Journal of Education and Learning; Vol. 2, No. 1; 2013
ISSN 1927-5250 E-ISSN 1927-5269
Published by Canadian Center of Science and Education

Assessing the Clinical Skills of Dental Students: A Review of the Literature

Carly L. Taylor[1], Nick Grey[1] & Julian D. Satterthwaite[1]

[1] School of Dentistry, University of Manchester, UK

Correspondence: Carly Taylor, School of Dentistry, School of Dentistry, University of Manchester, Coupland III, Coupland St, M19 1PL, UK. Tel: 44-161-275-6624. E-mail: carly.taylor@manchetser.ac.uk

Received: June 13, 2012 Accepted: September 17, 2012 Online Published: January 9, 2013

doi:10.5539/jel.v2n1p20 URL: http://dx.doi.org/10.5539/jel.v2n1p20

Abstract

Education, from a student perspective, is largely driven by assessment. An effective assessment tool should be both valid and reliable, yet this is often not achieved. The aim of this literature review is to identify and appraise the evidence base for assessment tools used primarily in evaluating clinical skills of dental students.

Methods: MEDLINE was searched for all relevant articles from January 1950- January 2011 published in the English language. References of the articles were then hand searched.

This review begins with a brief outline of the student learning process and the aim of assessment. The tools available for both formative and summative assessments are discussed, with particular reference to those used in assessing dental students' clinical ability. The problems of subjectivity and assessor variability associated with traditional teacher-led assessments are highlighted. Methods which have attempted to overcome these problems, such as the use of checklists and training are then discussed. The benefits and shortcomings of the use of students as assessors, both in self and peer assessment are reviewed. Finally, the use of objective assessment methods involving opto-electronic and haptic technology is considered.

Keywords: assessment methods, assessment tools, peer assessment

1. Introduction

1.1 Student Learning

To assess effectively, an understanding of the learning process is required. Different epistemological theories have conflicting views about how we acquire knowledge, stemming from various philosophical viewpoints. Theories aligned with empiricism believe that learning results from direct exposure to events, which forms and subsequently strengthens cognitive associations. This results in the individual recognising and responding to that pattern and ultimately applying this to other situations (Mc Guire 1983). Rationalism assumes that the individual works out their environment by reasoning, in an attempt to make sense of new experiences (Mc Guire 1983). Finally, socioculturalism centres on the belief that people learn according to the society in which they are placed. The individual learns the rules of that society, and their values and practices stem from the societies constraints and norms (Hjorland 2000).

Clinical professions are often concerned not just with knowledge acquisition, but achievement of skills and their application. Miller's pyramid (1990), attempts to explain how students in professions such as Medicine and Dentistry develop such skills.

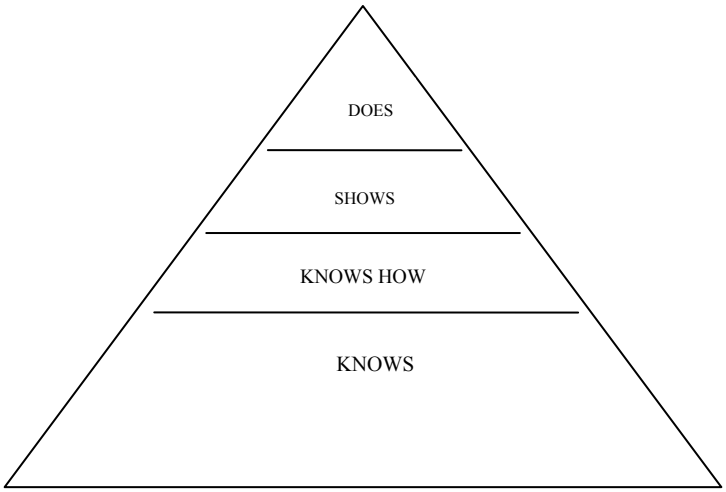

Figure 1. Miller's pyramid

Starting at the base of the triangle, the individual first assimilates knowledge only. At this stage they simply know but cannot apply the information. Progression to the "knows how" level is achieved when the individual can use that information and apply it to a situation. Further progression up the pyramid is achieved when the individual can demonstrate this ability, thus they are deemed to be competent at that particular procedure. When the individual achieves the tip of the pyramid they can perform the procedure.

This theory was further developed to produce the Cambridge Model (Rethans, Norcini & Baron-Macdonald, 2002) which focuses on performance. The authors argue that competency can be assessed in simulated clinical conditions but that performance is actual clinical practice. In these circumstances other factors not accounted for in Miller's model can impact upon an individual's performance. The Cambridge Model only considers the upper two tiers of Miller's pyramid; "shows how" and "does". It describes three factors which may determine performance; competence, the influence of factors of the individual and of the system (Rethans et al., 2000).

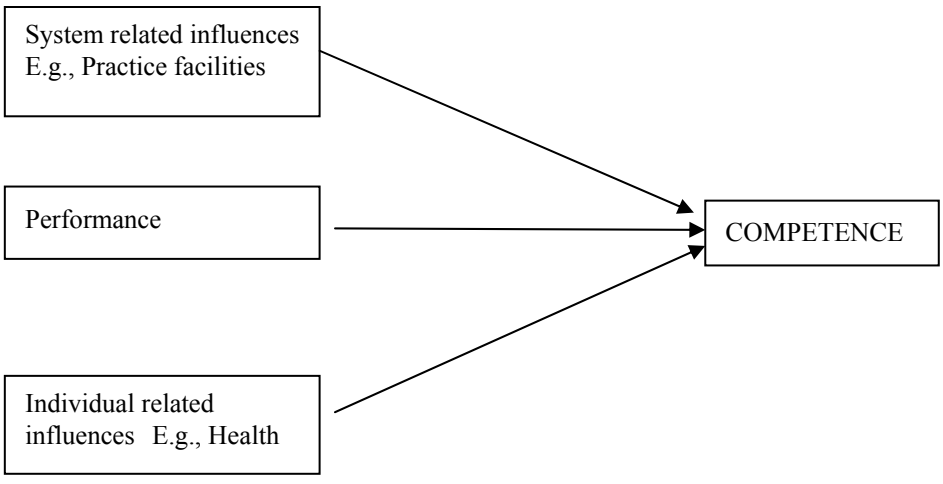

Figure 2. The Cambridge Model

2.2 What is Assessment?

Assessment is an integral part of the educational process at any level and in any discipline. It is a process during which "consideration is given to the amount, level, worth, value or quality of outcomes or products of the learning process" (Topping 1998). The process involves the assessor drawing inferences and making estimates about the value of that product (Manogue et al., 2002). An assessment is designed to evaluate the level of attainment of knowledge, behaviours or skills of students. They can be used to facilitate learning and provide

information to the student about their performance in addition to formal recognition of attainment of knowledge or skills. Assessments are usually the main focus for students, and the driving force for them to engage in the learning process.

2.2.1 Types of Assessment

Assessments can take a multitude of formats and can be classified in many ways; broadly speaking educational assessments are usually classified as summative or formative. Summative assessments are designed to evaluate knowledge and provide formal recognition (Orsmond, Merry & Reiling, 2000). They are usually used at the end of a course or unit and often used to determine student progression. Formative assessments are used as more of a diagnostic tool to provide feedback about the student's progression, which can be reflected upon in order to make any required improvements. Formative assessments are usually not used for formal recognition, but to aid the learning process (Orsmond et al., 2000).

Assessment formats may be written, practical or verbal. They may be carried out individually or as a group. Assessment can take place in a controlled academic setting, or at a distant site involving submission of coursework or online activities. Regardless of the format, it should be effective. This requires the assessment to be valid, reliable, practicable, and ideally objective (Brown 1930). Defined skills should be assessed which are directly related to intended learning outcomes (Manogue et al., 2002).

2.3 Problems with Assessments

An ideal assessment tool would have the following facets: reliability, validity, accountability, flexibility, comprehensiveness, feasibility, timeliness and relevance (Turnbull, Gray & MacFayden 1998, Vargis 2010). In this context, reliability refers to an indication of the consistency of scores over time. If an assessment method is reliable, then the same results should occur regardless of who administers it and when. Validity refers to the ability of the assessment to measure what it is supposed to. In many situations it is not possible to satisfy all of these requirements and many assessment tools used do not fulfil many of these criteria. For example, when considering correlation between three common assessment methods used in pre-clinical and clinical dental there is only a weak correlation between an Objective Structured Clinical Examination (OSCE) and patient assessment and virtually no correlation between the latter and preparation of a typodont model (Curtis, Lind, Brear & Finzen 2007). Preparation of typodont teeth is an almost universal competency assessment of students prior to allowing them to perform procedures on patient, yet this indicates possible poor reliability and validity of such assessments. However, assessment problems can be further compounded by the assessors themselves and the results of such studies could be confounded by variability amongst the assessors which are often not reported.

2.3.1 Assessors

Traditionally, teachers and staff have designed, administered and examined assessments. This can introduce a variety of problems from the choice of assessment, its design, how it is administered, interpreted and evaluated. In addition to the choice of assessment, the subjectivity associated with evaluation by assessors can affect the reliability and credibility of the assessment process. Disparities and contradictions from different assessors can lead to confusion and frustration for students (Natkin & Guild 1967). This can result in a negative impact on learning, as students devalue the process feeling that it is arbitrary, as the grade given does not (in the opinion of the student) relate to the quality of the work.

Assessor's beliefs and practices are not always in agreement with current evidence relating to assessment methods. Staff in restorative departments of UK dental schools have been shown to have values in agreement with current evidence, however practices that are not: they place "high importance on OSCE's, self and peer assessment, communication, feedback and portfolio based learning, however, the most commonly used assessment methods were glance and grade marking and target setting which are not perceived as valuable" (Manogue, Brown & Foster 2001).

2.3.2 Inter-Examiner Agreement

Many studies have focussed on measuring agreement amongst examiners, both in dentistry and other academic disciplines. In studies relating to clinical and laboratory assessments in dentistry, inter-examiner agreement scores have ranged from 0.012-0.94 (Dhuru 1978, Fuller 1972, Lilley, ten Bruggen Cate, Holloway, Holt & Start 1968, Gaines, Bruggers & Rasmussen1974, Goepferd and Kerber 1980, Houpt and Kress 1973, Salvendy et al., 1973, Paskins et al., 2010, Quinn, Keogh, McDonald & Hussey 2003, Satterthwaite & Grey 2008).

In a typical study by Jenkins et al. (1996), assessment of Class II cavities was considered. Variability of up to seven marks (in a thirteen point grading system) was noted and although some of the preparations had been carried out by staff and represented the "technically ideal" preparation, full marks were still not awarded. The

authors related this to staff being reluctant to award high grades if they felt it had been carried out by a student. Similar findings have been noted with assessment of endodontic procedures, for example a study by Natkin and Guild (1967) highlighted that 45% of grades varied by more than four grades (nine point grade range), with only 8 % being within one grade of each other. Even when grades were similar, the reasons for grade assignment were very different. The authors felt that this was due to "instructors struggling to recognise errors and assigning an appropriate level of severity to them". Other reasons suggested for such variability include examiner experience, internal rater bias, interpretation and design of rating scales, knowledge, training, and severity of standards set by the instructor (Fuller 1982, Helft et al., 1987, Houpt & Kress 1973, Natkin & Guild 1967).

2.3.3 Intra-Examiner Variability

In addition to variability between examiners, inconsistencies can also occur within examiners. Studies typically show that intra-examiner variability is less than inter-examiner (Lilley et al. 1968, Deranleau et al. 1983), and may even not be present (Sharaf, Abdel Aziz & El Meligy 2007). However, most studies suggest that agreement is still not good: for example Satterthwaite and Grey (2008) found intra-examiner agreement of two experienced assessors to be 0.53 when assessing typodont preparations, similar to Goepferd and Kerber (1980) who found similar results with intra-examiner agreement scores of 0.62-0.68 (compared to 0.3-0.47 between assessors) – in relation to borderline decisions, this can represent a wide range of a 5-22% difference in pass/fail grades when examiners re-assess preparations (Jenkins et al. 1996).

2.3.4 Training

Due to the large variation of agreement scores, studies have investigated if training yields any improvements in consistency, however, the length, content, type and number of training sessions varied considerably. Lilley et al. (1968) provided training to three assessors between the first and second assessments of thirty seven cavity preparations and their subsequent restoration with amalgam. This consisted of a conference to discuss criterion for each grade, classify faults and their subsequent severity. The authors did not specify the duration of the conference. They found a significant improvement in agreement for the cavity preparations but not for any other stages, nor was there much difference in pass/fail disagreement following the conference. Similarly, assessment scores using two different systems have been compared (Goepferd and Kerber 1980). The traditional "glance and grade" method was first used following annual faculty training. Training was then given for use of an analytical checklist using exemplars, pictures and explicit definitions. There was an improvement in agreement using this method following the training. Rather than being due to training however, this may be due to the new assessment tool, or mere repetition of assessing the same teeth. Similarly, training between successive assessments of cavity preparations which involved cavity assessment with immediate feedback of the "correct" mark, as determined by senior staff has shown that training improves total scores but not individual criterion this was not, however, statistically significant (Houpt and Kress 1973). Dental staff, students and nurses received training in assessment and training improved agreement amongst nurses the most, which is not surprising as they had no operative experience. These results should be interpreted with caution, as the bench mark was grades assigned by senior staff and as highlighted above such individuals may also suffer from inter- and intra-examiner variability. Conversely, other studies have reported little or no effect of training on examiner agreement and consistency (Hinklemann and Long 1973; Fuller, 1972).

2.3.5 Assessor Experience

Studies have also investigated the effect of examiner experience on agreement scores. Several studies have used assessors with varying levels of experience, as this more accurately reflects the situation in dental schools. Some studies have failed to show a correlation between examiner experience and agreement scores (for example see Deranleau et al., 1983, Hinkleman & Long 1973, Natkin and Guild 1967), but many studies highlight the influence of experience. Nurses exhibit the lowest levels of agreement, with students and instructors having similar levels of inter and intra-examiner agreement (Houpt & Kress 1973; Mast and Bethart 1978). However, it has been demonstrated that experience improves inter and intra-examiner agreement, with statistically significant differences between senior and junior students and between staff and both levels of students (Mast & Bethart 1978). Less experienced examiners have been shown to be more inconsistent than senior examiners, however, pass/fail scores do not relate to seniority with more experienced assessors still varying by 17% (Jenkins et al. 1996). Similarly, Fuller (1972) found instructors with only a year's experience had intra-rater reliability scores of 0.472, compared to 0 .831 for the most experienced assessors.

2.4 Assessment Tools

A wide range of examiner variability has been reported, with conflicting findings regarding potential reasons for this. Researchers therefore have focussed on the tools used for assessment to try and improve consistency and

agreement. As Manogue et al. (2001) discovered, tools widely used for assessment in dental education such as the "glance and grade" method are not regarded highly by assessors, this method is, however, widely used. The predictive validity of dental instructors has been assessed during a nine month pre-clinical skills course (Chambers, Leknius & Woodson 1997); instructors assessed students in seven behavioural categories in addition to the routine practical examinations. The authors found the instructor assessments to be as accurate at predicting future clinical performance as the practical test results. The most significant domain for predicting future performance appeared to be "purposefulness of action". This was defined as "organisation, efficiency in motion and proceeding in a logical fashion". No instructor calibration of training was given suggesting an innate ability of the assessors to identify levels of student performance.

2.4.1 Global Versus Analytical

Several papers have compared results using "glance and grade" methods which provide a global mark, with analytical methods utilising individual criterion to assess dental preparations. Vann, Machen & Hounshell (1983) compared global assessment with a checklist and a checklist and criterion to clarify the work of Goepferd and Kerber (1980). Results indicated that no method improved inter-examiner agreement, although the criterion and checklist did improve intra-examiner agreement. Although Vann et al. (1983) used the same grades and descriptors with similar numbers of examiners and preparations as Goepferd and Kerber (1980), their findings were different. Goepferd and Kerber (1980) found an improvement in both intra and inter-examiner agreement using the criterion and checklist. This could be due to the order of assessments, as Goepferd and Kerber (1980) used the "glance and grade" system first, therefore improvement may have been due to repeating the process rather than the rating system. Vann et al. (1983) alternated the assessment methods.

Fuller (1972) compared "glance and grade" assessment with the use of preparation models and a checklist for sixty seven Class II cavity preparations assessed by eight examiners. Correlation coefficients showed no significant differences between either method. Sharaf et al. (2007) also compared "glance and grade" assessment with checklist and criterion for two hundred and forty cavity preparations assessed by three examiners. They found statistically significant inter-examiner variability in 87% of cases using both assessment methods.

2.4.2 Checklists

Problems associated with rating scales have long been recognised. When assessing the personalities of senior dental students using checklists with criteria utilised but no objective statements Brown (1930) did not find any inter-examiner agreement using the checklist (although no raw data or statistical analysis was presented). Checklists for cavity preparations have also been compared, for example Gaines et al. (1974) compared two checklists; the first consisted of six assessment domains each scoring 0-5. The second checklist additionally contained objective statements for each score in each domain. Inter-examiner agreement using the first checklist was 0.26, this increased to 0.56 with the second checklist. The study involved seven examiners with unstated levels of experience, assessing only eight preparations. Similarly, Helft et al., (1987) also used a five point rating scale to assess the marginal adaptation and thickness of cemented crowns on extracted teeth. They discovered significant inter-examiner variability, with assessors either being very critical or more lenient, however their study used a poorly designed rating scale which had no objective criteria or exemplars, possibly causing mis-interpretation.

In relation to a checklist and criterion when assessing direct and indirect tooth preparations, it has been suggested that rater bias is the most significant factor in marking variance, followed by incorrect interpretation of the rating scale (Feil 1982). Paskins et al. (2010) assessed the use of a criterion based checklist designed to assess the management of simulated respiratory and cardiac emergencies. Two assessors used the tool and showed inter-examiner agreement of >0.9. Such high agreement could have been related to the simplicity of the situation being assessed, with limited opportunity for mis-interpretation of the checklist. The authors concluded that the tool was valid, as it yielded a statistically significant difference in scores related to the three different experience levels of the students.

2.4.3 Point Scales

Other research has focussed on the number of points in a rating scale, to assess if this impacts upon examiner agreement. It is not surprising given that fewer options are more likely to result in higher agreement, for example Houpt and Kress (1973) used a two point scale (incorrect/correct), five point scale with only the upper and lower limits specified, and a five point scale with detailed descriptions of each level. They discovered the best agreement with the two point scale, but concluded the five point scale with descriptions would be of more value in teaching. Similarly Deranleau et al. (1983) compared checklists with two and three point scales for evaluation of cavity preparations and crown wax ups using five assessors. Agreement was similar with both methods;

however the authors chose the three point scale as it provided greater student feedback.

Natkin and Guild (1967) compared a nine point grade scale with a new system for endodontic assessment. This procedure involved the preparation starting with an "A" grade. A list of errors categorised for severity were formulated, and the preparation downgraded for each one present. This did result in a reduction in variability, with the mean grade range being reduced from 4.16 to 3.34. Although there was a reduction in variability using the new assessment tool, it may not be practical to implement, as specific errors would need to be assimilated for every procedure undertaken.

For all the studies assessing examiner agreement and rating scales in dental education, similar study designs have been used. None of the authors make reference to the sample sizes or reasons for number of assessors studied. This could have influenced the findings. Due to the variation in study designs, lack of explicit methodologies reported and different statistical tests used, it would not be possible to combine the findings of them to give more substantive evidence about these issues.

3. Students as Assessors

The idea of using students as assessors is not new, and has been studied in many academic disciplines. Harris and Miller (1990) used senior medical students as patients and examiners for OSCE examinations involving junior colleagues. Although no data was presented, the authors reported good inter-examiner agreement in addition to a favourable response from candidates and assessors. Other reports of staff/student agreement during assessments have varied. Mast and Bethart (1978) reported a 99% agreement of restoration assessments with senior students and instructors. However, Burnett and Linden (1988) found statistically significant differences between junior and senior students and instructors. The authors also found that intra-examiner agreement had a positive correlation with experience level. 45% of instructors had agreement of >0.75, compared to only 9% of junior students.

3.1 Self Assessment

A previous literature review including qualitative studies revealed that "high achievers tend to under mark themselves, whereas poor students tend to over rate their performance" (Falchikov 1986). Orsmond, Merry and Reiling (1997) compared self assessment and tutor marks of presentations given by first year Biology students. There was disagreement of marks awarded in 86% of cases, with 56% of students over marking and 30% under marking; although most varied by only one mark (Orsmond et al., 1997): the authors found poor students have a tendency to over mark whilst better students under mark themselves.

The ability to accurately self assess is crucial in health care professions. Clinicians should be able evaluate their performance realistically in order to determine their future training needs. Falchikov and Boud (1989) performed a meta-analysis of forty eight quantitative self assessment studies from the arts, science and social sciences. They discovered that agreement of self and tutor marks was influenced by "the quality of the study design, level of the course and area of study". More advanced level courses and scientific disciplines showed higher agreement. Meta-analysis showed agreement ranges of -0.05 up to 0.82.The authors concluded that students rate their performance based upon the amount of effort involved, whereas staff grade the final product. There are, however, several problems with this meta-analysis. Firstly, studies were included with poor methodology on the premise that exclusion would result in too few studies. Included studies utilised various study designs, assessment tools and statistical analyses. Although the authors categorised them as "high or low" quality, the results from the meta-analysis could be misleading.

Davis et al. (2006) conducted a systematic review comparing physician self assessments and external competency measures. Seventeen studies were included showing twenty assessments. Seven of the assessments showed positive associations of the two measures, with the other thirteen having little, none or inverse relationships with the external measure. The authors admitted that the studies showed heterogeneity regarding statistical analyses and choice of comparisons, and concluded although the quality of the studies was poor, physicians had a limited ability to self assess. This paper also substantiates the findings of previous studies that poor students were over confident and worse at self assessment.

Self assessment can, however, improve with practice. Curtis, Lind, Dellinges, Setia and Finzen (2008) evaluated self assessment scores of seventy seven students with marks given by one academic assessor for two successive tooth set ups. The authors wanted to investigate if students could improve at self assessment. Intra-examiner agreement for the assessor was 0.77, and correlation between self assessment and instructor scores did improve for the second assessment, indicating an improvement in self assessment ability.

3.2 Peer Assessment

Peer assessment has been researched in many academic disciplines to try and determine its reliability. Topping (1998) defines it as "an arrangement in which individuals consider the amount, quality, level or success of products or outcomes of learning of peers of similar status". Often peer assessment involves groups of students assessing their colleague's achievements; however, some studies have used the term to describe individual students performing the assessment. Peer assessment may be used to either assess a product, such as a piece of written work, or a process, such as contribution to a discussion (Falchikov 1986).

3.2.1 Philosophy

Falchikov and Goldfinch (2000) suggested that peer assessment can be related to several theories of learning. Firstly that of active learning; where the responsibility of learning rests with the student. Secondly androgogy, which Smith (1996) defines as "the study of the adult education process". Social constructionist theory may also apply, as this relates to the development of phenomena as a product of social contexts.

3.2.2 Benefits

Numerous advantages of peer assessment have been suggested over traditional assessment methods. These include "ownership of learning, deeper learning, student motivation, active involvement in the learning process, interchange of ideas and more directed and effective learning" (Kwan & Leung 1996). Topping (1998) thought peer assessment provided "increased amounts of feedback, allows realisation of knowledge gaps, norm referencing and the ability to give and accept criticism". Wider benefits such development of transferable interpersonal skills, increased student autonomy, improvement of negotiation, communication and diplomacy skills, increasing critical thinking and responsibility to the group, have also been suggested (Heylings & Stefani 1997, Kwan & Leung 1996, Orsmond et al., 1996,Van Rosendal & Jennett 1994). Such skills could benefit the individual beyond the educational process and potentially be applied to other situations. Heylings and Stefani (1997) also proposed that "participation in peer assessment can promote independent, reflective learners who are more able to adapt to future changing attitudes and knowledge throughout their career."

Certainly student feedback from participants involved in peer assessment has generally been favourable. Student responses included a perceived improvement in the quality of their work, insight into the assessment process, development of transferable skills, increased ability to critique their own work, ability to think in a more structured way, in addition to being challenging yet enjoyable (Heylings & Stefani 1997, Kwan & Leung 1996, Orsmond et al., 2000,Van Rosendal & Jennett 1994). Van Rosendal and Jennett (1994) suggested peer assessment may give a more realistic view of student behaviour, as students may act differently around tutors. They studied medical interns and discovered that peer and instructor scores for operational skills were similar, but scores relating to patient interaction and behaviour were significantly different.

Peer assessment can be used in virtually all academic disciplines and some higher education institutions use this tool in every faculty (Loddington 2008). Peer assessment has the advantage of reducing the marking burden for staff. Several organisations have also produced computer software to allow electronic peer assessment to be carried out. This has the advantage of anonymity for the assessor and allows students time to consider the feedback or grade they shall award.

3.3.3 Limitations

There have been reports of limitations associated with peer assessment. The importance of the assessment may also impact on scores, as in high stakes situations, students may increase the marks awarded and similarly existing friendships can also influence grades as students may have "ulterior motives when marking friends or competing colleagues" (Norcini 2003). In Van Rosendal and Jennett's (1994) study, residents were initially reluctant to participate, due to concerns that the process may affect peer relationships. Participant reluctance was also reported in Heylings and Stefani's (1997) paper for fear of criticising their colleagues. Fallows and Chandramohan (2001) discovered that students questioned the responsibility, and emphasised the importance of student- tutor trust. The authors used peer assessment for presentations and highlighted the importance of clear instructions about what is being assessed. They reported that students assigned marks for "eye contact, perceived effort spent and memory of material, rather than content".

For peer assessment to be effective, each member of the group should have input into the process. Problems may arise with strong personalities who may dominate the process. Students also need to feel that the assessment is important, for this reason, some investigators have used peer assessment in summative situations (Fallows & Chandramohan 2006, Heylings & Stefani 1997). Falchikov & Goldfinch (2000), advised that "peer assessment should be conducted in small groups in an academic setting, using a global mark with well understood criterion".

Norcini (2003) advocated five steps to implement peer assessment. Initially, state the purpose of the assessment in writing. Criteria should be developed in conjunction with students and training should be given. Results of the assessment should be given along with any feedback.

3.3.4 Agreement with Other Methods

Falchikov and Goldfinch (2000) found a mean overall agreement of 0.69 for the forty eight studies included in their meta analysis. Topping (1998) reviewed twenty five peer assessment studies. He found 75% of them to have high peer /tutor agreement. Orsmond et al. (2000) found no statistically significant differences between peer, self and tutor assessment marks, although peers were more likely to under mark and self assessments had higher marks.

Peer assessment has also been studied in dental students. Brehm (1972) used peer and self assessment for bridge preparations on typodont teeth. Rather than using groups of students in the peer assessment, preparations were randomised and marked by a single colleague. Although no raw data is presented, Brehm concluded "students were more critical than faculty, providing more candid remarks. Agreement was higher for excellent and poor work, but less so for average preparations". Denehy and Fuller (1972) used two groups for peer assessment of crown wax ups. The agreement between groups was 0.744, whilst for peer and tutor marks it was 0.715, with peer grades used for summative assessment if they showed >0.75 agreement with the instructor. Satterthwaite and Grey (2008) also compared peer and tutor assessment marks of crown preparations on typodont teeth. They found peer/tutor agreement to be 0.356-0.497. Evans, Leeson & Petrie (2007), used self and peer assessment in evaluation of third molar surgery amongst post graduate students. Agreement between the two tutors was 0.91, peer and tutor score agreement was 0.83 when using a global mark, but only 0.58 with a checklist. Self assessment and tutor agreement was only 0.55 for both methods (Evans et al., 2007). The results indicate moderate to excellent agreement for both methods, however, the authors suspected collusion.

4. Assessment of Tooth Preparations

Much of the assessment research in dentistry has involved examination of tooth preparations.

4.1 Objective Assessment Methods

Given the subjectivity associated with human assessors, attention has been given to developing more objective methods of assessment. Schiff, Salvendy, Root, Ferguson and Cunningham (1975), designed the pulpal floor measuring instrument. This comprised of a platform holding a tooth mounting device. A probe attached to a recording device was then introduced into the cavity as the operator moved the platform in either bucco-lingual or mesio-distal directions to allow the probe to traverse the floor of the cavity and record its contour. The device did prove to be operator sensitive, as incorrect manipulation could result in the probe contacting the walls of the cavity, not the floor, resulting in erroneous recordings. The authors stated the test: re-test reliability to be 0.81-0.99 using the instrument, compared to 0.66-0.89 for intra-rater reliability. The latter figures, however, were taken from another study, thus may not be directly comparable.

4.1.1 Prepassistant

The Prepassisant (Kavo, Germany) is a CCD optical scanner, designed to objectively assess typodont teeth. The device scans model teeth by photographing them from different angles and light projections. The Prepassistant can be used to scan an unprepared tooth, the instructor's preparation and student preparations. The major advantage for teaching, is its ability to compare the student's preparation with that of the instructor. The software provides visual 3D images of both preparations, which can be rotated. The operator can choose a specific plane through each tooth along with measurement points. The device then calculates the deviation in millimetres of the student preparation from the instructors. The device can also measure the taper of the instructor preparation in any desired plane and the student's deviation from this. The major limitations of the Prepassistant however, are its lack of ability to assess surface roughness and continuity of the finish line. Additionally, the output is only a series of discreet measurements rather than an overall assessment of the preparation. Despite this, it can provide objective feedback and useful visual comparisons for the student.

Kournetas et al. (2004) conducted a pilot study to assess the reliability and repeatability of the device. Four prepared and four unprepared teeth were scanned several times both with and without repositioning the tooth in the device (reproducibility and repeatability). The authors aimed to detect the magnitude of variation produced by the device and if this would be visible to the human eye. The authors stated that changes of 100-200μm would be detectable. The eight teeth were scanned in six planes, resulting in one hundred and twenty measurements per tooth. They found repeatability measurements to be more accurate than reproducibility. The main reason given for this was the mounting device allowing the tooth to be positioned in more than one way. The authors reported

the mean accuracy of the device to be 89μm and acceptable for educational purposes.

Cardosa, Barbosa, Fernandes, Silva and Pinho (2006) used the Prepassistant to calculate 70% of the grade awarded to students in a pre-clinical skills course. Parameters such as axial reduction and inclination and occlusal reduction as calculated by the device were weighted according to perceived importance. The remaining 30% of the grade was assigned for surface roughness and geometry of the finish line as judged by the instructor. These grades were then compared to traditional global grades assigned by one instructor. The authors found that the mean, maximum and minimum scores using both methods were the same, however assessment scores using the Prepassistant had a lower standard deviation. The authors also found that the lowest scores given by the Prepassistant were those relating to taper. There are several problems with this study. Firstly the sample size only consisted of twenty five preparations. Only three slices per tooth were measured, with only sixteen points being measured per tooth. This could give erroneous results, as only measuring five or six points per slice would not allow for an accurate representation of the preparation. Additionally, the authors chose slices which they claimed were areas most prone to incorrect preparation. These were mesio-distal slices through the buccal cusp tips, central fissure and close to the lingual wall. These areas however, will not assess the bucco-lingual taper of the preparation, which tends to have increased taper (Ohm and Silness 1978).

4.1.2 Opto-Electronic Devices

As technology has advanced, more sophisticated machines have been developed for training dental students. DentSim (Image Navigation Ltd, USA) is an example of such a device. This system can be used in conjunction with a traditional phantom head simulation unit. The software provides the student with simulated patient information, online visual tracking, real time feedback and evaluation. This technology has the advantage that students are still in a simulated clinical environment, thus get the opportunity to practice patient positioning, ergonomics and four handed dentistry. As teeth are being prepared in jaws, adjacent teeth and occlusal contacts must also be managed. Although DentSim can provide objective tracking of a preparation, the final assessment still has to be carried out by a member of staff which still leaves the problem of subjectivity associated with the assessment. In a study, a group who trained on the DentSim received slightly higher grades than those who trained on traditional phantom heads, however no statistically significant difference was found (Jasinevicius, Landers, Nelson & Urbankova 2004).

Another important technological development which has an impact on dental education, is the field of haptic technology. Haptic literally means “sense of touch”. Haptic devices afford the user tactile interaction with a device, which provides continuous feedback. Such devices have been developed for use in training medical and dental students. Haptic devices used in dentistry, commonly consist of a hand piece aligned to a computer screen, which displays a virtual tooth. Such devices allow unlimited practice without additional resources or staffing. Virtual teeth have been modelled using the physical properties of enamel and dentine, to provide a realistic sensation when preparing the virtual teeth (Konukseven et al, 2010). Cone Bean Computed Tomography has also been used to model teeth with anatomical variation, thus providing a new dimension to the experiential learning of the student. All data produced by the user is stored and objective feedback provided, including constant feedback relating to the pressure applied and angulation of the handpiece. A comparison of grades awarded by such a device for preparations carried out by experienced prosthodontists and inexperienced students found a statistically significant difference in the grades, concluding that the device was able to accurately assess (Suebnukam, Phatthanasathiankul, Sombatweoje, Rheinmora & Haddawy 2000). Imbers et al (2003), also used such a device in predicting poorly performing students. Participants carried out preparations on the device prior to starting a clinical skills course. They found that ten of the thirteen students who performed poorly on the device, also performed poorly in the end of course operative test. No raw data was presented for the study however, and only small sample sizes were studied. The major limitations to use are the cost and space of providing enough for every student (Konukseven et al., 2010). Additionally, as a single tooth is projected onto a screen, no clinical simulation is provided, thus limiting the aspects of learning when compared to the DentSim or traditional phantom head unit.

5. Conclusion

This review has discussed the broad methods available for assessment, with particular focus on assessing dental students. All methods have their shortcomings, and the techniques that have been utilised in an attempt to overcome these have been highlighted. Although subjectivity is a significant problem with traditional methods, newer electronic devices which aim to provide objectivity still require development before an ideal assessment tool is created.

References

Brown, R. (1930). Researching in the use of a rating scale as a means of evaluating the personalities of senior dental students. *Journal of Dental Research, 10*(3), 271-280. http://dx.doi.org/10.1177/00220345300100030501

Burnett, A. C., & Linden, G. J. (1988). The reproducibility of the assessment of restorations by dental students and their teachers. *Journal of Dental Education, 52*(10), 568-570.

Cardosa J. A., Barbosa, C., Fernandes, S., Silva, C. L., & Pinho, A. (2006). Reducing subjectivity in the evaluation of pre-clinical dental preparations for fixed prosthodontics using the Kavo Prep Assistant. *European Journal of Dental Education, 10*(3), 149-156. http://dx.doi.org/10.1111/j.1600-0579.2006.00409.x

Chambers, D. W., Leknius, C., & Woodson, R. (1997). Predictive validity of instructor judgment in preclinical technique courses. *Journal of Dental Education, 61*(9), 736-740.

Clergen, K. (1985). The Social Constructionist Movement in Modern Psychology. *American Psychologist, 40*(3), 266-275. http://dx.doi.org/10.1037/0003-066X.40.3.266

Curtis, D. A., Lind, S. L., Brear, S., & Finzen, F. C. (2007). The correlation of student performance in preclinical and clinical prosthodontic assessments. *Journal of Dental Education, 71*(3), 365-372.

Curtis, D. A., Lind, S. L., Dellinges, M., Setia, G., & Finzen, F. C. (2008). Dental students' self-assessment of preclinical examinations. *Journal of Dental Education, 72*(3), 265-277.

Davis, D. A., Mazmanian, P. E., Fordis, M., Van Harrison, R., Thorpe, K. E., & Perrier L. (2006). Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. *JAMA, 296*(9), 1094-1102. http://dx.doi.org/10.1001/jama.296.9.1094

Deneby, G., & Fuller, J. (1974). Student peer evaluation: An adjunct to Preclinical Laboratory Evaluation. *Journal of Dental Education, 38*(4), 200-203.

Deranleau, N. J., Feiker, J. H., & 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.

Dhuru, S. (1978). Criterion-oriented grading system for preclinical operative dentistry laboratory course. *Journal of Dental Education, 42*(9), 528-531.

Evans, A. W., Leeson, R. M. A., & Petrie, A. (2007). Reliability of peer and self-assessment scores compared with trainers' scores following third molar surgery. *Medical Education, 41*(9), 866-872. http://dx.doi.org/10.1111/j.1365-2923.2007.02819.x

Falchikov, N. (1986). Product comparison and process benefits of collaborative peer group and self assessment. *Assessment and Evaluation in Higher Education, 11*(2), 146-166. http://dx.doi.org/10.1080/0260293860110206

Falchikov, N., & Boud, D. (1989). Students self assessment in Higher Education: A meta analysis. *Review of Educational Research, 59,* 395-430.

Falchikov, N., Goldfinch J. (2000). A Meta- analysis comparing peer and teacher marks. *Review of Educational Research, 70*(3), 287-322.

Fallows, S., & Chandramohan, B. (2001). Multiple approaches to assessment: Reflections on the use of tutor, peer and self-assessment. *Teaching in Higher Education, 6*(2), 229-246. http://dx.doi.org/10.1080/13562510120045212

Fuller, J. (1972). The effects of training and criterion models on interjudge reliability. *Journal of Dental Education, 36*(4), 19-22.

Gaines, W., Bruggers, H., & Rasmussen, R. (1974). Reliability of ratings in Preclinical Fixed Prosthodontics: Effect of objective scaling. *Journal of Dental Education, 38*(12), 672-675.

Goepferd, S. J., & Kerber, P. (1980). A comparison of two methods for evaluating Class II cavity preparations. *Journal of Dental Education, 44*(9), 537-541.

Harris, I., & Miller, W. (1990). Feedback in an Objective Structured Clinical Examination by medical students serving as patients, Examiners and Teachers. *Academic Medicine, 65*(7), 433-434. http://dx.doi.org/10.1097/00001888-199007000-00002

Helft, M., Pilo, R., Cardash, H. S., & Baharav, H. (1987). Reliability of ratings in evaluation of crowns. *Journal of Prosthetic Dentistry, 58*(5), 647-649. http://dx.doi.org/10.1016/0022-3913(87)90402-1

Heylings, D. J., & Stefani, L. A. (1997). Peer assessment feedback marking in a large medical anatomy class. *Medical Education, 31*(4), 281-286. http://dx.doi.org/10.1111/j.1365-2923.1997.tb02926.x

Hinkleman, K.W., & Long, N. (1973). Method for Decreasing Subjective Evaluation in Preclinical Restorative Dentistry. *Journal of Dental Education, September,* 13-17.

Hjorland, B. (2000). Library and Information Science: Practice, Theory and Philosophical basis. *Information Processing and Management, 36,* 501-531. http://dx.doi.org/10.1016/S0306-4573(99)00038-2

Houpt, M. I., & Kress, G. (1973). Accuracy of measurement of clinical performance in dentistry. *Journal of Dental Education, 37*(7), 34-46.

Image Navigation. (2011). *DentSim Overview.* Retrieved from http://www.denx.com/DentSim/overview.html (28th February, 2011)

Imbers, S., Shapira, G., Gordon, M., Judes, H., & Metzer, Z. (2003). A virtual reality dental simulator predicts performance in an operative dentistry manikin course. *European Journal of Dental Education, 7,* 160-163. http://dx.doi.org/10.1034/j.1600-0579.2003.00299.x

Jasinevicius, T., Landers, M., Nelson, S., & Urbankova, A. (2004). An evaluation of two dental simulation systems: Virtual reality versus contemporary non-computer-assisted. *Journal of Dental Education, 68*(11), 1154-1162.

Jenkins, S. M., Dummer, P. M., Gilmour, A. S., Edmunds, D. H., Hicks, R., & Ash, P. (1998). Evaluating undergraduate preclinical operative skill; use of a glance and grade marking system. *Journal of Dentistry, 26*(8), 679-684. http://dx.doi.org/10.1016/S0300-5712(97)00033-X

Kournetas, N., Jaeger, B., Axmann, D., Groten, M., Lachmann, S., Weber, H., & 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.

Kwan, K. P., & Leung, R. W. (1996). Tutor versus peer-group assessment of student performance in a simulation training exercise. *Assessment and Evaluation in Higher Education, 21*(3), 205-214. http://dx.doi.org/10.1080/0260293960210301

Lilley, J. D., ten Bruggen Cate, H. J., Holloway, P. J., Holt, J. K., & Start, K. B. (1968). Reliability of practical tests in operative dentistry. *British Dental Journal, 125*(5), 194-197.

Loddington, S. (2008). *Peer assessment of group work: A review of the literature.* Retrieved from http://webpaproject.lboro.ac.uk/files/WebPA_Literature%20review%20.pdf

Macdonald, J., Williams, R., & Rogers, D. (2003). Self-assessment in a simulation-based surgical skills training. *American Journal of Surgery, 185,* 319-322. http://dx.doi.org/10.1016/S0002-9610(02)01420-4

Mackenzie, I. (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., & Foster, H. (2001). Clinical assessment of dental students: Values and practices of teachers in restorative dentistry. *Medical Education, 35*(4), 364-370. http://dx.doi.org/10.1046/j.1365-2923.2001.00733.x

Manogue, M., Kelly, M., Bartakova, Masaryk, S., Brown, G., Catalanotto, F., Choo-Soo, T., Delap, E., Godoroja, P., Morio, I., Rotgans, J., & Saag, M. (2002). 2.1 Evolving methods of assessment. *European Journal of Dental Education, 6*(Suppl 3), 53-66.

Mast, T. A., & Bethart, H. (1978). Evaluation of clinical dental procedures by senior dental students. *Journal of Dental Education, 42*(4), 196-197.

McGuire, J. (1983). A Contexualist Theory of Knowledge: Its implications for innovation and reform in Psychology. *Advances in Experimental Social Psychology, 16,* 1-47. http://dx.doi.org/10.1016/S0065-2601(08)60393-7

Miller, G. (1990). The assessment of Clinical Skills/Competence/Performance. *Academic Medicine, 65*(9), 563-567.

Myers, S. (1977). Beliefs of dental faculty and students about effective clinical teaching behaviours. *Journal of Dental Education, 41*(2), 68-76.

Natkin, E., & Guild, R. E. (1967). Evaluation of preclinical laboratory performance: A systematic study. *Journal of Dental Education, 31*(2), 152-161.

Norcini, J. J. (2003). Peer assessment of competence. *Medical Education, 37*(6), 539-543. http://dx.doi.org/10.1046/j.1365-2923.2003.01536.x

Ohm, E., & Silness, J. (1978). The convergence angle in teeth prepared for individual crowns. *Journal of Oral Rehabilitation, 5,* 371-375.

Orsmond, P., Merry, S., & Reiling, K. (1996). The importance of marking criteria in the use of peer assessment. *Assessment and Evaluation in Higher Education, 21*(3), 239-250. http://dx.doi.org/10.1080/0260293960210304

Orsmond, P., Merry, S., & Reiling, K. (1997). A study in self assessment: Tutor versus students perceptions of performance criteria. *Assessment and Evaluation in Higher Education, 22*(4), 357-368. http://dx.doi.org/10.1080/0260293970220401

Orsmond, P., Merry, S., & Reiling, K. (2000). The use of student derived marking criteria in peer and self-assessment. *Assessment and Evaluation in Higher Education, 25*(1), 23-38. http://dx.doi.org/10.1080/02602930050025006

Paskins, Z., Kircaldy, J., Allen, M., Macdougall, C., Fraser, I., & Peile, E. (2010). Design, validation and dissemination of an undergraduate assessment tool using Sim Man in simulated medical emergencies. *Medical Teacher ,32,* e12-e17. http://dx.doi.org/10.3109/01421590903199643

Quinn, F., Keogh, P., Mc Donald, A., & Hussey, D. (2003). A pilot study comparing effectiveness of conventional training and virtual reality simulation on skills acquisition of junior dental students. *European Journal of Dental Education, 7,* 13-19. http://dx.doi.org/10.1034/j.1600-0579.2003.00264.x

Rethans, J., Norcini, J., & Baron-Macdonald, M. (2002). The relationship between competence and performance: Implications for assessing practical performance. *Medical Education, 36*(10), 901-909. http://dx.doi.org/10.1046/j.1365-2923.2002.01316.x

Salvendy, G., Root, C., Cunningham, P. R., Ferguson, G. W., Hinton, W. M., Baum, S., & Khan, L. (1973). Skills analysis of cavity preparations: Class 1 in mandibular right first molar. *Journal of Dental Education, 37*(10), 11-18.

Satterthwaite, J. D., & 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. http://dx.doi.org/10.1111/j.1600-0579.2008.00509.x

Schiff, A. J., Salvendy, G., Root, C. M., Ferguson, G. W., & Cunningham, P. R. (1975). Objective evaluation of quality in cavity preparations. *Journal of Dental Education, 39*(2), 92-96.

Sharaf, A. A., Abdel Aziz, A. M., & 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.

Smith, M. (1996). "Androgogy", the encyclopedia of informal education. Retrieved from http://www.infed.org/lifelonglearning/b-andra.htm

Stefani, L. (1994). Peer, self and tutor assessment: Relative reliabilities. *Studies in Higher Education, 19*(1), 69-75. http://dx.doi.org/10.1080/03075079412331382153

Suebnukam, S., Phatthanasathiankul, N., Sombatweoje, S., Rhienmora, P., & Haddawy, P. (2000). Process and outcome measures of expert/novice performance on a haptic virtual reality system. *Journal of Dentistry, 37,* 658- 665. http://dx.doi.org/10.1016/j.jdent.2009.04.008

Topping, K. (1998). Peer assessment between students in Colleges and Universities. *Review of Educational Research, 68,* 249-276.

Turnbull, J., Gray J., & MacFayden, J. (1998). Improving in-training evaluation programmes. *General Internal Medicine, 13*(5), 317-323. http://dx.doi.org/10.1046/j.1525-1497.1998.00097.x

Van Rosendaal, G. M., & Jennett, P. A. (1994). Comparing peer and faculty evaluations in an internal medicine residency. *Academic Medicine, 69*(4), 299-303. http://dx.doi.org/10.1097/00001888-199404000-00014

Vann, W. F., Machen, J. B., & Hounshell, P. B. (1983). Effects of criteria and checklists on reliability in preclinical evaluation. *Journal of Dental Education, 47*(10), 671-675.

Vargis, A. (2010). Principles of Assessment: A primer for Medical Education in the Clinical Years. *The Internet Journal of Medical Education, 1*.