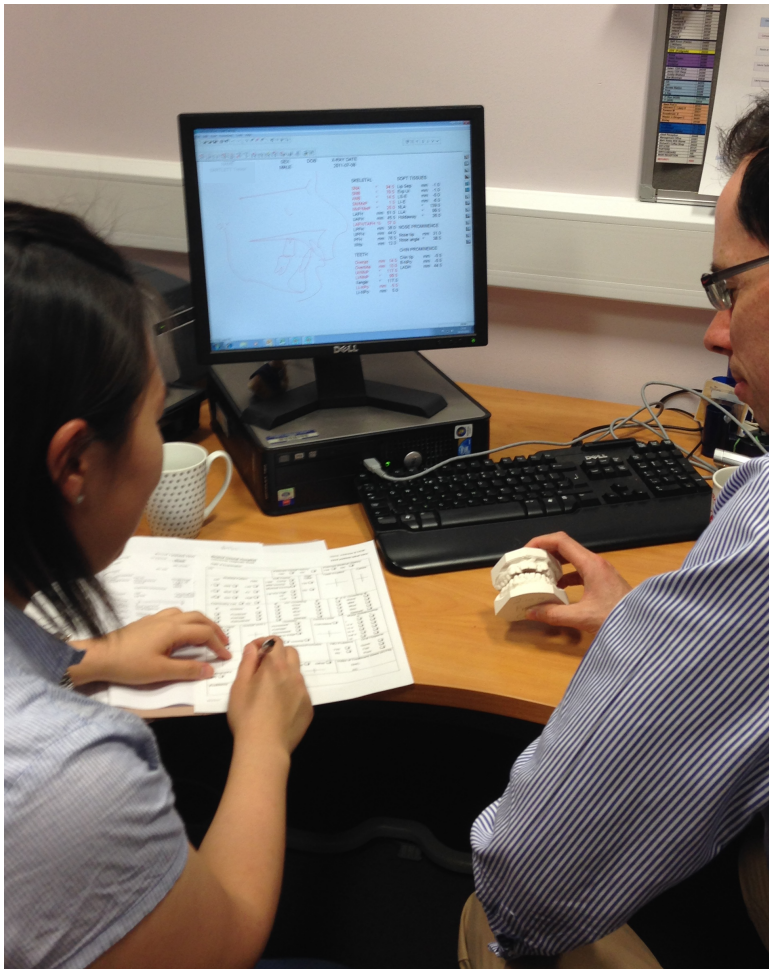
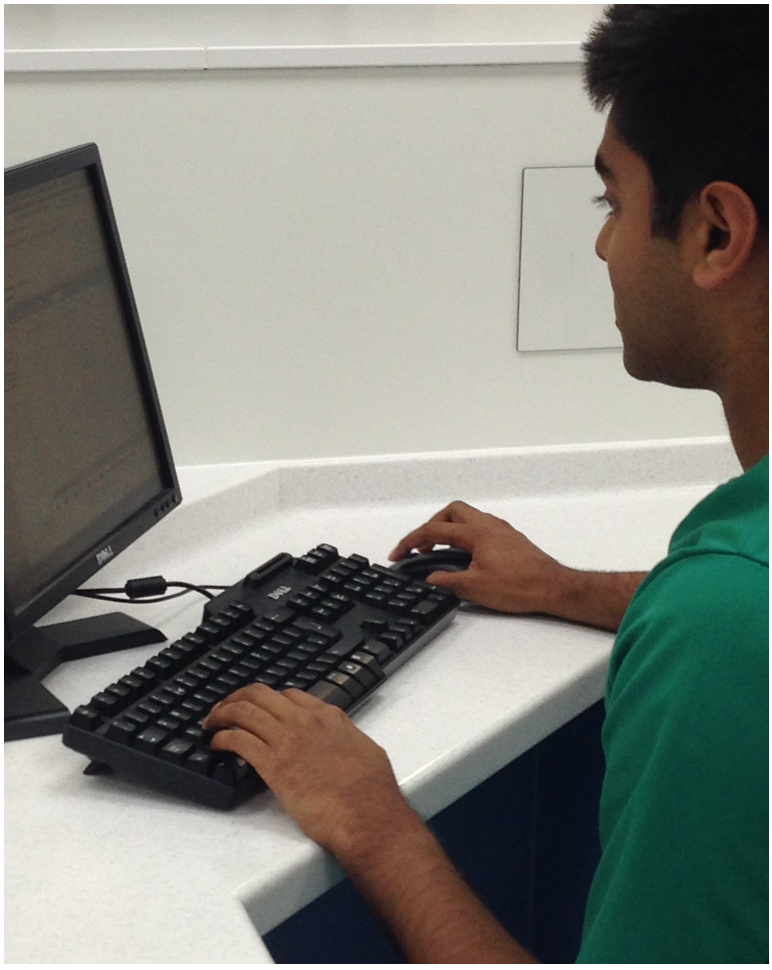


A Guide to Assessment in Dental Education



A Guide to Assessment
in
Dental Education

Authors: Julie Williams, Sarah Baillie, Susan Rhind, Sheena Warman, Jonathan Sandy and Anthony Ireland

November 2015

Version 1

Table of Contents

Introduction

Following the successful publication of "A Guide to Assessment in Veterinary Medical Education", 2nd edition, it was felt that a similar guide to assessment in Dental Education would be timely and valuable. A systematic review, which underpinned the initial document, is described in Rhind *et al.* (2008) and a review of the dental literature was undertaken by Willliams in 2015. It is anticipated that synthesis of the literature in an accessible format will be useful for both new and experienced dental educators at both undergraduate and postgraduate levels.

Authors' Background and Contacts

Julie Williams BDS, MOrth, DDS, MA, DPDS
Academic Clinical Lecturer in Orthodontics, School of Oral and Dental Sciences, University of Bristol, Bristol, BS1 2LY
Julie.Williams@bristol.ac.uk

Sarah Baillie BVSc, CertCHP, MSc, PhD, PFHEA, MRCVS
Professor of Veterinary Education and Veterinary Programme Director, School of Veterinary Sciences, University of Bristol, Langford, Bristol BS40 5DU
sarah.baillie@bristol.ac.uk

Susan Rhind BVMS, PhD, FRCPath, PFHEA, MRCVS
Chair of Veterinary Medical Education and Director of the Veterinary Medical Education Division at the Royal (Dick) School of Veterinary Studies, University of Edinburgh susan.rhind@ed.ac.uk

Sheena Warman BSc, BVMS, DSAM, DipECVIM-CA, SFHEA, MRCVS
Senior Clinical Fellow in Small Animal Medicine, School of Veterinary Sciences, University of Bristol, Langford, Bristol BS40 5DU Sheena.Warman@bristol.ac.uk

Jonathan Sandy BDS, MSc, PhD, MOrth, FDSRCS, FDSRCSEd, FFDRCS, FMedSci
Professor of Orthodontics and Dean of Health Sciences, University of Bristol
Jonathan.Sandy@bristol.ac.uk

Anthony Ireland BDS, MSc, PhD, MOrth, FDSRCS, FHEA
Professor of Orthodontics, School of Oral and Dental Sciences, University of Bristol
tony.ireland@bristol.ac.uk

Acknowledgements

Thanks to all who have given of their time and expertise to help with the development of this guide including Dr Alaa Daud, Dr David Dymock, Dr Gordon Gray, Dr Ross Hobson, Dr Jane Luker, Dr Alasdair Millar, Dr Chris Vernazza and Dr Andrea Waylen. Thanks also to Salisha Amin, Dr Joo Ming Cheong, Dr Gordon Gray, Dr Tim Jones, Akhil Patel and Dr Stefanie Tan for permission to use their images.

Thanks also to Claire Painton, Audrey Biddell and Robin Widdowson for their assistance with formatting and proofreading services.

Finally thanks to our mentors, our teachers and our students for all they have taught us about assessment.

"I've learned that people will forget what you said, people will forget what you did, but people will never forget how you made them feel." Maya Angelou

Principles of Assessment
Authors: Julie Williams and Susan Rhind

The Purpose of Assessment

The term assessment has been used to refer to the systematic determination of student/learner achievement and performance (Schuwirth *et al.*, 2011). The term derives from the Latin "*assidere*" meaning to sit beside, suggesting that the assessor and the student of dentistry travel together side by side on the journey to aid learning. Training the dental practitioner also requires evaluation against a series of agreed standards. Assessment is therefore distinct from evaluation of a student for the purpose of certification, yet one process certainly informs the other and the principle of triangulation uses data from multiple sources to determine the student's readiness for practice. Additionally assessment should be sustainable, if it is to foster lifelong learning. Sustainable assessment has been described as meeting the needs of the *present* without compromising the ability of the students to meet their own *future* learning needs (Boud, 2000). The same author states

" We owe it to ourselves and our students to devote at least as much energy to ensuring that our assessment practices are worthwhile, as we do to ensuring that we teach well".

Selecting the ideal assessment tool to identify the student's strengths and weaknesses in addition to how and when that tool is used and by whom, should inform how best to support the learning and continued professional development of the practising dentist. This process of accumulating sufficient information from multiple sources to allow professional judgement of the student dentist remains a challenging part of both assessment practice and curriculum planning.

The Transition from Student to Practitioner

The Dreyfus five-stage model for the acquisition of skills within the adult learner (Figure 1) has been successfully applied to the development of other health care professionals (Benner, 1982) and is helpful in positioning competence as one more stage on the road to expert status.

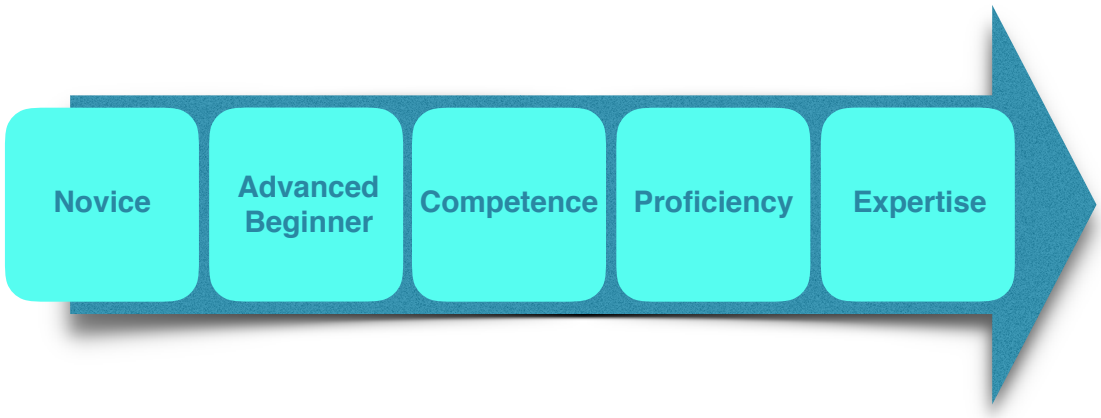


Figure 1. Dreyfus' Five-stage model of Adult Skill Acquisition, cited in Benner, 1982.

The student dentist (*Novice*) learns facts, figures and rules often outside the clinical context whilst the *Advanced Beginner* stage develops after repeated opportunities to apply these rules in clinical settings. Once the student has developed a good working knowledge and learns to correctly apply a hierarchical procedure of decision-making, they start to become responsible for the choice and execution of

their treatment plans. They may then be judged as *competent* against a set of external standards usually set by a Registration body such as the General Dental Council within the U.K.

Life-long professional development encourages the ability to respond to work-based patterns without separating problems into component parts. This stage still requires active decision-making albeit using greater intuition to become *proficient*. Stage 5 or *expertise* can be seen when this competence becomes unconscious and when asked, the *expert* will often say they are doing "what normally works". Not all dentists will necessarily reach expert status in all domains and the General Dental Council (GDC) **"Preparing for practice"** document refers to a cycle of self-regulation, training and validation to create all round proficiency rather than necessarily expertise (Figure 2).

Figure 2. GDC overview of registration and life-long learning from GDC "Preparing for practice" reproduced with permission.

Models of the Development of Competence

One of the most commonly cited models relating to assessment in medical education is that of Miller's Pyramid, originally described by Miller in 1990 (Figure 3). This is a conceptual model which encompasses the elements required for clinical competence – from the underpinning cognitive levels of knowledge and application of knowledge (Knows and Knows How) to the behavioural levels of practical competence, perhaps demonstrated on a model (Shows) and how a doctor (or dentist) actually performs in practice with patients (Does) (Miller, 1990).

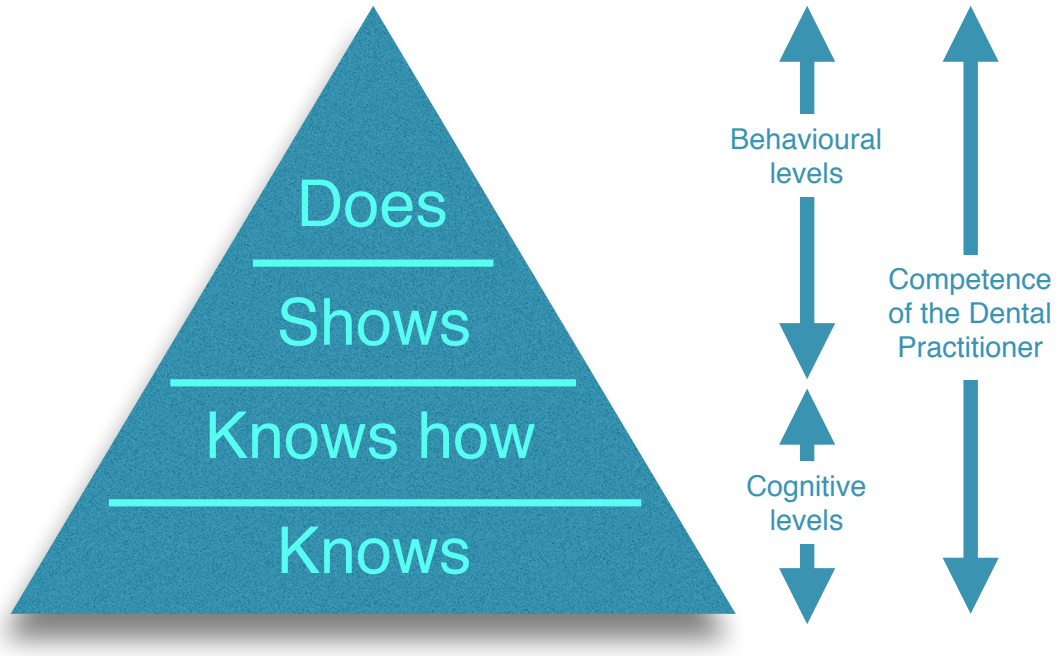


Figure 3. Miller's Pyramid

Although widely used, some cite limitations for this model since it implicitly assumes that competence predicts performance which may not be the case when the doctor is in a "real world" setting (Rethans *et al.* 2002). Although dental education has rightly learnt much from developments in medical education, not least in assessment, it is relevant here to emphasise some key differences between the two disciplines. Dental undergraduates (in common with veterinary students) perhaps differ from their medical colleagues in that, with less post-graduate specialisation, there is increased expectation of new graduates to be able to undertake a wide range of procedures, often with minimal supervision. Dental graduates have "early one on one responsibility for the practical delivery of complex, irreversible treatment with a focussed exposure to broader medical skills" (Bennett *et al.*, 2010).

Competence at the "does" level can therefore be seen as the minimum desired outcome of adequate education and training particularly for a newly qualified dentist or veterinary surgeon. This could explain the trend for assessment methods that were originally used in postgraduate medical education, being used in the final stages of the assessment of competence of dental and veterinary undergraduate students. (Figure 3).

The following is an example of how an undergraduate dental student might progress through the stages of Miller's pyramid, in the context of restoring a tooth. A dental student may first learn tooth morphology, then learn how to recreate this morphology using a material such as composite, then demonstrate that they can show this skill in a simulated setting such as on a dental mannequin followed by being able to complete the same task for a patient. Performance of the task for a patient will usually involve skills such as: communication, teamwork with the dental surgery assistant, gentle anaesthesia of the tooth, protection of the soft tissues, careful tooth preparation, checking the occlusion with the opposing tooth and post-operative advice of how to care for the restoration, in addition to the technical task described.

Miller's pyramid model for medical competence overlaps Bloom's taxonomy of educational objectives (Bloom, 1984) which was revised by Anderson *et al.* (2001) (Figures 4-6). This taxonomy was devised to improve communication between those working in assessment and to permit comparison and study of programmes of education. Bloom and his colleagues found that learning objectives could be classified into one of three domains – cognitive (knowledge based), affective (attitudinal) and psychomotor (skills). The GDC's required learning outcomes for competence can also be similarly classified as cognitive, affective and psychomotor although there is considerable overlap between the three.

Traditionally assessment methods have been developed to test learning objectives within the cognitive domain and less commonly the psychomotor domain. The cognitive domain can be seen in the revised form (Figure 3) as a sequence of six hierarchical categories that reflect a theory of progressive contextualization of knowledge as the student progresses.

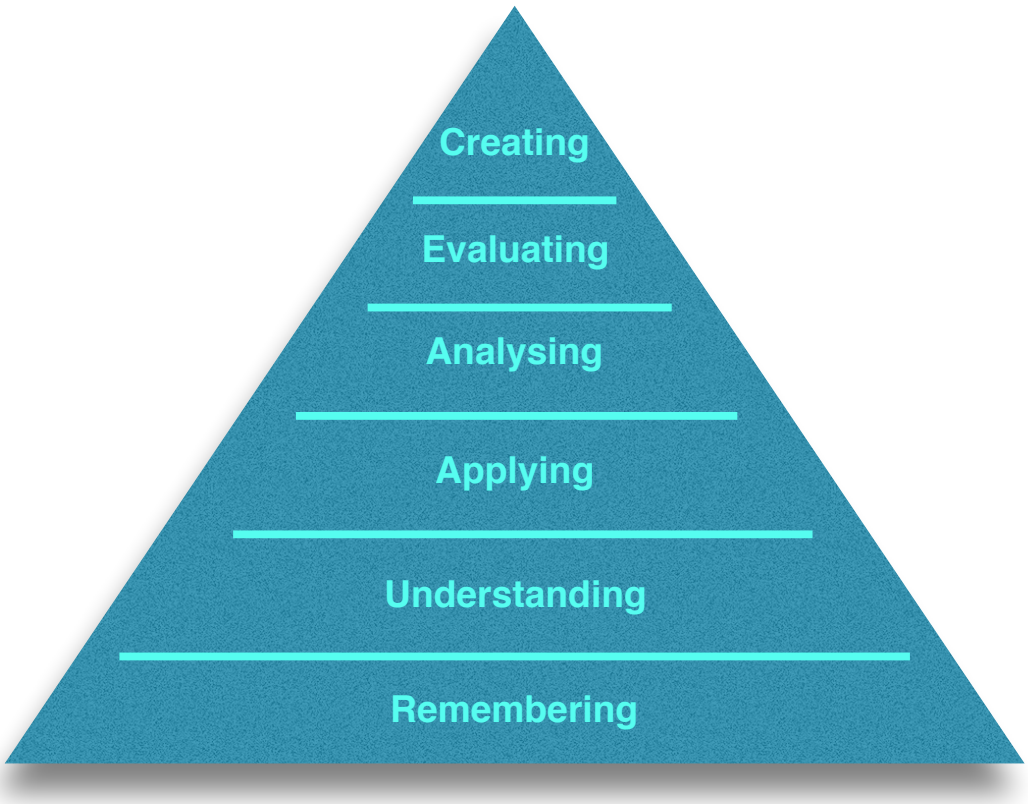


Figure 4. Bloom's Cognitive domain from Taxonomy of Educational Objectives Handbook I, as revised by Anderson *et al.* 2001

Following the usefulness of the first handbook, a similar hierarchical progression was outlined for the affective domain as shown in Figure 5. Creating learning outcomes and assessments within the affective domain appear more challenging, for example to assess whether a student puts patient's interests first and act to protect them (6.1, GDC learning outcomes from "Preparing for Practice"). There are two concerns with assessment within the affective domain. Firstly, on a practical basis it is difficult to construct a clearly and carefully worded multiple-choice question, an essay title or even a communication exercise with a simulated patient that will test an affective objective such as outlined above. Secondly with values-based objectives is the student response merely acquiescence for the purpose of the exam or is the response what they will actually do, in a real life scenario.

Acknowledging these difficulties is important since it is easy for objectives at the "Does" level or within the affective domain to be eroded away to become merely that which can be assessed. Once the student has become aware of the new concept or material and has willingly responded, this should lead to an enjoyment of the process for its own sake. It is at this stage that the student then starts to **value** the material before prioritising this material within daily activities such as caring for patients. At the most complex level the student becomes so committed to the material that it becomes a routine approach, a philosophy.

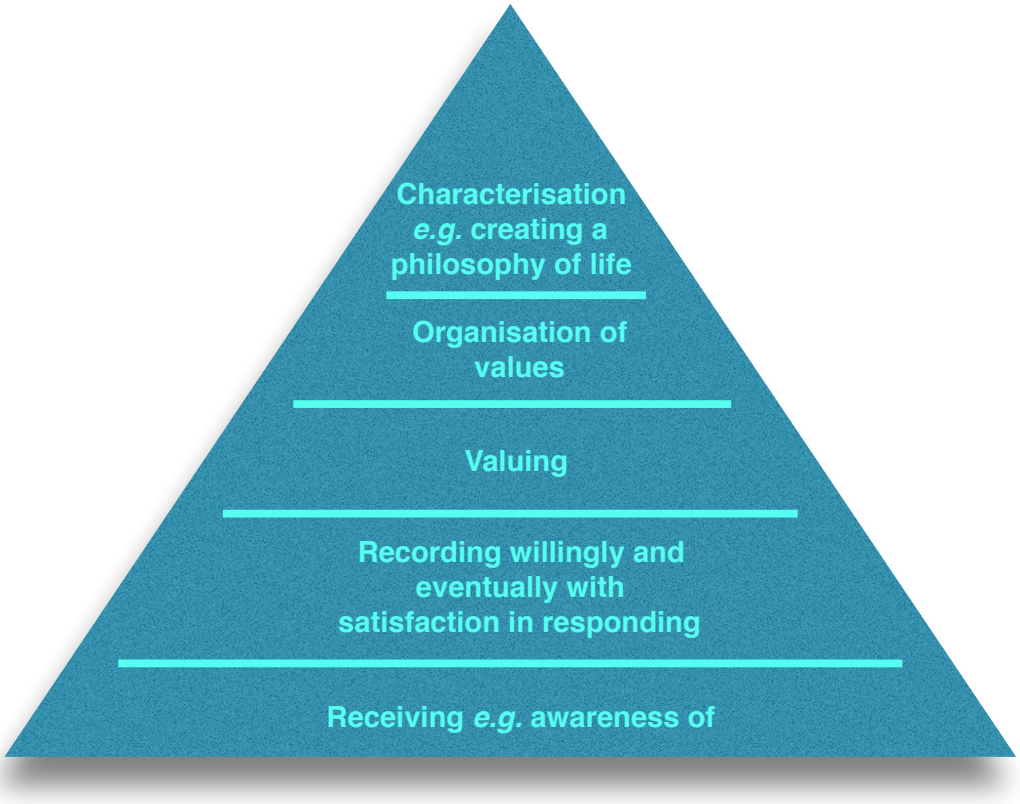


Figure 5. Affective Domain from Taxonomy of Educational Objectives, Handbook II (Krathwohl *et al.* 1964)

Bloom's concept of the Psychomotor domain was further developed as shown in Figure 6 (Dave, 1970). The first three stages are straightforward namely *imitation*, *manipulation* (seen as completing the task without concurrent demonstration) and *precision* (shown when the clinician is independent from tuition). *Articulation* confirms that the clinician is able to adapt the skill to a non-standard setting. In the context of dentistry it is possible to reach precision on a mannequin but in the authors' view,

articulation would refer to the clinical setting. *Naturalisation* refers to the skills of the unconsciously competent practitioner and in terms of Miller's model, whilst this could map to the "shows" level it would seem to be more likely to be observed as part of "does" or performance.

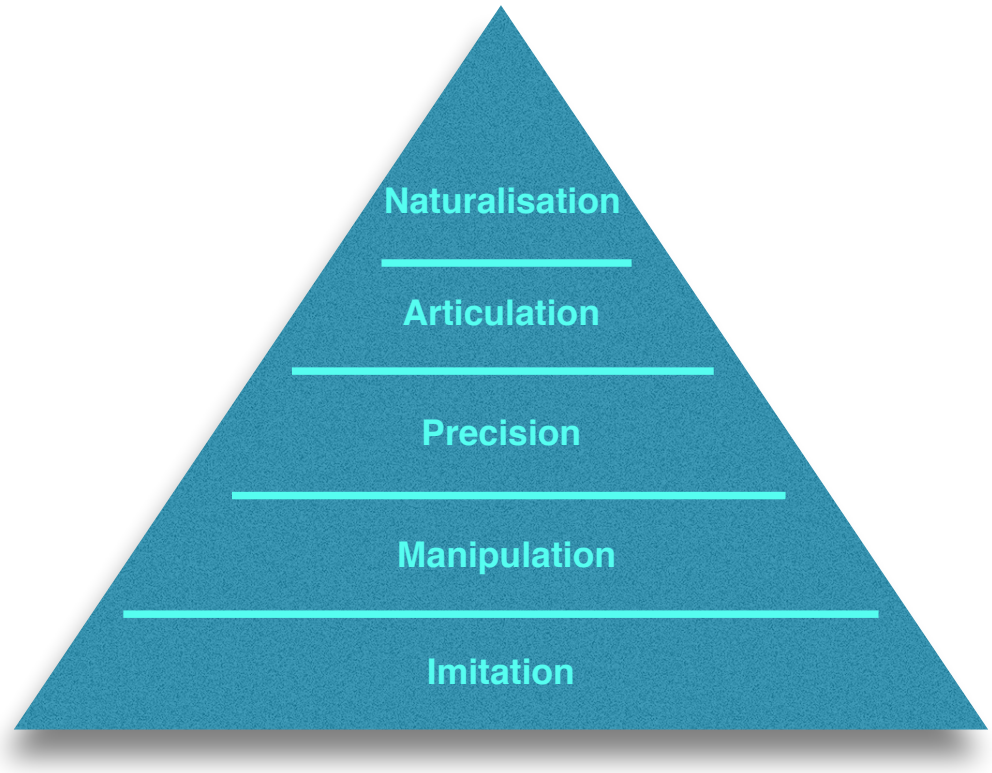


Figure 6. Psychomotor Domain from Taxonomy of Educational Objectives (revised by Dave 1970)

Assessment Tools and Terminology

The case for a specific toolbox of assessment instruments that can be used to assess dental students has been ably made by the American Commission on Change and Innovation in Dental Education (ADEA CCI) and has informed this guide (Kramer *et al.*, 2009). Albino's review of the literature found that five methods have been key for student assessment within dental schools (Albino *et al.* 2008). These are multiple-choice exams, laboratory based "practicals", completion of specified numbers of procedural requirements, daily grades and clinical competency assessments during patient care. **Summative** assessment ("assessment-of-learning") is usually associated with a mark or grade and often occurs towards the end of a course. Summative assessment is important for high-stakes decisions to be made relating to e.g. progression of the student through the course. **Formative** assessment, ("assessment-for-learning") includes **feedback** to help students improve their future performance.

Section 1 of this guide describes specific assessment methods as they map to Miller's pyramid and outlines practical considerations relating to their use. Section 2 provides updated summaries of key topics linked to, or directly involved in, assessment. We note that throughout the international assessment literature there are potential areas of confusion resulting from discipline specific or local use of terminology to describe certain types of assessment. It is therefore a further aim of this guide to provide some clarity in assessment terminology (a brief glossary of

terms is provided on p60).Throughout Section 1, we refer to the terms reliability and validity with the following definitions.

Reliability is defined as the reproducibility and accuracy of results – in assessment science this is often calculated as a reliability coefficient between 0 and 1. Two commonly reported measures of reliability are Cronbach's alpha and KR20 (Kuder–Richardson Formula 20). Reliability is now considered an important contributor to validity in that it determines the upper limit of the validity of an assessment. In this document the heading "reliability" refers to the scores that are obtained from that test, not the test itself.

Validity addresses the question of whether an assessment measures what it is supposed to measure. Validity has previously been considered as one of the characteristics of individual assessment instruments (Van Der Vleuten, 1996) within an 'assessment formula' which emphasises the other key factors that need to be considered i.e.:

Utility of an assessment = reliability x validity x educational impact x acceptability x cost.

We will discuss in Section 2 how recent concepts of validity are more detailed and encompassing such that validity replaces utility on the left hand side of this equation. Nevertheless, this formula neatly encapsulates the many factors that can influence decisions on assessment and also serves to highlight why decisions regarding programmes of assessment are heavily influenced by local context.

Blueprinting

One aspect of validity, which we highlight in this introduction to aid the considerations presented in Section 1, is that of content validity (often referred to as blueprinting). Blueprinting refers to the process of ensuring that assessment is a true reflection of the taught content. It can be performed simply using a spreadsheet to map assessment questions to the course content on a pro rata basis or using more complex curriculum mapping tools.

References and Further Reading

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 pre-doctoral dental education. *Journal of Dental Education,* 72, 1405-35.

Anderson, L. and Krathwohl, D. (2001). *A taxonomy for learning, teaching, and assessing*: A Revision of Bloom's Taxonomy of Educational Objectives 1st ed. New York: Longman.

Assessment in undergraduate medical education. (2009). 1st ed. [ebook] General Medical Council. Available at: http://www.gmc-uk.org/Assessment_in_undergraduate_medical_education_0211.pdf_48902978.pdf [Accessed 5 May 2015].

Benner P. (1982). From novice to expert. *The American Journal of Nursing,* 82(Cooney *et al.*), 402-7. Epub 1982/03/01.

Bennett, J., Freeman, A., Coombes, L., Kay, L. and Ricketts, C. (2010). Adaptation of medical progress testing to a dental setting. *Medical Teacher,* 32, 500-2.

Bloom, B. (1984). Taxonomy of Educational Objectives. In: D. McKay, ed., *The Cognitive Domain*, 1st ed. New York: Company Inc.

Dave, R.H. (1970). Psychomotor levels. In R.J. Armstrong (Ed.), Developing and writing educational objectives, 33-4. Tucson AZ: Educational Innovators Press.

Epstein R.M. (2007). Assessment in medical education.*The New England Journal of Medicine*. 356(4), 387-96. Epub 2007/01/26.

General Dental Council (2014) Preparing for practice. Dental team learning outcomes for registration Available from https://www.gdc-uk.org/newsandpublications/publications/publications/gdc%20learning%20outcomes.pdf [Accessed 12 May 2015]

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, 12-35.

Krathwohl D.R., Bloom B.S., Masia B.B. (1964). Taxonomy of educational objectives: the classification of educational goals: handbook II: affective domain: [S.l.]: Longman.

Miller, G. (1990). The assessment of clinical skills / competence/ performance. *Academic Medicine*, 65(9), 63-7.

Rethans J-J., Norcini J.J., Baron-Maldonado M., Blackmore D., Jolly B.C., LaDuca T., Lew, S., Page, G.G., and Southgate, L.H. (2002). The relationship between competence and performance: implications for assessing practice performance. *Medical Education*, 36(10), 901-9. Epub 2002/10/23.

Rhind, S., Baillie, S., Brown, F., Hammick, M. and Dozier, M. (2008). Assessing Competence in Veterinary Medical Education: Where's the Evidence? *Journal of Veterinary Medical Education*, 35, 407-11.

Schuwirth, L., Colliver, J., Gruppen, L., Kreiter, C., Mennin, S., Onishi, H., Pangaro, L., Ringsted, C., Swanson, D., Van Der Vleuten, C. and Wagner-Menghin, M. (2011). Research in assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. *Medical Teacher*, 33, 224-233.

Schuwirth, L. and Van der Vleuten, C. (2012). Programmatic assessment and Kane's validity perspective. *Medical Education*, 46(1), 38-48.

Schuwirth, L. and Van der Vleuten, C. (2013). How to design a useful test: The principles of assessment. In: Swanwick T. *Understanding medical education: evidence, theory and practice,* Oxford, Wiley-Blackwell, 18, 243-254.

Van der Vleuten, C. (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., Scheele, F., Driessen, E. W. and Hodges, B. (2010). The assessment of professional competence: building blocks for theory development. *Best Pract Res Clin Obstet Gynaecol,* 24, 703-19.

SECTION 1: ASSESSMENT METHODS

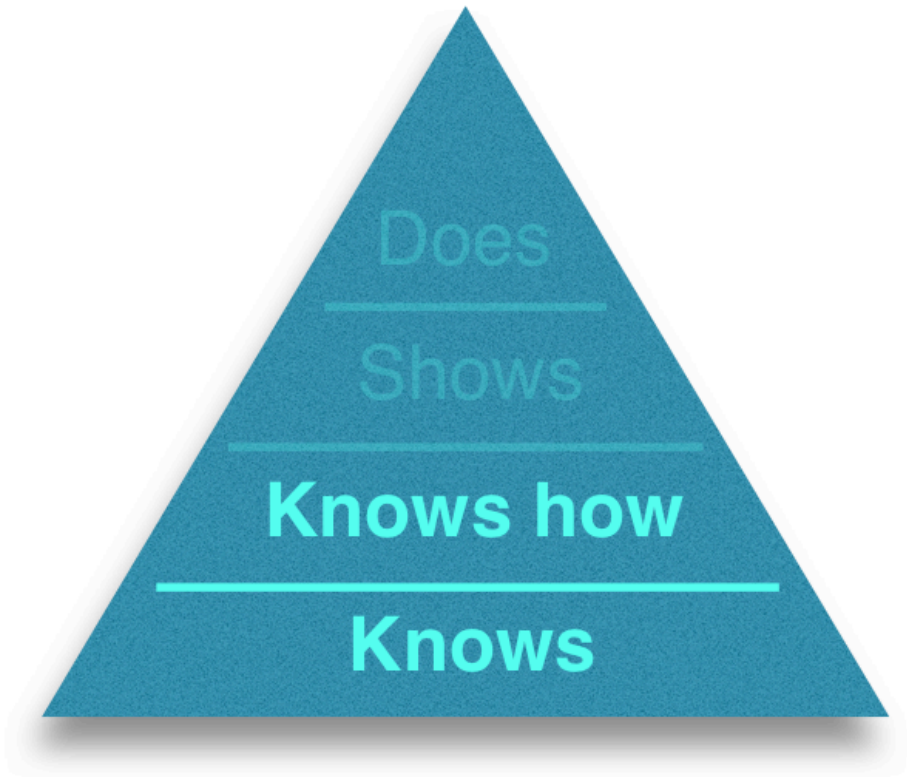
Authorship: Many of the following short summaries were originally written by Sarah Baillie and Susan Rhind for 'A Guide to Assessment in Veterinary Medicine' (2008) with updates in the 2nd Edition (2014) by Sarah Baillie, Susan Rhind and Sheena Warman. The current version of Section 1 has been updated and placed in the context of dental education by Julie Williams and Tony Ireland with input from the original authors.

Careful application of assessments permits testing the dental undergraduate or postgraduate student at different levels of Miller's pyramid. The first third of this section covers those assessments that test "knows" and "knows how", followed by those that test "shows" whilst the third part comprises tests for performance or "does".

Assessment methods always comprise a stimulus i.e. the task presented to the candidate and a response i.e. how the answer is captured (van der Vleuten *et al.* 2010). The task (or stimulus) may be written or practical and the response may be captured in many forms such as multiple-choice, essay or by direct observation and a checklist. Assessment methods are usually classified in terms of response, as below, although it is likely that the format of the task determines the validity of the assessment rather than the format of the response (van der Vleuten *et al.* 2010).

a) Miller's Pyramid 'Knows' and 'Knows How'

Focussing on the cognitive 'levels' of Miller's Pyramid, it is a reasonable aim to strive to examine at the 'Knows How' level, even in fixed response questions such as MCQs, and to use the cognitive domain of Blooms taxonomy (Figure 4, p8) at the highest levels where possible. These structural frameworks can help focus the minds of question authors on the thought processes they wish to examine when writing assessment questions.



Multiple Choice Questions (MCQs)

Knowledge Assessed: Depending on the question, this can range from "Knows → Knows How" levels of Miller's pyramid and "remembering → evaluating" levels of Bloom's taxonomy.

Description: MCQs are the most common written tests at all levels of medical education. The most commonly used template of MCQs consists of a lead-in question or statement (stem) followed by a list of options (usually five) from which the examinee selects one answer. At the most basic level, only one of the options is correct. At higher levels, examinees are asked to choose the 'best answer', with several options being potentially correct, but with one being a better match to the stem than the others (by a clear margin). This type of MCQ is called the Single Best Answer type or SBA.

There are other types of MCQ format and scoring including True/False, sentence completion, assertion reasoning, matching questions, negative marking (True-False-Abstain), elimination scoring and confidence scoring. However, these are no longer recommended for a number of reasons. These include the chance of answering correctly by guessing, and the tendency for these formats to test at the lower levels of Miller's pyramid and Bloom's taxonomy. These formats have largely been replaced by SBAs, and are consigned to the 'graveyard' in Case and Swanson's guide on Constructing Written Test Questions For the Basic and Clinical Sciences (Case and Swanson, 2002).

MCQs are used to test knowledge (factual recall) objectively and efficiently (computer-marked). They can be structured to test higher order skills and levels of cognition such as understanding, application of knowledge and evaluation of information particularly when the question stem takes the form of a clinical vignette. The tests can be used formatively (in-training) as an indicator of progress, as well as summatively. Writing items to test higher cognitive levels such as creating or evaluating is time consuming and may require non-content experts (Tractenberg *et al.* 2013).

MCQs are extensively used in dental undergraduate assessments, especially during pre-clinical years. MCQs, along with extended matching questions (EMQs) and short answer questions (SAQs) are used by some medical schools and at least one dental school for 'progress testing' - a longitudinal exam with regular sampling throughout the course (Bennett *et al.*, 2010). The improvement in students' scores can be used to monitor progress although the absence of a large data bank of questions, such as those that our medical colleagues have in their possession, may make this challenging to construct within a small faculty. The MCQ exam can be presented in a paper-based format or on-line : a mixed-methods study of dental undergraduate students' perceptions and performance concluded that online assessment was perceived as both fair and acceptable, even in high-stakes examinations (Escudier *et al.*, 2011). Computer-marking results in considerable savings in tutor marking time compared with the marking of free text responses. For a given amount of time, MCQs give a wider and therefore improved coverage of the examinee's knowledge of a subject area compared with other methods e.g. essays. A robust standard-setting process (see Section 2), whilst time-consuming, is essential.

Considerations:

Question Format. The MCQ format may encourage students to take a superficial approach to learning if a correct answer depends purely on factual recall rather than understanding. For improved authenticity in terms of testing clinical competence, it is preferable for SBAs to be based on clinical vignettes, requiring candidates to use

their knowledge base to make a diagnosis, or choose an appropriate investigation or treatment, thus engaging higher-order thinking (clinical decision making). The development of the large number of good quality test items required for an exam is time consuming.

Cueing. In MCQs, and similar exam formats, cueing effects can mean that examinees are able to eliminate wrong answers and recognise the correct answer, rather than needing to work out the answer. Questions should be designed to avoid cueing. Guidance on good MCQ question writing and how to avoid some of the common pitfalls is provided in 'Case and Swanson' (2002). High numbers of non-functioning distractors can lead to an over-estimation of student knowledge (McMahan *et al.*, 2013).

'Good Practice'. When writing questions, the first thing to do is establish the "testing point"; precisely which bit of knowledge or skill are you testing? The question must be clear, and not contain superfluous information. In most cases it should be possible to arrive at an answer without looking at the options (the "cover-up test"). All distractors (*i.e.* incorrect or unlikely options) should be homogeneous (*e.g*. all are muscles, diagnoses, drugs, *etc*.); plausible and attractive to the uninformed; similar to the correct answer in construction and length; and grammatically consistent and logically compatible with the stem. Try to avoid negatively phrased questions *e.g*. "which of the following statements is NOT TRUE" or "each of the following statements is correct EXCEPT"; this style of question inevitably fails the "cover-up test" and should only be used when there is no other way of addressing the testing point of the question. If unavoidable, ensure that the negative element is emphasised in the text either in upper case or bold typeface.

The Test-Wise Student. There are a number of ways a test-wise student can gain an advantage based on the way MCQs are written and there are several ways to minimise this effect. Avoid grammatical cues *e.g.* do all the answer options follow grammatically from the question? Avoid absolute terms such as "always" or "never" in answer options *(these are unlikely to be the correct answer and are ruled out by the test-wise student).* Avoid vague terms in the answer options *e.g*. "rarely", "usually". Is the correct answer obviously different to the rest *i.e*. correct answer is longer, more specific, or more complete than other options? Avoid word repeats, where a word or phrase is included in the question (stem) and in the correct answer. Beware convergence strategy where the correct answer includes the most elements in common with the other options.

Students with specific learning disabilities: One study has found that the format of certain types of MCQ does not entail systematic bias against learning-disabled medical students (Ricketts *et al.*, 2010).

Reliability: The reliability should be monitored with a target coefficient (Cronbach's alpha) in excess of 0.7 - 0.8.

Key Points:

- High reliability
- Computer marking saves time and resources
- Feedback often limited to overall score or score in different sections (due to question security)
- Easy to blueprint comprehensively to the syllabus
- Requires significant staff training and quality assurance
- Standard setting is time consuming

Example of Multiple Choice Questions

Case, S. and Swanson, D. (2002). Constructing Written Test Questions for the Basic and Clinical Sciences. 3rd ed. [ebook] Philadelphia: National Board of Medical Examiners. Available at: http://www.nbme.org/PDF/ItemWriting_2003/2003IWGwhole.pdf [Accessed 2 March 2015]

References and Further Reading

Anderson, J. (2004). Multiple-choice questions revisited. *Medical Teacher*, 26(2), 110-3.

Bennett, J., Freeman, A., Coombes, L., Kay, L. and Ricketts, C. (2010). Adaptation of medical progress testing to a dental setting. *Medical Teacher*, 32, 500-2.

Escudier, M. P., Newton, T. J., Cox, M. J., Reynolds, P. A. & Odell, E. W. 2011. University students' attainment and perceptions of computer delivered assessment; a comparison between computer-based and traditional tests in a 'high-stakes' examination. *Journal of Computer Assisted Learning,* 27**,** 440-447.

Jolly, B. (2014). Written assessment In: Swanwick T. *Understanding medical education: evidence, theory and practice,* Oxford, Wiley-Blackwell, 19, 264-265.
McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: a literature review. *Medical Teacher*, 26(8), 709-12.

Ricketts, C., Brice, J. and Coombes, L. (2010). Are multiple choice tests fair to medical students with specific learning disabilities? *Advances in Health Sciences Education: Theory and Practice,* 15, 265-75.

Tractenberg, R., Gushta, M., Mulroney, S. and Weissinger, P. (2013). Multiple choice questions can be designed or revised to challenge learners' critical thinking. *Advances in Health Sciences Education*, 18(5), 945-61.

Waugh, C.K. and Gronlund, N.E. (2013). 10th ed. Assessment of Student Achievement Boston, [Mass]; London, Pearson.

Extended Matching Questions (EMQs)

Knowledge Assessed: Depending on the question can range from "Knows → Knows How" levels of Miller's pyramid and "remembering → evaluating" levels of Bloom's taxonomy.

Description: The EMQ format has four components and starts with a title or theme statement defining the subject area *e.g*. 'Teeth:- Dates of eruption'. The title is followed by the list of 'options' or answers which may be numbered or lettered. There is then a lead in statement which provides instructions and links the list of answers (options) to the question/s (item/s), which may take the form of clinical vignettes or clinical images. The examinee has to respond to each question by selecting the best answer from a large list (range from 5 up to 20 or more), where one or more answers are potentially correct. Where there are several questions under one title, each answer can be used once, more than once or not at all. Ordering the list of answers alphabetically helps to minimise cueing. Usually 1 to 2 minutes is allowed per question.

Considerations: as for MCQs.

Reliability: The reliability should be monitored with a target coefficient (Cronbach's alpha) in excess of 0.7 - 0.8.

Key Points:

- Potentially higher reliability than MCQs
- Writing items that will test higher cognitive levels is time consuming
- Linked items can reduce the choice of topics and therefore reduce sampling across the curriculum
- Feedback is often limited to overall score or score in different sections (due to maintaining question security)
- Good discriminators at higher levels of ability

Example of Extended Matching Questions

Case, S. and Swanson, D. (2002). *Constructing Written Test Questions for the Basic and Clinical Sciences*. 3rd ed. [ebook] Philadelphia: National Board of Medical Examiners. Available from http://www.nbme.org/PDF/ItemWriting_2003/2003IWGwhole.pdf [Accessed 6 May 2014]

References and Further Reading

Beullens, J., Damme, B., Jaspaert, H. and Janssen, P. (2002). Are extended-matching multiple-choice items appropriate for a final test in medical education? *Medical Teacher*, 24(4), 390-5.

Jolly, B. (2014). Written assessment In: Swanwick T. *Understanding medical education : evidence, theory and practice,* Oxford, Wiley-Blackwell, 19, 265-267

van Bruggen, L., Manrique-van Woudenbergh, M., Spierenburg, E. and Vos, J. (2012). Preferred question types for computer-based assessment of clinical reasoning: a literature study. *Perspectives on Medical Education*, 1(4), 162-71.

Short-Answer Questions (SAQs)

Knowledge Assessed: Depending on the question can range from "Knows → Knows How" levels of Miller's pyramid and "remembering → creating" levels of Bloom's taxonomy.

Description: A written test consisting of a series of questions that require students to supply or formulate an answer rather than choose from a list of options (as in MCQs). The answer format is quite heterogeneous. At one end of the spectrum a short and quite specific answer is required *e.g*. one word (fill in the blank) or completion of a sentence. Alternatively, an SAQ may require the examinee to construct a short response (several sentences, a plan or a diagram) and in some contexts write a short or structured version of an essay. Questioning can be directed to test a specific objective or area. The question format may be based on a case scenario or set of data and may include additional information *e.g*. images. Sometimes several SAQs are written as a linked series covering a particular topic area. Compared to MCQ/EMQ, there is no cueing effect, as examinees are not presented with the correct answer amongst a number of other choices.

Considerations: Considerable resources are required for marking – this is mainly done 'by hand', although computer marking can be used for single word and short phrase answers. Basic factual knowledge is generally more efficiently examined using computer-based or computer-marked alternatives (MCQs, EMQs). Compared with essays, SAQs are easier to write and mark and have the potential to be more objective, although questions need to be worded carefully to elicit the desired answer. In linked SAQs, question design should ensure the examinee's progression through the answer is not blocked by an incorrect response early on. There is limited evidence that using SAQs may improve dental student academic achievement perhaps by creating a more challenging examination (Pinckard *et al.*, 2012).

Reliability: Reliability is affected by marker subjectivity with regard to what constitutes an acceptable answer, which is more of a problem the longer and less structured the answer format.

Key Points:

- Resource intensive marking compared to MCQ/EMQ
- Heterogeneity in interpretation of the term SAQ
- Reliability improved if structured marking schemes, clear outline answers and independent double scoring employed
- Has the advantage of no cueing effect
- Provision of written feedback possible but time consuming
- Context provided by the question

Example of short answer questions

Royal College of Edinburgh Membership in Orthodontics Written MSA question Example 2009 Available at http://www.rcsed.ac.uk/examinations/exam-details-page.aspx?calId=552e0372-ad67-4864-8b41-ce941f807f17&locid=1 [Accessed 6 November 2015]

References and Further Reading

Pinckard, R. N., Mcmahan, C. A., Prihoda, T. J., Littlefield, J. H. and Jones, A. C. (2012). Short-answer questions and formula scoring separately enhance dental student academic performance. *Journal of Dental Education* 76, 620-34.

Rademakers, J., ten Cate, T. and Bar, P. (2005). Progress testing with short answer questions. *Medical Teacher*, 27(7), 578-82.

Schuwirth, L. and Van der Vleuten, C. (2004). Different written assessment methods: what can be said about their strengths and weaknesses? *Medical Education*, 38(9), 974-79.

Essays

Knowledge Assessed: Depending on the question can range from "Knows → Knows How" levels of Miller's pyramid and "remembering → creating" levels of Bloom's taxonomy.

Description: An essay has been described as 'a short literary composition on a particular theme or subject, usually in prose and generally analytic, speculative, or interpretative.'[a] Sometimes also referred to as 'long answer' or 'extended answer' questions. A variation is the modified essay question, which may include *e.g.* an element of data handling. It should be clear to students whether the essay is being assessed / marked as a structured argument or is being used as a means of testing knowledge. For the latter, more efficient alternatives are preferable.

Considerations: Marking is labour intensive. Techniques to detect plagiarism should be considered. Not recommended for high stakes assessment in terms of either its ability to test higher order cognitive functioning or its validity (Hift, 2014).

Reliability: Reliability is low as sampling across content tends to be low due to lengthy testing time if a large number of essays are used. Essay marking is susceptible to rater (examiner) and candidate bias.

Key Points:

- Resource intensive marking
- Low reliability
- Double marking recommended to improve reliability
- Heterogeneity in interpretation of the word 'essay' which can be confusing for students and make comparison as a 'method' confusing.
- Provision of written feedback possible but time consuming
- Not recommended for high stakes assessment

References and Further Reading

[a]Dictionary.com, (2015). *Dictionary.com*. [online] Available at: http://dictionary.reference.com [Accessed 10 March 2015].

Hift, R. J. (2014). Should essays and other "open-ended"-type questions retain a place in written summative assessment in clinical medicine? *BMC Medical Education,* 14, 249.

Jolly, B. (2014). Written assessment In: Swanwick T. *Understanding medical education : evidence, theory and practice,* Oxford, Wiley-Blackwell, 19, 257-262

Palmer, E. J., and Devitt, P. G. (2007). Assessment of higher order cognitive skills in undergraduate education: Modified essay or multiple choice questions? *BMC Medical Education, 7*, 49. Available at http:// www. biomedcentral. com/ 1472-6920/ 7/ 49 [Accessed 9th March 2015]

Schuwirth, L. and van der Vleuten, C. (2004). Different written assessment methods: what can be said about their strengths and weaknesses? *Medical Education*, 38(9), 974-9.

Schuwirth, L. (2003). ABC of learning and teaching in medicine: Written assessment. *British Medical Journal*, 326(7390), 643-5.

Structured Oral/ Viva / Viva Voce

Knowledge Assessed: Depending on the question can range from "Knows → Knows How" levels of Miller's pyramid and "remembering → creating" levels of Bloom's taxonomy. Oral defence (*viva*) is often part of the assessment of a project or thesis. In this context, there is an important element of authenticating the work as belonging to a given student. If designed appropriately, this method can also be used to assess clinical reasoning and decision-making, oral communication skills and professionalism.

Description: The *viva* format involves the examinee being questioned by one or more examiners using an interview or discussion-like format typically to ascertain knowledge of a subject area or the ability to solve a clinical problem. This is followed by discussion and questioning aimed at probing the examinee's depth and breadth of knowledge, understanding, reasoning, and decision-making process. A *viva* can be used to explore ethical issues, assess professionalism, attitudes and communication skills. As with several other forms of assessment, there is considerable variation in the format and use of this type of assessment, with standardised content and structure (Structured Viva) more commonly used with the aim of reducing bias.

Considerations: If used as part of routine examinations for all students, the time and resources required are considerable. This is even more of a problem when the number of questions or cases presented is increased (as one way of trying to improve reliability). *Vivas* (as well as other one-to-one encounters) can be subject to "Halo effects" *i.e.* the effect whereby a judgement on one aspect is influenced by an overall impression of the person or where the judgement is influenced by the performance of previous candidates in contrast to the current candidate.
These issues mean that the use of oral examinations in any form of high stakes assessment setting is not recommended.

Reliability: Reliability is often low due to a lack of standardisation of questioning and marking, and the possibility of examiner bias (use of favoured and / or irrelevant questions), and 'halo effects'. Reliability can be improved by using the same questions for all students (but this will require corralling to prevent later candidates being advantaged), a structured marking system, examiner training, increasing the number of *vivas* per examinee and having a total minimum testing time of four 20-minute oral examinations, each with two examiners (Wass *et al.* 2003).

Key Points:

- Heterogeneity in interpretation of the method
- Low reliability unless multiple examiners, multiple cases and large testing time
- Often seen as having high authenticity to examiners
- Needs careful planning of subject matter and appropriate coverage of the syllabus
- Resource intensive
- Immediate face to face feedback can be built in to the process
- A good exam case should involve a relatively common presenting problem with several plausible diagnoses and should primarily test the student's problem solving skills

References and Further Reading

Davis, M. and Karunathilake, I. (2005). The place of the oral examination in today's assessment systems. *Medical Teacher*, 27(4), 294-7.

Muzzin, L. and Hart, L. (1985). Oral Examinations. In: V. Neufeld and G. Norman, ed., *Assessing Clinical Competence*, 1st ed. New York: Springer Publishing Company, 71-93.

Ryding, H. A. and Murphy, H. J. (1999). Employing oral examinations (viva voce) in assessing dental students' clinical reasoning skills. *Journal of Dental Education,* 63, 682-7.

Wass, V., Wakeford, R., Neighbour, R. and van der Vleuten, C. (2003). Achieving acceptable reliability in oral examinations: an analysis of the Royal College of General Practitioners membership examination's oral component. *Medical Education*, 37(2), 126-31.

The 'Spotter' Test

Knowledge Assessed: Depending on the question can range from "Knows → Knows How" levels of Miller's pyramid and "remembering → creating" levels of Bloom's taxonomy.

Description: This format has been a traditional assessment format in many UK dental schools, particularly for examination of disciplines such as anatomy and pathology. However, the format is increasingly being replaced, in part or completely, by computerised assessments using high quality images. Various local terms are used to describe this type of assessment including 'Spot', 'Steeplechase', 'Timed stations' or 'Bell-ringer'. However, there are few references in the literature to the method and it should not be confused with methods assessing at the 'Shows' level of Miller's pyramid. The format usually has examinees moving around a series of stations consisting of *e.g*. a specimen, a labelled dissection or radiograph. The answer may be one word or involve a response that requires some level of deduction or diagnostic skill *i.e*. similar to that described under the category of short answer questions. As for SAQs, therefore, the same reliability issues exist, which can be improved using structured marking schemes.

Considerations: Resources required to set up the stations, run the exam and marking. Often the same knowledge could be tested more efficiently and reliably by using images within an MCQ or SAQ test.

Reliability: Reliability will be compromised if the number of items is small and when marking is not structured.

Key Points:

- Has been in common use but is being replaced by computerised assessment and marking (where possible)
- Little published in literature on description
- Heterogeneity in interpretation of the term
- Reliability improved if structured marking schemes employed
- Provision of written feedback possible but time consuming
- Consider using images within more reliable and evidence-based forms of assessment

References and Further Reading

Note: Literature searching to date for further information on this method has found no specific papers related to this type of test.

Script Concordance Test (SCT)

Knowledge Assessed: Depending on the question can range from "Knows → Knows How" levels of Miller's pyramid and "remembering → evaluating" levels of Bloom's taxonomy.

Description: Designed to assess decision-making and clinical reasoning skills. In everyday work experienced clinicians, largely sub-consciously, refer to 'scripts' when using their knowledge to make decisions. These scripts are built up over years in clinical practice and can be about diagnosis, investigation or treatment. The SCT investigates the organisational structure of an examinee's knowledge when presented with a situation where a decision needs to be made using information or data about a clinical case.
The SCT is a written exam that starts with a clinical scenario or vignette that summarises the case (*e.g*. a patient presents with dental pain). This is followed by a proposed diagnosis or suggested treatment or action. Examinees have to rate the effect of further information (*e.g*. the pain gets worse with hot and better with cold) or findings on the probability of the diagnosis / treatment being: more certain / likely, unchanged or less certain / likely, using a 5-point scale. The answers are compared to those of a panel of experts. The marking system usually takes into account the variation in expert opinion, with answers being weighted accordingly *i.e*. an answer the same as the majority of experts scores highest but answers that correspond to those chosen by some experts still receive some credit. Alternatively, there is an agreed single best answer.

Considerations: Considerable time and practice required to develop suitable test items. The numbers of experts required make this a challenging format to develop in dentistry but an example of its use in restorative dentistry found it to be reliable in distinguishing experts from those in training (Hahn P. *et al.*, 2012). This assessment has some similarities with the Situational Judgement Test used in medicine and job-selection interviews. These tests may be used to test knowledge or knowledge of what should be done rather than as a measure of what will be done or actual performance. The situational judgement test also comprises a clinical or management scenario but then provides several options which need to be ranked as the most appropriate action in response to the details provided.

Reliability: High if there are sufficient questions. Formulation of up to 5 questions per case has been shown to be an efficient way to optimize the reliability of SCT score (Gagnon *et al.* 2009).

Key Points:

- Written test based on scenario / vignette
- Significant training required for item writing
- Considerable time needed to produce each question
- Good reliability if sufficient questions are used (Meterissian *et al*. 2007)
- Requires suitably qualified panels of examiners to produce scoring system
- Feedback often limited to overall score or score in different sections (due to question security)

Example of a Script Concordance Test

Charlin, B., Roy, L., Brailovsky, C., Goulet, F. and van der Vleuten, C. (2000). The Script Concordance test: a tool to assess the reflective clinician. *Teaching and Learning in Medicine*, 12(4), 190.

References and Further Reading

Charlin, B., Roy, L., Brailovsky, C., Goulet, F. and van der Vleuten, C. (2000). The Script Concordance test: a tool to assess the reflective clinician. *Teaching and Learning in Medicine*, 12(4), 189-95.

Gagnon, R., Charlin, B., Lambert, C., Carriere, B. and van der Vleuten, C. (2009). Script concordance testing: more cases or more questions? *Advances in Health Sciences Education*, 14, 367-75.

Hahn P., Kapaun C., Raden L., Fabry G. and Raden, L. (2012). Development of a Script Concordance Test to Assess Competence in Clinical Data Interpretation in Restorative Dentistry *Poster presentation, 38th Annual Meeting of the Association for Dental Education in Europe.* Lyon, France: Blackwell Publishing.

Jolly, B. (2014). Written assessment In: Swanwick T. *Understanding medical education : evidence, theory and practice,* Oxford, Wiley-Blackwell, 19, 267-269.

Lubarsky S., Dory V., Duggan, P., Gagnon, R. and Charlin, B. (2013). Script concordance testing. AMEE Guide No. 75. *Medical Teacher,* 35, 184-93.

Meterissian, S., Zabolotny, B., Gagnon, R. and Charlin, B. (2007). Is the script concordance test a valid instrument for assessment of intraoperative decision-making skills? *The American Journal of Surgery*, 193(2), 248-51.

Triple Jump Exercise (TJE)

Knowledge Assessed: Depending on the question can range from "Knows → Knows How" levels of Miller's pyramid and "remembering → applying" levels of Bloom's taxonomy.

Description: Originally developed at McMaster University in 1974 as a method of testing the critical thinking abilities of the individual on an undergraduate programme with an emphasis upon group work and problem-based learning (PBL). Each stage (jump) of the assessment is scored to produce a cumulative score. There are several variations:

i) Clinical TJE comprises a written patient history and examination, followed by a write-up of the findings in the format of subjective data, objective data, assessment and plan (SOAP) to include evidence from the literature (the second "jump"). The write-up is submitted to the member of the faculty who observed the first "jump" (i.e. history and examination) and this same faculty member also conducts the third "jump", an oral examination of the student to cover the pathophysiology, diagnosis, and treatment of the patient including appropriate research evidence.

ii) Pre-clinical TJEs uses a written patient scenario for which the student must identify key issues and write a research question in the PICO format (Problem, Intervention, Comparison and Outcome). Researching the literature forms the second "jump" with a report of their findings including the answer to the research question and a critical appraisal of the quality of the evidence to complete the third "jump".

Skills assessed: The individual's ability to search for evidence to support their clinical practice with appropriate analysis and application to health care problems *i.e*. reasoning and learning skills, two of the four "problem-based learning process skills. " The other two PBL skills, namely group work and feedback, are not tested in the TJE.

Considerations: The original format of this assessment relies on one assessor working with the student throughout the process in order to provide feedback on each "jump" which is resource intensive. It has been used for assessment of dental students in a small scale study where it met with student satisfaction (Navazesh *et al.*, 2014).

Reliability: Initially found to be poor in its original form since it was totally oral and graded subjectively, in a similar manner to other *viva* examinations. There is limited evidence in the literature of adequate reliability or validity, although the training of assessors is seen as essential. The subjectivity inherent in the examination can be reduced using standardisation of the examination and use of set criteria for the first step of the assessment (Smith, 1993).

Key points:

- Requires a significant amount of faculty time to conduct and score
- Student performance may be case dependent
- More experienced students have been successfully used as examiners
- No evidence to confirm whether a good performance in this examination will be indicative of a similar approach in the work place
- Usually used as a formative rather than summative assessment

Example of Triple Jump Exercise

Availlable online at http://www.iupui.edu/~idd/active_learning/5_17d.html [Accessed 6 June 2015]

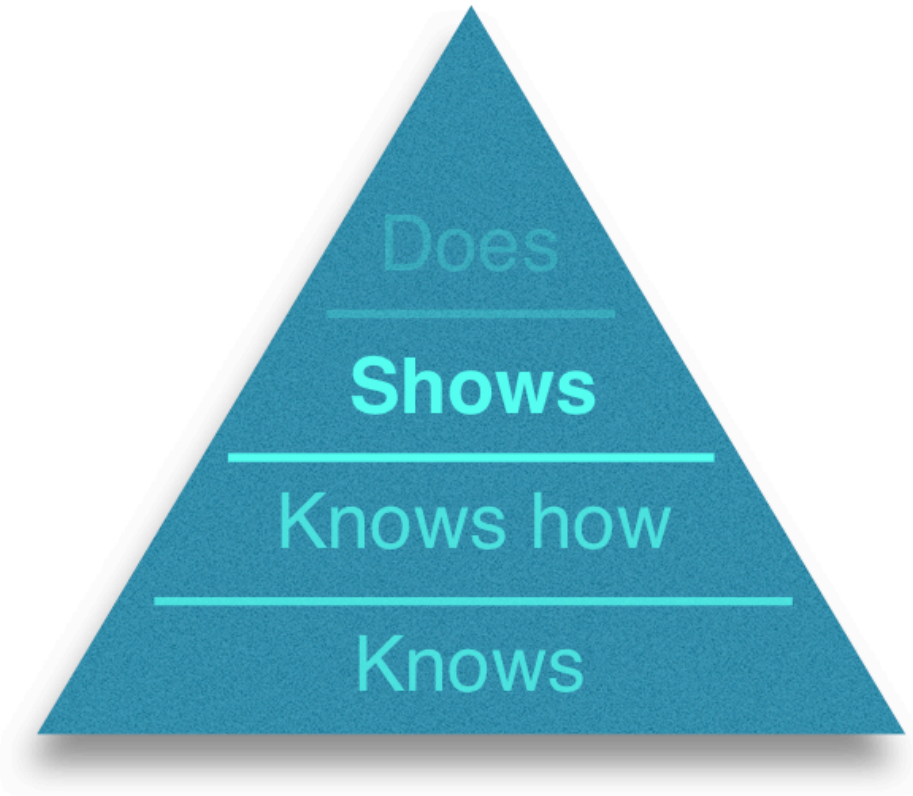
References and Further Reading:

Navazesh, M., Rich, S. K. and Keim, R. G. (2014). Triple jump examination evaluation of faculty examiners by dental student examinees. *Journal of Dental Education,* 78, 714-22.

Smith, R. M. (1993). The triple-jump examination as an assessment tool in the problem-based medical curriculum at the University of Hawaii. *Academic Medicine*, 68, 366-72.

b) Miller's Pyramid 'Shows'

The 'Shows' level of Miller's pyramid can be considered as assessing practical ability or competence when performing a task '*in vitro'* . It may also be an opportunity to assess the student dentist's learning objectives from the psychomotor domain or skills (see Figure 6, p10). Whilst several skills can be performed together at the "shows" level (articulation), careful assessment is required to ensure that deficiency in one skill (*e.g.* cavity preparation) is not masked by proficiency in another (*e.g*. shaping or carving occlusal anatomy). It is also important to remember that there will be essential underpinning of the "shows" level with knowledge.



Objective Structured Clinical Examination (OSCE)

Description: The Objective Structured Clinical Examination (OSCE) was introduced in medical education nearly 40 years ago as a more standardised, objective and reliable way of assessing certain clinical skills (Harden *et al.* 1975) and is now in widespread use.
The exam consists of multiple mini-stations (typically 10 to 20), which the examinees rotate round in sequence, completing a variety of tasks. Each station in the circuit lasts the same amount of time; from 5 to 6 minutes for basic practical skills *e.g*. gloving and up to 20 minutes when embracing multiple aspects of a patient interaction *e.g*. history taking, physical examination, diagnosis and treatment plan. The examinee reads the scenario, then enters the station and undertakes the task. The OSCE is now widely adopted in dental education and typical stations test *e.g*. history taking (gathering information from a client), dental charting, recording a patient's blood pressure, taking consent for a procedure or scoring the index of orthodontic treatment need (IOTN) from a set of study models. The station set-up varies and can include: simulated patients, models, part-task trainers, laboratory equipment and simulated work stations. The selection of stations should be representative of, and mapped (blueprinted) to, the taught course. With the more holistic OSCE (15 to 20 minute patient interactions) blueprinting needs to consider several dimensions of competence within each station including stages in a clinical case and body systems.

Marking: Detailed Checklists
Originally, OSCEs were marked using a very detailed checklist often with 15 to 25 items that the examinee did or did not complete / undertake. Each item can be equally weighted *i.e*. 1 or 0, although some critical steps (*e.g*. fatal errors, a break in sterility, *etc*.) may carry a heavier weighting (more marks) or be a requirement to pass the station. The pass mark is usually calculated via a borderline regression or borderline group method using both the checklist score and the global rating (see Boursicot *et al.* 2007).

Marking: Global Rating Scales (GRS)
Checklists are usually accompanied by a global rating scale for the examiner to make a more subjective judgement (selecting one of 4 to 7 categories with descriptors across the spectrum from a bad fail to an excellent pass).

Traditionally, detailed checklists were considered to be more objective and reliable than global rating scales. However, research has challenged this view and there is evidence that GRS are more reliable and able to measure increasing levels of expertise (Cunnington *et al.* 1996, Regehr *et al.* 1998, Hodges and McIlroy 2003). Thus, in recent years, GRS have grown in popularity.

Skills Assessed: Clinical practical, technical and diagnostic skills, treatment planning, and communication skills.

Considerations: Considerable resources are required (costs of equipment, consumables, and personnel / staff time) to develop and set up the stations, and to run the OSCE. However, the checklists and rating scales can be computer marked. When developing OSCE stations a team is required to write the scenarios and itemised checklists. Examiner training and briefing sessions are also important. Early implementation of this type of testing was shown in a small-scale study to change dental students' learning strategies. This led to greater achievements of specific clinical competencies when students were advised that an exam would be OSCE style rather than conventional essays (Schoonheim-Klein *et al.* 2006)

Reliability: Reliability is usually high if there are enough stations and 15 to 20 stations are recommended (Brown *et al.*, 1999). However, examiners need to be trained and station and inter-rater (examiner) reliability should be monitored. The exam is fair and objective as the same scenarios are presented to all examinees and the same marking criteria are applied.

Key Points:

- High reliability compared to a few long cases or individual clinical examinations
- Potential to compromise validity by excessively deconstructing tasks
- Resource intensive to establish, set up and run
- Can provide detailed specific feedback

Example of OSCE Stations

RCPSG MFDS examination examples of OSCE stations. [online] Available at: https://www.rcpsg.ac.uk/~/media/Files/Examinations/Dentists/MembershipOfTheFacultyOfDentalSurgeryParts1and2/Sample%20Questions%20Part%202%20OSCE.pdf [Accessed 2 March 2015].

References and Further Reading

Boursicot, K., Roberts, T. and Pell, G. (2007). Using borderline methods to compare passing standards for OSCEs at graduation across three medical schools. *Medical Education*, 41(11),1024-1031.

Boursicot, K., Roberts, T. and Burdick, W. (2014). Structured assessment of clinical competence. In: Swanwick T. *Understanding medical education: evidence, theory and practice,* Oxford, Wiley-Blackwell, 21, 293-303.

Brown, G., Manogue, M. and Martin, M. (1999). The validity and reliability of an OSCE in dentistry. *European Journal of Dental Education,* 3, 117-25.

Cunnington, J., Neville, A. and Norman, G. (1996). The risks of thoroughness: reliability and validity of global ratings and checklists in an OSCE. *Advances in Health Sciences Education*, 1, 227-33.

Harden, R., Stevenson, M., Downie, W. and Wilson, G. (1975). Assessment of clinical competence using objective structured examination. *British Medical Journal*, 1(5955), 447.

Hodges, B. (2006). The objective structured clinical examination: Three decades of development. *Journal of Veterinary Medical Education*, 33(4), 571-577.

Hodges, B. and McIlroy, J. (2003). Analytic global OSCE ratings are sensitive to level of training. *Medical Education*, 37(11), 1012-1016.

Khan K.Z., Gaunt, K., Ramachandran S. and Pushkar, P. (2013). The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part I: an historical and theoretical perspective. *Medical Teacher,* 35(9), 1437-46. Part II: organisation and administration. *Medical Teacher,* 35(9), 1447-63.

Ma, I., Zalunardo, N., Pachev, G., Beran, T., Brown, M., Hatala, R. and McLaughlin, K. (2012). Comparing the use of global rating scale with checklists for the assessment of central venous catheterization skills using simulation. *Advances in Health Sciences Education*, 17(4), 457-470.

Norman, G., van der Vleuten, C. and De Graaff, E. (1991). Pitfalls in the pursuit of objectivity: issues of validity, efficiency and acceptability. *Medical Education*, 25(2), 119-126.

Regehr, G., MacRae, H., Reznick, R. and Szalay, D. (1998). Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Academic Medicine*, 73(9), 993-7.

Schoonheim-Klein, M. E., Habets, L. L., Aartman, I. H., 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, 226-35.

Van der Vleuten, C., Norman, G. and De Graaff, E. (1991). Pitfalls in the pursuit of objectivity: issues of reliability. *Medical Education*, 25(2), 110-118.

Practical Test in Simulated Clinical Setting or Laboratory

Description: The student undertakes a defined piece of practical work (*e.g*. Class II resin composite restoration) in the absence of a patient. This usually involves a phantom head/dental manikin and synthetic teeth, although may also involve the construction of laboratory items such as dentures, crowns or inlays using a variety of materials. Assessment may be at each stage of the process (*e.g*. caries removal, cavity preparation, matrix band placement, restoration placement, finishing) or for the overall result. Marking is usually by an expert clinician using a check list of the required standards with a global rating such as ideal, satisfactory, borderline or unsatisfactory. In this way it is similar to an OSCE but the students are all usually undertaking the same test rather than a series of stations. An alternative method of marking is to use expert opinion to inform software development for haptic simulators which can then provide either summative or formative feedback for the practical test.

Skills assessed: Manual dexterity and handling of dental materials in a quasi-clinical setting such as a clinical skills laboratory. The global rating reflects that the outcome is underpinned by an appropriate level of knowledge and professionalism which may be tested during the examination with structured questioning and observation, for example as part of the Overseas Registration Examination (ORE). High frequency feedback and repetitive skills training has been found to improve early procedural skills in medical students (Bosse *et al.* 2015)

Considerations: It is time consuming to construct suitable marking criteria and requires marking by a trained clinician. An example of a marking sheet for this type of assessment is referenced below.

Reliability: Inter-examiner variability has the potential to influence the reliability of these tests and methods to minimise this include clinical checklists, examiner training, and the use of multiple examiners. Haptic simulators have been used to successfully predict preclinical operative dentistry performance (Urbankova *et al.*, 2013) and have been shown to successfully distinguish between experts and novice participants in endodontics (Suebnukarn *et al.*, 2014).

Key points:

- Clinical checklists, examiner training and the use of multiple examiners improve reliability but will increase the resource requirements
- The scores for each student from two or more examiners can be compared to check correlation and hence reliability of scoring for a particular assessment

Examples of a Practical Examination and assessment criteria

Example of a practical examination can be found at:
http://www.orepart2.org.uk/images/GDC/PDFs/dmguidance.pdf [Accessed on 16 March 2015]

Example of assessment criteria for a practical examination can be found at:
http://www.adc.org.au/documents/Practical%20Examination%20Handbook%202015%20Jan_v2.pdf [Accessed on 19 March 2015]

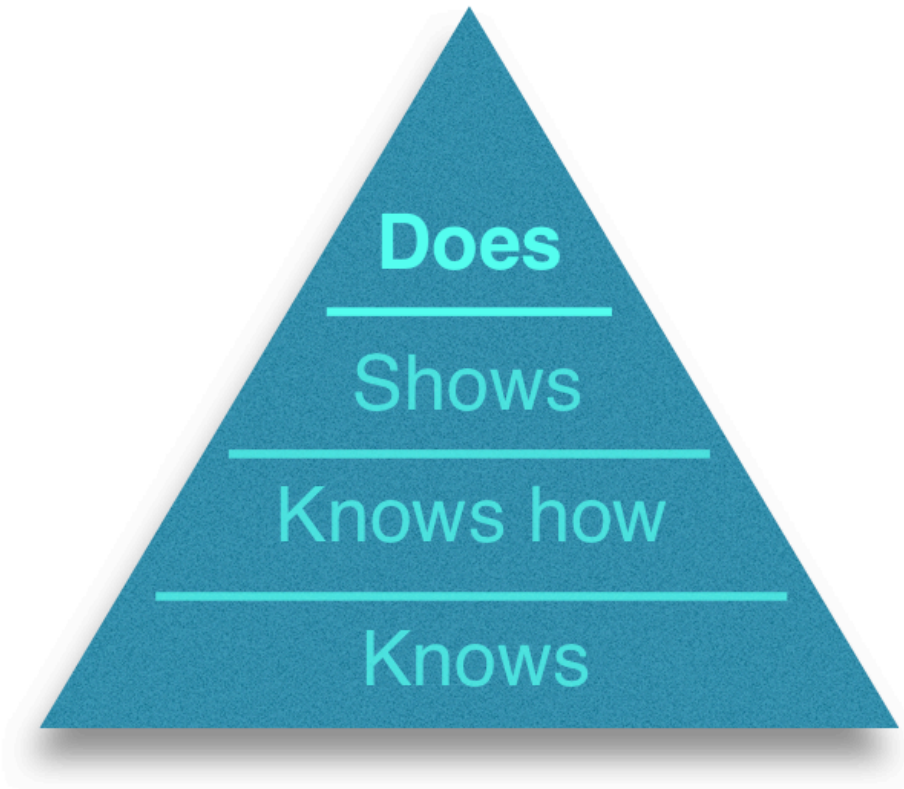
References and Further Reading:

Bosse H.M., Mohr J., Buss B., Krautter M., Weyrich P., Herzog W., Jünger J., and Nikendei C. (2015). The benefit of repetitive skills training and frequency of expert feedback in the early acquisition of procedural skills. *Medical Education,* 15(1), 286.

Suebnukarn, S., Chaisombat, M., Kongpunwijit, T. and Rhienmora, P. (2014). Construct validity and expert benchmarking of the haptic virtual reality dental simulator. *Journal of Dental Education,* 78, 1442-50.

Urbankova, A., Eber, M. and Engebretson, S. P. (2013). A complex haptic exercise to predict preclinical operative dentistry performance: a retrospective study. *Journal of Dental Education*, 77, 1443-50.

c) Miller's Pyramid 'Does'



Assessment at the top level ('Does') of Miller's pyramid is often seen as the holy grail of clinical assessment. In contrast to performance assessment *'in vitro'* discussed in the previous section, assessment at this level can be considered as performance assessment '*in vivo' i.e*. in the workplace.

Whilst some elements of Bloom's affective domain (Figure 5, page 9) can be observed during assessment at the "Show" level of Miller's pyramid, attainment at the upper levels of the affective domain are often assessed in the workplace setting. For example, "putting patient's interests first" (6.1, GDC Preparing for practice) may need the dentist to try to restore a useful tooth with a lengthy complex procedure such as endodontics rather than remove the tooth which could be simpler and more cost-effective, especially in the short-term. Whilst the dentist may express this value formally within a written examination, it is more important that their action once within the workplace reflects these values. A well-trained supportive dental team can easily evaluate how much value the dentist places upon the patient's interests. Similarly a patient who unexpectedly attends the practice suffering with dental pain may need intervention at the end of a busy clinic. The dentist who is observed undertaking the intervention despite the lateness of the hour will be more likely to be perceived as placing the patient's interests first and acting in a professional manner. Defining which element of professional values is to be tested and how best to measure this element informs the current literature regarding the development of a professional identity.

It is important to recognise that there is no single preferred method of measuring professionalism, which is an essential component of assessing performance at this level of the pyramid. Indeed, 9 different clusters of assessment tools have been described within medical education that have relevance for

assessment in this area (Wilkinson *et al.* 2009). It therefore follows that the use of multiple methods of assessment is desirable to allow 'triangulation' of information assessing different aspects of professional behaviour (van Mook *et al.* 2009). There has been increased use of workplace-based assessment tools within health care although comprehensive training in their use has been shown to be essential for maximal efficacy (Kirton *et al.*, 2013). A systematic review of the impact of work-based assessments upon doctors' performance and education found some subjective reporting of educational benefits but no objective evidence of an improvement in doctors' performance (Miller and Archer, 2010). Furthermore it remains untested whether allowing sufficient time for a relationship of trust to develop between the assessor and the dentist-in-training could have an impact on the educational value to be gleaned from the exercise.

Workplace-based assessment tools can be utilised during both undergraduate and postgraduate programmes. It should be noted that postgraduate dental training in the UK has changed in recent years. Competences are now organised within four interlinked domains: clinical knowledge, communication, professionalism, and leadership and management (Kalsi *et al.*, 2013). Historically, newly qualified dentists were informally mentored through the early years of general dental practice by, in particular, practice owners. General dental practitioners initiated and paid for the first voluntary vocational training scheme through income top-slicing which has now evolved via vocational dental training to become mandatory dental foundation training. This is mandatory to acquire a "performer" number required to work within the National Health Service and dentists can still work privately without this number. The newly qualified dentist pairs with a trainer, usually within a dental practice, who is appointed by the Deanery. The trainer role has evolved from clinical mentorship to take on the responsibilities for both assessment and evaluation of the Foundation Dentist, often using work-place assessment tools. The document outlining the curriculum for postgraduate dental foundation training envisages assessment to be at the level of whole performance and states:

''It is recognised that each foundation dentist begins Dental Foundation Training with their own individual strengths, weaknesses and levels of prior experience with respect to practice. Whilst the validity of assessment will require that the cases assessed within the workplace are representative of all major areas of the competency framework, a prescribed 'list' of areas for assessment is not provided" (COPDEND 2015, p5).

References

Committee of Postgraduate Dental Deans and Directors (COPDEND) Dental Foundation Training Curriculum 2015 [Draft as at November 2015] Available at www.copdend.org. [Accessed 26 October 2015]

Kalsi, H. K., Kalsi, J. S. and Fisher, N. L. (2013). An explanation of workplace-based assessments in postgraduate dental training and a review of the current literature. *Br Dent J*, 215, 519-24.

Kirton, J. A., Palmer, N. O., Grieveson, B. and Balmer, M. C. (2013). A national evaluation of workplace-based assessment tools (WPBAs) in foundation dental training: a UK study. Effective and useful but do they provide an equitable training experience? *British Dental Journal,* 214, 305-9.

Norcini, J. (2014). Workplace assessment In: Swanwick T. *Understanding medical education : evidence, theory and practice,* Oxford, Wiley-Blackwell, 20, 279-392.

Miller, A. and Archer, J. (2010). Impact of workplace based assessment on doctors' education and performance: a systematic review. *British Medical Journal,* 341, c5064.

Prescott, L. E., Norcini, J. J., Mckinlay, P. and Rennie, J. S. (2002). Facing the challenges of competency-based assessment of postgraduate dental training: Longitudinal Evaluation of Performance (LEP). *Medical Education,* 36, 92-7.

van Mook, W., van Luijk, S., O'Sullivan, H., Wass, V., Schuwirth, L. and van der Vleuten, C. (2009). General considerations regarding assessment of professional behaviour. *European Journal of Internal Medicine*, 20(4), 90-95.

Wilkinson, T., Wade, W. and Knock, L. (2009). A blueprint to assess professionalism: results of a systematic review. *Academic Medicine*, 84(5), 551-8.

Mini-Clinical Evaluation Exercise (Mini-CEX)

Description: Direct observation of a trainee by one examiner during a clinical encounter with a real patient in the normal work setting *e.g.* on a ward or in a dental clinic. The mini-CEX evolved from the original clinical evaluation exercise (CEX), developed to replace *vivas* for the assessment of clinical competency. The CEX is no longer used since its focus on a relatively long (typically 2 hour) pre-planned single patient encounter in a clinical setting immediately causes problems in terms of assessment reliability and resources.
In the mini-CEX, the "snap-shot" observation lasts 15 to 20 minutes and is followed by immediate feedback from the examiner. Typically, multiple mini-CEXs are used with a variety of patients. The observation is marked using a standardised tick box form that is used to record information about the case, setting, trainee and examiner (for an example of a marking sheet see Norcini, 2005). Performance is rated for a list of skills as: at, above or below expectation. This is usually against the standard expected for certification, but may be also against the standard of the student's current year of training. Mini-CEX are primarily used formatively with feedback to produce an action plan that is structured to support the trainee's learning.

Skills Assessed: History taking, physical examination, diagnostic, clinical judgement, decision making, communication and time management.

Considerations: With a certain amount of planning, the mini-CEX is feasible and can be fitted into routine clinical training. The patient/s chosen should be typical of the trainee's case load.

Reliability: Reliability increases with the number of encounters with 6 to 8 mini-CEXs giving acceptable reliability. Assessor training is also important for reliability and to improve the quality of feedback.

Key Points:

- High authenticity
- Reliability increases with number of examinations (mini-CEXs) performed
- Having an assessor who is also a trainer may compromise the trainer's role
- Students may be reticent in requesting assessments that could highlight training needs if they form part of a gateway process for progression

Example of mini-CEX form

Available online at http://www.uccdentsoc.com/wp-content/uploads/2013/06/DFT_Information.pdf [Accessed 2 June 2015]

References and Further Reading

Norcini, J. (2005). The Mini Clinical Evaluation Exercise (mini-CEX). *The Clinical Teacher*, 2(1), 25-30.

Directly Observed Procedural Skills (DOPS)
Also previously known as Structured Clinical Operative Tests (SCOTs)

Description: Directly Observed Procedural Skills (DOPS), also referred to as Direct Observation of Procedural Skills, are designed specifically to assess practical skills in a workplace setting. A student is observed and scored by an assessor while performing a routine practical procedure during his / her normal clinical work. The assessor uses a standard DOPS form to score the technique (for an example of a DOPS form see: Wilkinson *et al.* 2008). The student is deemed either competent or incompetent. For any particular skill the student usually has to pass a number of repeated assessments (typically six) to be signed off as competent at that skill. Alternatively the student may also request this type of assessment when they have judged that they have developed the required level of competency. DOPS have been used as part of peer assessment where dental students have been able to detect improvement in the performance of their peers over time (Tricio *et al.*, 2014).

Skills Assessed: Practical / technical ability and manual dexterity in a work place setting.

Considerations: DOPS are run during normal clinical work and, with a certain amount of planning and organisation, this represents a feasible way of assessing the key procedures and practical skills required for particular disciplines / specialties. They have been used in undergraduate settings but there have been some concerns that there are insufficient suitable patients within a dental school environment for this type of testing. This has been thought to be partly due to patients being referred for more complex problems and partly due to accessibility issues such as car parking (Blinkhorn, 2002). This assessment has therefore been conducted with appropriately trained assessors within community clinics, which could be seen to increase the authenticity of this assessment.
Students can request to take these examinations when they feel ready. A small study of dental students has shown that those who choose when to take practical tests have been found to have better results than those who set the test on an assigned date (Bakker *et al.* 2015 - Poster at ADEE to be referenced when published).

Reliability: Use in medical specialties indicates that six observations *i.e*. DOPS exams are required per procedure for a reasonable level of reliability.

Key Points:
- High authenticity
- Multiple assessments of the same skill
- Present a valuable opportunity for formative feedback with written marking sheet
- Criterion referenced marking
- Emphasis upon testing psychomotor skills
- Resource intensive to conduct the assessment and need suitable cases

Example of a DOPS form

Available online at https://www.iscp.ac.uk/static/public/sdops_form.pdf

References and Further Reading

Blinkhorn, F. A. (2002). Evaluation of an undergraduate community-based course in Family Dentistry. European Journal of Dental Education, 6, 40-4.

Durham, J. A., Moore, U. J., Corbett, I. P. and Thomson, P. J. (2007). Assessing competency in dentoalveolar surgery: a 3-year study of cumulative experience in the undergraduate curriculum. *European Journal of Dental Education,* 11, 200-7.

Magnier, K., Dale, V. and Pead, M. (2012). Workplace-Based Assessment Instruments in the Health Sciences. *Journal of Veterinary Medical Education*, 39(4), 389-95.

McLeod, R., Mires, G. and Ker, J. (2012). Direct observed procedural skills assessment in the undergraduate setting. *The Clinical Teacher*, 9(4), 228-32.

Rolland, S., Hobson, R. and Hanwell, S. (2007). Clinical competency exercises: some student perceptions. *European Journal of Dental Education*, 11, 184-91.

Tricio, J., Woolford, M., Thomas, M., Lewis-Greene, H., Georghiou, L., Andiappan, M. and Escudier, M. (2014). Dental students' peer assessment: a prospective pilot study. *European Journal of Dental Education*. doi: 10.1111/eje.12114.

Wilkinson, J., Crossley, J., Wragg, A., Mills, P., Cowan, G. and Wade, W. (2008). Implementing workplace-based assessment across the medical specialties in the United Kingdom. *Medical Education*, 42(4), 364-73.

Dental Evaluation of Performance Test (ADEPT)

Description: A combination of DOPS and mini-CEX currently in use as part of UK dental postgraduate foundation training, for which it was specifically designed. The evaluator observes the foundation dentist's performance within a patient encounter and rates their performance against a 6-point scale across several broad criteria such as treatment planning, communication or time management and organisation. The evaluator's opinion of the standard expected upon completion of Dental Foundation Training (DFT) is used as a reference point *i.e.* safe, independent practice. The clinical competencies covered are recorded for each case and following feedback, the evaluator also records how closely the foundation dentist's insight into their own performance, matches the evaluator's opinion, on the 6-point scale.

Considerations: Trainer and foundation dentist satisfaction was high as reported in one internet-based survey (Kirton *et al.* 2013). The median time for observing the foundation dentist to complete this test was 40 minutes and feedback was recorded as requiring a median time of 20 minutes. 68% of foundation dentists who responded to the survey felt that the grades that they had been awarded for this test were an accurate reflection of their abilities.

Reliability: Reliability depends in part on the assessor's training in use of the form and giving feedback. There are no published studies reporting the reliability of this test but Kirton *et al.* (2013) reported some foundation dentists' dissatisfaction with the variability in different trainer assessments of different foundation dentists. This is particularly relevant when only one rater is involved as there is potential for bias and variable reliability.

Key points:

- Wide range of cases should be assessed, to cover all major competencies of the clinical domain
- Used in the UK once a month for newly qualified (foundation) dentists for the first year in supported practice

Example of an ADEPT form

Available online from http://www.bristol.ac.uk/media-library/sites/dentalpg/migrated/documents/d-epassessmenttool.pdf [Accessed 11 April 2015]

References and Further Reading

Kirton, J. A., Palmer, N. O., Grieveson, B. and Balmer, M. C. (2013). A national evaluation of workplace-based assessment tools (WPBAs) in foundation dental training: a UK study. Effective and useful but do they provide an equitable training experience? *British Dental Journal,* 214, 305-9.

360º (Multi-source Feedback, MSF)

Description: Involves collecting information about a clinician's performance in the workplace from those working with that individual. Feedback is gathered using a structured form or questionnaire (for an example of a 360º assessment form see: Wood *et al.* 2006). Different members of the clinical team assess the individual's performance and particularly his or her professional behaviour. Those 'assessing' the individual include staff who are more senior, more junior and peers; and representatives of all groups in the clinician's daily working environment (not just co-professionals) *e.g*. patients. The feedback is used as part of appraisals and to help clinicians gain insight into their professional development. In most cases the student selects "raters" and must select a range of colleagues including fellow health professionals such as dental technicians and dental nurses in addition to administrative support staff and sometimes patients. Patients may be asked to complete a purpose- made patient feedback form or questionnaire. The written feedback, which may be free text or using a scale, is then collated and anonymised. It may be also used by a mentor or supervisor to assess insight, by comparing the rater responses with the student's self-rating.

Skills Assessed: Communication, team working, professionalism and possibly insight.

Considerations: It is feasible for those working with the trainee to participate in this form of assessment as it is based on observations made during everyday work. Each rater fills out a short form that takes 5 to 10 minutes to complete. Longer rating scales with scores of 1-9 rather than 1-3 or 1-4 are associated with a lower proportion of trainees being awarded problem scores (Hassell *et al*., 2012).

Reliability: Reliability depends on feedback from a wide enough range of team members (from all levels) and sufficient raters (usually 8 to 12). An important part of 360º is making good use of the feedback.

Key Points:

- Allows feedback from range of individuals (a variety of staff +/- patients)
- Resource intensive
- Useful information may be gained about professional behaviour and insight

Example of a Multi-Source Feedback form

Available online from https://www.iscp.ac.uk/static/public/msf_self_form.pdf [Accessed on 23 June 2015]

References and Further Reading

Evans, R., Elwyn, G. and Edwards, A. (2004). Review of instruments for peer assessment of physicians. *British Medical Journal*, 328(7450), 1240.

Hassell, A., Bullock, A., Whitehouse, A., Wood, L., Jones, P. and Wall, D. (2012). Effect of rating scales on scores given to junior doctors in multi-source feedback. *Postgradraduate Medical Journal,* 88, 10-4.

Wood, L., Hassell, A., Whitehouse, A., Bullock, A. and Wall, D. (2006). A literature review of multi-source feedback systems within and without health services, leading to 10 tips for their successful design. *Medical Teacher*, 28(7), 185-191.

Wood, L., Wall, D., Bullock, A., Hassell, A., Whitehouse, A. and Campbell, I. (2006). 'Team observation': a six-year study 1 of the development and use of multi-source feedback (360-degree assessment) in obstetrics and gynaecology training in the UK. *Medical Teacher*, 28(7), 177-84.

Case-based Discussion (CbD)

Description: A formal discussion between a trainee and an assessor about a case that the trainee has managed and for which they have been directly responsible. Case-based discussions are primarily used for formative assessment ('in-training'). During the discussion, the trainee refers to the case records. The assessor will probe the trainee's depth of understanding, decision-making and clinical judgement. The trainee has the opportunity to talk about any issues that arose and explain decisions. The assessor can also determine the quality of various aspects of case management *e.g*. synthesising information, prioritising, planning and record keeping.

A structured assessment form is used to record basic case details and rate the key skill areas (for an example of an assessment form see References and Reading). The discussion is followed by a short feedback session.

Choosing a challenging case enables the trainee to maximise the benefits of discussing and reflecting on a case with a more senior clinician. The format is broadly similar to that described in medicine for 'Chart Stimulated Recall' where a doctor's own cases are used as the basis for a structured oral examination.

Skills Assessed: Application of knowledge, decision making, clinical judgement, professionalism.

Considerations: The discussion lasts about 20 minutes with 5 to 10 minutes for feedback. Typically the assessment is performed several times per placement and over that time should cover a range of cases that are typical for the particular speciality. Although undertaken during workplace training the assessment is not carried out during a clinical encounter but in an office or meeting room setting. Both trainers and foundation dentists have reported finding CbD to be the most beneficial of the workplace based assessments with 92% of foundation dentists who responded to the survey reporting an improvement to their patient care following feedback in this form (Kirton *et al.* 2013).

Reliability: Reliability depends in part on the assessor's training in use of the form and giving feedback. However, as only one rater is involved there is potential for bias and variable reliability. Essentially as this is a structured oral it suffers from the same problems of reliability as other orals described earlier.

Key Points:

- High authenticity
- Standardised rating system
- Mostly used formatively
- Low reliability

Example of CBD forms and guidance

Available at: https://www.iscp.ac.uk/static/public/cbd_guidance.pdf [Accessed 6 May 2015]

References and Further Reading

Cunnington, J., Hanna, E., Turnhbull, J., Kaigas, T. and Norman, G. (1997). Defensible assessment of the competency of the practicing physician. *Academic Medicine*, 72(1), 9-12.

Guidance notes and example forms for CBD. (2010). 1st ed. [ebook] The Royal College of Surgeons of England.

Kirton, J. A., Palmer, N. O., Grieveson, B. and Balmer, M. C. (2013). A national evaluation of workplace-based assessment tools (WPBAs) in foundation dental training: a UK study. Effective and useful but do they provide an equitable training experience? *Br Dent J,* 214, 305-9.

Setna, Z., Jha, V., Boursicot, K. and Roberts, T. (2010). Evaluating the utility of workplace-based assessment tools for speciality training. *Best Practice and Research. Clinical Obstetrics and Gynaecology*, 24(6), 767-82.

Observation on Clinics or Rotations

Description: Students are observed and assessed during clinical work *i.e*. on intramural and extramural rotations / clerkships. This type of assessment has sometimes been referred to as "Longitudinal Evaluation of Performance (LEP)". The assessment is based on performance over a period of time (days to weeks) and a number of skills can be rated from basic factual knowledge to technical skills as well as other aspects of professional behaviour. The method of marking and assigning grades varies considerably. Students are often assigned a grade at the end of the rotation / placement, which can be derived from a global rating form that includes general categories of professional and clinical ability *e.g*. knowledge, clinical skills, communication skills, case responsibility, preparation and professionalism.

The assessment may be undertaken by one tutor or several members of the team. It may involve grading a log book, frequently electronically, and taking a mean of the week-by-week grades to give an overall grade at the end of a term or placement. If individuals other than clinicians are involved, the assessment style approaches the 360º multi-source feedback evaluations described earlier.

Skills Assessed: Knowledge, application of knowledge, clinical/practical skills, diagnostic skills, clinical reasoning, communication skills, attitudes and professionalism.

Considerations: As the assessment is embedded in day-to-day work there are potentially relatively low demands on resources. However it does require an engaged assessor who has the time to observe the student and provide detailed feedback.

Reliability: Reliability tends to be low as the assessment often lacks standardisation *e.g*. observational frequency varies, marking can be very subjective as it is often based on 'clinical impressions', can be affected by 'halo effects', and inter-rater reliability is poor. Additionally, tutors are sometimes reluctant to fail students. The objectivity and reliability can be improved if checklists are used and the frequency and breadth of assessment is increased.

Key Points:

- Based on observation of students in routine practice
- Low reliability
- Subjective and prone to 'halo effects'
- Can provide useful opportunity for feedback

References and Further Reading

Miller, G. (1990). The assessment of clinical skills / competence / performance. *Academic Medicine*, 65(9), 63-7.

Prescott-Clements, L., van der Vleuten, C., Schuwirth, L., Hurst, Y. and Rennie, J. (2008). Evidence for validity within workplace assessment: the Longitudinal Evaluation of Performance (LEP). *Medical Education*, 42(5), 488-95.

Turnbull, J. and Van Barneveld, C. (2002). Assessment of clinical performance: in-training evaluation. *Springer International Handbooks of Education*, 7, 793-810.

Portfolios

Description: This is a collection of work developed as a cumulative 'body of evidence' to demonstrate the student's learning and achievements. **It is not an examination format in its own right**, rather a receptacle containing a mixture of materials, each piece assessable to predefined marking criteria which may be graded or pass/fail.

Hence although included in this section on the 'Does' level of Miller's pyramid, in real terms the portfolio itself contains evidence relating to 'Does'. The content, which can be paper-based or in an electronic format (e-portfolio), is collected during day-to-day activities and is typically quite diverse *e.g*. written assignments, reports, feedback, case studies and projects. Supplementary material such as photographs, videos and curriculum vitae may be included.

A portfolio can also be used to plan learning needs and to monitor progress *e.g*. with checklists of skills and activity logs. Evidence of the student's reflections on learning is a valuable aspect of a portfolio. Portfolios have been used in nurse training in the UK for many years.

The approach to the assessment of portfolios and the criteria applied are quite variable and depend on content. Assessment is often an on-going process which can be formative and/or summative. Interviews provide an opportunity to determine how well the portfolio reflects the student's achievements. Portfolios are not always formally assessed, instead the requirement being for the provision of evidence that certain tasks have been completed with a grade for engagement with the process.

Skills Assessed: Knowledge, knowledge application and interpretation, case recording and interpretation, attitudes and professionalism (skills not always easy to assess using other methods).

Considerations: Staff time is a major consideration as portfolios are labour intensive to supervise and mark, and to provide feedback, although the workload may be spread throughout the year. Student perceptions of value vary from being seen as providing a useful framework for learning, to having a low return relative to the time and effort expended. Uptake and engagement vary and are affected by learner type and maturity, and by tutor enthusiasm and support. Using a framework to align portfolio content with curriculum or course outcomes will help students produce a representative and comprehensive 'body of evidence'. Portfolio scores from one U.S. study have been found to correlate with scores in the National Board Dental Examination Parts I and II (Gadbury-Amyot *et al.*, 2014) although the heterogeneity of the portfolio system elsewhere including the U.K., limits the generalisability of this study.

Reliability: Achieving reliability can be difficult and is affected by the diverse content of a portfolio and the subjective aspects of the evaluation particularly if only one examiner is involved. Reliability can be improved using rating scales and having more than one marker. Assessing the student's process of reflection is not straightforward if that is deemed desirable for a given context.

Key Points:

- Heterogeneity in meaning – covers many different formats
- Resource intensive
- Assessing reflection is difficult and controversial

References and Further Reading

Challis, M. (1999). AMEE Medical Education Guide No. 11 (revised): Portfolio-based learning and assessment in medical education. *Medical Teacher*, 21(4), 370-86.

Davis, M., Harden, R., Howie, P., Ker, J. and Pippard, M. (2001). AMEE Medical Education Guide No. 24: Portfolios as a method of student assessment. *Medical Teacher*, 23(6), 535-551.

Davis, M. and Ponnamperuma, G. (2005). Portfolio assessment. *Journal of Veterinary Medical Education*, 32 .279-284.

Friedman Ben-David, M., Davis, M., Harden, R., Howie, P., Ker, J. and Pippard, M. (2001). AMEE Medical Education Guide No. 24: Portfolios as a method of student assessment. *Medical Teacher*, 23(6), 535-51.

Gadbury-Amyot, C. C., Mccracken, M. S., Woldt, J. L. and Brennan, R. L. (2014). Validity and reliability of portfolio assessment of student competence in two dental school populations: a four-year study. *Journal of Dental Education,* 78, 657-67.

Vernazza, C., Durham, J., Ellis, J., Teasdale, D., Cotterill, S., Scott, L., Thomason, M., Drummond, P. and Moss, J. (2011). Introduction of an e-portfolio in clinical dentistry: staff and student views. *European Journal of Dental Education*, 15, 36-41.

SECTION 2: CONCEPTS / TERMINOLOGY 'HEADLINES'

This section includes expanded descriptions of key concepts relating to validity, standard setting, feedback, psychometrics and developing an assessment strategy.

Validity – Modern Concepts
Author: Susan Rhind

Description: Validity addresses the question of whether a test measures what it is supposed to measure. It is about the true meaning of the test scores. While validity[1] has been considered as one of the characteristics of specific assessment instruments (Van Der Vleuten, 1996) together with reliability, educational impact, acceptability and cost, the latest concept of validity is more overarching. This "unitary" view of validity was initially described and developed by Messick (1989, 1995) and further developed by Kane (2001, 2006) and Messick (2014). It is becoming increasingly accepted as a foundation stone for evaluating assessment tools or whole programmes of assessment (Schuwirth and van der Vleuten, 2012)

The fundamental concept of validity is whether the decisions made on the basis of particular tests can be reasonably defended. Therefore, there are certain criteria and evidence which need to be documented and presented to support the decisions that are made as a result of any test (examination). These criteria should be particularly considered for every test that has summative impact on candidates' lives (i.e. high-stakes tests), such as progression from one year to another in undergraduate education, graduation, or certification for postgraduate degrees.

Considerations: The following criteria need to be considered as having an impact upon this unified concept of validity. Although these criteria (Messick) have been listed separately they are clearly linked and in some cases complementary and should be viewed as all contributing to the validity of a test:

1. Test content: Refers to the purpose of the test and how it is defined i.e. Does the test content appropriately reflect the learning objectives of the course/module? Is the assessment task both relevant and representative of the work a dentist will do? Blueprinting is key to this aspect of validity.

2. Response process: Refers to the type of testing formats being used i.e: chosen or constructed. Chosen is where the candidates choose from a list of answers offered within the test whereas constructed is where the candidates generate answers for themselves. There is more scope for error with constructed responses as these currently cannot be electronically scored and require examiners to read and mark. Certain criteria should therefore be met to minimize error i.e:

- Clearly set-out outline answers with scoring rubrics
- Blinded double marking
- Clear process for moderation where there is a difference in score between examiners
- Rater training and trainee familiarity with the format of the test

3. Internal structure: Refers to how a test is constructed and includes the following criteria:

- Number of items (in a written test) or number of stations (in a practical test)
- Format of the items
- Whether the format is appropriate for the domain of skill being tested (e.g. MCQs for knowledge tests, OSCEs for clinical skills)

- Sufficient sampling for the tests to be reliable
- Scrutiny of the psychometric analyses of the test i.e. reliability coefficients
- Item analysis data (test score correlation, facility indices, etc.)
- Weighting of certain parts of the test or equal weighting of all parts
- Presence of a system of compensation
- Method of standard setting applied to determine the pass mark

4. Relationship to other tests: How do the results of a test compare to the results of other tests taken by the same candidates?

5. Effects/outcomes: Consider the implications and consequences of decisions made on the basis of each test, e.g.

- Effect on student learning
- Impact of failing - on students, on parents, on tutors, on remediation and support staff
- Impact of passing - on students, on patients, relevant health authorities, university reputation and the regulatory body.

Key Points:

- Modern concepts of validity are all-encompassing and do not consider validity as a property of an individual assessment instrument.
- Evidence needs to be gathered against a range of criteria to ensure an overall programme of assessment is valid for the purposes intended. This process should be on-going for each test.

References and Further Reading

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38(4), 319-342.

Kane, M. (2006). Validation. In: R. Brennan, ed., *Educational Measurement*, 1st ed. Westport, CT: Praeger, p7-64.

Messick, S. (1995). Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning. *American Psychologist,* 50, 741-749.

Messick, S. (1996). Standards-based Score Interpretation: Establishing Valid Grounds for Valid Inferences. In: *Proceedings of the Joint Conference on Standard Setting for Large Scale Assessments, Sponsored by National Assessment Governing Board and the National Center for Education Statistics*. Washington, DC: Government Printing Office.

Messick, S. (2014). Validity. In: R. Linn, ed., *Educational Measurement*, 3rd ed. New York: Macmillan, 13-103.

Schuwirth, L. and van der Vleuten, C. (2011). General overview of the theories used in assessment: AMEE Guide No. 57. *Medical Teacher*, 33(10), 783-797.

Schuwirth, L. and van der Vleuten, C. (2012). Programmatic assessment and Kane's validity perspective. *Medical Education*, 46(1), 38-48.

Van der Vleuten, C., Norman, G. and De Graaff, E. (1991). Pitfalls in the pursuit of objectivity: issues of reliability. *Medical Education*, 25(2), 110-18.

Van der Vleuten, C. (1996). The assessment of professional competence: developments, research and practical implications. *Advances in Health Sciences Education*, 1(1), 41-67.
Van der Vleuten, C. (2000). Validity of final examinations in undergraduate medical training. *British Medical Journal*, 321(7270), 1217.

NB. For more succinct definitions of different terms of validity, as often used within the literature, please see the Glossary.

Standard Setting
Author: Susan Rhind

Description: Standard setting is the process whereby decisions are made about boundaries or 'cut-off points' between groups of students. Most commonly this decision focuses on those who pass and those who fail but the process can also be applied to other boundaries *e.g.* those who gain distinction or other form of credit and those who do not. It is a systematic way of gathering value judgements, reaching consensus and expressing that consensus as a score on a specific test (Norcini, 2003).

Standards can be described as

- relative (norm referenced),
- absolute (criterion referenced) or as a
- compromise.

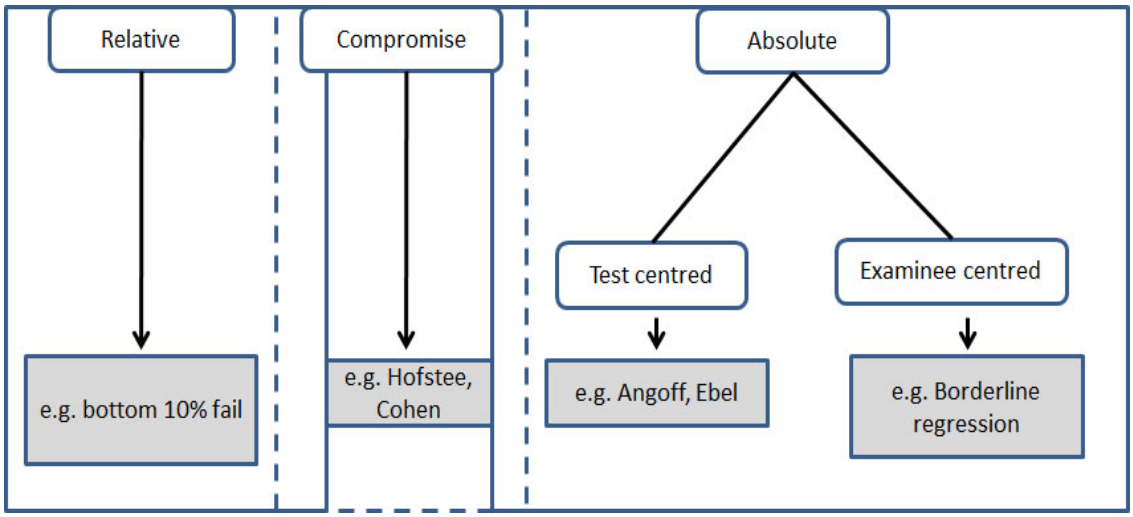


Figure 7. Diagram to show the relationship between different methods of standard setting.

Relative Standards
The performance of candidates is reported relative to each other. Relative standards may be used for ranking of candidates e.g. for courses which may be competitive or in admissions.

Absolute Standards
A decision is made before the test is taken about the difficulty of the test and the requirements for passing. In theory, using absolute methods, all candidates could pass and all could fail. Such absolute standards are most appropriate for tests of competence when we want to be assured that candidates are 'safe' either to move to the next phase of the curriculum or out into practice. Absolute standards can be further considered as either 'test-centred' or 'examinee-centred'.

Test-Centred Absolute Methods
Two of the best known 'test-centred' methods for establishing an absolute standard on MCQ assessments are the Angoff and Ebel method. Both these methods rely on judges estimating the performance of a hypothetical group of 'borderline' candidates in the context of the assessment they are setting the standard for.

Examinee-Centred Absolute Methods
In these methods (which are common in OSCEs), the standard takes into account the performance of individual candidates based on overall criteria or overall test performance. A commonly used example is the borderline regression method where candidates are marked on a checklist and then given an additional overall global rating. The checklist score is then regressed on the global rating which allows calculation of the checklist passing score.

Compromise Standards
In these methods, elements of both relative and absolute standards are incorporated. The best example of this is the Hofstee method where decisions are made in advance about the tolerance rates for failure and also the minimum and maximum acceptable cut-point for the given assessment. A further example has been suggested by Cohen-Schotanus *et al.* (1996) which uses a conventional, pre-fixed cut-off score with high performers as a relative point of reference e.g. setting the standard at 60% of the highest achiever's score in the test.

Considerations: Although the rationale for absolute standard setting is clear, in practice several factors need to be considered:

1. Number of judges. Standard setting panels typically require 6-8 judges if there is a discussion phase in the process and 10 if there is no discussion phase (Fowell *et al.*, 2008). Whilst this can be easy to achieve e.g. in a final examination of competence when tutors may feel comfortable with the concept of a borderline student, this may be more difficult to achieve at earlier stages in the curriculum or in small disciplines. The faculty resource required means it is a costly process.
2. It is well recognised that for judges, conceptualising the borderline student can be challenging.
3. Where absolute methods are used, examination boards should ensure that the process produces a credible result and should have a well described pre-published strategy to deal with unexpected results.

Key Points:

- Standards can be relative or absolute or compromise
- The method chosen should relate to the purpose of the assessment and should be defensible
- Standard setting is resource intensive and may, at least initially, be conceptually challenging
- Rather than complete standard setting for e.g. each item on a checklist, the methods may be modified to provide a more holistic judgement of the required standard e.g. at the level of an OSCE station
- Significant staff development needs to be implemented to ensure a robust standard setting process

References and Further Reading

Angoff, W. (1971). Scales, norms and equivalent scores. In: R. Thorndike, ed., *Educational Measurement*, 2nd ed. Washington, DC: American Council on Education, 508-600.

Bandaranayake, R. (2008). Setting and maintaining standards in multiple choice examinations: AMEE Guide No. 37. *Medical Teacher*, 30(9-10), 836-45.

Boursicot, K., Roberts, T. and Pell, G. (2007). Using borderline methods to compare passing standards for OSCEs at graduation across three medical schools. *Medical Education*, 41(11), 1024-31.

Cohen-Schotanus, J and Van der Vleuten, C. (2010). A standard setting method with the best performing students as point of reference:Practical and affordable. *Medical Teacher,* 32, 154-160.

De Gruijter, D. (1985). Compromise models for establishing examination standards. *Journal of Educational Measurement*, 22(4), 263-69.

Downing, S., Tekian, A. and Yudkowsky, R. (2006). Procedures for Establishing Defensible Absolute Passing Scores on Performance Examinations in Health Professions Education. *Teaching and Learning in Medicine*, 18(1), 50-57.

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 Science Education: Theory and Practice*, 13, 11-24.

Friedman Ben-David, M. (2000). AMEE Guide No. 18: Standard setting in student assessment. *Medical Teacher*, 22(2), 120-30.

Kramer, A., Muijtjens, A., Jansen, K., Dusman, 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.

McKinley D.W. and Norcini J.J. (2014). How to set standards on performance-based examinations: AMEE Guide No. 85. *Medical Teacher,* 36(2), 97-110.

Norcini, J. (2003). Setting standards on educational tests. *Medical Education*, 37(5), 464-69.

Reid, J. (1991). Training Judges to Generate Standard-Setting Data. *Educational Measure: Issues Practice*, 10(2), 11-14.

Puryer, J. and O'Sullivan, D. (2015). An introduction to standard setting methods in dentistry. *British Dental Journal,* 219, 355-358

Feedback
Author: Sheena Warman

Description: It is well recognised (for example in results of the UK National Students' Survey) that students in professional educational programmes often complain of a lack of feedback. On the other hand, teachers are frustrated that their efforts to provide feedback go unrecognised and do not seem to effect significant change.
Whilst summative assessments such as examinations may give students an indication of their performance relative to their peers, there is an increasing drive for more effective formative feedback to become part of the culture. Traditionally feedback has been considered to be a teacher-led process whereby the student is given information about their performance. However, modern approaches encourage a dialogue between trainer and trainee, with trainees taking a more active role in seeking and using feedback. Additionally, an effective feedback process should help develop the student's self-evaluation skills. Feedback in the preclinical environment might include for example, discussion of a piece of coursework with a tutor, or peer feedback within a small group setting. Feedback in the clinical environment commonly takes the form of "in-the-moment" feedback during routine clinical work, regular progress discussions with a tutor, or written feedback at the end of a clinical placement. It is important that both positive and negative aspects of a students' performance are discussed in a timely, accurate, non-judgemental manner, using specific examples, and that the student is supported to engage with and act upon the feedback. Staff training in techniques for increasing the effectiveness of feedback, and student training in seeking and using feedback, can be invaluable in improving the "feedback culture" within the teaching environment (Warman *et al.* 2014).

Considerations / Practicalities: Feedback takes time, and staff members need to be encouraged to prioritise feedback discussions with students. This requires a culture of feedback within the teaching environment, and training of staff to increase their confidence and skills in feedback dialogue. Techniques such as the "One-minute teacher" (Neher and Stevens, 2003) are described for maximising teaching opportunities and feedback in clinics without disrupting a busy clinical workload. Frameworks such as "Pendleton's rules" (Pendleton *et al.* 2003) encourage the student to take responsibility for evaluating their own performance within the feedback dialogue. Feedback given in an inappropriate manner can be ineffective or indeed harmful.
Staff can find it hard to give critical feedback, particularly when it relates to issues of professionalism rather than knowledge or skills. Training, and an increased expectation of a feedback dialogue by students, may help to overcome this. Students need to seek, recognise and act on feedback, which may require some proactive student training.

Key Points

- Feedback is essential for the effective development of professionals, and development of an effective "feedback culture" is paramount
- Written or verbal feedback should be timely, accurate, specific, objective, non-judgemental and balanced
- Feedback dialogue should encourage the student to identify and reflect on the strengths and weaknesses of their own performance, supported by a skilled tutor
- Every feedback interaction should generate a plan for the student's improvement

References and Further Reading

Boud, D. and Molloy E. (2013). Feedback in Higher and Professional Education: Understanding it and doing it well. Routledge, Oxon. ISBN 978-0-415-69228.

Neher, J. and Stevens, N. (2003). The one-minute preceptor: shaping the teaching conversation. *Family Medicine-Kansas City*, 35(6), 391-93.

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.

Nicol, D. (2010). From monologue to dialogue: improving written feedback processes in mass higher education*. Assessment and Evaluation in Higher Education,* 35(5), 501-17.

Pendleton, D., Schofield, T., Tate, P., and Havelock, P. (2003). The consultation: An approach to learning and teaching: Oxford: Oxord University Press.

Ramani, S. and Kracko, S. K. (2012). Twelve tips for giving feedback effectively in the clinical environment. *Medical Teacher,* 34, 787-791.

Van de Ridder, J.M., Stokking, K.M., McGaghie, W.C. and ten Cate, O.T. (2008). What is feedback in clinical education? *Medical Education,* 42, 189-197.

Warman S.M., Laws, E., Crowther E., and Baillie S (2014). Initiatives to immprove feedback culture in the final year of a veterinary programme. *Journal of Veterinary Medical Education,* 41(2), 162-171.

Wood, D. (2014). Written assessment In: Swanwick T. *Understanding medical education: evidence, theory and practice,* Oxford, Wiley-Blackwell, 23, 317-328.

Psychometrics
Author: Sheena Warman

Description: Psychometrics in this context refers to the application of statistical methods to assessment data to ensure that the assessment process is accurate *i.e*. reliable and valid. Validity is discussed above; here we will focus on reliability *i.e*. the reproducibility of the results. An assessment process cannot be valid if it is not reliable; however reliability does not guarantee validity (Hecker and Violato, 2009). Measurement errors arise from examinee factors such as carelessness or fatigue, from test factors such as poor directions and thirdly from marking factors such as non-uniform guidelines or marking errors. Three main approaches to modelling responses to assessment have been developed (Schuwirth and van der Vleuten, 2011; McManus, 2010).

Classical test theory: This is the most widely used theory and assumes that a candidate has a true ability (true score) but that the actual observed score is influenced by measurement errors. It is of most use in multiple choice tests when all students are given the same set of questions at the same time. Commonly analysed statistics include the p-value (the proportion of candidates answering the question correctly), the item-total correlation (the discriminatory power of the individual item), the standard error of measurement (Torok *et al.*) and Cronbach's α (the internal consistency of the test). It is generally considered that Cronbach's α should be >0.8 in a high stakes assessment, and that if it is >0.9 it is likely that there is some redundancy in the test (*e.g*. the test may contain more items than necessary for reliability or is repeatedly sampling the same knowledge base). For further discussion of the limitations of Cronbach's α see Schuwirth and van der Vleuten (2011) and Tavakol and Dennick (2011).

Generalisability theory: This is more useful when there is the potential for multiple sources of measurement error within an assessment *e.g*. clinical or OSCE style assessments where not all candidates are seen by all examiners, or may not all see the same patient. It can be used to identify variability due to different examiners (*e.g*. hawks and doves), and also allows the examining team to answer questions such as "How would the reliability be affected by having *e.g*. fewer stations or fewer examiners?"

Item-response theory: This requires pre-testing and large data sets and is best used for testing carried out at a large-scale level (*e.g*. national level testing rather than at Faculty level). It calculates the difficulty of items as well as the discriminative value and the likelihood of the candidate guessing rather than knowing the answer. It estimates item difficulty and student ability independently of each other. It requires complex mathematical modelling and significant input from a psychometrician.

Considerations: Psychometrics is a discipline in its own right; expert input from a psychometrician is extremely valuable when decisions are made around a particular score, *e.g*. a cut score for pass/fail decisions. It is valuable to calculate the standard error of measurement (Torok *et al.*) in order to establish 95% confidence intervals around this cut score. (*Note that this then raises the issue that, for students whose scores fall within these confidence intervals, it is not possible to conclude (with a p≤0.05) whether or not these students have passed the test. Some assessment teams will then raise the pass mark to account for this uncertainty, in order to have confidence in the reliability of a passing score representing a true pass).*

Key Points:

- Psychometric principles are increasingly being adopted as a standard part of professional programme's assessment protocols to evaluate and continually refine/ improve assessments.
- Classical test theory is currently the most widely used and dental educators are increasingly developing expertise in its use.
- Familiarisation with the basics is straightforward, but users should also familiarise themselves with the limitations and underlying assumptions in order to be confident in their interpretation of the results.

References and Further Reading

Hecker, K. and Violato, C. (2009). Validity, reliability, and defensibility of assessments in veterinary education. *Journal of Veterinary Medical Education*, 36, 271-75.

McManus, C. (2010). Focus on: The measurement of reliability. In: T. Swanwick (Ed), ed., *Understanding Medical Education: Evidence, Theory and Practice*, 1st ed. Chichester: Wiley-Blackwell.

Schuwirth, L. and van der Vleuten, C. (2011). General overview of the theories used in assessment: AMEE Guide No. 57. *Medical Teacher*, 33(10), 783-97.

Tavakol, M. and Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2, 53-55.

Developing an Assessment Strategy for a Curriculum
Authors: Sarah Baillie and Jonathan Sandy

Description: As well as selecting individual assessment methods for each component within the curriculum, it is important to develop an overarching assessment strategy for the whole programme. The strategy aims to describe the assessment methods used for each domain of competence, year by year throughout the curriculum, as a student progresses towards graduation and entering the profession. It provides an overview (and often includes a table or diagram e.g. Figure 8) that demonstrates how the 'whole' represents the 'sum of its parts' (assessments) which are combined in such a way as to achieve the programme outcomes.

Domain of Competence →	Scholar and Scientist (Bloom: Cognitive)	Clinical & Practical Skills (Bloom: Psychomotor)	Professionalism (Bloom: Affective)
Year of Curriculum ↓	Assessment type/s	Assessment type/s	Assessment type/s
Year 1	MCQ SAQ	OSCEs	Portfolio
Year 2	MCQ	OSCEs	Portfolio Work-based assessment

And so on, for each year of the curriculum and domain of competence until graduation

Figure 8: Example 'Assessment Strategy Overview' representing typical domains of competence and some possible assessment methods as used year by year in the curriculum

Benefits of Adopting an Assessment Strategy: There are a number of benefits in having a strategy that summarises all assessments and is accessible to a range of different stakeholders (accrediting bodies, curriculum managers, tutors, staff and students). It demonstrates that assessments are appropriately aligned all domains or competences to be covered. The overview provided can also illustrate how students are developing and building their competences year on year and when and how decisions are made about progression.

The strategy document is an extremely useful management tool and can help support a coordinated and consistent approach to all aspects of assessment including planning, blueprinting, standard setting, psychometrics, feedback, etc. It can also be used to facilitate regular monitoring and updating of assessment practices including the adoption of methods supported by current evidence and aligning the most suitable assessment type/s with the relevant domain of competence. When designing a new curriculum (or reviewing an existing one) the development of an assessment strategy is particularly valuable and should be an integral part of the process.

The strategy should include both 'assessment of learning' (summative and for decision making) and 'assessment for learning' (formative) within a culture of feedback and mentoring. Providing opportunities for comprehensive and high quality feedback and encouraging a dialogue-type approach is extremely valuable for students but gathering such feedback is often resource intensive requiring commitment from both the tutors and the students. Adopting a programmatic

approach has the potential to improve the signposting of all such formative opportunities to students and highlight examples of best practice among staff/tutors.

Considerations: Once a programmatic overview of assessment has been developed the impact on learning should be monitored. Additionally, the strategy should be seen as a dynamic document and reviewed regularly and updated as required.

Key Points:

- An assessment strategy aims to:
- Provide a clear and comprehensive overview of all assessments throughout the curriculum
- Demonstrate how assessments align with each domain of competence and graduate outcomes (GDC Preparing for Practice)
- Include both assessment of learning (summative) and for learning (formative)

References and Further Reading

Bok, H.G.J., Teunissen, P.W., Favier, R.P., Rietbroek, N.J., Theyse, L.F.H., Brommer, H., Haarhuis, J.C.M., van Beukelen, P., van der Vleuten, C.P.M. and Jaarsma, A.D.C. (2013). Programmatic assessment of competency-based workplace learning: when theory meets practice. *BMC Medical Education,* 13, 123.

General Dental Council (2014). Preparing for practice. Dental team learning outcomes for registration Available from https://www.gdc-uk.org/newsandpublications/publications/publications/gdc%20learning%20outcomes.pdf [Accessed 12 May 2015].

General Medical Council (2009). Assessment in undergraduate medical education. http://www.gmc-uk.org/Assessment_in_undergraduate_medical_education_1114.pdf_56439668.pdf [Accessed 12 May 2015].

van der Vleuten C.P., Schuwirth L.W., Driessen E.W., Govaerts M.J.B. and Heeneman, S. (2015). Twelve Tips for programmatic assessment. *Medical Teacher*, 37:641-46.

Glossary of Selected Terms

Blueprint: Indicates the content / areas covered for an exam, and relates to the learning objectives of the course. The relative amounts / approximate number of questions in each area can be defined.

Criterion referencing: Assessment is linked to achievement of outcomes regardless of the performance of other students i.e. measured against a defined criterion, absolute requirement or objective. Theoretically all students could pass or all could fail.

Cueing effects: In MCQs, and similar exam formats, examinees are able to eliminate wrong answers and recognise the correct answer, rather than needing to work out the answer. Questions should be designed to avoid cueing.

Formative Assessment: Sometimes referred to as 'assessment for learning' and provides information and feedback to the student on their performance. This allows the student to use the feedback to inform and guide future learning.

Global rating scales: These differ from checklists as the rater(s) assess each skill on a scale with categories that represent a range of performance e.g. from unsatisfactory to above expected performance levels. The forms usually include assessment of a range of skills such as: technical ability, consultation skills, knowledge, history taking, professionalism, team working and communication. They may include areas for the examiner to provide feedback comments. Global rating scales are used in a number of assessment methods e.g. OSCEs and mini-CEX.

Halo effects: Can be used to describe:
a) the effect whereby a judgement on one aspect is influenced by an overall impression of the person.
b) the effect whereby a judgement is influenced by the performance of previous candidates in contrast to the current candidate i.e. after one or more consecutive poor candidates the subsequent candidate, even if average, would appear good and be 'over scored'.

Hawks and Doves: Two categories of examiners. Hawks tend to mark examinees 'down', while doves are lenient and award higher marks than the average across the board. When a hawk and a dove are together, the hawk tends to dominate.

Norm referencing: Refers to assessment which aims to discriminate between students by ranking them or 'grading on a curve'. The achievement of one student is relative to the rest of the students in that cohort.

Summative Assessment: Usually associated with a mark or grade and often occurs towards the end of a course. Important for enabling high stakes decisions relating to e.g. student progression.

Validity: As referred to in the text, these terms have often been superceded in more recent texts with a global approach to validity but have been included here as these forms are still described within the literature.

Face validity: the assessment method, on first impression, appears to measure the intended competency.

Content validity: refers to the content of the assessment and how representative it is of the desired learning objectives. In practice, ensuring content validity typically involves the creation of a blueprint or spread sheet to facilitate mapping of the assessment to the learning objectives.

Construct validity: refers to whether the scores on an assessment align with the trait the assessment is intended to measure

Criterion-related validity: refers to how well the assessment relates to some other criterion. This may be predictive (where the criterion of interest is future performance) or concurrent (where the criterion of interest is another criterion measured at the same time)

Consequential validity: refers to the impact that the assessment may have on student behaviour