



A Guide to Assessment Methods in Veterinary Medicine



A Guide to Assessment Methods in Veterinary Medicine

Authors: Sarah Baillie and Susan Rhind

September 2008

Version 1.1

A 'Blue Sky' project funded by the Royal College of Veterinary Surgeons Trust



Contents

Introduction	1
Overview of Current Assessment Challenges: Towards Common Standards	2
Section 1: Assessment Methods used in Veterinary Medicine	4
Multiple Choice Questions (MCQs)	5
Extended Matching Questions (EMQs)	7
Short-answer Questions (SAQs)	8
Essays	9
Viva / Viva Voce / Oral	10
Practical Assessment (the 'Spot' test)	11
Objective Structured Clinical Examination (OSCE)	12
Long Case	14
Observation on Rotations	15
Portfolios	16
Section 2: Additional Assessment Methods used in Medicine	18
Clinical Evaluation Exercise (CEX)	19
Mini-clinical Evaluation Exercise (mini-CEX)	20
Longitudinal Evaluation of Performance (LEP)	21
Directly Observed Procedural Skills (DOPS)	22
360° (Multi-source Feedback)	23
Case-based Discussion and Chart Stimulated Recall Oral Examination (CSR)	24
Script Concordance Test (SCT)	25
References and Bibliography	26
Glossary: Definition of Terms	27
Acknowledgements	29
Appendices	30

Introduction

This document is part of a 'Blue Sky' project 'Evidence Based Development of a Common Final Examination for Veterinary Undergraduates' which was funded by the Royal College of Veterinary Surgeons Trust. The aim of this part of the project was to develop a guide to assessment methods in a quick reference format that would provide useful information for those involved in the development and delivery of examinations for veterinary undergraduates and postgraduate certification.

The document contains two sections. The first section describes methods commonly used in UK veterinary schools.¹ The second section extends the list by describing some additional methods used in medical education that may be of relevance to the veterinary field. In the second section, particular emphasis has been placed on methods that assess the clinical competencies appropriate to the modern veterinary graduate. Methods described include those suitable for use in both the undergraduate and postgraduate context as the distinction between these two phases, at least in terms of assessment, is becoming increasingly blurred. Also, it is worth noting that a combination of methods will give best coverage for the range of skills required.

Methods have also been described at a level that we believe captures the essence of the nature of the interaction between examiner and examinee rather than including all variants and modifications e.g. the use of models versus live cases or simulated cases.

For each assessment method there is: a short description; a list of skills assessed; the practical considerations for running such assessments; the reliability and validity issues; a list of 'key-points'; and some references for further reading.

Note:

This document is intended as a guide and in its current state is 'work in progress'. Titles and terminology used are in the tradition of Higher Education in the UK and may not be directly equivalent to those used in other countries. It is also acknowledged that there are variations of many of the methods described and there are additional methods that have not been included.

Veterinary assessment is a field that continues to change and evolve. Therefore, it is anticipated that this guide will evolve over time. We intend to produce updated versions that will be informed by research and developments in assessment science. We would also appreciate feedback from readers to help update and improve the booklet. As the document is updated and extended, a PDF of the latest version will be available for download at:

<http://www.link.vet.ed.ac.uk/beme/> or http://www.live.ac.uk/html/activities_assessment.html

The guide and further information will also be available in the Veterinary Education section of WikiVet (www.wikivet.net). Registration for WikiVet is straightforward and we would welcome comments and discussion on the guide and assessment in veterinary education in general.

Author contact information:

Dr Sarah Baillie, The Royal Veterinary College, University of London, Hawkshead Lane, North Mymms, Hatfield, Hertfordshire, AL9 7TA, UK. Email: sbaillie@rvc.ac.uk
Prof Susan Rhind, The Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush Veterinary Centre, Roslin, Midlothian, EH25 9RG, UK. Email: susan.rhind@ed.ac.uk

¹ Based on information gathered at the project launch workshop (09/2006) and from e-mail requests to all UK veterinary schools

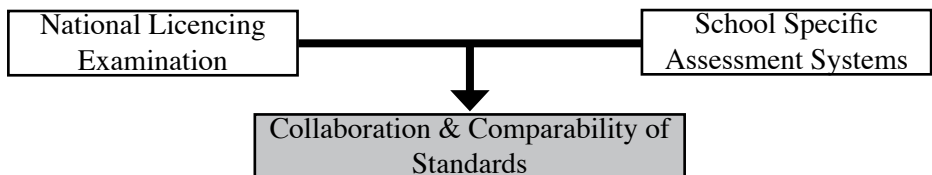
Overview of Current Assessment Challenges: Towards Common Standards

Assessment processes and procedures in veterinary schools in the United Kingdom are audited by the Royal College of Veterinary Surgeons (RCVS) during their rolling cycle of visitations in association with the European Association of Establishments of Veterinary Education (EAEVE). In addition, some UK schools are also accredited by the American Veterinary Medical Association (AVMA) who have their own requirements in terms of audit of assessment process. In the wider Higher Education context in the UK, the main quality assurance procedure is the external examiner system. All veterinary graduates from RCVS accredited schools are licensed to practice in the UK although they may have been examined in very differing ways at the 'point of the entry' into the profession - the 'final' examination.

This heterogeneity in assessment systems has also been recognised in medical education. Based on data indicating that pass standards across different medical schools may be different (Boursicot *et al.* 2007), some are now calling for consideration of a common final year assessment as exists, for example, in North America. A report from the General Medical Council (GMC) education committee 'Strategic Options for Undergraduate Medical Education' (2006), produced following consultation with a range of stakeholders, highlights the differing views on this subject. Whilst the "need for consistency in outcomes between medical schools and between students" was acknowledged, overall, stakeholders were not in favour of a common final year assessment.

Setting examinations and ensuring rigorous standards is increasingly resource intensive. Furthermore, the external examiner system is non-uniform, and anecdotally becoming increasingly difficult to support given the time involved and limited remuneration. Whilst there is general agreement that this issue is of crucial importance for veterinary schools, there appears currently to be little enthusiasm for a common final year assessment across the schools.

A possible way forward is illustrated below where the two 'extremes' are a national licensing examination and non-collaborative school-specific systems.



Such a collaboration with the aim of comparing standards can only be achieved on the background of a sound understanding of assessment instruments. In order to focus discussions on how to achieve 'the best of both worlds' in this context (Schuwirth 2007), this document has been prepared to provide an up to date overview of assessment methods, their utility and limitations.

Assessment Terminology Overview

More specific detail on some of the terminology is given in the glossary (towards the end of the booklet). However, our inclusion of comments on reliability and validity in forthcoming sections necessitates a brief overview of these aspects here.

Reliability is defined as the reproducibility and accuracy of results – in assessment science, this is often computed as a reliability coefficient between 0 and 1.

Validity addresses the question of whether a test measures what it is supposed to measure. This is a complex area and more details on validity as it relates to assessment are given in Hopkins (1998), and Schuwirth and van der Vleuten (2004). However, for the purposes of each assessment method presented here, we will refer to the model of Miller's pyramid (Miller 1990) as a useful guide for selecting an assessment method which is valid for the competency to be tested (Fig 1).

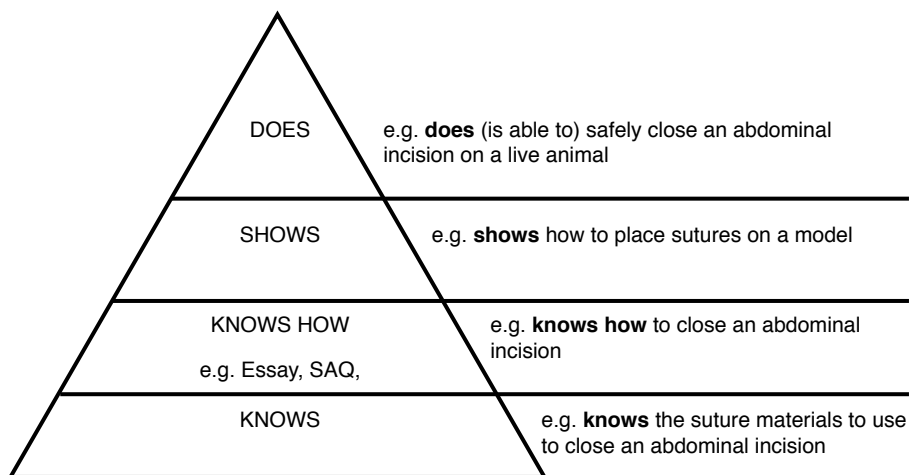


Figure 1. Miller's Pyramid aligned with a specific example of the stages in the acquisition of knowledge and skills pertaining to a specific task

Further Subdivision of the Cognitive (Knowledge) Domain

When considering some forms of assessment, it is also useful to refer to Bloom's taxonomy (Bloom 1984). This taxonomy categorises knowledge into 6 domains (knowledge, comprehension, application, analysis, synthesis and evaluation) reflecting progressive contextualization of knowledge. This structural framework is relevant when considering the more complex types of tests described in Section 2 such as the script concordance test but is also useful to consider when writing items e.g. for Multiple Choice Questions (MCQs) and Short Answer Questions (SAQs).

Finally, it is also worth noting 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.

Section 1: Assessment Methods used in Veterinary Medicine

The following section describes assessment methods commonly used in UK veterinary schools to assess students in the preclinical, paraclinical and clinical phases of undergraduate training² and as part of postgraduate certification.

Examples of questions and / or marking sheets for some of the assessment methods included in Section 1 are given in Appendices 1 to 4.

Section 2 includes examples of other assessment methods used in medicine, particularly in the postgraduate context and for assessment in the workplace, which may be relevant to and / or are beginning to be trialled in veterinary schools.

² The UK undergraduate programme is a 5 year course

Multiple Choice Questions (MCQs)

Description: A multiple choice question (MCQ) 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 one being a better match to the stem than the others. MCQs are used to test knowledge (factual recall) objectively and efficiently (computer-marked). MCQs can be structured to test higher order skills and levels of cognition such as understanding, application of knowledge and evaluation of information, when the question stem may take the form of a clinical vignette. Clearly MCQs will involve an element of guessing based on partial knowledge, and care should be taken not to cue the answer in the question. The tests can be used formatively (in-training) as an indicator of progress, as well as summatively. The MCQ format may encourage students to take a superficial approach to learning as a correct answer may depend purely on factual recall rather than understanding. Variations of the MCQ test format include negative marking where the correct answer gains a mark, the wrong answer loses a mark and no response scores zero. Negatively marked MCQs are known to be stressful and affected by the student's willingness to 'take a risk'. MCQs are extensively used in veterinary assessments. One example of a computer-based, large (360 question), high stakes MCQ is the NAVLE (North American Veterinary Licensing Examination). MCQs, along with Extended Matching Questions (EMQs) and Short-answer Questions (SAQs) are used by some medical schools for 'progress testing' - a longitudinal exam with regular sampling throughout the course. The improvement in students' scores can be used to monitor progress. MCQs are the most common written test at all levels of medical education.

Skills Assessed: Factual knowledge / knowledge recall, +/- understanding, application and interpretation.

Practical Considerations: The MCQ exam can be presented in a paper-based format or on a computer. Both can be computer-marked resulting in considerable savings in staff marking time compared with other methods e.g. essays (if these methods are used to test knowledge only). However, the development of the large number of test items (questions) required for an exam is both time consuming and challenging particularly when designing questions to assess higher order skills. MCQs give better coverage of the examinee's knowledge of a subject area than other methods e.g. essays.

Validity and Reliability: Appropriate at Miller's pyramid level/s: 'knows', +/- 'knows how'. The reliability should be monitored with a target coefficient (Cronbach's alpha) in excess of 0.8. For any item (question) the reliability indicates the generalisability of that item: the student's score should correlate with the performance on other related items. Training for those writing MCQs helps to improve the quality and reliability. If questions provide good coverage of the subject area and are correctly designed to test 'knows' or 'knows how', then test validity will be high. MCQ marks may show a gender bias with males outperforming females when compared to results in other test formats.

Key Points:

- High reliability
- Computer marking saves time and resources
- Writing items to test higher cognitive levels is time consuming

Further Reading

A North American Study of the Entry-Level Veterinary Practitioner: A Job Analysis to Support the North American Veterinary Licensing Examination (NAVLE®).

Report: <http://nbvme.taopowered.net/?id=13&page=2003+NAVLE+Job+Analysis+Report>

Anderson J. Multiple-choice questions revisited. *Med Teach* 2004;26(2):110-3.

Case SM, Swanson DB. Constructing Written Tests For the Basic and Clinical Sciences National Board of Medical Examiners, USA, 3rd Edition, 2002.

http://www.nbme.org/PDF/ItemWriting_2003/2003IWGwhole.pdf

McCoubrie P. Improving the fairness of multiple-choice questions: a literature review. *Med Teach* 2004;26(8):709-12.

Extended Matching Questions (EMQs)

Description: EMQs are designed to test more complex understanding than MCQs and have been reported to test clinical reasoning. The EMQ format has four components and starts with a title or theme statement defining the subject area e.g. 'Equine Surgery - Colic' (and an example of such an EMQ is shown in Appendix 1). The title is followed by the list of 'options' (numbered or lettered) - the possible answers to the question/s or 'item/s' that follow. A lead in statement then provides instructions and links the list of answers (options) to the question/s (item/s), which often take the form of a clinical vignette. The examinee has to respond to each question by selecting the best answer from a large list (range from 5 up to 20+), 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 cuing. Usually 1 to 2 minutes is allowed per question.

Skills Assessed: Factual knowledge / knowledge recall, understanding and interpretation, clinical reasoning.

Practical Considerations: Similar to MCQ, time per question is short and the exam can be computer marked. Although, generally used for testing higher order skills, question writing may take more time and require more training.

Validity and Reliability: Appropriate at Miller's pyramid level/s: 'knows', 'knows how'. EMQs have also been shown to have validity for assessing clinical reasoning. Reliability: similar to MCQ.

Key Points:

- Reduced chance of guessing the correct answer
- Questions can be written to test clinical reasoning
- Potentially high reliability
- Question writing can be time consuming

Further Reading

Beullens J, Struyf E, van Damme B. Do extended matching multiple-choice questions measure clinical reasoning? *Med Ed* 2005;39(4):410-7.

Tomlin J, Pead MJ, May SA. Veterinary student attitudes towards the assessment of clinical reasoning using extended matching questions. In press *J Vet Med Educ* 2008;35(4).

Tomlin J, Pead MJ, May SA. Attitudes of veterinary faculty to the assessment of clinical reasoning using extended matching questions. In press *J Vet Med Educ* 2008;35(4).

Wilson RB, Case SM. Extended Matching Questions: An Alternative to Multiple-choice or Free-response Questions. *J Vet Med Educ* 1993;20(3):75-81.

Short-answer Questions (SAQs)

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, a SAQ may require the examinee to construct a short response (several sentences, a plan or a diagram) and in some contexts write a short 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 cuing effect as examinees are not presented with the correct answer amongst a number of other choices.

Skills Assessed: Knowledge, understanding and application of knowledge.

Practical Considerations: Considerable resources required for marking - 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 / computer-marked alternatives (MCQs/EMQs). Compared with essays, SAQs are easier to write and mark and are 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.

Validity and Reliability: Appropriate at Miller's pyramid level/s: 'knows', 'knows how'. Reliability 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. Reliability improved if marking sheets are used and the test is of adequate length.

Key Points:

- Resource intensive marking compared to MCQ/EMQ (unless computer-markable)
- Heterogeneity in interpretation of the term
- Reliability improved if structured marking schemes employed
- No cuing effect

Further Reading

Rademakers J, Cate ThJ ten, Bar PR. Progress testing with short answer questions Med Teach 2005;27(7):578-82.

Schuwirth LWT, van der Vleuten C. ABC of learning and teaching in medicine: Written assessment. BMJ 2003;326:643-5.

Schuwirth LWT, van der Vleuten C. Different written assessment methods: what can be said about their strengths and weaknesses? Med Ed 2004;38(9):974-9.

Essays

Description: 'a short literary composition on a particular theme or subject, usually in prose and generally analytic, speculative, or interpretative.'^a Essays can be used in-course and completed over several days/weeks or under timed exam conditions. Sometimes essays are 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.

Skills Assessed: Knowledge, understanding, integration of knowledge, ability to go beyond taught material, writing skills. In a clinical context can be used as a test of ability to present and discuss a case 'with' a professional colleague.

Practical Considerations: Marking is labour intensive but provision of a model or outline answer will reduce the amount of time required. Consider computer-based / computer-marked alternatives for certain skills e.g. MCQs for factual knowledge and understanding. Techniques to detect plagiarism should be considered.

Validity and Reliability: Appropriate at Miller's pyramid level/s: ('knows'), 'knows how'. Reliability has potential to be quite low and is affected by the number of examiners, a lack of inter-rater reliability due to examiner subjectivity and the range of assessment criteria applied by the different examiners. Try to standardise allocation of marks, apply objective marking criteria, provide examiners with marking scheme and instructions and, where appropriate, a model answer.

Key Points:

- Resource intensive marking
- Low reliability
- Double marking recommended to improve reliability
- Heterogeneity in interpretation of the term

Further Reading

^a <http://dictionary.reference.com/> Accessed: 05/03/2008.

Schuwirth LWT, van der Vleuten C. ABC of learning and teaching in medicine: Written assessment. *BMJ* 2003;326:643-5.

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

Viva / Viva Voce / Oral

Description: The examinee is 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 to probe 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.

Skills Assessed: Knowledge, application and understanding of knowledge, problem solving, clinical reasoning, communication skills, professionalism.

Practical 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 are a more feasible option (resource-wise) when used only to examine pass/fail or borderline cases but the reliability and validity issues (see below) can be problematic in such high stakes examinations.

Validity and Reliability: Appropriate at Miller's pyramid level/s: 'knows', 'knows how'. Valid for exploring the examinee's knowledge and understanding of a subject area and for assessing communication skills and aspects of professionalism. Reliability is low due to a lack of standardisation of questioning and marking, and the possibility of examiner bias and 'halo effects'. Reliability can be improved when using the same questions for all students, having a structured marking system, increasing the number of vivas per examinee, having a testing time of 4 hours or more and with improved examiner training. Variations on the oral include 'case based discussions' and 'chart stimulated recall' which are used in medicine (see Section 2).

There is debate about the use of vivas in pass/fail situations as there are issues about the level of reliability and validity especially when the skills assessed can, or already have been, tested using more reliable methods e.g. factual knowledge examined using MCQs. For these reasons, an increasing number of institutions / examining bodies are reviewing policies around the use of vivas.

Key Points:

- Heterogeneity in interpretation of the term
- Low reliability unless multiple examiners, multiple cases and large testing time
- Often seen as having high face validity
- Resource intensive

Further Reading

Davis MH, Karunathilake I. The place of the oral examination in today's assessment systems. *Med Teach* 2005;27(3):294-7.

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

Practical Assessment (the ‘Spot’ test)

Description: This category is included as it has been a traditional assessment format in many UK veterinary schools – particularly in disciplines such as anatomy and pathology. Various local terms are used to describe this type of assessment including ‘Spot’, ‘Steeplechase’, ‘Timed stations’ or ‘Bellringer’. However, there are few references in the literature to the method. 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. There is also a similarity with OSCE type examinations especially the organisational aspect of timed stations.

Skills Assessed: Knowledge, application of knowledge, interpretation.

Practical Considerations: Resources required to set up the stations, run the exam and marking (although depending on the answer format some elements can be computer-marked).

Validity and Reliability: Appropriate at Miller’s pyramid level/s: ‘knows’, ‘knows how’, +/- ‘shows’ e.g. Write a report for referring vet on...

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

Key Points:

- Still in common use but little published in literature on description
- Heterogeneity in interpretation of the term
- Reliability improved if structured marking schemes employed

Further Reading

Note: Literature searching to date for further information on this method has found no specific papers (picking up only variants of the OSCE examination).

Objective Structured Clinical Examination (OSCE)

Description: The Objective Structured Clinical Examination (OSCE) was introduced in medical education over 30 years ago as a more standardised and objective way of assessing certain clinical skills (Harden *et al.* 1975) and is now in widespread use. OSCEs have recently been introduced in the veterinary field and are sometimes called OSPVE (Objective Structured Practical Veterinary Examination) or OSPRE (Objective Structured PRactical Examination). These terms reflect the inclusion of a broader content including paraclinical material such as pathology or veterinary public health.

The exam consists of multiple mini-stations (typically 10 – 20) commonly lasting 5 (but up to 15) minutes. Examinees rotate round stations in sequence completing a variety of tasks which test a range of skills. The examinee reads the scenario, then enters the station and undertakes the task. Stations might test: aseptic gloving; bandaging a wound; placement of an intravenous catheter; discussing a case with a client. The station set-up varies and can include: live animals, models, part-task trainers, laboratory equipment, simulated clients. Each station is marked using a checklist, with or without a global rating scale for the examiner to make a more subjective assessment. The examiner acts in a purely observational role ticking a checklist but not asking questions in the traditional sense. Sometimes ‘must pass’ OSCEs are used for certain ‘critical’ skills or as a ‘rite of passage’ to parts of the course. In Appendix 2 there is an example of a veterinary OSCE marking checklist for the task of hand washing.

Skills Assessed: Clinical practical, technical & diagnostic skills, communication skills.

Practical Considerations: Considerable resources are required (costs of equipment and personnel / staff time) to develop and set up the stations, and to run the OSCE exam. However, the checklists can be computer-marked.

Validity and Reliability: Appropriate at Miller’s pyramid level/s: ‘shows’, (not ‘does’ as assessment completed in an artificial setting).

Validity is high for specific clinical practical/technical skills. It is however low for skills that depend on more complex interactions, which are difficult to assess in the short time frame. The OSCE format also tends to fragment tasks into individual component skills. Reliability is usually high. 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. Using OSCEs to assess certain non-technical skills e.g. communication or attitudes is more challenging and tends to have lower reliability and poor generalisability (although these factors can be improved if a number of stations are used).

Key Points:

- High reliability compared to longer individual clinical examination stations
- Potential to compromise validity by excessively deconstructing tasks
- Resource intensive to establish, set up and run

Further Reading

Bark H, Cohen R. Use of an objective, structured clinical examination as a component of the final-year examination in small animal internal medicine and surgery. *JAVMA* 2002;221(9):1262-5.

Bark H, Shahar R. The Use of the Objective Structured Clinical Examination (OSCE) in Small-Animal Internal Medicine and Surgery. *J Vet Med Educ* 2006;33(4):588-92.

Davis MH. OSCE: the Dundee experience. *Med Teach* 2003;25(3):255-61.

Davis MH, Ponnampereuma GG, McAleer S, Dale VHM. The Objective Structured Clinical Examination (OSCE) as a Determinant of Veterinary Clinical Skills. *J Vet Med Educ* 2006;33(4):578-87.

Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of Clinical Competence using Observed Structured Examination. *BMJ* 1975;1(5955):447-51.

Hodges BD. The Objective Structured Clinical Examination: three decades of development. *J Vet Med Educ* 2006;33(4):571-7.

Long Case

Description: The examinee spends a period of time (typically 30 minutes to an hour) with a client and patient, gathers a history, performs a clinical examination and then reports the findings to the examiner/s. In some instances, the examiner is present throughout and will ask the examinee to 'think aloud' during the client / patient encounter. Although representing the real clinical setting more closely than most other assessment methods, the long case has been largely replaced by other more structured methods e.g. OSCEs (and in post graduate medicine by the mini-CEX and DOPS – see Section 2) because of reliability and practicality issues. Variations on the long case include direct observation during the clinical examination and evaluation of video footage. A more objective version of the long case is the Objective Structured Long Examination Record (OSLER) where examiners use a structured record or mark sheet to assess each examinee.

Skills Assessed: Clinical competence, communication skills, history taking, physical examination, problem solving.

Practical Considerations: Very resource intensive, which renders long cases impractical when faced with the assessment of large numbers of examinees.

Validity and Reliability: Appropriate at Miller's pyramid level/s: 'shows', +/- 'does'. The face validity is high as the exam closely represents everyday encounters between clinician and client/patient. However, the reliability is low particularly if there is only one examiner and only one or a very limited number of cases as performance on one case does not reliably predict performance on another case. Therefore, if using long cases, the reliability can be improved by having multiple examiners and many cases (for example 10 cases of 20 minutes each). Other issues include lack of standardisation of cases and content between examinees, and inadequate sampling of a subject area to test competence.

Key Points:

- High face validity
- Resource intensive
- Low reliability unless multiple examiners, multiple cases and large testing time

Further Reading

Gleeson F. AMEE Medical Education Guide No 9: assessment of clinical competence using the Objective Structured Long Examination Record (OSLER). *Med Teach* 1997;19(1):7-14.

Newble DI. The observed long-case in clinical-assessment. *Med Ed* 1991;25(5):369-73.

Wass V, van der Vleuten C. The long case. *Med Ed* 2004;38(11):1176-80.

Observation on Rotations

Description: Students are observed and assessed during clinical work i.e. on intramural and extramural rotations / clerkships. 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. If used summatively, students may be inhibited from asking questions, not wanting to 'look stupid'. The assessment may be undertaken by one member of staff or several members of the team. If individuals other than the clinicians are involved the assessment approaches the 360° evaluations used in medicine (see Section 2).

An example of a rotational feedback form is given in Appendix 3. An online database evaluation system 'VOLES' has been developed at UC Davis School of Veterinary Medicine for clinicians to score students on rotations using a range of predefined criteria.³

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

Practical Considerations: As the assessment is embedded in day-to-day work there are relatively low demands on resources.

Validity and Reliability: Appropriate at Miller's pyramid level/s: 'shows', 'does'. Validity is high as the assessment occurs 'on the job' in the clinical environment. 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, staff 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:

- High face validity
- Low reliability
- Subjective and prone to 'halo effects'

Further Reading

Miller GE. The assessment of clinical skills/ competence/ performance. Acad Med 1990;65(9):S63-7.

Turnbull J, Van Barneveld C. Assessment of Clinical performance: In-Training Evaluation. International Handbook of Research in Medical Education. Kluwer Academic Publishers; 2002;793-810.

³ Personal communication from Jan Ilkiw, UC Davis

Portfolios

Description: A portfolio is a collection of work developed as a cumulative 'body of evidence' to demonstrate the student's learning and achievements. It is not an assessment method in its own right, rather a receptacle containing a mixture of materials. Each piece may be assessed individually and/or a mark or grade is awarded to the portfolio as a whole. 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 (Royal College of Veterinary Surgeons Professional Development Phase - PDP)⁴ and activity logs. Evidence of the student's reflections on learning is a valuable aspect of a portfolio. Portfolios have been used in veterinary 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 ongoing process, can be formative and/or summative, and in an ideal situation involves more than one marker. 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.

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

Practical 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 can be affected by issues such as: learner type and maturity; 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'.

Validity and Reliability: Appropriate at Miller's pyramid level/s: 'shows', evidence of 'does'.

The validity of a portfolio can be high if the material (evidence) included accurately reflects real-life activities that provide evidence that the required learning outcomes are being achieved. 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. Reliability of self-assessment can be poor. Additionally, assessing the student's process of reflection is not straightforward.

⁴ Royal College of Veterinary Surgeons (RCVS) Professional Development Phase (PDP)
<http://www.rcvs.org.uk/Templates/Internal.asp?NodeID=94970&int1stParentNodeID=94964>

Key Points:

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

Further Reading

Friedman Ben David M, Davis MH, Harden RM, Howie PW, Ker J, Pippard MJ. AMEE Medical Education Guide No. 24: Portfolios as a method of student assessment. *Med Teach* 2001;23(6):535-51.

Challis M. AMEE medical education guide no. 11 (revised): Portfolio-based learning and assessment in medical education. *Med Teach* 1999;21(4):370-86.

Davis MH, Ponnampetuma GG. Portfolio assessment. *J Vet Med Educ* 2005;32(3):279-84.

Section 2: Additional Assessment Methods used in Medicine

In this section, we describe a number of assessment methods used in medicine and other health professions. These methods were developed to try to address the challenges of assessing performance in a valid, reliable and feasible way when the trainee is engaged in clinical work.

Most of these methods have been developed primarily for use in the postgraduate period and this should be considered before deciding if and when the method is appropriate to the veterinary context. However, it is anticipated that as veterinary educators seek to improve assessment practices, information from medical assessment will continue to be useful and instructive. Furthermore, in order to address assessment at the highest level of Miller's pyramid ('Does') and to assess performance in authentic clinical situations, it is highly relevant to consider these methods.

In Appendix 5 there is a list of papers that include examples of marking sheets used for these assessment methods.

Clinical Evaluation Exercise (CEX)

Description: The original clinical evaluation exercise (CEX) was introduced several decades ago to replace orals used in the assessment of clinical competency. The exam format involves a relatively long (typically 2 hour) pre-planned single patient encounter in a clinical setting. A clinician observes the trainee taking a history and performing a physical examination. The trainee presents the findings, decides on the diagnosis and the treatment plan. Additionally, a written patient report is produced. The examiner gives feedback. For reasons of practicality and reliability, the CEX has been superseded by the mini-CEX. **

Skills Assessed: History taking, physical examination, problem solving, clinical reasoning, communication (oral and written).

Practical Considerations: Time consuming. Prohibitively so since the number of trainees has risen and the number of CEXs and examiners is increased to improve reliability.

Validity and Reliability: Appropriate at Miller's pyramid level/s: 'does'.

Validity is reasonable although the trainee has much more time than in a real clinical situation. Results from one CEX fail to generalise i.e. one clinical case is a poor indicator of performance on other clinical cases. Reliability is low but can be improved by increasing the number of CEXs, using several assessors per exam and different assessors at subsequent CEXs.

Key Points:

- High face validity
- Time consuming
- Low reliability

Further Reading

Kroboth FJ, Hanusa BH, Parker S, Coulehan JL, Kapoor WN, Brown FH, *et al.* The inter-rater reliability and internal consistency of a clinical evaluation exercise. *J Gen Intern Med* 1992;7:174-9.

Woolliscroft JO, Stross JK, Silva J Jr. Clinical competence certification: a critical appraisal. *J Med Ed* 1984;59:799-805.

** Although the CEX has now been superseded by methods that are more practical and feasible to implement, it has been included by way of a general introduction to methods that assess clinical competency, and the range of skills that entails, in real patient-based situations.

Mini-clinical Evaluation Exercise (mini-CEX)

Description: The mini-CEX involves 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 an out-patient clinic. The exam lasts 15 – 20 minutes and is followed by immediate feedback from the examiner. Typically, 4 to 6 mini-CEX exams per year are taken and in a variety of situations. The exam 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. Primarily used formatively with feedback and an action plan is produced which is structured to support the trainee's learning. Mini-CEX forms may be included in a trainee's portfolio. A veterinary version, the Veterinary Clinical Assessment Tool (V-CAT), based on the mini-CEX has been developed and trialled at the Faculty of Veterinary Medicine, University of Glasgow. The initial use has been in the context of the Small Animal General Practice rotation and an example of the V-CAT is shown in Appendix 4.

Skills Assessed: History taking, physical examination, problem solving, clinical reasoning, communication.

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

Validity and Reliability: Appropriate at Miller's pyramid level/s: 'does'. Validity is high as the encounters occur in the real working environment while examining real patients. Reliability increases with the number of encounters i.e. mini-CEXs, with 4 – 6 per year as optimal (increased reliability versus practicality). Assessor training is also important for reliability and to improve the quality of feedback.

Key Points:

- High face validity
- Reliability increases with number of examinations (mini-CEXs) performed

Further Reading

Kogan JR, Bellini LM, Shea JA. Implementation of the mini-CEX to evaluate medical students' clinical skills. *Acad Med* 2002;77(11):1156-7.

Norcini JJ. The Mini Clinical Evaluation Exercise (mini-CEX). *Clin Teach* 2005;2(1):25-30.

Longitudinal Evaluation of Performance (LEP)

Description: This is similar in format to the mini-CEX, with an observer using a rating form to assess a variety of aspects of a trainee's performance in the workplace, including clinical skills (technical, knowledge and reasoning) and aspects of professionalism. As with the mini-CEX, it is used primarily formatively and with feedback. However, there are some key differences between LEP and mini-CEX. Evaluations are performed more frequently, typically at weekly intervals and the mark given at any point in time is referenced against the standard expected on completion of a particular part of clinical training. The assessment provides a record of the trainee's progress towards the required standard. The LEP is intended to be flexible and does not require coverage of a specific list of cases. Instead, a range of cases are chosen by the trainee's assessor / supervisor, guided by a wide-ranging list of competencies. Integral to the LEP is the feedback process, when marks are discussed and used to identify areas for improvement. For an example of a marking sheet / rating scale see Prescott-Clements *et al.* 2008.

Skills Assessed: Technical skill, knowledge, clinical reasoning, professionalism, communication.

Practical Considerations: The observation and feedback process takes about 30 minutes and this, combined with the frequency of 'sampling', necessitates a commitment to supporting and resourcing such an assessment process. Also training is required to ensure examiners understand how to use the rating scale and give feedback.

Validity and Reliability: Appropriate at Miller's pyramid level/s: 'does'. Validity is high as observations are made regularly during real patient encounters and a range of cases are covered. As frequency of testing is high, reliability should also be reasonable provided that assessors are adequately trained.

Key Points:

- Major focus is on feedback
- High face validity

Further Reading

Prescott LE, Norcini JJ, McKinlay P, Rennie JS. Facing the challenges of competency-based assessment of postgraduate dental training: Longitudinal Evaluation of Performance (LEP). *Med Ed* 2002;36(1):92-7.

Prescott-Clements L, van der Vleuten C, Schuwirth LWT, Hurst Y, Rennie JS. Evidence for validity within workplace assessment: the Longitudinal Evaluation of Performance (LEP). *Med Ed* 2008;42(5):488-95.

Directly Observed Procedural Skills (DOPS)

Description: Directly Observed Procedural Skills (DOPS), also referred to as Direct Observation of Procedural Skills, is designed specifically to assess practical skills in a workplace setting. A trainee 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.* 2005). For any particular skill the trainee has to pass a number of assessments (typically six) to be signed off as competent at that skill.

Skills Assessed: Practical / technical skills.

Practical 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.

Validity and Reliability: Appropriate at Miller's pyramid level/s: 'does'. Validity high as observation is made during normal working day while trainee performs the procedure on a real patient. 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 face validity
- Multiple assessments of the same skill

Further Reading

Wilkinson JR, Crossley JGM, Wragg A, Mills P, Cowan G, Wade W. Implementing workplace-based assessment across the medical specialties in the United Kingdom. *Med Ed* 2008;42(4):364-73.

Wragg A, Wade W, Fuller G, Cowan G, Mills P. Assessing the performance of specialist registrars. *Clin Med* 2003;3(2):131-4.

360° (Multi-source Feedback)

Description: This is also called multi-source feedback (MSF) and 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 360° assessment form see: Wood, Wall *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; representatives of all groups in the clinician's daily working environment (not just co-professionals); and sometimes from patients and their families. The feedback is used as part of appraisals and to help clinicians gain insight into their professional development.

Skills Assessed: Communication, team working, professionalism.

Practical 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 – 10 minutes to complete.

Validity and Reliability: Appropriate at Miller's pyramid level/s: 'does'.

Validity is high as information about performance is based on observations during daily clinical practice from a range of members of the clinical team and other stakeholders. 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 360° is making good use of the feedback.

Key Points:

- High face validity
- Allows feedback from range of individuals (a variety of staff +/- patients)
- Resource intensive

Further Reading

Evans R, Elwyn G, Edwards A. Review of instruments for peer assessment of physicians. *BMJ* 2004;328:1240–3.

Wood L, Wall D, Bullock A, Hassell A, Whitehouse A, Campbell I. 'Team observation': a six-year study of the development and use of multi-source feedback (360-degree assessment) in obstetrics and gynaecology training in the UK. *Med Teach* 2006;28(7):e177-84.

Wood L, Hassell A, Whitehouse A, Bullock A, Wall D. A literature review of multi-source feedback systems within and without health services, leading to 10 tips for their successful design. *Med Teach* 2006;28(7):e185-91.

Case-based Discussion and Chart Stimulated Recall Oral Examination (CSR)

Description: A formal discussion between a trainee and an assessor about a case that the trainee has managed and been directly responsible for. 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: Intercollegiate Surgical Curriculum Website <http://www.iscp.ac.uk/Assessment/WBA/CBD.aspx>). The discussion is followed by a short feedback session. Case-based discussions are primarily used for formative assessment ('in-training'). 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 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.

Practical 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.

Validity and Reliability: Appropriate at Miller's pyramid level/s: 'shows' / 'does'. Validity is high as the discussion is centred on a case 'owned' and managed by the trainee. 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 (see Section 1).

Key Points:

- High face validity
- Mostly used formatively
- Low reliability

Further Reading

Cunningham JPW, Hanna E, Turnbull J, Kaigas TB, Norman GR. Defensible Assessment of the Competency of the Practicing Physician. *Acad Med* 1997;72(1):9-12.

Intercollegiate Surgical Curriculum Website. Case Based Discussion <http://www.iscp.ac.uk/Assessment/WBA/CBD.aspx> Accessed: 05/08/2008.

Jennett P, Affleck L. Chart audit and chart stimulated recall as methods of needs assessment in continuing professional health education. *J Cont Ed Health Prof* 2005;18(3):163-71.

Script Concordance Test (SCT)

Description: The Script Concordance Test (SCT) is a tool designed to assess decision-making and clinical reasoning skills. In everyday work, experienced clinicians refer to 'scripts' when using their knowledge to make decisions. These scripts are built up over years in clinical practice. 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. This is followed by a proposed diagnosis or suggested treatment or action. Examinees have to rate the effect of further information 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 correct answer.

Skills Assessed: Knowledge, interpretation, decision making and clinical reasoning

Practical Considerations: The main considerations are the time and practice required to develop suitable test items and the time required to produce the scoring system for each answer, which is based on responses from a panel of experts.

Validity and Reliability: Appropriate at Miller's pyramid level/s: 'knows how' [Note however, that this test assesses at the higher levels of Bloom's taxonomy – analysis, synthesis and evaluation].

Validity depends on the test item design, which should present an authentic case and require an element of reasoning and degree of experience. Well designed SCTs have been validated with scores increasing with clinical experience. Reliability has also been found to be high if there are sufficient questions (in excess of 50).

Key Points:

- Written test based on scenario/ vignette
- Significant training required for item writing
- Good reliability if sufficient questions are used

Further Reading

Charlin B, Roy L, Brailovsky C, Goulet F, van der Vleuten C. The Script Concordance Test: A Tool to Assess the Reflective Clinician. *Teach Learn Med* 2000;12(4):189-95.

Meterissian S, Zabolotny B, Gagnon R, Charlin B. Is the script concordance test a valid instrument for assessment of intraoperative decision-making skills? *Am J Surg* 2007; 193(2):248-51.

A veterinary example of a Script Concordance Test is available at:

http://www.veteducation.org/resources/view_workshop2006_SCT_examples.pdf

References and Bibliography

- Bloom BS. Taxonomy of Educational Objectives. In: The Cognitive Domain, David McKay Company Inc, 1984.
- Boursicot KAM, Roberts TE, Pell G. Using borderline methods to compare passing standards for OSCEs at graduation across three medical schools. *Med Ed* 2007;41(11):1024–31.
- Crosby J. Assessment of the student practitioner. In: *Effective Learning & Teaching in Medical, Dental & Veterinary Education*; Sweet J, Huttly S, Taylor I, editors. 2003. p. 71-89.
- General Medical Council: Strategic Options for Undergraduate Medical Education Consultation. Final Report. General Medical Council Education Committee 2006. http://www.gmc-uk.org/education/documents/strategic_outcomes_final_report_jun_2006.pdf Accessed: 07/08/2008.
- Hopkins KD. Test Validity. In: *Educational and psychological measurement and evaluation*. 8th edition. Allyn and Bacon, Needham Heights, MA, 1998.
- Jackson N, Jamieson A, Khan A. (Eds) *Assessment in Medical Education and Training - A practical guide*. Radcliff Publishing Ltd. 2007
- Miller GE. The assessment of clinical skills/ competence/ performance. *Acad Med* 1990;65:s63-s67.
- Norcini J, Burch V. Workplace-based assessment as an educational tool: AMEE Guide No. 31 *Med Teach* 2007;29(9):855-71.
- Rhind SM. Competence at graduation: implications for assessment. *J Vet Med Educ* 2006;33(2):172-5.
- Rhind SM, Baillie S, Brown F, Dozier M, Hammick M. Assessing Competence in Veterinary Medical Education – Where’s the Evidence? In press *J Vet Med Educ* 2008;35(3)**
- Schuwirth L. The need for national licensing examinations. *Med Ed* 2007;41(11):1022–3.
- Schuwirth LWT, van der Vleuten C. How to design a useful test: the principles of assessment. Association for the Study of Medical Education, Understanding Medical Education series, 2004.
- Streiner DL, Norman GR. Biases in responding. In: *Health Measurement Scales Third edition*. Oxford University Press, 2003.
- Swing S, Bashook PG. *Toolbox of Assessment Methods: A product of the joint initiative of the ACGME Outcome Project of the Accreditation Council for Graduate Medical Education (ACGME), and the American Board of Medical Specialties (ABMS)*. Version 1.1 September 2000. <http://www.acgme.org/Outcome/assess/Toolbox.pdf>
- van der Vleuten C. Validity of final examinations in undergraduate medical training. *BMJ* 2000;321:1217-9.
- van der Vleuten C, Schuwirth LWT. Assessing professional competence: from methods to programmes. *Med Ed* 2005;39(3):309-17.
- Wass V, McGibbon D, Van der Vleuten C. Composite undergraduate clinical examinations: how should the components be combined to maximize reliability. *Med Ed* 2001;35(4):326-30.

** N.B. Part 1 of the RCVS Trust 'Blue Sky' funded project involved a systematic review of the evidence base around clinical and professional assessment in veterinary education and is reported in the *Journal of Veterinary Medical Education (in press)*. The review process followed the BEME (Best Evidence Medical Education) protocol. More details of the review process and findings are available at: <http://www.vet.ed.ac.uk/beme/>

Glossary: Definition of Terms

Introduction

This section attempts to define words and terms that readers will encounter in the wider assessment literature. It is intended that the list will be edited and extended as a result of feedback from readers and as part of the ongoing development of the booklet (i.e. in Version 1.2 and so on).

Validity: Does the assessment method measure what it is supposed to measure? Is the assessment method fit for purpose? If the exam is being used to test a particular skill, is that skill actually being assessed? For example, an MCQ would not be a valid assessment method to use to test a clinical skill. Ideally several pieces of evidence should be used when validating a test. Validity can be subdivided into different categories which include:

Face validity: the assessment method, on first impression, appears to measure the intended competency (Hopkins 1998)

Content validity: refers to the content of the assessment and how representative it is of the desired learning objectives

Reliability: How reproducible are the results? Reliability is measured and expressed as a coefficient ranging from 0 to 1 i.e. from zero reliability to complete reliability. Reliability can be affected by and expressed in terms of the consistency of results across examiners (inter-rater reliability), for the same examiner at different times (intra-rater reliability), between examinees, and between test items.

For further reading on validity and reliability, see:

- Hopkins (1998)
- Schuwirth and van der Vleuten (2004)
- van der Vleuten and Schuwirth (2005)

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.

Summative Assessment: Usually associated with a mark or grade and often occurs towards the end of a course. There is clearly overlap between these two categories as results and feedback from summative tests can also be used formatively.

Global rating forms: These differ from other forms or checklists as the rater/s assess a range of skills or abilities such as: technical ability, consultation skills, knowledge, history taking, professionalism, team working and communication. Each skill is rated on a scale with categories that represent the range from unsatisfactory to above expected performance levels. The forms usually include areas for the examiner to provide feedback comments and may include basic case / patient information. Global rating forms are used in a number of assessment methods e.g. mini-CEX, LEP, at the end of a clinical rotation.

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 (Streiner and Norman 2003)
- 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'.

Cuing 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 cuing.

Blueprint: Indicates for an exam the content / areas covered, and the relative amounts / approximate number of questions typically expected in each area.

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.

Standard Setting: Variety of different methods available e.g. Angoff's, Hofstee's. For further information:

Norcini, J. J. (2003). Setting standards on educational tests. *Medical Education*, 37, 464-469.

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 whole cohort.

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.

Acknowledgements

The authors wish to thank the following people for advice, contributions to and assistance during the preparation of this guide:

- Stephen May, Peter Nunn, Jane Tomlin, Adrian Boswood and Tierney Kinnison from the Royal Veterinary College, University of London
- Marshall Dozier, Fiona Brown and Catriona Bell from the Royal (Dick) School of Veterinary Studies, University of Edinburgh
- Carol Gray, University of Liverpool and Liz Mossop, University of Nottingham

And to thank those who supplied additional material for Appendices:

- Jane Tomlin from the Royal Veterinary College, University of London (Appendix 1)
- John Mosley from the Royal (Dick) School of Veterinary Studies, University of Edinburgh (Appendix 2)
- Danielle Gunn-Moore from the Royal (Dick) School of Veterinary Studies, University of Edinburgh (Appendix 3)
- Jenny Hammond and Jim Anderson from the Faculty of Veterinary Medicine, University of Glasgow (Appendix 4)

Appendices

List of Appendices:

1. Example of an Extended Matched Question (EMQ)
2. Example of an Objective Structured Clinical Examination (OSCE)
3. Example of a Rotation Feedback Form
4. Veterinary Clinical Assessment Tool (V-CAT) - a Veterinary mini-CEX
5. A Reference List of Papers - that include marking sheets for assessment methods described in Section 2

Appendix 1: Example of an Extended Matched Question (EMQ)

An example EMQ supplied by Jane Tomlin from The Royal Veterinary College, University of London, that has been used as an 'example for publication'.*

I. EQUINE SURGERY - COLIC - Diagnosis

Options

- A. Acute grass sickness
- B. Anterior enteritis
- C. Colitis
- D. Large colon torsion
- E. Nephrosplenic entrapment
- F. Parturition
- G. Peritonitis
- H. Pelvic flexure impaction
- I. Rhabdomyolysis
- J. Strangulating lipoma

For each case described below, choose the single most likely diagnosis from the above list of options

Question 1: A 22 year old pony gelding presents with moderate colic of 6 hours duration. The heart rate is 50 beats/minute, the respiratory rate is 12 breaths/minute and the rectal temperature is 37.8°C. Mucous membranes are congested with a capillary refill time of 2 seconds. Gastrointestinal sounds are reduced in all four quadrants. Rectal examination reveals multiple tube-like structures of approximately 8cm diameter. Nasogastric intubation yields 3 litres of reflux. Abdominocentesis yields fluid with a total protein concentration of 40g/l.

Answer: J Strangulating Lipoma

II. SMALL ANIMAL – DERMATOLOGY – Investigation

Options

- A. adrenocorticotrophic hormone stimulation test
- B. bacterial culture & sensitivity
- C. bile acid stimulation test
- D. dermatohistopathology

* J Tomlin, M Peard and S May. The Royal Veterinary College, University of London

- E. fungal culture
- F. high dose dexamethasone stimulation test
- G. intradermal test
- H. serum thyroxine level
- I. serum thyroxine and thyroid stimulating hormone level
- J. sarcoptes serology

For each case described below, choose the single most appropriate diagnostic step from the above list of options

Question 1. An 8 year old female neutered Jack Russell terrier weighing 8kg is presented with a 5-week history of a bald patch involving the right pinna. The owners do not report any pruritus. General examination is unremarkable. Examination of the skin shows alopecia, mild erythema and scaling of the right pinna, and a circular patch of alopecia and mild scaling on the right shoulder. A coat brushing, skin scraping and hair plucks are unremarkable. Wood's lamp examination shows no fluorescence.

Answer E: fungal culture

III. FARM ANIMAL – REPRODUCTION - Treatment

- A. Equine chorionic gonadotrophin (eCG)
- B. Gonadotrophin Releasing Hormone (GnRH) analogue
- C. Gonadotrophin Releasing Hormone (GnRH) followed by Prostaglandin F2 alpha (PGF) 14 days later
- D. Human chorionic gonadotrophin (hCG)
- E. Intramuscular cephalosporin injection
- F. Intrauterine cephalosporin infusion
- G. Oestradiol benzoate injection
- H. Prostaglandin F2 alpha (PGF) injection
- I. Prostaglandin F2 alpha (PGF) – 2 injections 11 days apart
- J. Progesterone releasing device (PRID) for 12 days







For each case described below, choose the single most appropriate treatment from the above list of options

Question 1. A cow has calved 35 days ago and has not been seen in oestrus. The vagina is normal. Trans-rectal ultrasound reveals the uterine horns are less than 3cm in diameter; both ovaries are less than 1cm in diameter and no follicles more than 6mm are visible.

Answer: J Progesterone releasing device (PRID) for 12 days

Appendix 2: Example of an Objective Structured Clinical Examination (OSCE)

An example of a veterinary OSCE marking sheet (Task: hand washing). Supplied by John Mosley, Royal (Dick) School of Veterinary Studies, University of Edinburgh.

CANDIDATE NAME:		General Marks	Compulsory To pass
I			
Hands wet before soap dispensed		<input type="checkbox"/>	
Sufficient soap dispensed to produce lather over both hands			<input type="checkbox"/>
	Palm-to-palm	<input type="checkbox"/>	
	Minimum 5 repetitions	<input type="checkbox"/>	
	R over left	<input type="checkbox"/>	
	Minimum 5 repetitions	<input type="checkbox"/>	
	L over R	<input type="checkbox"/>	
	Minimum 5 repetitions	<input type="checkbox"/>	
	Palm-to-palm with fingers interlaced.....	<input type="checkbox"/>	
	Minimum 5 repetitions	<input type="checkbox"/>	
	Backs of fingers to palms	<input type="checkbox"/>	
	Minimum 5 repetitions	<input type="checkbox"/>	
	Rotational rubbing of thumb		
	Right thumb	<input type="checkbox"/>	
	Minimum 5 repetitions	<input type="checkbox"/>	
	Left thumb	<input type="checkbox"/>	
	Minimum 5 repetitions	<input type="checkbox"/>	
	Backward and forwards rotational rubbing (finger-tips)		
	R fingers to L palm	<input type="checkbox"/>	
	Minimum 5 repetitions	<input type="checkbox"/>	
	L fingers to R palm	<input type="checkbox"/>	
	Minimum 5 repetitions	<input type="checkbox"/>	
Hands rinsed thoroughly		<input type="checkbox"/>	
Wipe and dry hands with disposable paper towels		<input type="checkbox"/>	
Turn off taps using disposable paper towel/elbow control		<input type="checkbox"/>	
Dispose of paper towel in clinical waste bin			<input type="checkbox"/>
Waste bin not touched with hands			<input type="checkbox"/>
Additional comments by assessor:			

Appendix 3: Example of a Rotation Feedback Form

An example of a feedback form used to assess students when 'Observed on Rotations'. Supplied by Danielle Gunn-Moore, the Royal (Dick) School of Veterinary Studies, University of Edinburgh.

Royal Dick School of Veterinary Studies FINAL YEAR ROTATION – FORMATIVE FEEDBACK

Name of Rotation:

Name of Student:

Week of rotation:

Attendance:

	Monday	Tuesday	Wednesday	Thursday	Friday
Morning					
Afternoon					

Performance (NOTE: not all areas of performance will be assessed in every rotation)

	Unsatisfactory	Satisfactory	Good
Professional appearance (ie cleanliness, tidiness, dress and appearance)			
Veterinary knowledge + application of knowledge			
Animal handling abilities			
Quality of clinical case care			
Problem solving abilities			
Manual skills			
Communication skills			
Contribution to clinical discussions			
Objective assessment			

Additional feedback:

Overall score for rotation (circle)

FAIL

PASS

Signed: Date

Score descriptors: *PASS* – displayed professional attributes and behaviour which at all times were wholly appropriate to the practice of veterinary medicine. *FAIL* - has areas of professional attitudes or behaviour which are not entirely compatible with the practice of veterinary medicine.

Appendix 4: Veterinary Clinical Assessment Tool (V-CAT) - a Veterinary mini-CEX

35

A veterinary version of the mini-CEX. This has been recently trialed at, and is supplied by Jenny Hammond and Jim Anderson, the Faculty of Veterinary Medicine, University of Glasgow

Small Animal Clinical Studies Course: Veterinary Clinical Assessment Tool (V-CAT)

Please complete the questions or circle the appropriate response. Further information on the V-CAT is provided in the "Guide for Assessors" and "Guide for Students" provided.

Student name _____ Student ID No _____

Assessor _____ Rotation _____

Clinical setting: Consultation Wards Theatre Other

Main clinical problem(s) _____

Complexity of case: Low Average High

Focus of encounter: History Examination Diagnosis Management Explanation

Please grade the following areas using the scale below		Unsatisfactory			Highly satisfactory			U/C*
1. History taking		1	2	3	4	5	6	7
2. Animal handling		1	2	3	4	5	6	7
3. Physical examination		1	2	3	4	5	6	7
4. Communication		1	2	3	4	5	6	7
5. Practical skills		1	2	3	4	5	6	7
6. Clinical judgement		1	2	3	4	5	6	7
7. Professionalism		1	2	3	4	5	6	7
8. Organisation / Efficiency		1	2	3	4	5	6	7
9. Overall clinical care		1	2	3	4	5	6	7

U/C: Please mark this if you have not had an opportunity to observe this behaviour

Anything especially good?	Suggestions for Development
Agreed action:	

	Not at all						Highly	
	1	2	3	4	5	6	7	
Student satisfaction with V-CAT								
Assessor satisfaction with V-CAT	1	2	3	4	5	6	7	

Assessor's signature _____ Date _____

Time taken for observation (mins) _____ Time taken for feedback (mins) _____

Appendix 5: Reference List of Papers that Include Marking Sheets for Assessment Methods Described in Section 2

Mini-clinical Evaluation Exercise (mini-CEX)

Norcini JJ. The Mini Clinical Evaluation Exercise (mini-CEX). Clin Teach 2005;2(1):25-30.

Longitudinal Evaluation of Performance (LEP)

Prescott-Clements L, van der Vleuten C, Schuwirth LWT, Hurst Y, Rennie JS. Evidence for validity within workplace assessment: the Longitudinal Evaluation of Performance (LEP). Med Ed 2008;42(5):488-95.

Directly Observed Procedural Skills (DOPS)

Wilkinson JR, Crossley JGM, Wragg A, Mills P, Cowan G, Wade W. Implementing workplace-based assessment across the medical specialties in the United Kingdom. Med Ed 2008;42(4):364-73, Supporting Information: S1 p7.

360° (Multi-source Feedback)

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

Case Based Discussion and Chart Stimulated Recall Oral Examination (CSR)

Surgical Curriculum Website. Case Based Discussion
http://www.iscp.ac.uk/Documents/cbd_form.pdf

Script Concordance Test (SCT)

Medical Example:

Meterissian S, Zabolotny B, Gagnon R, Charlin B. Is the script concordance test a valid instrument for assessment of intraoperative decision-making skills? Am J Surg 2007; 193(2):248-51.

Veterinary Example:

http://www.veteducation.org/resources/view_workshop2006_SCT_examples.pdf

Further examples of forms used for a variety of workplace-based assessment methods can be found in:

Norcini J, Burch V. Workplace-based assessment as an educational tool: AMEE Guide No. 31 Med Teach 2007;29(9):855-71.