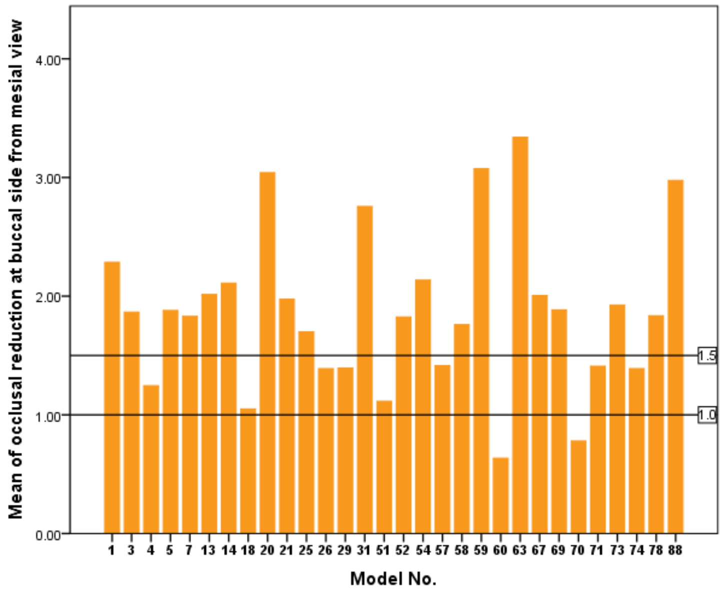
b.



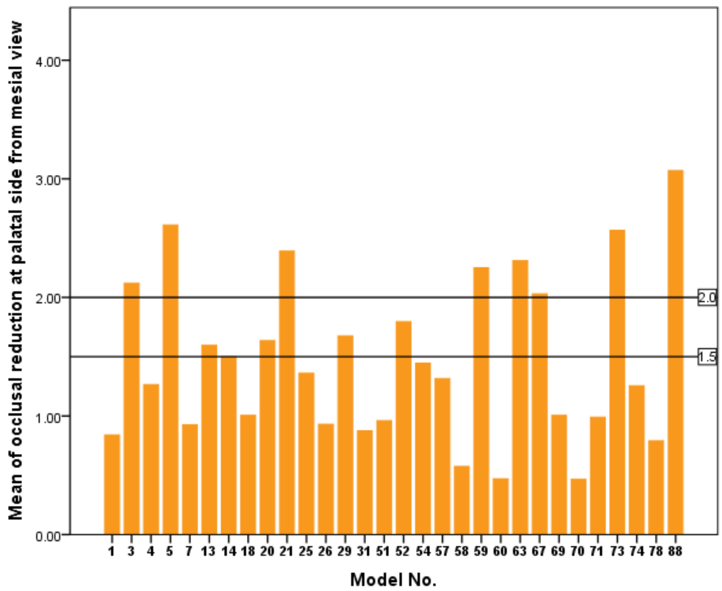
Mean for **a**. measurements (mm) of the occlusal reduction at the mesial side from buccal view (Bucco-mesial cusp) and **b.** measurements (mm) of the occlusal reduction at the distal side from buccal view (Bucco-distal cusp), using ImageJ software for each full veneer gold shell crown preparation and compared to recommended measurements in the literature. The horizontal lines at 1.0 and 1.5 mm represent the range of acceptable occlusal reduction on the non-functional cusps reported in the literature

Figures showed the measurement of occlusal reduction at mesial and distal side of non-functional (buccal) cusps of upper first molar teeth. These figures demonstrated that the occlusal reduction on the mesial side (bucco-mesial cusp) was more than on the distal side (bucco-distal cusp). The number of models which lie within recommended measurement on the bucco-mesial cusp was less than for the bucco-distal cusp. There were only 6 models which had ideal and acceptable occlusal reduction on both cusps.

a.



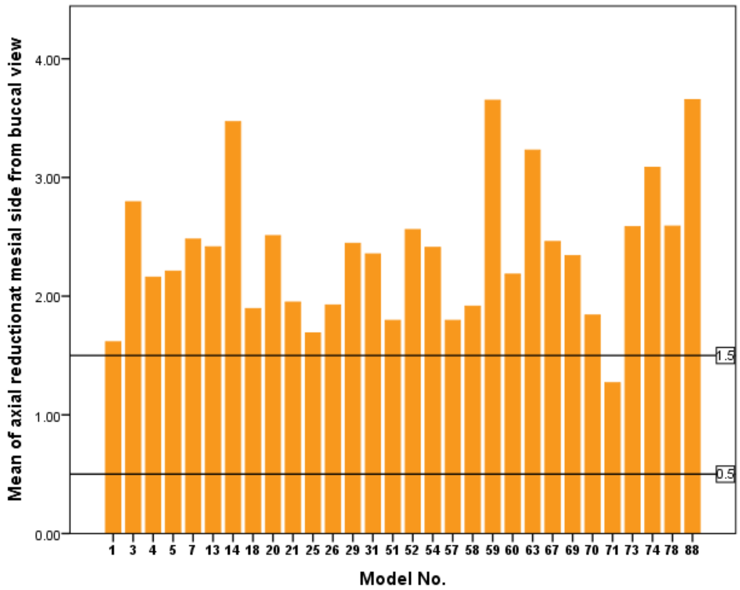
b.



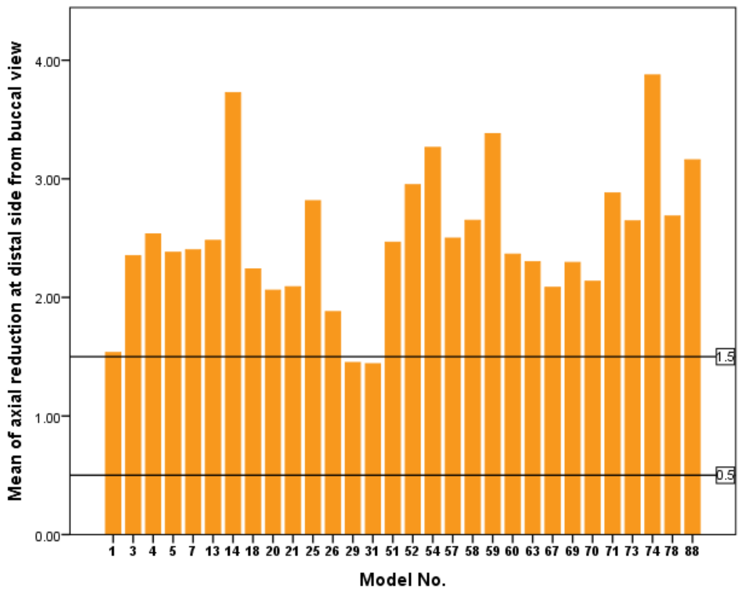
Mean for **a**. measurements (mm) of the occlusal reduction at the buccal side from mesial view (Mesio-buccal cusp) and **b.** measurements (mm) of the occlusal reduction at the palatal side from mesial view (Mesio-palatal cusp) for each full veneer gold shell crown preparation and compared to recommended measurements in the literature. The horizontal lines at (1.0 – 1.5 mm) and (1.5 – 2.0 mm) represent the range of acceptable occlusal reduction on the Mesio-buccal cusp and Mesio-palatal cusp respectively which reported in the literature

Non-functional (buccal) and functional (palatal) cusps were also measured from mesial view. Figures illustrated that the number of models which lie within the recommended measurement for both non-functional cusp (buccal cusps) and functional cusp (palatal cusps) from mesial view. 27% of the occlusal reduction at the buccal side from mesial view for full veneer gold shell crown preparations fell within the recommended range in the literature while 13% of the occlusal reduction at the palatal side from mesial view for full veneer gold shell crown preparations fell within the recommended range in the literature.

a.



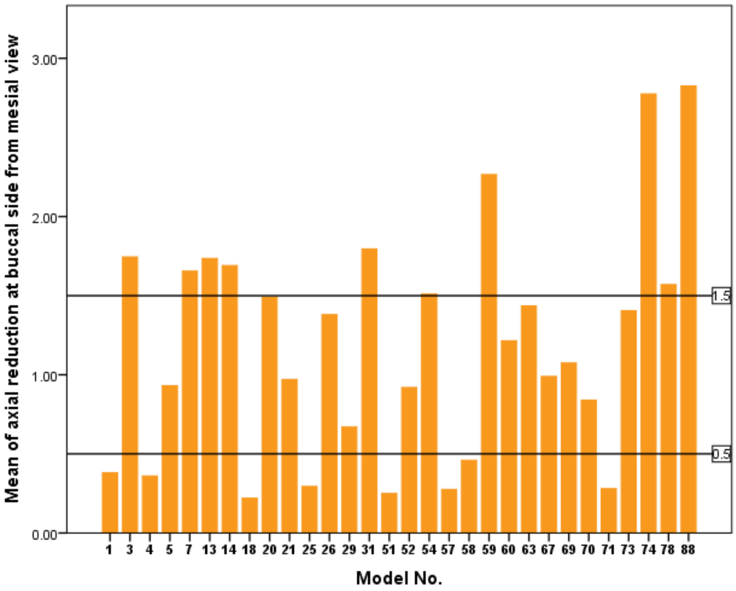
b.



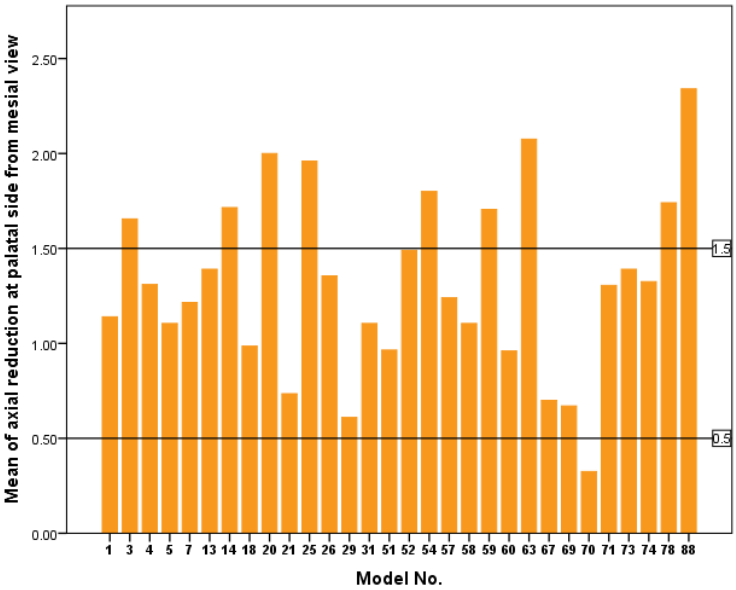
Mean for **a**. measurements (mm) of the axial reduction at the mesial side from buccal view and **b.** measurements (mm) of the axial reduction at the distal side from buccal view for each full veneer gold shell crown preparation and compared to recommended measurements in the literature. The horizontal lines at 0.5 and 1.5 mm represent the range of acceptable axial reduction reported in the literature

Figures showed the number of models which lie within recommended measurement of the axial reduction at mesial and distal sides. There was only one model which had recommended measurements for the axial reduction at the mesial side and two on the distal side. Achieving ideal or acceptable reduction on both sides was very difficult and no preparation achieved this. Most of the prepared teeth had been over-reduction on both sides.

a.



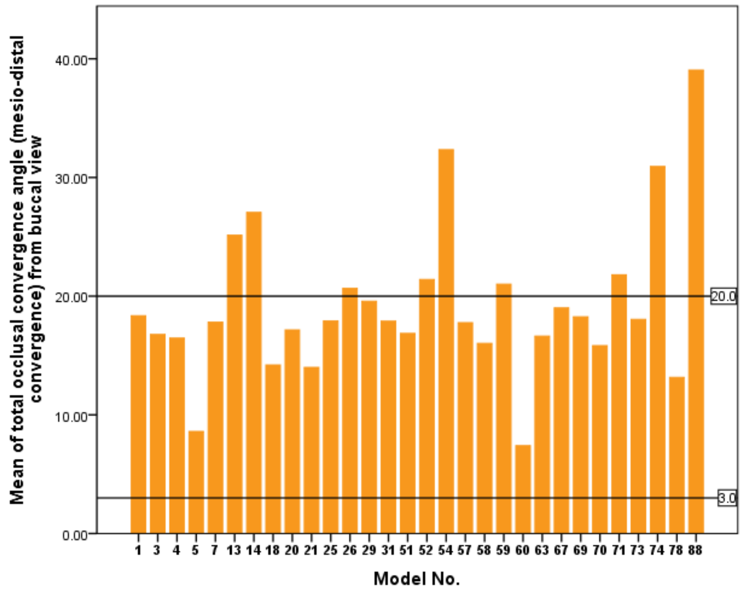
b.



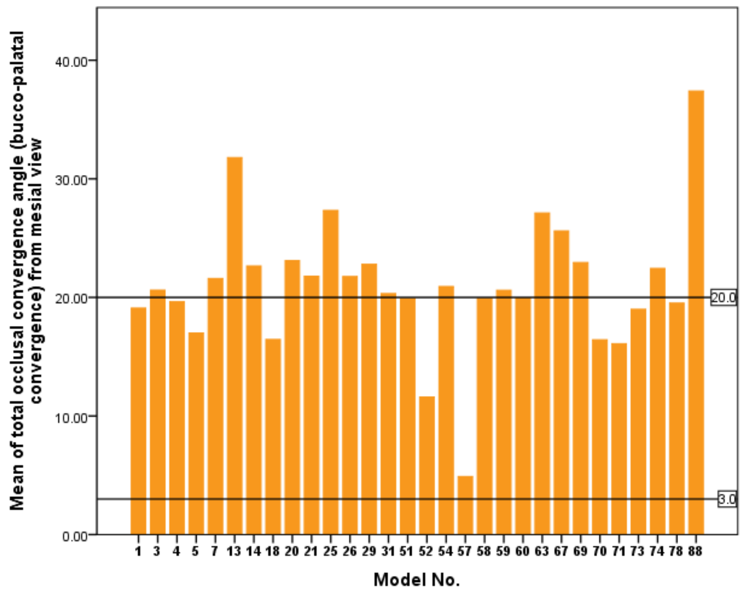
Mean for **a**. measurements (mm) of the axial reduction at the buccal side from mesial view and **b.** measurements (mm) of the axial reduction at the palatal side from mesial view for each full veneer gold shell crown preparation and compared to recommended measurements in the literature. The horizontal lines at 0.5 and 1.5 mm represent the range of acceptable axial reduction reported in the literature

Figures demonstrated the number of models which lie within the recommended measurement for axial reduction of the buccal and palatal sides. 40% of the axial reduction at the buccal side from mesial view for full veneer gold shell crown preparations fell within the recommended range in the literature while 67% of the axial reduction at the palatal side from mesial view for full veneer gold shell crown preparations fell within the recommended range in the literature.

a.



b.



Mean for **a**. total occlusal convergence angle from buccal view and **b.** total occlusal convergence angle from mesial view for each full veneer gold shell crown preparation and compared to recommended angle in the literature. The horizontal lines at 3° and 20° convergence angles represent the range of total occlusal convergence angles reported in the literature.

Figures demonstrated the total occlusal convergence from buccal and mesial sides. 70% of total occlusal convergence angle from buccal view for full veneer gold shell crown preparations fell within the recommended range in the literature while 43% of total occlusal convergence angle from mesial view for full veneer gold shell crown preparations fell within the recommended range in the literature.

APPENDIX 9: Steps to determine the grades of the best examiner and whether or not they can be selected as the Developed Standard grades/scores at Dundee Dental School

Steps to determine the best examiner grades:

*Identify '**The best examiner grades**' according to reliability*

1. Intra-examiner agreement (Kappa) according to **Grades** of feedback sheet
2. Inter-examiner agreement (ICC) according to **Grades** of feedback sheet
3. Intra-examiner agreement (Kappa) according to **the number of negative points**
4. Inter-examiner agreement (ICC) according to **the number of negative points**
5. Correlation between **Grades** and **the number of negative points** for all examiners (Spearman Correlation) *on round one, two and both round one and two together.*

*How do these **Grades** of' **the best examiner'** relate to 'what is generally considered correct'?*
(e.g. do **grades 4 and 5** of '**the best examiner'** represent the ideal and acceptable dimensions of tooth preparation?)

What are the dimensions of tooth preparation can be measured easily?
Evaluation of tooth preparation:

- Easier to measure is **'features can be assessed objectively' (SAFMs)**
- Harder to measure or cannot measure is **'features can be assessed subjectively' (SAFs)**

*What is measurable for a tooth preparation **(SAFM)**?* • Width and Depth (mm) • Angulation (˚) • Volume (mm^3) • Roughness • Other	*What is non-measurable feature for a tooth preparation **(SAF)**?* • Retention form • Clearance • Damage • Other
What researcher can measure (SAFM) or (the capacity of researcher to measure)? This capacity comes from: • **Feedback sheet** • **Literature** Researcher can measures specific anatomical feature (SAFM) of tooth preparation	*What researcher can evaluate (SAF) or (the capacity of researcher to evaluate)?* This capacity comes from: • **Feedback sheet** • **Literature** Researcher can evaluates specific anatomical feature (SAF) of tooth preparation
What is available and reliable method(s) can be used to measure all tooth preparations (SAFM)? • **Digital callipers** (twice measured x2) • **MeshLab software** (twice measured x2) • **ImageJ software** (twice measures x2) • **etc…** Find reliable SAFMs by using sample paired t-test, (ICC) and Bland and Altman tests.	*What is available method(s) can be used to evaluate all tooth preparations (SAF)?* • **Subjectively,** the best agreed criterion (exhibiting the best agreement) of SAF among the examiners is selected

*Can researcher link '**what is generally considered correct**' to **grades** awarded by '**the best examiner'**?*

1. Create graphs to compare specific anatomical feature measurements (SAFMs) of **researcher** and measurements from **literature** to determine which feature is passing and which one is failing
2. Compare the highest agreement criterion for each specific anatomical feature (SAF) among the examiners with acceptable feature from literature to determine which feature is passing and which one is failing
3. Convert all **SAFMs** and **SAFs** of tooth preparation to **Final Pass/Fail scores** and compare them according to Knight's recommendation
4. Convert **the best examiner grades** into **Pass/Fail scores** (i.e. grades 1, 2, 3 =Fail and grades 4, 5= Pass)
5. Compare the **Final scores** from SAFMs and SAFs with **the best examiner scores** (agreement)

Finally,** answer this question; Are grades of '**the best examiner**' the gold standard grades for your sample?   Answer: **NO

APPENDIX 10: The grades and converted scores of specific additional tools for each of the three examiners on two occasions and the Developed Standard scores from SAFMs and SAFs evaluations for the class II amalgam cavities

Examiner 1									
Model number	**Condenser**				**Bur**				**Developed Standard scores**
	Occasion 1		Occasion 2		Occasion 1		Occasion 2		
	Grade	Score	Grade	Score	Grade	Score	Grade	Score	
5	2	Fail	2	Fail	2	Fail	2	Fail	**Fail**
8	3	Fail	3	Fail	3	Fail	3	Fail	**Fail**
15	4	Pass	4	Pass	4	Pass	4	Pass	**Fail**
16	4	Pass	3	Fail	4	Pass	3	Fail	**Fail**
36	5	Pass	5	Pass	5	Pass	5	Pass	**Pass**
39	4	Pass	4	Pass	4	Pass	4	Pass	**Fail**
40	3	Fail	3	Fail	4	Pass	3	Fail	**Fail**
41	3	Fail	4	Pass	4	Pass	4	Pass	**Fail**
43	3	Fail	4	Pass	3	Fail	3	Fail	**Fail**
46	5	Pass	5	Pass	5	Pass	5	Pass	**Fail**
53	3	Fail	4	Pass	3	Fail	4	Pass	**Fail**
54	3	Fail	5	Pass	3	Fail	5	Pass	**Fail**
57	2	Fail	2	Fail	2	Fail	2	Fail	**Fail**
62	3	Fail	3	Fail	3	Fail	3	Fail	**Fail**
73	2	Fail	2	Fail	2	Fail	2	Fail	**Fail**
78	3	Fail	3	Fail	3	Fail	3	Fail	**Fail**
80	3	Fail	4	Pass	3	Fail	4	Pass	**Fail**
83	4	Pass	4	Pass	3	Fail	4	Pass	**Fail**
85	3	Fail	3	Fail	3	Fail	2	Fail	**Fail**
87	4	Pass	5	Pass	4	Pass	5	Pass	**Fail**
88	2	Fail	3	Fail	2	Fail	2	Fail	**Fail**
94	2	Fail	2	Fail	2	Fail	2	Fail	**Fail**
109	3	Fail	4	Pass	3	Fail	4	Pass	**Fail**
111	3	Fail	2	Fail	3	Fail	2	Fail	**Fail**
120	5	Pass	4	Pass	5	Pass	4	Pass	**Fail**
138	2	Fail	2	Fail	2	Fail	2	Fail	**Fail**

Examiner 2									
Model number	**Condenser**				**Bur**				**Developed Standard scores**
	Occasion 1		Occasion 2		Occasion 1		Occasion 2		
	Grade	Score	Grade	Score	Grade	Score	Grade	Score	
5	4	Pass	3	Fail	4	Pass	3	Fail	**Fail**
8	3	Fail	3	Fail	3	Fail	3	Fail	**Fail**
15	4	Pass	3	Fail	4	Pass	3	Fail	**Fail**
16	5	Pass	4	Pass	5	Pass	4	Pass	**Fail**
36	5	Pass	5	Pass	5	Pass	5	Pass	**Pass**
39	3	Fail	4	Pass	4	Pass	4	Pass	**Fail**
40	3	Fail	3	Fail	3	Fail	4	Pass	**Fail**
41	3	Fail	3	Fail	3	Fail	3	Fail	**Fail**
43	3	Fail	3	Fail	3	Fail	3	Fail	**Fail**
46	4	Pass	5	Pass	4	Pass	5	Pass	**Fail**
53	3	Fail	3	Fail	3	Fail	3	Fail	**Fail**
54	3	Fail	3	Fail	3	Fail	3	Fail	**Fail**
57	3	Fail	4	Pass	4	Pass	4	Pass	**Fail**
62	3	Fail	3	Fail	3	Fail	3	Fail	**Fail**
73	2	Fail	2	Fail	2	Fail	2	Fail	**Fail**
78	3	Fail	3	Fail	3	Fail	3	Fail	**Fail**
80	3	Fail	4	Pass	3	Fail	4	Fail	**Fail**
83	3	Fail	4	Pass	3	Fail	4	Fail	**Fail**
85	3	Fail	3	Fail	3	Fail	3	Fail	**Fail**
87	4	Pass	5	Pass	4	Pass	5	Pass	**Fail**
88	3	Fail	3	Fail	3	Fail	3	Fail	**Fail**
94	3	Fail	3	Fail	3	Fail	3	Fail	**Fail**
109	3	Fail	3	Fail	3	Fail	3	Fail	**Fail**
111	3	Fail	3	Fail	3	Fail	3	Fail	**Fail**
120	5	Pass	4	Pass	5	Pass	4	Pass	**Fail**
138	2	Fail	2	Fail	2	Fail	2	Fail	**Fail**

Examiner 3									
Model number	**Condenser**				**Bur**				**Developed Standard scores**
	Occasion 1		Occasion 2		Occasion 1		Occasion 2		
	Grade	Score	Grade	Score	Grade	Score	Grade	Score	
5	3	Fail	3	Fail	3	Fail	2	Fail	**Fail**
8	3	Fail	3	Fail	2	Fail	3	Fail	**Fail**
15	3	Fail	4	Pass	4	Pass	4	Pass	**Fail**
16	2	Fail	3	Fail	2	Fail	3	Fail	**Fail**
36	5	Pass	5	Pass	5	Pass	5	Pass	**Pass**
39	5	Pass	5	Pass	5	Pass	5	Pass	**Fail**
40	5	Pass	5	Pass	4	Pass	4	Pass	**Fail**
41	3	Fail	3	Fail	3	Fail	3	Fail	**Fail**
43	3	Fail	3	Fail	3	Fail	3	Fail	**Fail**
46	3	Fail	5	Pass	3	Fail	5	Pass	**Fail**
53	3	Fail	3	Fail	3	Fail	3	Fail	**Fail**
54	4	Pass	4	Pass	3	Fail	4	Pass	**Fail**
57	2	Fail	3	Fail	2	Fail	3	Fail	**Fail**
62	3	Fail	3	Fail	3	Fail	3	Fail	**Fail**
73	2	Fail	2	Fail	2	Fail	2	Fail	**Fail**
78	3	Fail	4	Pass	3	Fail	4	Pass	**Fail**
80	4	Pass	4	Pass	4	Pass	4	Pass	**Fail**
83	4	Pass	5	Pass	4	Pass	5	Pass	**Fail**
85	2	Fail	3	Fail	2	Fail	3	Fail	**Fail**
87	3	Fail	4	Fail	4	Pass	4	Pass	**Fail**
88	2	Fail	2	Fail	2	Fail	2	Fail	**Fail**
94	2	Fail	2	Fail	2	Fail	2	Fail	**Fail**
109	4	Pass	5	Pass	4	Pass	5	Pass	**Fail**
111	2	Fail	2	Fail	2	Fail	2	Fail	**Fail**
120	4	Pass	4	Pass	4	Pass	4	Pass	**Fail**
138	2	Fail	2	Fail	2	Fail	2	Fail	**Fail**

APPENDIX 11: The grades and converted scores of specific additional tools for each of the three examiners on two occasions and the Developed Standard scores from SAFMs and SAFs evaluations for the full veneer gold shell crown preparations

Examiner 1									
Model number	**Bur**				**Impression index**				**Developed Standard scores**
	Occasion 1		Occasion 2		Occasion 1		Occasion 2		
	Grade	Score	Grade	Score	Grade	Score	Grade	Score	
1	4	Pass	4	Pass	3	Fail	4	Pass	**Fail**
3	5	Pass	3	Fail	5	Fail	4	Pass	**Fail**
4	3	Fail	3	Fail	3	Fail	3	Fail	**Fail**
5	3	Fail	4	Pass	3	Fail	4	Pass	**Fail**
7	4	Pass	4	Pass	4	Pass	5	Pass	**Fail**
13	3	Fail	3	Fail	4	Pass	3	Fail	**Fail**
14	5	Pass	5	Pass	5	Pass	5	Pass	**Fail**
18	3	Fail	3	Fail	3	Fail	3	Fail	**Fail**
20	5	Pass	5	Pass	5	Pass	5	Pass	**Fail**
21	5	Pass	5	Pass	4	Pass	5	Pass	**Fail**
25	4	Pass	4	Pass	4	Pass	4	Pass	**Fail**
26	2	Fail	2	Fail	2	Fail	2	Fail	**Fail**
29	3	Fail	4	Pass	3	Fail	3	Fail	**Fail**
31	5	Pass	5	Pass	4	Pass	4	Pass	**Fail**
51	5	Pass	5	Pass	4	Pass	4	Pass	**Fail**
52	4	Pass	2	Fail	2	Fail	2	Fail	**Fail**
54	3	Fail	3	Fail	4	Pass	3	Fail	**Fail**
57	5	Pass	4	Pass	4	Pass	4	Pass	**Fail**
58	3	Fail	3	Fail	3	Fail	3	Fail	**Fail**
59	2	Fail	2	Fail	2	Fail	2	Fail	**Fail**
60	4	Pass	3	Fail	3	Fail	4	Pass	**Fail**
63	3	Fail	3	Fail	4	Pass	3	Fail	**Fail**
67	4	Pass	5	Pass	4	Pass	5	Pass	**Fail**
69	5	Pass	4	Pass	5	Pass	5	Pass	**Fail**
70	4	Pass	4	Pass	4	Pass	4	Pass	**Fail**
71	5	Pass	4	Pass	3	Fail	5	Pass	**Fail**
73	4	Pass	3	Fail	4	Pass	3	Fail	**Fail**
74	4	Pass	3	Fail	4	Pass	3	Fail	**Fail**
78	5	Pass	5	Pass	4	Pass	5	Pass	**Fail**
88	3	Fail	3	Fail	3	Fail	3	Fail	**Fail**

Examiner 2									
Model number	**Bur**				**Impression index**				**Developed Standard scores**
	Occasion 1		Occasion 2		Occasion 1		Occasion 2		
	Grade	Score	Grade	Score	Grade	Score	Grade	Score	
1	4	Pass	3	Fail	3	Fail	3	Fail	**Fail**
3	4	Pass	4	Pass	4	Pass	4	Pass	**Fail**
4	4	Pass	3	Fail	4	Pass	3	Fail	**Fail**
5	3	Fail	4	Pass	3	Fail	4	Pass	**Fail**
7	5	Pass	4	Pass	4	Pass	4	Pass	**Fail**
13	3	Fail	3	Fail	4	Pass	3	Fail	**Fail**
14	4	Pass	4	Pass	4	Pass	4	Pass	**Fail**
18	3	Fail	3	Fail	3	Fail	3	Fail	**Fail**
20	5	Pass	4	Pass	4	Pass	4	Pass	**Fail**
21	4	Pass	4	Pass	4	Pass	4	Pass	**Fail**
25	5	Pass	5	Pass	4	Pass	4	Pass	**Fail**
26	3	Fail	3	Fail	3	Fail	3	Fail	**Fail**
29	4	Pass	4	Pass	4	Pass	4	Pass	**Fail**
31	4	Pass	4	Pass	4	Pass	4	Pass	**Fail**
51	5	Pass	4	Pass	4	Pass	4	Pass	**Fail**
52	4	Pass	4	Pass	4	Pass	4	Pass	**Fail**
54	3	Fail	3	Fail	4	Pass	3	Fail	**Fail**
57	5	Pass	5	Pass	4	Pass	4	Pass	**Fail**
58	3	Fail	4	Pass	3	Fail	3	Fail	**Fail**
59	3	Fail	2	Pass	3	Fail	2	Fail	**Fail**
60	3	Fail	3	Fail	3	Fail	3	Fail	**Fail**
63	4	Pass	4	Pass	4	Pass	4	Pass	**Fail**
67	4	Pass	3	Fail	4	Pass	4	Pass	**Fail**
69	4	Pass	4	Pass	4	Pass	4	Pass	**Fail**
70	3	Fail	4	Pass	3	Fail	4	Pass	**Fail**
71	4	Pass	5	Pass	4	Pass	4	Pass	**Fail**
73	4	Pass	4	Pass	4	Pass	4	Pass	**Fail**
74	4	Pass	4	Pass	4	Pass	4	Pass	**Fail**
78	4	Pass	4	Pass	4	Pass	4	Pass	**Fail**
88	3	Fail	3	Fail	3	Fail	3	Fail	**Fail**

Examiner 3									
Model number	**Bur**				**Impression index**				**Developed Standard scores**
	Occasion 1		Occasion 2		Occasion 1		Occasion 2		
	Grade	Score	Grade	Score	Grade	Score	Grade	Score	
1	3	Fail	3	Fail	3	Fail	3	Fail	**Fail**
3	2	Fail	3	Fail	3	Fail	3	Fail	**Fail**
4	2	Fail	2	Fail	2	Fail	2	Fail	**Fail**
5	2	Fail	3	Fail	3	Fail	3	Fail	**Fail**
7	4	Pass	5	Pass	3	Fail	4	Pass	**Fail**
13	2	Fail	3	Fail	2	Fail	3	Fail	**Fail**
14	4	Pass	5	Pass	3	Fail	4	Pass	**Fail**
18	3	Fail	3	Fail	3	Fail	3	Fail	**Fail**
20	4	Pass	5	Pass	4	Pass	4	Pass	**Fail**
21	3	Fail	3	Fail	3	Fail	3	Fail	**Fail**
25	3	Fail	4	Pass	3	Fail	3	Fail	**Fail**
26	2	Fail	3	Fail	2	Fail	3	Fail	**Fail**
29	3	Fail	3	Fail	3	Fail	3	Fail	**Fail**
31	2	Fail	3	Fail	2	Fail	3	Fail	**Fail**
51	3	Fail	4	Pass	3	Fail	3	Fail	**Fail**
52	3	Fail	4	Pass	3	Fail	4	Pass	**Fail**
54	2	Fail	3	Fail	3	Fail	3	Fail	**Fail**
57	4	Pass	4	Pass	3	Fail	3	Fail	**Fail**
58	3	Fail	2	Fail	3	Fail	2	Fail	**Fail**
59	3	Fail	2	Fail	2	Fail	2	Fail	**Fail**
60	3	Fail	4	Pass	3	Fail	4	Pass	**Fail**
63	3	Fail	4	Pass	3	Fail	4	Pass	**Fail**
67	3	Fail	3	Fail	3	Fail	3	Fail	**Fail**
69	4	Pass	3	Fail	3	Fail	3	Fail	**Fail**
70	3	Fail	3	Fail	3	Fail	3	Fail	**Fail**
71	5	Pass	4	Pass	3	Fail	4	Pass	**Fail**
73	4	Pass	4	Pass	3	Fail	4	Pass	**Fail**
74	4	Pass	4	Pass	3	Fail	4	Pass	**Fail**
78	3	Fail	3	Fail	3	Fail	3	Fail	**Fail**
88	2	Fail	3	Fail	2	Fail	3	Fail	**Fail**

APPENDIX 12: The Developed Standard scores from SAFMs and SAFs evaluations converted to Developed Standard grades from the grades (mode) of 'Gray feedback sheet' for each of the three examiners on two occasions for the class II amalgam cavities.

Model number	**Examiner 1**		**Examiner 2**		**Examiner 3**		**Developed Standard scores**	**Developed Standard grades** (mode)
	Gray feedback sheet		Gray feedback sheet		Gray feedback sheet			
	Occasion 1	Occasion 2	Occasion 1	Occasion 2	Occasion 1	Occasion 2		
5	2	2	~~4~~	3	3	2	Fail	**2**
8	3	3	3	3	2	3	Fail	**3**
15	~~4~~	~~4~~	~~4~~	3	~~4~~	~~4~~	Fail	**3**
16	~~4~~	~~4~~	~~5~~	~~4~~	2	3	Fail	**3**
36	5	5	5	5	5	5	Pass	**5**
39	~~5~~	~~5~~	~~4~~	~~4~~	~~5~~	~~5~~	Fail	**Excluded**
40	~~4~~	~~5~~	~~4~~	~~4~~	~~4~~	~~4~~	Fail	**Excluded**
41	~~5~~	~~5~~	3	3	3	3	Fail	**3**
43	~~5~~	3	3	3	3	3	Fail	**3**
46	~~5~~	~~5~~	~~4~~	~~5~~	~~5~~	~~5~~	Fail	**Excluded**
53	3	2	3	3	2	3	Fail	**3**
54	3	~~5~~	3	3	~~4~~	~~4~~	Fail	**3**
57	3	2	~~4~~	~~4~~	2	2	Fail	**2**
62	3	3	3	3	3	3	Fail	**3**
73	2	2	2	2	2	2	Fail	**2**
78	3	3	3	3	3	~~4~~	Fail	**3**
80	3	~~4~~	~~4~~	~~4~~	~~4~~	~~4~~	Fail	**3**
83	~~5~~	~~5~~	3	~~4~~	~~5~~	~~5~~	Fail	**3**
85	3	3	3	3	2	2	Fail	**3**
87	~~5~~	~~5~~	~~4~~	~~5~~	~~4~~	~~4~~	Fail	**Excluded**
88	2	2	3	3	2	2	Fail	**2**
94	2	2	3	3	2	2	Fail	**2**
109	3	~~5~~	3	3	~~5~~	~~5~~	Fail	**3**
111	2	2	3	3	2	2	Fail	**2**
120	~~5~~	~~4~~	~~5~~	~~5~~	~~4~~	~~4~~	Fail	**Excluded**
138	2	2	2	2	2	2	Fail	**2**

The grade in like this cell was excluded because it is not representative the Developed Standard score.

APPENDIX 13: The Developed Standard scores from SAFMs and SAFs evaluations converted to Developed Standard grades from the grades (mode) of 'Mhanni feedback sheet' for each of the three examiners on two occasions for the full veneer gold shell crown preparations

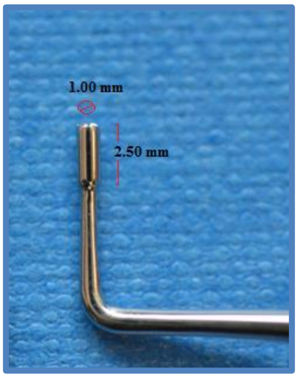
Model number	Examiner 1		Examiner 2		Examiner 3		Developed Standard scores	Developed Standard grades (mode)
	Mhanni feedback sheet		Mhanni feedback sheet		Mhanni feedback sheet			
	Occasion 1	Occasion 2	Occasion 1	Occasion 2	Occasion 1	Occasion 2		
1	3	~~4~~	3	3	3	3	Fail	**3**
3	~~4~~	~~4~~	3	3	3	~~4~~	Fail	**3**
4	3	3	3	3	2	2	Fail	**3**
5	3	~~5~~	3	~~4~~	3	3	Fail	**3**
7	~~4~~	~~5~~	~~4~~	~~4~~	3	~~4~~	Fail	**3**
13	3	3	~~4~~	3	2	~~4~~	Fail	**3**
14	~~5~~	~~5~~	~~4~~	~~4~~	3	3	Fail	**3**
18	3	3	3	3	3	3	Fail	**3**
20	~~5~~	~~5~~	3	~~4~~	~~4~~	3	Fail	**3**
21	~~5~~	~~5~~	~~4~~	~~4~~	3	3	Fail	**3**
25	3	~~4~~	~~4~~	~~5~~	3	3	Fail	**3**
26	2	2	3	2	2	2	Fail	**2**
29	3	3	~~4~~	~~4~~	3	3	Fail	**3**
31	~~5~~	~~4~~	~~4~~	~~4~~	2	3	Fail	**3**
51	~~5~~	~~5~~	~~4~~	~~4~~	3	3	Fail	**3**
52	~~4~~	3	3	~~4~~	3	~~4~~	Fail	**3**
54	3	3	3	3	3	3	Fail	**3**
57	~~5~~	~~5~~	~~4~~	~~5~~	3	3	Fail	**3**
58	3	3	3	3	3	2	Fail	**3**
59	2	2	2	2	2	2	Fail	**2**
60	3	3	3	3	3	3	Fail	**3**
63	3	3	3	~~4~~	3	~~4~~	Fail	**3**
67	~~4~~	~~5~~	~~4~~	~~4~~	3	3	Fail	**3**
69	~~4~~	~~4~~	~~4~~	~~4~~	3	3	Fail	**3**
70	3	3	3	~~4~~	3	3	Fail	**3**
71	~~5~~	~~5~~	~~4~~	~~5~~	3	3	Fail	**3**
73	~~4~~	3	~~4~~	~~4~~	3	3	Fail	**3**
74	3	~~4~~	3	~~4~~	3	~~4~~	Fail	**3**
78	~~5~~	~~5~~	~~4~~	~~4~~	3	3	Fail	**3**
88	2	2	2	2	2	3	Fail	**2**

[crossed cell] The grade in like this cell was excluded because it is not representative the Developed Standard score.

APPENDIX 14: Instructions for how to use new the new checklist for class II amalgam cavity preparation with tools

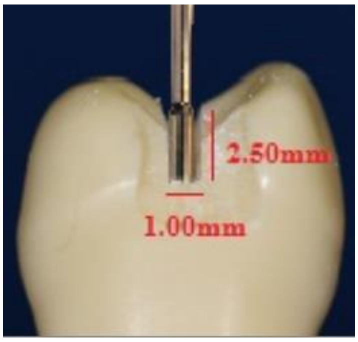
<u>Instructions for "nCIIpc"</u>

This **tool** is used with "**nCIIpc**":

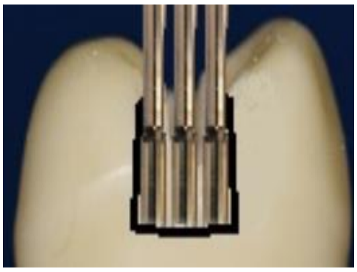


Amalgam condenser

- **The cavity occlusal floor** is at the correct **depth (at isthmus area)** - which was defined as 1.50 to 2.00 mm – i.e. more than half of the length of the amalgam condenser.



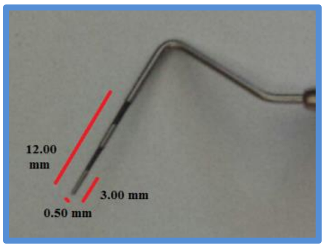
- **The cavity occlusal floor** is at the correct **width (at isthmus area)** - which was defined as 1.50 to 2.00 mm – i.e. one and half or two thickness (dimeters) of the amalgam condenser.
- **The cavity occlusal walls** (buccal wall and palatal or lingual wall of the cavity) are **converged or parallel.**
- **The marginal ridge** is at correct **thickness** – which was defined as 1.00 mm or more – i.e. more than one thickness (dimeter) of the amalgam condenser.
- **The proximal box floor** is at the correct level **with the contact point** - which was defined as 2.50 to 4.00 mm – i.e. the whole length to more than length of the amalgam condenser.



- **The proximal box floor width** is at the correct width - which was defined as 3.00 to 4.00 mm – i.e. three to four thickness or dimeters of the amalgam condenser.
- **The proximal box floor depth** is at the correct width - which was defined as 1.00 to 1.50 mm – i.e. one dimeter to one and half thickness (dimeter) of the amalgam condenser.
- **The proximal box walls** (buccal wall and palatal or lingual wall of the cavity) are converged or parallel.
- For **damage to adjacent tooth**:
  - No damage visible to the naked eye or under magnifying glass or a sound surface with regular curved proximal surface without loss of contour. **(None)**
  - Slight damage visible to the naked eye and identifiable with a magnifying glass; *<u>Scratches:</u>* narrow, shallow score-lines, usually multiple with a consistent orientation. *<u>Indentation:</u>* a regular defect without an orientation, roughly circular or irregular in shape. **(Minor)**
  - Obvious damage Extensive: damage involving a large area of the proximal surface. **(Moderate or severe)**

APPENDIX 15: Instructions for how to use the new checklist for full veneer gold shell crown preparation with tools.

Instructions for new Gold Shell Crown preparation checklist "nGSCpc"

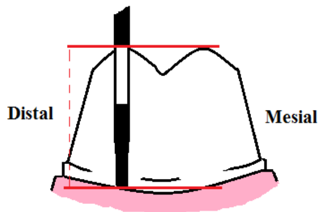
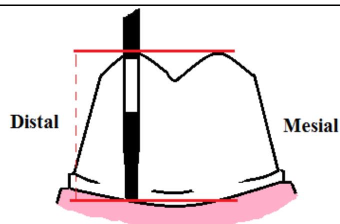


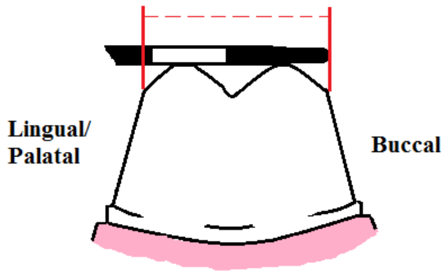
Periodontal probe

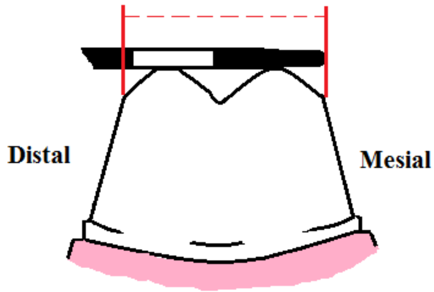
Rround-ended, chamfer, diamond bur

Impression index

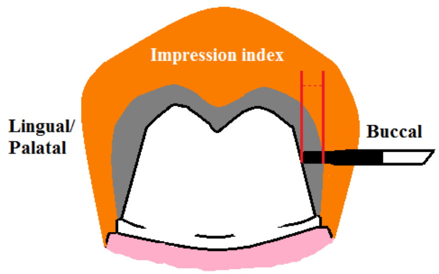
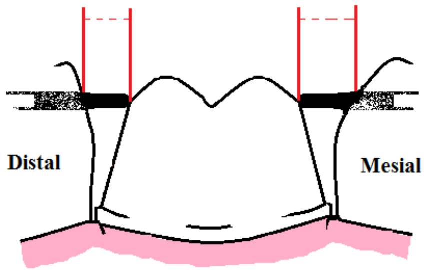
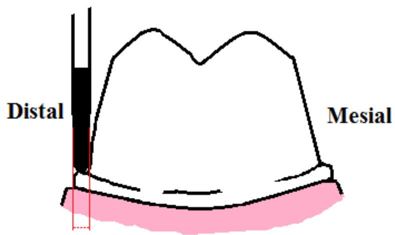
1. Objective features

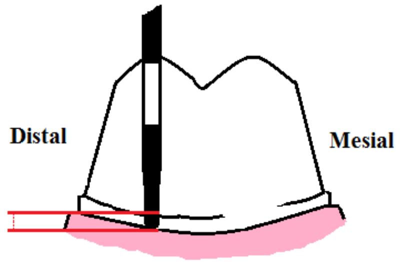
1.1 Occlusal reduction	
1.1.1 Functional cusps	
Tools	Pick up periodontal probe CP12 or round-ended, chamfer, diamond bur
Measure	From gingival margin to both functional cusps 
Criteria	○ Underprepared (>6.50 mm) ○ Adequate with symmetrical bevel (6.00 – 6.50 mm) ○ Acceptable without symmetrical bevel (6.00 – 6.50 mm) ○ Overprepared (<6.00 mm) **Functional cusps reduction is** **underprepared (>6.50 mm)** **adequate with symmetrical bevel (6.00 - 6.50 mm)** **acceptable with non-symmetrical bevel (6.00 - 6.50 mm)** **overprepared (<6.00 mm)**
Record	The worst result (Overprepared > Underprepared > Adequate without symmetrical bevel > Adequate with symmetrical bevel) The worst ---→ The best
1.1.2 Non-functional cusps	
Tools	Pick up periodontal probe CP12 or round-ended, chamfer, diamond bur
Measure	From gingival margin to both functional cusps 
Criteria	○ Underprepared (>7.00 mm) ○ Adequate (6.50 – 7.00 mm) ○ Overprepared (<6.50 mm) **Non-functional cusps reduction is** **underprepared (>7.00 mm)** **adequate (6.50 - 7.00 mm)** **overprepared (<6.50 mm)**
Record	The worst result (Overprepared > Underprepared > Adequate) The worst ---------------------→ The best

<table>
<tr><th colspan="3">1.2 Total occlusal convergences</th></tr>
<tr><td colspan="3">1.2.1 Bucco-lingual/palatal convergence</td></tr>
<tr><td>Tools</td><td colspan="2">Pick up periodontal probe CP12 or round-ended, chamfer, diamond bur</td></tr>
<tr><td>Measure</td><td>The widest bucco-lingual/palatal aspect of occlusal part of preparation (marginal ridge)</td><td>
</td></tr>
<tr><td>Criteria</td><td colspan="2">○ Underprepared <0° (>9.00 mm)
○ Adequate 0° - 20° (7.00 – 9.00 mm)
○ Overprepared >20° (<7.00 mm)

Bucco-lingual/palatal occlusal convergence is
underprepared < 0° (>9.00 mm)
adequate 0° - 20° (7.00 - 9.00 mm)
overprepared > 20° (<7.00 mm)</td></tr>
<tr><td>Record</td><td colspan="2">The worst result
(Overprepared > Underprepared > Adequate)
The worst ---------------------→ The best</td></tr>
<tr><td colspan="3">1.2.2 Proximal convergence</td></tr>
<tr><td>Tools</td><td colspan="2">Pick up periodontal probe CP12 or round-ended, chamfer, diamond bur</td></tr>
<tr><td>Measure</td><td>The widest mesial-distal aspect of occlusal part of preparation (marginal ridge)</td><td>
</td></tr>
<tr><td>Criteria</td><td colspan="2">○ Underprepared <0° (>9.00 mm)
○ Adequate 0° - 20° (7.00 – 9.00 mm)
○ Overprepared >20° (<7.00 mm)

Proximal occlusal convergence is
underprepared < 0° (>9.00 mm)
adequate 0° - 20° (7.00 - 9.00 mm)
overprepared > 20° (<7.00 mm)</td></tr>
<tr><td>Record</td><td colspan="2">The worst result
(Overprepared > Underprepared > Adequate)
The worst ---------------------→ The best</td></tr>
<tr><td colspan="3"></td></tr>
<tr><td colspan="3"></td></tr>
<tr><th colspan="3">1.3 Axial surface reduction</th></tr>
<tr><td colspan="3">1.3.1 Bucco-palatal/lingual axial reduction</td></tr>
<tr><td>Tools</td><td colspan="2">Pick up periodontal probe CP12 or round-ended, chamfer, diamond bur and impression index</td></tr>
</table>

Measure	The greatest distance between fit surface of the index and buccal-palatal/lingual prepared axial surface 
Criteria	○ Underprepared (<0.50 mm) ○ Adequate (0.50 – 1.50 mm) ○ Overprepared (>1.50 mm)
1.3.2 Proximal axial reduction	
Tools	Pick up periodontal probe CP12 or round-ended, chamfer, diamond bur
Measure	The greatest distance between proximal marginal ridges of prepared tooth to adjacent marginal ridge of unprepared tooth on both sides 
Criteria	○ Underprepared (<0.50 mm) ○ Adequate (0.50 – 1.50 mm) ○ Overprepared (>1.50 mm) **Axial surfaces reduction is** **underprepared (<0.50 mm)** **adequate (0.50 – 1.50 mm)** **overprepared (>1.50 mm)**
Record - *for axial surfaces reduction*	The worst result (Overprepared > Underprepared > Adequate) The worst ----------------------→ The best
1.4 Depth of finish line	
Tools	Pick up periodontal probe CP12 or round-ended, chamfer, diamond bur
Measure	Use the width of probe (0.50 mm) to estimate the depth of finish line all around preparation 
Criteria	○ Underprepared (<0.50 mm) ○ Adequate (0.50 – 1.00 mm) ○ Overprepared (>1.00 mm) **Depth of finish line is** **underprepared (<0.50 mm)** **adequate (0.50 – 1.00 mm)** **overprepared (>1.00 mm)**
Record	The worst result from all four sides of tooth (Overprepared > Underprepared > Adequate) The worst ----------------------→ The best

1.5 Level of finish line	
Tools	Pick up periodontal probe CP12 or round-ended, chamfer, diamond bur
Measure	The greatest distance from finish line to gingival margin on all four sides of tooth 
Criteria	○ Underprepared (supragingival) (>1.00 mm) ○ Adequate (supragingival) (0.50 – 1.00 mm) ○ Acceptable (at to supragingival) (0 - <0.50 mm) ○ Overprepared (subgingival) (<0 mm) **Finish line location is** **underprepared (supragingival) (>1.00 mm)** **adequate (supragingival) (0.50 - 1.00 mm)** **acceptable (at or supragingival) (0 - <0.50 mm)** **overprepared (supragingival) (<0mm)**
Record	The worst result from all four sides of tooth (Overprepared > Underprepared > Acceptable > Adequate) The worst ------------------------------------→ The best

2. Subjective features

2.1 Type of finish line	
Tools	Human eyes
Ask	What is the type of finish line?
Criteria	○ Knife edge ○ Chamfer ○ Chamfer with small lip ○ Shoulder **Type of finish line is** **Knife edge** **Chamfer** **Chamfer with small lip** **Shoulder**
Record	The worst result from all four sides of tooth (Shoulder > Knife edge>Chamfer with lip > Chamfer) The worst ------------------------------→ The best

2.2 Contact area with adjacent teeth	
Tools	Human eyes
Ask	Are the contact areas between the teeth cleared?
Criteria	○ Not cleared on one or both sides ○ Cleared on both sides (0.50 mm) ○ Cleared on one or both sides (> 0.50mm) **Contact area with adjacent teeth is** **not cleared on one or both sides** **cleared on both sides (0.50 mm)** **cleared on one or both sides (> 0.50mm)**
Record	The worst result (Cleared on one or both sides (> 0.50mm) > Not cleared on one or both sides > Cleared on both sides (0.50 mm)) The worst --→ The best

2.3 Damage of the adjacent teeth	
Tools	Human eyes
Ask	Is there damage visible on the adjacent teeth?
Criteria	○ Not damage ○ Minor damage for one or both teeth ○ Moderate to severe damage for one or both teeth **Adjacent teeth have** **not damaged** **minor damage for one or both teeth** **moderate or severe damage for one or both teeth**
Record	The worst result (Moderate to severe damage for one or both teeth > Minor damage for one or both teeth> Not damage) The worst --→ The best

Note:

In relation to degree and pattern of damage to adjacent teeth, the following scale must be indicated to assess adjacent teeth damage:

- No damage visible to the human eye or under magnifying glass. Adjacent teeth have a sound surface with regular curved proximal surface without loss of contour. **(No damage)**
- Slight damage visible to the human eye and identifiable with a magnifying glass. This damage is scratches (i.e. narrow, shallow score-lines, usually multiple with a consistent orientation) and/or indentation (i.e. a regular defect without an orientation, roughly circular or irregular in shape). **(Minor damage)**
- Obvious damage. The pattern of damage is groove (i.e. a deeper defect, length greater than width with a vertical or horizontal orientation) and/or extensive damage (i.e. damage involving a large area of the proximal surface). **(Moderate or severe damage)**