Review of the Examination Processes for Membership and Fellowship of the Australian College of Veterinary Scientists

Report to the Australian College of Veterinary Scientists

Gabrielle Matters
Australian Council for Educational Research
January 2008
List of tables
Table 1: Evaluation of quality of College oral examinations.................................5
Table 2: Fellowship examination 2007, by question choice (of and within) ...............9
Table 3: Consistency of questions with subject guidelines, membership examinations, 2007.......................................................................................10
Table 4: Analysis of state of standardisation of College oral examinations ............13
Table 5: Requirements of the examination-setting process......................................18
Table 6: Fellowship examination questions, by frequency of instructional word .................................................................19
Table 7: Setting numerical boundaries for grades......................................................27
Table 8: Five perspectives on validity........................................................................29
Table 9: Comparison of College issues with some validity perspectives ...............30
Table 10: Three aspects of reliability..........................................................................31
Table 11: Comparison of College issues with some reliability aspects .................31
Table 12: Results of hypothesis testing.....................................................................34

List of figures
Figure 1: Alignment of essential elements in a good examination question ............20
Figure 2. Score distribution (Display 1).................................................................35
Figure 3. Sex effects (Display 2)............................................................................36
Figure 4A. Sex effect on scores (Display 3)............................................................37
Figure 4B. Sex effect on scores (Display 4)............................................................38
Figure 5. Item scoring (Display 5)..........................................................................39
Figure 6. Calculating component scores (Display 6)............................................40
INTRODUCTION

Background to the review
This paper deals with examinations administered by the Australian College of Veterinary Scientists (ACVSc) (herein referred to as ‘the College’) to veterinarians at two levels – membership and fellowship. Membership examinations are set at a level that denotes an above-average level of expertise of the candidate in a particular field, signifying advisory status but no specialist status. Fellowship examinations are set at a very high standard and are intended to signify specialist status.

The system of examinations administered by the College involves two components – written and oral/practical. Candidates, for both memberships and fellowships, sit the written examination in early June in locations across Australia and New Zealand (and several other countries). Candidates take the oral/practical examination in the first week of July at a venue on the Gold Coast, Queensland.

The College commissioned a rigorous review of their examinations processes by an independent person trained in educational theory. Although constantly reviewing its examination procedures, and having involved educational expertise in this process in the past, the College’s specific requirements for this current review of the examination processes included monitoring oral examinations and identification of areas in need of change or further development.

Terms of reference
In summary, the terms of reference were as follows.

1. Review procedures and written information for examinations.
2. Review processes for preparing examination papers and oral/practical material.
3. Review the conduct of examination and assessment procedures (including grading).
4. Appraise procedures and processes against College objectives and guidelines, and against accepted standards for professional assessment.
5. Recommend improvements or alternatives.

Outline of the method
The review consisted of a real-time natural study located in the experiences of the College executive, examiners and candidates. The method was iterative, descriptive, action research. It designed to produce well-grounded and informed opinion so that any procedural recommendations would be based on existing practice and theoretical notions of validity and reliability.
Acknowledgment
This section of the report, unlike other sections, is written in the first person. Therefore, I would like to take the opportunity to acknowledge the meticulous attention to detail and the good will of all concerned with the examinations for the MACVSc and FACVSc. I very much appreciated being granted a high level of accessibility to key people, and being provided with a large volume of documentation. The College is to be commended on the transparency of its procedures.

Period and rhythm of review
The ACVSc contracted the Australian Council for Educational Research (ACER) for 15 days of research time over the period April to September 2007, subsequently extended to October 2007. The reviewer requested extra time to collate and analyse data available from the College in paper form but required in electronic form before analyses could be carried out. The nature of, and access to, data required by the researcher (but not specifically mentioned in the research proposal) is indicative of the researcher’s underestimation of the time required for the final stages of this review. The delay, however, was not reflected in extra charges for the College, and the deliverables, although the same in name, are based on a richer dataset than originally envisaged.

Sources of information
The reviewer absorbed information, both formal and informal, through reading, listening, observing, and asking questions of people and data. Sources of information, many of which are discussed in detail later, are summarised in the list below.

- Briefings from Executive Officer and Chief Examiner;
- College reports, handbooks, specifications;
- Examination material (e.g. examination papers and examination timetable);
- Attendance at meetings of examiners;
- Observation of oral examinations;
- Examination records (e.g. forms for completion by examiners and records of marks awarded to candidates);
- Informal conversations with examiners; and,
- The lived experiences of the reviewer in the world of examinations, testing and assessment.

I received initial briefings from Dr Megan Parker (Executive Officer, ACVSc) and Dr Nigel Perkins (Chief Examiner) at the College’s office at Eight Mile Plains. I maintained telephone and e-mail contact throughout the project with Dr Parker and Dr Perkins, and also with Mrs Elaine Lowe, Administrative Officer, ACVSc.

I was supplied with copies of all relevant documentation for candidates and examiners, written examination papers before editing at the College, a sample of written papers after editing had occurred at the College, details of candidates, observers and examiners for the oral/practical examinations (including a timetable for these), templates of all forms associated with oral
examinations that required completion by examiners and observers, copies of marking schemes (or indicative responses) for oral examinations (in the cases where I requested same), templates of all forms associated with marking the written examinations, the complete set of data pertaining to marks awarded to candidates on questions, papers, and overall, by whom, and the final result expressed as a grade of Pass, Fail or Supplementary for each candidate.

I scanned the literature on the assessment of veterinary and medical competencies, especially oral examinations, and I re-visited the educational measurement literature on validity of assessment, reliability of results, and behaviour of examinees.

I drafted a letter to candidates, which was adapted and circulated over the signature of the Executive Officer, about the possibility of an external person being present during their oral examinations, and I was informed of the outcome of that correspondence.

I observed the processes of the oral examinations for membership and fellowship status at the Gold Coast International Hotel on three days (2–4 July 2007), and provided a status report orally to a gathering of the Board of Examiners during that same period.

Tasks undertaken
The review proceeded in overlapping stages to parallel the terms of reference for the review. At the micro-level, the reviewer undertook the following tasks, some of which have been referred to above.

- Conducted desk analysis of information provided to examiners and candidates (‘The Purple Book’, ‘The Blue Book’, and ‘The Red Book’);
- Reviewed the research literature;
- Studied written examination papers at two stages in their development;
- Observed a sample of oral examinations at membership and fellowship levels;
- Met with the Board of Examiners in the week of the oral/practical examinations in July at the Gold Coast International Hotel;
- Tested assumptions about practice of assessment at critical stages of the examination cycle (in particular, setting and refining of papers, and administration of orals);
- Received copies of ongoing correspondence (complaint) from a former candidate;
- Devised frameworks for the evaluation being carried out;
- Evaluated the overall validity and reliability of the dual-format approach to assessment;
- Considered salient features of the examination environment and its products (e.g. commitment required of examiners, success rate of candidates, and number of appeals if any); and,
- Used the data to test out hypotheses formulated during observation of the oral examinations and the grading process at its conclusion.

Frames of reference for analysis
Analyses conducted during this review were framed in terms of:

- Research literature on assessment in general (especially format and orals);
- Research literature on assessment of professional competencies;
- Best practice in setting examination papers – technical and editorial aspects;
- Validity (Cronbach, 1989) and Reliability (Moss, 1992).

---

1 Arriving at overall result
Structure of the report

The structure of the main body of this report is in the form of a commentary and critique, which covers the preparation and conduct of examinations, the marking of candidate responses, and the grading of candidates for membership and fellowship of the Australian College of Veterinary Scientists.

The review draws on the literature about the theory of educational assessment and the many facets of professional assessment in action. Because of the high degree of interrelatedness between them, it is not easy to treat particular facts of professional assessment in isolation. To a certain extent each one shapes the other. Thus, in the ensuing commentary and critique, facets are not always assigned to different compartments in the discussion or confined to only one compartment. Nor is it the case that the commentary and critique proceed in parallel to the chronological stages of the examination process – setting examinations, conducting examinations, marking examination responses, and grading candidates. Sometimes there is no direct link between sections; in fact, some sections could quite easily make sense if read out of sequence.

The sequence generally reflects the frames of reference listed above. The first major section is about assessment in general, especially examination format and oral examinations.

The second major section is about the assessment of professional competencies, especially communication within a veterinary context.

The third major section is about examination design, the fourth about marking and grading, and the fifth about validity and reliability.

Following the main commentary and critique are sections with a different emphasis: formulating and testing hypotheses; summarising conclusions; and presenting three overarching recommendations and ten operational recommendations.
Quality of oral examinations

The oral examinations administered by the College were evaluated by the application of the following six criteria.

Oral examinations should:

1. Add value when compared to other forms of examination;
2. Assess aspects of competence that are not assessed by written examinations;
3. Place the assessment in a valid context when compared with veterinary practice;
4. Involve clear and professional oral communication between examiner(s) and candidate;
5. Meet psychometric standards individually and in combination with written examinations;
6. Be able to deal with candidate anxiety.

The reviewer’s judgments about the College’s oral examinations against these six criteria are presented in Table 1. In the left-hand column the criteria are re-stated in short form. The second column (‘Check’) contains a tick (✓) when it is considered that there is evidence of the criterion in action (i.e. as a feature of the observed oral examinations). It contains a cross (✗) when it is considered that there is a lack of evidence of that criterion in action. This does not prove that there is evidence of a total lack of the feature just that there was no evidence of its presence. There is an elaboration of the evaluation in the right-hand column (‘Comment’) for some of the criteria.

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Check</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Adds value</td>
<td>✓</td>
</tr>
<tr>
<td>Criterion</td>
<td>Check</td>
<td>Comment</td>
</tr>
<tr>
<td>-----------</td>
<td>-------</td>
<td>---------</td>
</tr>
</tbody>
</table>
| 2 | Assesses different things | ✓ | The obvious ones are:  
• Listening skills;  
• Using vocabulary appropriate to a context (e.g. technical veterinary science terms when talking to a colleague; person-in-the-street terms when talking to animal owners);  
• Explaining to others orally; and,  
• Expounding a viewpoint (about a diagnosis perhaps).  
Less obvious ones are:  
• Empathising (i.e. appreciating the emotional reactions of clients or animal owners in certain situations);  
• Solving problems out loud, justifying a position in the company of others, coping with being assessed by colleagues or peers (these are examples of ‘deprivatising practice’); and,  
• Quoting the research literature (to demonstrate a scholarly approach and/or to cite some finding that is relevant to the discussion). |
| 3 | Contextually valid | ✓ | In same area of competence as written exam  
✓ | At same level of difficulty as written exam  
✓ | At a similar level of stress to real-life situation |
| 4 | Clear and professional communication between examiners(s) and candidate | ? | See hypotheses later about cultural and gender issues.  
✓ | Common understanding of the use of language conventions  
✓ | Clear, audible speech  
× | Commonly applied cue-ing required  
See section ‘Cracking the code’. |
| 5 | Psychometric soundness | × | See separate discussion. |
| 6 | Candidate anxiety | ? | Dealt with but not completely  
See comments below.  
Familiarising candidates with:  
• structure of examination questions | ✓ |
### Criterion Check

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Check</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>• examination set up</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>• time allocation</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Familiarising candidates with response expectations</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Thinking time is given during question time.</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Questions are repeated or clarified upon request.</td>
<td>x</td>
<td>Differential use of cues is a concern.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>See ‘Cracking the code’ (next section).</td>
</tr>
</tbody>
</table>

The significance of the one issue arising from Criterion 3 is self-explanatory. Issues arising from Criterion 5 are listed in a separate section of this report (within the section, ‘Observations and Conclusions’). Issues arising from Criteria 4 and 6 are of the same ilk; that is, they relate to candidates ‘cracking the code’, which is discussed in the next section.

### Cracking the code

The question asked of a candidate by an examiner indicates *what* the candidate has to do (interpret, talk about). It is like a command given in question form. An observation about examinations of any kind is that candidates are often left in limbo, feeling that there is further essential information that they need in order to decide where to direct their effort or focus their answer.

A general impression of this reviewer from the reactions of many candidates for the College’s oral examinations was that some candidates in some situations appeared to be unsure of the audience for, and purpose of, their responses. Assuming that they came into the oral examination realising that they are supposed to be engaging in scholarly conversation with their professional peers (albeit peers who are already members or fellows), some became obviously confused about what was expected of them and what was going to be valued by examiners. Further, many of those who became confused asked for clarification.

Whether the candidate asks for this information or whether the examiner intuits that the candidate needs the information, extra information given about *how* to answer is a cue – not unlike a prompt given to an actor in a play or a signal/nod given to a partner in a game of Bridge.

If the examiner goes further and gives a hint or a piece of advice or a reminder of a fact that the candidate needs to know in order to answer the question accurately, then this should be noted by the examiner and included with the question for any subsequent candidates and revisited in the marking notes for any previous candidates. Unlike a cue, the candidate might not understand why a hint is being given, much less what it means. But it certainly gives the recipient of the hint an advantage over the non-recipient.

Without putting too fine a point on the distinction between cue and hint, a cue is used to *clarify* the examiner’s intentions for candidates whereas a hint is used to provide a pointer or clue (clue not cue!) to the correct or best answer. A rule of thumb, in the opinion of this reviewer, is that:
• Hints should not be given during an oral examination;
• Cues, if given to one candidate, should be given to all other candidates; and, if this is impossible because it would require retrospective cue-ing, the advantage bestowed on the receiver of the cue should be taken into account in marking.

Cue gives clarity. This is OK. Hint gives pointer. This is not OK.

The point being made here is not that examiners were observed to be giving ‘hints’ and thereby conferring advantage. The point is that some candidates required ‘cues’ – because at that particular moment in the examination they needed to have affirmed that they were expected to ‘show off’ their professional knowledge and their ability to explain, justify, synthesise and so on. (What they might have been wondering at this stage was whether they were supposed to be role playing.) Furthermore, if this sort of clarification is needed by one candidate in a given subject, examiners should insert it into the questioning for the next candidate (if there is one) rather than wait for the ‘cue-able moment’. They should also retrospectively consider its possible effect on other candidates’ marks (if there were any earlier candidates).

The source of the candidate anxiety discussed (i.e. having to ‘crack the code’, the classic case of reading the examiner’s mind) relates to the topic of self-imposed difficulty, which is discussed in the next section.

Sources of difficulty
Examination questions (or test items or assessment tasks) are different in fundamental ways. Matters (1998) proposed a typology of difficulty to capture three fundamental differences – nature of the cognitive task; aspects of examination design; and candidate’s perception of task difficulty.

First, examination questions involve different kinds of thinking and different mental abilities, and they place different emphases on given topics. These differences contribute to what might be called the intrinsic difficulty of an examination question, and they are based in the cognitive demands of the question: How hard was the question?

Another type of difficulty originates in features of the question related to: the response medium (written, oral, practical, artefact etc.); the response format in the case of written examinations (multiple-choice, constructed response, extended writing); the level of preparation for the examination (beyond studying the content of the examination is the need to understand the ritual and protocols of that particular sort of examination); and the duration of the examination (limited time, continuing over days etc.). These different aspects of the design of examinations contribute to what might be called design-imposed difficulty: How hard did the form of the examination make it for the candidate to show what s/he knew and could do?

It is the third type of difficulty, self-imposed difficulty, which is particularly relevant to a discussion of the College examinations. Self-imposed difficulty is candidate-specific. It is a function of the particular candidate’s mindset on viewing the question (written paper) or hearing the question (oral examination) or receiving instructions, say, for a practical examination. A candidate’s perception of success is influenced by features of the stimulus material (e.g. radiograph, photograph, or scatter plot) such as content and context. Some such features can be perceived as a source of difficulty (read disadvantage) to that candidate. How hard did the context make it for the candidate to show what s/he knew and could do?

The notion of self-imposed difficulty is important for examiners to keep in mind when setting written papers that permit candidates a choice of questions to answer or when presenting material to candidates in oral examinations. Obviously, if the subject being examined is about veterinary radiology then radiographs for interpretation and comment would be expected. But, if the subject being examined is, say, Analysis of data in veterinary practice, then presenting all data in tabular form would not be fair (there should be a balance across tables, graphs, diagrams, maps,
photographs etc.). Nor would it be fair to have a selection of tables, graphs and diagrams all of which are set in the context of the cattle industry if the analytical techniques being examined were deemed to be general rather than animal-specific. In such cases, a candidate’s disposition can also be the factor that precipitates under-performance. For example, unless persistence is being examined, it is not fair to set up a case study where persistence with boring data conflates measures of performance on interpreting case study data.

The illustrations above relate more to the way material is presented to candidates in an oral examination. Of relevance to choice on written papers is the finding that candidates are often not capable of choosing the questions for which their answers would gain the most credit. This is discussed in the next section.

**Question choice in examinations**

Allowing candidates to choose the question(s) that they will answer from among several possible alternatives is often viewed as a mechanism for increasing fairness in certain types of assessments. The fairness of optional topic choice is not a universally accepted fact, however, and various studies have been done to assess this question. For example, Allen, Holland, and Thayer (2005) found that differential topic difficulty exists in real choice data, that it affects naive analyses of such data and masks the effects, positive or negative, of candidate choice, that the beneficial strength of candidate choice varies from case to case, and that while the benefits of choice in terms of average marks awarded are usually positive, there is a substantial amount of variation around these averages and it is not uncommon for inappropriate choices to be associated with higher examination performance.

Of the four fellowship subjects, Canine medicine, Small animal medicine, Small animal surgery, and Veterinary radiology, all eight papers allowed candidates to choose the questions they would answer. In Veterinary radiology Paper 2 and both papers of Small animal surgery, there was a constraint on that choice. In both papers of Canine medicine and both papers of Small animal medicine, there was further choice for candidates from options within questions (see Table 2). A similar breakdown for the membership subjects appears within Table 3 (together with other information).

**Table 2: Fellowship examination 2007, by question choice (of and within)**

<table>
<thead>
<tr>
<th>Subject</th>
<th>Question choice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canine medicine Paper 1</td>
<td>Answer 5 out of 6. Choice within q2, 5, 6.</td>
</tr>
<tr>
<td>Small animal medicine Paper 1</td>
<td>Answer 5 out of 6. Choice within q2, 5, 6</td>
</tr>
<tr>
<td>Small animal medicine Paper 2</td>
<td>Answer 5 out of 6. Choice within q3, 4, 6</td>
</tr>
<tr>
<td>Small animal surgery Paper 1</td>
<td>Answer 4 out of 6, including 2 from each of Sections 1 and 2.</td>
</tr>
<tr>
<td>Small animal surgery Paper 2</td>
<td>Answer 4 out of 6, including 2 from each of Sections 1 and 2.</td>
</tr>
<tr>
<td>Veterinary radiology Paper 1</td>
<td>Answer 4 out of 6.</td>
</tr>
<tr>
<td>Veterinary radiology Paper 2</td>
<td>Answer 4 out of 6, including ≥ 1 from Section A.</td>
</tr>
<tr>
<td>Subject</td>
<td>Item type</td>
</tr>
<tr>
<td>------------------------------</td>
<td>------------------</td>
</tr>
<tr>
<td>Animal nutrition (ruminant)</td>
<td>Paper 1</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Paper 2</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Equine dentistry</td>
<td>Paper 1</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Equine dentistry</td>
<td>Paper 2</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2 Given as verbs (commands) in original question; expressed as verbs without modifiers in other cases.
<table>
<thead>
<tr>
<th>Subject</th>
<th>Item type Question choice</th>
<th>Instructional words (key verbs)</th>
<th>Topic</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epidemiology Paper 1</td>
<td>Answer all 8. Alternatives within each</td>
<td>Write notes to explain</td>
<td>Overview Disease patterns Measurements Study design Study types Data Economics</td>
<td>Purple book: General</td>
</tr>
<tr>
<td>Epidemiology Paper 2</td>
<td>Answer all 3. Alternatives in q1 and q2.</td>
<td>Calculate (2) Explain (2) Describe (1) Evaluate (1)</td>
<td>Outbreak investigation analysis: application Anaplasmosis vaccine Johnes disease Survey paper</td>
<td>Considerable reading required Purple book: Specific scenarios</td>
</tr>
<tr>
<td>Veterinary anaesthesia and critical care Paper 1</td>
<td>Answer 4 of 6. Alternative in q6</td>
<td>Compare &amp; contrast (1) List (2) Discuss (64) Define (1) Explain (1) Describe (1)</td>
<td>Intravenous agents Equipment for monitoring Alpha 2 adrenergic agonist drugs Carbon dioxide levels Pregnancy Pain physiology</td>
<td>Purple book: General</td>
</tr>
<tr>
<td>Medicine of cats Paper 1</td>
<td>Answer 4 of 7.</td>
<td>Discuss (4) Describe (4) Explain (1) Write notes (2) Diagnose (1)</td>
<td>Feline infectious peritonitis Pancreatitis Feline nephron blood pressure in cat feline blood groups Lymphosarcoma Bronchial disease, left ventricular hypertrophy, diabetic ketoacidosis, Conn’s syndrome</td>
<td></td>
</tr>
<tr>
<td>Subject</td>
<td>Item type Question choice</td>
<td>Instructional words (key verbs)</td>
<td>Topic</td>
<td>Comment</td>
</tr>
<tr>
<td>---------</td>
<td>--------------------------</td>
<td>---------------------------------</td>
<td>-------</td>
<td>---------</td>
</tr>
<tr>
<td>Medicine of cats Paper 2</td>
<td>Answer 4 of 6, including q5 or q6. Choice within q1 and q2</td>
<td>Write notes (2) Discuss (3) List (1), Formulate (plan) (1)</td>
<td>Blood pathology data in q5 and q6</td>
<td></td>
</tr>
<tr>
<td>Small animal medicine Paper 1</td>
<td>Answer 4 of 6. Choice within q1, q3, q4, q5</td>
<td>Write notes (2) List (2) Describe (2) Discuss (3)</td>
<td>Action of drugs Liver failure Haemostasis Neurology Clinical tests Pleural fluid</td>
<td></td>
</tr>
<tr>
<td>Small animal medicine Paper 2</td>
<td>Answer 4 of 6, including q5 or q6. Choice within q3, q4</td>
<td>List (2) Discuss (4), Describe (1)</td>
<td>Hypoglycemia Pyrexia Diagnosis Management Vomiting cat; blood data</td>
<td>No q 6; q5 incomplete pathology data only but no actual question Purple book: q2, q4 general but cats and dogs not any small mammal</td>
</tr>
<tr>
<td>Small animal surgery Paper 1</td>
<td>See note below table.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Small animal surgery Paper 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Veterinary emergency &amp; critical care Paper 1 Pathophysiology</td>
<td>Answer 4 of 6.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Veterinary emergency &amp; critical care Paper 2 Clinical</td>
<td></td>
<td></td>
<td>Table to complete; pathology data</td>
<td></td>
</tr>
</tbody>
</table>

(NB: Table deliberately incomplete. Sufficient entries to verify consistency of questions with subject guidelines.)
Standardised assessment
A standardised examination, by definition, is one that is the same for all candidates; that is, common questions for all candidates irrespective of place of residence or hours of study, taken under the same conditions, and marked according to a commonly applied marking scheme.

Based on this definition, the College’s oral examination is a standardised examination. It is patient-based and intended to assess directly a candidate’s data interpretation (citing of the research literature), analytical skills, and communication skills, especially in a veterinary situation in handling the patient/client appropriately.

As illustrated by the following table, the College oral examinations have all the hallmarks of a standardised assessment instrument. The note in the last row of the table points to the fact that standardisation of itself does not automatically deliver reliable results.

Table 4: Analysis of state of standardisation of College oral examinations

<table>
<thead>
<tr>
<th>‘Commonly applied’ requirement</th>
<th>Oral</th>
<th>Comment (where necessary)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same examination irrespective of place of residence</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Same set of questions for all candidates</td>
<td>✓</td>
<td>As in same set of questions offered although different candidates in the same subject might have chosen different questions on written papers where choice is permitted</td>
</tr>
<tr>
<td>Same conditions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>∙ Venue</td>
<td>✓</td>
<td>Although reliability of multi-media equipment not guaranteed from room to room at venue</td>
</tr>
<tr>
<td>∙ Examiners</td>
<td>✓</td>
<td>Actually same pair of examiners, or trio of examiners if the number of candidates is large</td>
</tr>
<tr>
<td>∙ Physical environment</td>
<td>✓</td>
<td>Same introduction from examiners</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Equipment usually dependable</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Supply of drinking water is noted as an element of the environment vital to, and appreciated by, candidates.</td>
</tr>
<tr>
<td>∙ Duration</td>
<td>✓</td>
<td>Same amount of time available to all, even if some candidates took the opportunity to use every available minute whereas others stopped short</td>
</tr>
<tr>
<td>‘Commonly applied’ requirement</td>
<td>Oral</td>
<td>Comment (where necessary)</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>------</td>
<td>---------------------------</td>
</tr>
<tr>
<td><em>Same marking scheme</em></td>
<td>✓</td>
<td>But this does not automatically deliver consistent applications of marking scheme (standards) by examiners individually or in teams (pairs or trios). Marker monitoring is briefly discussed later in this report.</td>
</tr>
</tbody>
</table>
Assessing content, process and skills
Recall of cognitive information is not the primary purpose of an examination, whether written multiple-choice, constructed response or extended writing or oral/practical. Recall of facts should be assessed only in the context of application of factual knowledge and/or where the incorrect recall would create a damaging situation.

Whereas a written examination can attest to a candidate’s knowledge of the field (facts, concepts and procedures) or, in simple terms, knowing and understanding the what and knowing and understanding about the how, an oral examination should test for the presence of those qualities and attributes which are fundamental to performance in the subject, in this case veterinary science as advisor or specialist veterinarian.

This means that the candidate is required to provide evidence orally and in the form of immediate feedback of the following attributes or skills:

- Soundness of judgment and rationality of thought in making and applying decisions;
- Assimilation and analysis of data;
- Defining priorities, suggesting treatment plans, recognising potential complications; and planning for contingencies;
- Adaptability as in incorporating additional and/or updated information/data into a response; and,
- Communicating effectively, through speaking, about issues of specific relevance to the corresponding field of veterinary science.

A certain amount of factual information is required to answer questions in both written and oral mode, whatever the subject. Higher-order thinking skills such as analysing, extrapolating and judging can be examined in both written and oral mode, whatever the subject. In the professions, the oral examination is usually designed to test the candidate’s judgment, application of knowledge, clarity of expression, and adaptability to changing, sometimes unexpected, circumstances (e.g. as the candidate answers questions about the management of the animal (patient)).

Rationale for assessing communication skills
A performance dimension in the College examinations is communication skills, both written and oral. It is claimed that adequate oral communication skills and, to a lesser extent, interpersonal skills, are associated with enhanced client satisfaction and improved clinical outcomes, as well as with a decreased risk for malpractice litigation. It is therefore necessary for the examinations to be objective and structured when assessing skills such as the following.

- Noting results of physical examination (indirectly not hands-on; this is what is called second-order testing);
- History taking (based on a simulation of sorts);
- Making a clinical judgment based on notes;
- Communicating with an animal’s owners;
- Breadth and depth of knowledge;
• Ability to call forth summaries and findings of research in the area; and
• Ability to make a differential diagnosis, or plan treatment.

Effect of format
The literature on comparing and contrasting formats for assessment, although prolific, does not appear to provide definitive answers to some fundamental questions: Do the same candidates find multiple-choice questions or constructed-response items to be more difficult? Is difficulty a function of the person or the examination question or both? Do the same candidates find it more difficult to argue a case in writing or face-to-face orally? Is difficulty a function of the person (e.g. confidence in front of an ‘audience’) or the examination question (i.e. topic being examined) or both? Any analysis of format, mode or medium of assessment is invariably transformed into a litany of pros and cons for one or the other format, mode or medium.

And so objectivity and structure, as already mentioned, are the hallmarks of effective assessment of communication skills or of a person’s ability to communicate adequately with the owner of a sick animal or with a veterinary colleague. The notion of structure is covered in two other sections, ‘Quality of oral examinations’ and ‘Examination design’. The notion of objectivity is discussed in the next section.

Objectivity and subjectivity
There is an old stereotype, which usually refers to school teachers and university lecturers. It could be transferred to the realm of College examinations. It goes like this.

End-of-semester exams and teachers are in the shared staffroom marking candidate scripts.

   English teacher: ‘I wish I were a Maths teacher. They don’t have to read mountains of work. They just have to mark things right or wrong.’

   Maths teacher: ‘Well at least my marking is objective. Yours is so subjective.’

Who is wrong? Both of them are. There is a pervasive view in the literature that links subjectivity to writing and speaking tasks/tests with human judges at one extreme, objectivity to multiple-choice tests with computer marking at the other, and written constructed responses marked by human judges according to a highly prescriptive rubric somewhere in between.

Actually that Maths teacher’s subjectivity was on display when she set the paper in the first place; that is, when she sampled from the domain, decided on the difficulty levels of the questions on the paper, and decided that there would be a traditional examination-type paper rather than, say, continuous assessment throughout the course.

And it is possible for the English teacher to be more objective than some multiple-choice testing if s/he has set a well-crafted old-fashioned essay with unambiguous cues for candidates and a clear marking scheme.

The point to be made here is that subjectivity/objectivity does not reside in the format but in the way an assessment instrument (be it a test item or examination paper) is constructed and scored (marked/graded).

   At every stage in the design and administration of any objective test constructed …, subjective judgments are involved. The [examiner] has to decide on the subject matter to include, the behaviours to sample, the complexity and difficulty of proposed asks, the item format, and the wording and mode of presentation. The process is objective only at the very last stage, which is deciding on the correctness of an answer. So-called objective assessment consists of a chain of subjective decisions, with one final objective link. Unfortunately the essential objectivity of the end point and the fact that the outcome of the final step is often expressed in numerical form (which, to many people, is the hallmark of objectivity) obscures the subjectivity inherent in all the steps leading up to it.

   (Sadler, 1986)
Enhancing assessment in oral mode

It can be concluded from the preceding sections about structure and objectivity that the desirable features of any examination are a function of:

- Rigorous procedures at the setting stage regarding decisions about content for inclusion, behaviours to be sampled, difficulty of questions to be asked, how the question is to be worded and so on; and
- Rigorous procedures at the marking/grading stage regarding the application of the criteria and standards for assessment response, reference to an exemplar response during marking, monitoring of markers, rules for combining scores and so on.

Screening of candidates to Tel Aviv University Medical School requires a biographical questionnaire, descriptions of dilemmas that require candidates to make judgments, and simulations for which there are eight behavioural stations. The simulations aim to test (i) communication skills, (ii) initiative and responsibility, (iii) handling of stress, and (iv) consciousness and self-awareness (Cohen, 2007). Examiners at the behavioural stations observe these four attributes and score each of them on a scale of 1–6. The weighting of the simulation score is 60% of the overall score. According to the measurement literature, inter-rater reliability, an estimate of agreement between examiners, is 0.7–0.8 for simulations. For the Tel Aviv simulations it was reported to be 0.72; and test-retest reliability was reported to be 0.70.

Strengths lie in its reliability, validity, social messages, and change to composition of ‘successful’ candidate group. Weaknesses include the length and difficulty of the validation process, potential cultural biases, and financial cost.

There is some diversity of opinion concerning the use of oral examinations for the MACVSc and FACVSc examinations. Despite the inherent authenticity and demonstrated feasibility of the orals, there could be other features that are not considered in such positive terms.

Of relevance to the College from the recent work in Israel are the possibilities presented by simulations. To that end, the oral format could be enhanced with the use of simulations, obviously not of the type introduced into the assessment of medical candidates where an actor can pretend to be a (human) patient but of the type where an actor can pretend to be an owner of the animal-patient(s) or a government department head looking for advice or whatever. Such examinations could then assess clinical decision-making alongside the application or use of veterinary science knowledge with simulated clients. (This reviewer’s imagination does not extend into the sphere of simulated patients of veterinarians although it might well be possible.)

The point to be made here is that the College examinations like any other examinations require multiple methods, multiple assessments, multiple raters, and raters’ training. The current examination model does have training for examiners (although it could be more extensive), multiple examiners (2 or 3), multiple assessments (if the written and oral/practical examinations are considered separate occasions), and multiple methods (written examinations and orals). If the oral examination were to be eliminated then the requirement for multiple methods would not be met. The choices are, therefore, to:

1. Retain the oral examination in its present form;
2. Adapt the oral examination according to recommendations made in this report;
3. Enhance the oral examination with the use of simulations.

Option 1 is not really an option (the desire to enhance the present model is shared, and the process of enhancement is not complicated). Option 3 is unwieldy and expensive and perhaps not desirable. Option 2 is worthy of further consideration.
EXAMINATION DESIGN

Requirements of the examination-setting process
Instructional words or key verbs
Fellowship examination questions, by frequency of instructional word
Alignment of essential elements in a good examination question
Technical considerations
Editorial considerations
Further development

Setting an examination is a process of design. Some features of the design process that are pertinent to the College examinations are illuminated in Table 5. The first column lists examination requirements from the assessment literature. The second column, ‘ACVSc’, summarises the way that the College fulfils each of the requirements listed in Column 1. The third column gives a judgment about the College response to each of the requirements.

Table 5: Requirements of the examination-setting process

<table>
<thead>
<tr>
<th>Requirement</th>
<th>ACVSc response</th>
<th>Adequacy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detailed specifications</td>
<td>Flowcharts</td>
<td>✓ ✓</td>
</tr>
<tr>
<td>Expert assessors</td>
<td>Experienced and highly selected examiners – selected veterinarians representing the variety of subjects being examined</td>
<td>✓ ✓</td>
</tr>
<tr>
<td>Examination setting</td>
<td>Scheme of submitting questions to Executive and then revising on the basis of editorial feedback</td>
<td>✓ ×</td>
</tr>
<tr>
<td>Incline of difficulty</td>
<td>Easier item(s) at the beginning</td>
<td>✓ ✓</td>
</tr>
<tr>
<td>Examination supervision</td>
<td>Chief Examiner&lt;br&gt;Observer&lt;br&gt;Standardised forms&lt;br&gt;Close and continuous cooperation with Executive</td>
<td>✓ ✓</td>
</tr>
<tr>
<td>Set and marked with objectivity</td>
<td>Explicit rationale for inclusion of certain topics&lt;br&gt;Exemplar responses and partial solutions but not universally</td>
<td>? ✓ ×</td>
</tr>
<tr>
<td>Checks on consistency of markers</td>
<td>Post-hoc consensus between examiners but no analysis or monitoring of data</td>
<td>✓ ✓</td>
</tr>
</tbody>
</table>

Legend: (✓✓) = requirement met; response complete; (✓×) = requirement met; response incomplete; (?) = unknown to reviewer; does not necessarily mean requirement not met.

Instructional words
One of the highly regarded aspects of the College’s examination-setting process is the consistent use of instructional words (or key terms). These words are verbs of command (e.g. ‘compare’). The use of certain key terms means that what is required of candidates is made explicit in the question – the candidate knows exactly what type of answer to provide.

Some examples follow of the meaning of key verbs (or instructional terms).

Compare: Recognise similarities and differences and the significance of those similarities and differences
Discuss: Talk about (in a sustained way)
Describe: Represent, picture or portray; say what something is
Define: Identify/characterise
List: Identify relevant information and accurately and methodically write it down (record it) (in a column).
Outline (in words): Provide a structure for; give main points or features
Write notes on: Identify relevant information and present it accurately and methodically (e.g. under logical headings or within predetermined categories).

Of the four Fellowship subjects, Canine medicine, Small animal medicine, Small animal surgery, and Veterinary radiology, ‘discuss’ accounted for 56% of the instances of an instructional word in an examination question (see Table 6). It is to be hoped that candidates and examiners have a shared understanding of the required genre (e.g. discussion) for answering the question. A similar analysis of the membership papers appears within Table 3 (together with other information).

Table 6: Fellowship examination questions, by frequency of instructional word

<table>
<thead>
<tr>
<th>Subject</th>
<th>Instruction</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canine medicine Paper 1</td>
<td>Discuss</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>List</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Describe</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Write notes on</td>
<td>1</td>
</tr>
<tr>
<td>Canine medicine Paper 2</td>
<td>Discuss</td>
<td>6</td>
</tr>
<tr>
<td>Small animal medicine Paper 1</td>
<td>Discuss</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Describe</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Write notes on</td>
<td>1</td>
</tr>
<tr>
<td>Small animal medicine Paper 2</td>
<td>Discuss</td>
<td>7</td>
</tr>
<tr>
<td>Small animal surgery Paper 1</td>
<td>Define</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Describe</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Discuss</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>List</td>
<td>1</td>
</tr>
<tr>
<td>Small animal surgery Paper 2</td>
<td>Describe</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Discuss</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Define</td>
<td>1</td>
</tr>
<tr>
<td>Veterinary radiology Paper 1</td>
<td>Describe</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Discuss</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Outline</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Compare</td>
<td>1</td>
</tr>
<tr>
<td>Veterinary radiology Paper 2</td>
<td>Discuss</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Describe</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Outline</td>
<td>1</td>
</tr>
</tbody>
</table>

If candidates understand the terms before they sit for the examination they will not waste time wondering what is expected of them. On the other hand, if the terms are used loosely and/or interchangeably, there is really no point in specifying the type of answer required. If describing as opposed to defining is important in the veterinary specialties then examiners should reward candidates who actually do answer the question by doing one or either as required, as well as for the quality of the content in the answer.
Then there is an obvious link between what is demanded of candidates and what is rewarded in their answers: The instructional word used in the question and the content area of the question should be aligned with the features of the candidate’s answer that are rewarded in the marking criteria as discussed below.

**Alignment**

In a good examination question, there is an effective interplay of three simple elements – what is set for study (the intentions of the course), what is assessed in the examination question (knowledge, skills/dispositions in the domain being sampled), and what is rewarded (high-quality performance on the criteria set down in the marking scheme).

Figure 1 illustrates the alignment of what is set to be studied with what is assessed in the examination question with what is rewarded in the marking scheme.

![Figure 1: Alignment of essential elements in a good examination question](image)

**Technical considerations**

Below is a checklist of technical matters to be addressed in developing an examination question/paper. Some are more pertinent to the College examinations.

**Expert**

- Ensure that the questions (and marking scheme) are academically sound.
- Produce your own answer to the question and/or write an indicative solution (or model answer).
- Is the marking scheme valid, usable and precise?

**Equity considerations**

- Cultural (e.g. accommodate varied social, cultural and ethnic backgrounds)
- Gender and social justice (e.g. acknowledge the gender-based differences in examination-taking strategies and problem-solving skills)
- Content (e.g. acknowledge the effects of geography such as isolation).

**Construct/design**

- Use a variety of response mediums.
- Use a variety of question structures.
- Acknowledge varying familiarities with different thinking patterns (e.g. abstract).
- Monitor the use/preponderance/dearth of visuals.
•

*Space, time, difficulty*

- Is the examination of appropriate length?
- Is the intrinsic difficulty of questions appropriate based on the nature of the cognitive task involved?
- What would be the general effect, psychological and intellectual, on a candidate?
- Does the examination use a variety of language styles?
- Does the marking scheme cover the likely range of answers?

**Editorial considerations**
Below is a checklist of editorial matters to be addressed in developing an examination question/paper. Some are more pertinent than others to the College examinations.

**Rhetorical**

- Do the key terms (instructional words) clearly identify the task and the frame of reference?
- Do the cues (if any) clearly indicate how the candidate is to respond?
- Is the question in the appropriate voice?
- Does the question explicitly state what the question is that the candidate is to answer?
- If the introductory material in the questions has been adapted or extracted from an original, is it still accurate? Understandable?

**Layout**

- Is the design of the question effective and efficient?
- Is the response area adequate and indicative of the response required?
- Are tables, graphs and figures accurate, clear and correctly labelled? Referred to as in the text?
- Are legends clear? Properly positioned?
- Is the use of fonts and spacing consistent with a College-produced style guide?

**Language**

- Are grammar, spelling and punctuation used correctly?
- Is language used as efficiently as possible?
- Is the language current? (This question is not always appropriate.)
- Are unfamiliar terms defined?

**Further development**

It is suggested that the procedures for setting papers, devising marking criteria, preparing exemplar responses be further developed. The time spent on this would be off-set by a faster decision making and recording process at the end of the examination cycle.

It is acknowledged that the College examination procedures have to be run to a tight budget. It is still worth considering, however, to pay examiners a fee for setting and marking the papers *but at the same time* expect from them a tightening of the discussions between them as a group and as examining ‘partners’. This would increase the validity and reliability of the examinations simultaneously. (There are not enough papers and examiners to undertake traditional quality control processes’ associated with the marking of high-stakes examinations.)

---

1 Statistical analyses such as inter-marker agreement as measured by concordance and weighted kappa; check marking; analyses of effect of different combinations of markers to identify markers who are out-of-step and so on
MARKING AND GRADING

The College’s marking model

Analytic marking
Holistic marking
Combining results

Summary of arithmetical operations currently undertaken by examiners
Setting numerical boundaries for grades

This section presents the theoretical background to marking and grading. It sets the scene for asking the boxed question and then answering in the negative.

Does the College really need to collect on the marking sheets all the information as presently required?

The argument for not needing to collect all the information that is presently required for the marking sheets is based on the following observations and opinions.

1. Expert judges (College examiners) possess a tacit understanding of what performance ‘makes the grade’.
2. Examiners have already decided on a candidate’s result before quantifying candidate performance on the oral examination.
3. Groups of examiners (in pairs or trios) reach consensus on a candidate’s result well before they complete the marking sheets.
4. Examiners (and others) subconsciously view impression marking (holistic judgment) as subjective and therefore not rigorous enough for College examinations.
5. If examiners really believed that analytic marking is to occur first followed by grading based on the aggregate score, they would use more detailed marking schemes and indicative solutions.

In the opinion of the reviewer, College examiners would be more than capable of using the wine-judging model rather than the minutissima of the version of the analytic marking endorsed by the College (although some good reasons for this are acknowledged). See Recommendations 2 and 3.

The concepts analytic marking, impression marking, holistic judgment, and the wine-judging model are explained in the next section.

The College’s marking model

Even though the marking model used by the College for the written and oral/practical examinations is considered to be analytic, the reality is that it is holistic.

The terms analytic and holistic scoring (marking) are most often used in the context of large-scale testing or marking operations where speed and reliability are paramount. Brookhart (1999) provides the following definitions.

- **Analytic scoring rubrics** allow for the separate assessment of each of several (two or more) criteria. Each criterion is scored on a different descriptive scale.
- **Holistic scoring rubrics** support broader judgments concerning the quality of the process or product. In the holistic scoring rubric, the criteria are considered together on a single descriptive scale.

**Analytic marking**

Analytic marking is where some sort of marking scheme is employed that gives guidance to the marker about what features s/he should look for and what weighting should be given to them (or what influence each feature’s mark should have on the overall mark).
It is claimed that analytic marking ensures intra-marker consistency and inter-marker reliability. It is claimed that impression marking can be very reliable if there is appropriate training before marking and continuous monitoring of markers during the marking operation.

Written constructed-response items lend themselves readily to analytic marking as can be seen in the following example of an analytic marking scheme (original question not included).

*Example of an analytic marking scheme*

A total of 10 marks is allowed for this question.

- 5 marks for demonstrating understanding of the basic concepts of two theories
  - For theory 1, candidates should be rewarded for making specific mention of developmental stages, adaptation as assimilation and accommodation, schemata, sensory-motor operations, concrete operations, formal operations.
  - For theory 2, candidates should be rewarded for making specific mention of personal constructs and construing, validation and core constructs.

- 5 marks for noting significant differences between the two theories, in particular:
  - Theory 1 is a developmental theory.
  - Theory 2 grew out of a concern with therapeutic change.
  - Theory 1 accounts for motivation in terms of ‘equilibration’.
  - Theory 2 accounts for motivation in terms of ‘a search for meaning’.
  - Theory 1 is logico-mathematical; theory 2 is not.

A possible problem with analytic marking is that, if it is too specific, it may be difficult to differentiate between excellent candidates and competent candidates. In deciding the credit given (marks awarded) to different parts of the answer, it is important to allow some flexibility in the marks to allow discrimination; for example, by having bonus marks awarded for overall quality or originality.

Sometimes it is impossible to separate a judgment into independent properties. When there is an overlap between the criteria identified for the assessment, an holistic scoring rubric may be preferable to an analytic scoring rubric.

*Holistic marking*

Holistic marking and impression marking are sometimes used to refer to the same phenomenon. Impression marking refers to situations where a marker (examiner), using his/her expert judgment, assigns a mark on the basis of an overall impression of the work’s worth. The word ‘impression’, with its overtones of impressionistic can give the wrong impression (no pun intended) about capturing the distinction in question. After all, an impression is, by definition, something that leaves a trace – such as the image (visual or auditory) of the candidate’s written, practical or oral work.

*The wine-judging model*

Impression marking is not unlike the classic wine-judging model. The case there is not that separate (analytic) scores are applied to three criteria (e.g. viscosity, nose, palate) and then aggregated. Rather, a well-informed judge trades these off in her/his head and makes an overall judgment. ‘Chateau Cardboard’ will always get a lower score than Grange Hermitage, provided the wine judge is not a maverick.

If suitably interrogated, an impression marker should be able to give some account of why marks were assigned as they were. S/he will be following some sort of private marking scheme with associated weightings or internalised trade-off rules and priorities. A perceived problem with
impression marking is that markers might differ considerably in their private marking schemes and thus give very different grades to a given candidate’s performance.

It is understandable that some examining veterinarians are wary of holistic grading because it is often associated with so-called subjective judgment. The notion of impression might go against the grain for a scientist who has possibly read that holistic assessment is less reliable than analytic without realising that this research usually refers to large-scale marking operations where the examiner (marker) may only have to read the script once. Some veterinarian examiners might feel comfortable with the word analytic because it has connotations of things clinical. Some examining veterinarians might have heard colleagues in other professions discussing the criteria for awarding literary prizes – interminable arguments about flair, structure and correct spelling, punctuation and grammar – and decided that they wanted another solution.

Nevertheless, because the criteria for assessment are considered together on a single scale (or dimension) the marking model used by the College can be classified as holistic. Implications of this finding for the College examinations are captured in Recommendation 2.

General and task-specific marking schemes

Scoring rubrics (marking schemes) may be designed for the assessment of a specific task or the assessment of a broader category of tasks (e.g. all the tasks that make up a 2-year assessment program). If the purpose of a given course is to develop a candidate’s knowledge and understanding of the subject, a general scoring rubric may be developed and used to assess each of the tasks done by that candidate. This approach would allow the candidates to use the feedback they acquired from the last presentation to improve their performance on the next presentation – which complements the principle that the fullest and latest information should be the basis for decisions about a final overall judgment.

If each task focuses upon a different veterinary concept or field of inquiry and the purpose of the assessment is to obtain evidence of the candidates’ knowledge of that particular topic, a general scoring rubric for assessing a sequence of tasks may not be adequate. Scientific topics or problems differ in both influencing factors and outcomes. In order to assess candidates’ factual and conceptual knowledge of these topics it would be necessary to develop separate scoring rubrics for each task. This is called a task-specific scoring rubric because it is designed to judge the quality of candidate performance on a single assessment event.

Scoring rubrics can be designed to contain both general and task-specific components. If the purpose of an oral examination is to assess candidates’ communication skills, say, and their knowledge of a particular veterinary topic or speciality, an analytic rubric could be used that contains both a general component and a task-specific component. The communication component of the rubric could consist of a general set of criteria developed for the assessment of communication skills; the task-specific component of the rubric could contain a set of criteria developed with the specific topic/speciality in mind.

The point being made here is that combining results on written and oral/practical examinations for a single result requires careful consideration of what is being required of candidates and what is being rewarded in the marking scheme (scoring rubric), and how the specific and general components of the results will map onto the features of candidate performance that are described as desirable in College statements.

Combining results

Do you add up the marks? Do you trade off across the criteria? Do you apply a given combination rule? These are three classic questions in educational measurement.

Assumed weightings in aggregation of scores

The College examiners appear to do the following (or variations of it): Adding up marks from Examiner 1 on a given paper. Adding up the marks from Examiner 2 on that same paper. Adding
the totals (aggregating) and then dividing by 2 (averaging). Sometimes this even occurs at the question level. There is a sequence of aggregating. First it might involve Written Paper 1 and Written Paper 2 (each with multiple markers) and then Written Examination score. Second it might involve Oral Examination score (from multiple markers). Third it might involve Written Examination score and Oral Examination score … finally producing an overall score.

Let us call the Oral Examination score (SOE) and the Written Examination score (SWE). It is necessary to calculate both of these to decide whether the candidate has met minimum requirements. (The requirements for a ‘Pass’ are that candidates achieve an average score across the oral and written components of at least 70; and at least 70 in one component and at least 55 in the other component.)

Let us call the average of SOE and SWE the overall examination score (SOO). It is necessary to have a value for this to decide how the candidate is to be graded.

**Summary of arithmetical operations currently undertaken by examiners**

For a given student,

SOE = \( \sum \text{SOE}_w / n \)

SWE = \( \sum \text{SWE}_o / n \)

Where \( n = \) no. of examiners, typically 2 or 3

Then SOO = \( (\text{SOE} + \text{SWE})/2 \).

**Conditions for a ‘Pass’**

That SOO \( \geq 70 \)

And that

Either SOE \( \geq 70 \) and SWE \( \geq 55 \)

Or SWE \( \geq 70 \) and SOE \( \geq 55 \).

Is it assumed that written and oral components ‘count’ equally because of the process of adding up and dividing by 2? If it is a College requirement that written and oral results be equally weighted, then the process of adding up and dividing by 2 does not guarantee such an outcome. An explanation follows.

Combining results of different assessments so that the different assessments (here written and oral components) count equally towards the total occurs only if the following condition is met: Both sets of results have the same standard deviation; that is, the spread of marks on the written component (based on two papers) and oral examinations (based on one ‘paper’) were equivalent. Say, for example, the examiners bunched up the scores on the written component and this spread of written scores turned out to be much less than the spread of the scores on the oral component. The consequence of this would be that the overall ranking of candidates was a function of performance on the orals. In other words, the relative weighting of the different assessments towards the overall score is a function of the ratio of the standard deviations of the two components.

The population of candidates in any given subject, being less than 30, is not large enough to have to deal with this artefact of combining marks.

A practical consequence of this, however, is in the algorithm applied at the end of the oral examinations. Assuming that the actual weighting really is 1:1 (written to oral), it becomes unnecessary to do the interminable sequence of adding up and dividing by 2 (as described earlier in this section of the report).
If only one number (an overall score) were required, it would be sufficient to add scores from Examiner 1 on all questions on both papers and scores from Examiner 2 on all questions on both papers, in any order, and divide the total thus obtained by 2, rounding off at this last step. But more than one number is required so that the 70, 55 rule can be applied. There could still be fewer steps than in the present algorithm.

If three numbers were required (overall score, written component score, oral component score), calculate the overall score first as in the sequence above.

Rounding-off rules would express each final score as two digits (e.g. 74 not 74·125). When rounding off, .5 goes ‘up’ if it is attached to an even number and goes ‘down’ if it is attached to an odd number (e.g. 74·5 becomes 75; 75·5 becomes 75). (The opposite could apply as long as some go ‘up’ and some go ‘down’; that is, 74·5 becomes 74; 75·5 becomes 76).

If the overall score thus obtained is less than 70, do not bother calculating the written and oral component scores separately because the candidate has already failed. If it is greater than 70, proceed to add scores from Examiner 1 on all questions on the oral component and scores from Examiner 2 on all questions on the oral component, in any order, and divide the total thus obtained by 2 ($T_1$). Do the same thing for scores on the written component ($T_2$). Check that one or other of $T_1$ or $T_2$ is greater than 70 while the other is greater than 55.

This might well be the algorithm applied by many examiners already. It is definitely not the algorithm applied by all examiners. A survey of examiners on how many times they divide by 2 in operating on scores for one candidate might reveal more information about preferred algorithms. It must be reiterated that the algorithms applied by the examiners did not produce incorrect outcomes; but they may have been more complex than necessary.

There would not seem to be any point in comparing marks from different examiners on individual questions unless at the final stage the examiners found that their final marks were significantly different and they wished to go back into individual questions to attempt to reconcile the apparent inconsistency in their judgment of the candidate. Of course, if a question-by-question analysis were undertaken retrospectively to discuss, for example, how different questions ‘worked’ then comparing marks across examiners or averaging marks assigned by different examiners would be a necessary calculation.

Finally, on the subject of location and spread of scores, it is worth noting that the mathematics underpinning the discussion on combining scores is transferable to another topic of interest in the College examinations—‘parity of esteem’ of subjects. In other words, are all subjects held in the same regard? Is it easier to gain membership or fellowship in one veterinary discipline than in another? Are all subjects of equal difficulty either intrinsically or in standards applied by examiners? An interesting exercise might be to interrogate existing data within and across years to calculate mean and standard deviation of all written scores in membership examinations and mean and standard deviation of all oral scores in membership examinations, compare their magnitude, and then reflect on possible explanations for similarities and differences in magnitude. It is not discussed any further in this report.

**Ways of trading off inconsistent performances**

What is the model used by the College to decide if a candidate passes or fails, especially if the candidate’s performance is not consistent across papers?

There are five ways to deal with making an on-balance judgment when the candidate gives a high performance on one criterion being assessed and low performance on another.
1. No trade-off allowed
2. Holistic marking
3. Maximum domain performance
4. Minimum domain performance
5. Analytic marking.

The first method does not allow for compensatory performance. For example, in the unlikely event that a candidate scored 99 on the written paper and 50 in the oral examination giving an overall result of 75, the final result would still not be a Pass because the outstanding performance on the written is not considered compensation for the below-par performance on the oral even though the total score exceeds 70.

The second method, holistic marking, has already been discussed in detail in this report. It assumes that expert judges make ‘private’ trade-offs (refer to wine-judging model).

Maximum domain performance is not relevant to our discussion. Minimum domain performance is what the College examinations purport to do because the 70, 55 rule for Pass/Fail defines minimum acceptable standards – an ‘at least’ model. Candidate has to obtain at least 55 on one paper and at least 70 on the other.

Analytic marking has already been discussed in detail in this report. It assumes that more than one criterion is assessed and that those multiple criteria are on different dimensions, which is not necessarily the case in the College examinations although there is some ‘pretence’ of marking schemes. The most sophisticated use of analytic marking has the scores on each criterion adjusted to that they are on the same scale before being combined.

Assigning grades
After candidates have sat the membership or fellowship examinations, they are awarded one of three available grades – Pass, Fail, or Supplementary. The College requirement for a pass is that candidates achieve an average score across the oral and written components of at least 70 and at least 70 in one component and at least 55 in the other component.

Theoretically there are various ways of assigning grades but only three of them are relevant here: numerical boundaries; combination rules, and qualities of a particular piece of work or performance. Each is now discussed in turn starting with setting numerical boundaries for grades.

Table 7: Setting numerical boundaries for grades

<table>
<thead>
<tr>
<th>Score range</th>
<th>Grade</th>
<th>The most commonly used grading scheme throughout the world is to award grades according to predetermined numerical ranges; that is, after setting numerical boundaries for the grades.</th>
</tr>
</thead>
<tbody>
<tr>
<td>90–100</td>
<td>A+</td>
<td></td>
</tr>
<tr>
<td>80–89</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>65–79</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>50–64</td>
<td>C</td>
<td></td>
</tr>
</tbody>
</table>

This scheme does not grade candidates against one another, but according to whether they reach pre-set standards as specified by the mark ranges. This is what underpins many examiners’ marking sheets. In some place at some time, the grade boundaries were set as policy.

For use of numerical boundaries (or ‘cut scores’) to be educationally sound and ethically defensible, there needs to be a great deal of attention to how the marks are generated; for example, it would require attention to subject matter and skill domains, sampling of the domain, validity of assessment items, and cut scores related to achievement in that subject.
A grading scheme that has become increasingly common throughout the world is to communicate to candidates at the beginning of a course what the criteria are for the examination. Candidates are informed about what the assessment program will be, its components, and how results will be combined. In the case of overall grade, the composition rule states how the results are to be combined and grades assigned. Here the criteria are the composition rules. In practice this often looks like an elaborate version of cut scores with mandatory minima on different components.

Nominating the criteria, however, does not tell the candidates anything at all about the standards.

Criteria for marking a single large piece of writing could be as follows (they refer to the quality of the writing):

- Relevance to question asked;
- Validity of argument, including logical development;
- Organisation of the answer, including clarity of expression; and,
- Presentation.

In the example above, it would not be sensible to have the assessment criteria written in the language of one instance of assessment. The standards issue in the marking of orals cannot satisfactorily be resolved by stipulating quality criteria for individual pieces of work or academic episodes (i.e. question-specific marking schemes). There must also be a general component.

**Conclusions about marking and grading**

Despite all the good intentions that no doubt accompanied the design and implementation of the clerical scoring rules, the situation is too complicated and too far removed from a few simple concepts; hence Recommendations 2 and 3.

Although a change in marking model has been suggested, it should be noted that candidate results from the present scoring rules would not be inaccurate even though the process is cumbersome. It should also be noted that holistic marking was recommended because College examiners are demonstrably capable of internalising standards and applying them in the manner that is necessary for making overall judgments.
VALIDITY AND RELIABILITY

Validity
To ask about validity is to ask about the degree to which an assessment instrument (examination) measures what it purports to measure: Does the examination or examination question represent the real thing we want to assess?

For the College-devised examinations, an important part of being valid is being authentic. The examinations are designed so that the questions are real (candidates’ examination experiences should not be unlike the environment of veterinary practice). Also, the range of response modes is not restricted to written multiple-choice questions but includes written constructed response and extended answer, and oral/practical components. This assessment model is authentic because it involves candidates in demonstrating relevant and useful knowledge.

Cronbach (1989) provides five perspectives for questioning the validity of tests. Table 8 summarises these perspectives. As an aid to interpreting the table, the summary of the first perspective is expanded: The functional perspective on validity requires us to look at the examination (or test) through a lens that focuses on the worth of the examination and, in drawing conclusions about worth, ask questions about the consequences of the examination on various players and about what sorts of behaviours might be observed before the examination. In a certain context the answers could be: As a result of this examination, certain people are denied access to high-status positions and that an antecedent to the examination is the pressure on candidates to attend a coaching college. Of course, positive scenarios also exist.

Table 8: Five perspectives on validity

<table>
<thead>
<tr>
<th>Perspective</th>
<th>Concerned with …</th>
<th>A story about …</th>
</tr>
</thead>
<tbody>
<tr>
<td>Functional</td>
<td>Worth</td>
<td>Antecedents and consequences</td>
</tr>
<tr>
<td>Political</td>
<td>Fairness and the disputatious community</td>
<td>Democracy and accountability</td>
</tr>
<tr>
<td>Operationist</td>
<td>Content</td>
<td>Demands &amp; delivery, Range &amp; balance</td>
</tr>
<tr>
<td>Explanatory</td>
<td>Interpretations</td>
<td>Adequacy and appropriateness</td>
</tr>
<tr>
<td>Economic</td>
<td>Statistics</td>
<td>Relevance and utility ('bangs for the buck')</td>
</tr>
</tbody>
</table>

In the following statements, terms that appear in Table 8 are italicised by way of illustrating one of the perspectives.

4 The terms ‘assessment’, ‘test’ and ‘examination’ are used interchangeably throughout this report even though, according to Calfee (1993), there is a hierarchy.
Absence of bias is a political essential as well as a prerequisite for public acceptance (at least within the veterinary community) and for the validity of the examination. Appropriate demands on candidates is a seldom formulated but absolutely central requirement with implications for question content as well as for layout, clarity of instructions etc. Content validity requires a balance of proportion of various dimensions and a suitable range of question characteristics.

Some of the entries in Tables 8 are now elaborated on. Of the five validity perspectives referred to above, Table 9 takes three that relate to topics of discussion about College examinations elsewhere in this report, and provides the cross-referencing.

Table 9: Comparison of College issues with some validity perspectives

<table>
<thead>
<tr>
<th>Perspective on validity (Cronbach)</th>
<th>Related topic discussed elsewhere in this report</th>
</tr>
</thead>
<tbody>
<tr>
<td>Political</td>
<td>Absence of bias</td>
</tr>
<tr>
<td>Economic</td>
<td>Analytic versus holistic marking</td>
</tr>
<tr>
<td>Operationist</td>
<td>Examination design</td>
</tr>
</tbody>
</table>

Reliability

To ask about reliability is to ask about the degree to which the results of assessment are consistent, dependable, or repeatable; traditionally between and within markers, over time: Would the candidate get the same score if/she did the examination again and gave the same performance? Would different examiners (or pairs of examiners) score it differently?

Standardised multiple-choice tests are reliable by design but their validity can be questionable. On the other hand, exhibitions and orals are valid but not necessarily reliable. So how are candidates protected from capricious, biased judgments?

An adequate response to the reliability challenge of the College-devised examinations would be an enhanced procedure for monitoring the judgments of examiners. There is an operational recommendation to this effect (Recommendation 12). It complements the overarching recommendation about the capture of data (Recommendation 3). The discussion leading up to this occurs in the section, ‘Hypothesis Testing’.

The psychometric model that ‘observed score = true score + error’ might readily suit notions of validity and reliability for omnibus multiple-choice testing, but it does not readily suit notions of validity and reliability for testing in the authentic way of College-devised assessment. In reality, for the College’s type of assessment involving examiners and candidates, some assumptions of the true-score model do not hold (e.g. infinite populations, examiners and questions sampled at random from a universe of examiners and questions). In the College setting, split-half reliability estimates are not possible and the practice of inter-rater agreement studies is beyond its resources. Here, the purposes served by reliability can be broadened to include the marking practice of contextualised judgment.

Moss (1994) has three aspects for reliability in which examiners’ contextualised judgments are in the foreground. Table 9 summarises these aspects.

---

5 In the literature these are the judgments of school or university teachers. Used here in the same sense -- examiner and candidate not necessarily being unknown to each other. In the case of internal assessment, the teacher or lecturer is known to the candidate.
Table 10: Three aspects of reliability

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Concerned with …</th>
</tr>
</thead>
<tbody>
<tr>
<td>Privileging contextualised judgments</td>
<td>Criteria and standards for marking</td>
</tr>
<tr>
<td>Generalising across tasks (questions)</td>
<td>Inconsistency: a puzzle to be solved</td>
</tr>
<tr>
<td>Generalising across readers (examiners)</td>
<td>Candidates become part of the dialogue by evaluating and challenging conclusions</td>
</tr>
</tbody>
</table>

Some of the entries in Table 10 are now elaborated on. Three reliability aspects referred to above (left-hand column) spawn seven artefacts (unevenly distributed over three rows in the right-hand column). Four of these seven artefacts are highlighted in Table 10.) These relate to topics of discussion about College examinations that can be found elsewhere in this report (see Table 11).

Table 11: Comparison of College issues with some reliability aspects

<table>
<thead>
<tr>
<th>Aspect of reliability (Moss)</th>
<th>Related topic discussed elsewhere in this report</th>
</tr>
</thead>
<tbody>
<tr>
<td>Criteria and standards for marking</td>
<td>Analysis of state of standardisation</td>
</tr>
<tr>
<td>Inconsistency: a puzzle to be solved</td>
<td>Examining in pairs</td>
</tr>
<tr>
<td>Latitude in selecting products of assessments/examinations</td>
<td>Question choice</td>
</tr>
<tr>
<td></td>
<td>Combining written and oral results</td>
</tr>
<tr>
<td>Candidates become part of the dialogue by evaluating and challenging conclusions.</td>
<td>Appeals process</td>
</tr>
<tr>
<td></td>
<td>Disaffected candidate</td>
</tr>
</tbody>
</table>

Finally, a comment about objectivity (refer back to discussion of objectivity and subjectivity in the section, ‘Assessing professional competencies’): Objectivity is not a synonym for reliability. Objectivity is operationalised as the degree to which results are independent of examination location, supervisor, and examiner. This requirement refers to marking schemes and marking procedures and also to the construction of the examination paper.

Using these notions of validity and reliability

Cronbach’s tight notions of validity and Moss’s broad notions of reliability can be helpful when discussing situations where the pursuit of validity and reliability together necessitates a balancing act of sorts when setting and marking examinations – which is the case for the MACVSc and FACVSc examinations.

Post script on bias

While this review has dealt with issues such as question format, question choice, allocating and combining marks, clarifying ambiguities for candidates, and the incline of difficulty of questions on an examination paper, there would no doubt be some important issues that remain unidentified and therefore unexplored.

---

\(^6\) There is no glossing of terms that are not discussed within this report.
An issue that was identified but that was not able to be explored in the current review is the notion of bias due to socio-economic status or geographical location. The hypotheses on bias that were tested during this review, and which appear towards the end of this report, relate to sex differences. Socio-economic differences and place of residence/study require serious investigation: For example, is there an interaction of socio-economic status and field of study for membership or fellowship? Is there an interaction of geographical location and gender? How do these factors account for any variance in results obtained by the sub-groups? How does the interaction of sex, socio-economic status, and geographical location, say, account for variations in oral communication skills? Is there a significant difference in success rates for candidates from the country and candidates from the city? And what about candidates from outside Australia?

A concern that arose during observations of the oral examinations relates to the notion of parity of esteem of subjects, which has already been mentioned in this report (see ‘Combining results’). The notion of parity of esteem finds expression in this apparently simple question: Is it harder to get a membership/fellowship in some subjects than in others? Do the candidates perceive this to be the case? Does it matter? It might be the case that the College does not proclaim comparability between subjects to be an operating principle – there is no reason why it should do so. But what is important it that the phenomenon be acknowledged.

Not only do such questions require answers for reasons of fairness and accountability but also they require answers so that the relevance and usefulness of the qualifications, membership and fellowship of the Australian College of Veterinary Scientists, can be further enhanced.
During the review several hypotheses were formulated. After the complete dataset was compiled, five of these hypotheses, stated below, were put to the test. Six figures, Figures 2–7, present the evidence for supporting or not supporting each of the hypotheses. Table 12 summarises the results.

**Statement of hypotheses**

*Hypothesis 1 (H1)*
Examiners first decide whether a candidate has passed or failed, and then use marks to support this decision.

*Hypothesis 2 (H2)*
In the pairing of examiners for the membership examinations there is an assortment of examiners by sex.

*Hypothesis 3 (H3)*
There are sex-related differences in scores.

*Hypothesis 4 (H4)*
Examiners do not use a common scale for marking the oral examinations.

*Hypothesis 5 (H5)*
The current method for recording/calculating results is a potential source of discrepancies.

**Testing of hypotheses**
The analyses are presented within six displays. See Figures 2, 3, 4A, 4B, 5 and 6.

**Results of hypothesis testing**
See Table 12. The results from testing Hypothesis 1 relate to Display 1 and so on.

**Displays**
There are six displays presenting data and describing the analyses that lead to hypotheses being support or not as outlined in Table 12.

- Figure 2: Score distribution – Display 1
- Figure 3: Sex effects – Display 2
- Figures 4A & 4B: Sex effect on scores – Displays 3 & 4
- Figure 5: Scoring – Display 5
- Figure 6: Calculating component scores – Display 6
Table 12: Results of hypothesis testing

<table>
<thead>
<tr>
<th>H#</th>
<th>Supported?</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>One of the features of Display 1 is that the 70% ‘pass’ score is the single most frequent score in the distribution of overall scores while there is a big dip in scores either side of 70.</td>
</tr>
<tr>
<td>2</td>
<td>Yes</td>
<td>According to the tables in Display 2, the pairing of examiners does not appear to be random. This, of itself, is not necessarily a good or bad thing. It depends on whether the examiner pairings affect examination results. (We do not propose random assignment of examiners even though (unmeasured) examiner characteristics such as experience and compatibility seem to affect the sex mix.)</td>
</tr>
<tr>
<td>3</td>
<td>Maybe, maybe not</td>
<td>According to Display 3, there is an interaction of sex of candidate with examiner pairing type (sex mix). The effect is not the same for the four groups of candidates – male (oral), male (written), female (oral), and female (written) vary according to examiner pairing type (sex mix). Could be an artefact. According to Display 4, the correlation of written scores and oral scores is significantly lower for female candidates when the examiners are a mixed-sex pair.</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>According to Display 5, there is a lack of consistency (albeit only slight) between examiners in their use of marking scales. This would only be an issue if H1 were not supported (i.e. if examiners had not already made the pass/fail decision before using marks).</td>
</tr>
<tr>
<td>5</td>
<td>Yes</td>
<td>Display 6 identifies two potential sources of error, one associated with transcribing numbers, the other to rounding off numbers. Both sources could be eliminated (see Recommendation 2).</td>
</tr>
</tbody>
</table>

Disclaimers

All results should be treated as indicative only.

The following information should be read in conjunction with Table 12 and Displays 2–4:

- We do not suggest that these are ‘statistical analyses’, or that the results are statistically significant. Inferential statistics would not be useful in situations like this: We are not sampling from a theoretically infinite population of candidates and we are not assuming distributional characteristics of the variables in question. We are looking at a point-in-time census.

- We recognise that for some combinations of examiners the number of examiners is small and the results might be an artefact of this. But for other combinations of examiners where numbers are around 20, we are confident that the differences are worth commenting on – we refer to some differences as ‘slight’ and some as ‘marked’. For 2007, there is no large subject with a disproportionate number of candidates of one sex or a disproportionate number of examiners of one sex to distort the results.

- Not all analyses are between subjects or based on raw scores. For example, the oral–written correlation is within candidate and independent of subject.

The observed patterns may be due solely to this particular assortment of candidates and examiners. Therefore, scores from the 2008 examinations should be studied to see if similar patterns arise (incorporated in Recommendations 12 and 13).

Thus the objectivity of the assessment process can thus be continuously monitored.
The distributions of recorded individual examiners' scores for candidates for the oral and written components are shown at right. The requirements for a pass are that candidates achieve an average score across the oral and written components of at least 70% and at least 70% in one component and at least 55% in the other component.

Features of the display:
- the 70% 'pass' score is the single most frequent score for both components. This feature of score distributions is sometimes explained in terms of item selection in exams, a result of 'fudging the test', or a result of marking practices at the pass/fail threshold.
- the peak at 66% for both components and the dip between 65% and 70% (circled). One possible reason is that examiners adjust scores that are just below 70% (i.e., around 65%) up or down to give either a pass or a more obvious fail.
- the score distribution is an artefact of item scoring criteria.
- marks are higher for the written component across the range than for the oral component. The difference is of the order of 3 percentage points across the distribution.

Notes to the display:
The 70% score is indicated by the solid triangles in the display. The scores are not integers and have been 'binned' for display purposes so that, for example, the 70% score includes 70.2, 70.8 and 70.5.

Figure 2. Score distribution (Display 1)
Figure 3. Sex effects (Display 2)
Figure 4A. Sex effect on scores (Display 3)

Feature 1
There is a marked difference in median oral scores for female candidates between cases where the examiner pair is mixed sex and where the examiners are both males - but this pattern is not replicated for male candidates.

This effect - if it is a true effect and not an artefact - merits monitoring and further investigation.

Feature 2
Female candidates receive slightly lower written scores than do their male counterparts when the examiners are both male. Although the median written scores are the same for both sexes, the spread of scores is greater for males than females.

general
To examine whether pairing type had any effect on scores we can consider the six possible combinations of candidates with examiners. A candidate may have two examiners of the same sex as themselves, two examiners of the opposite sex, or a female and male examiner; and candidates may be female or male.

The expected pattern if there is no effect on candidate's scores of examiner type is something like we see in the 'male candidates - oral' panel. The median is similar across the three examiner types, and the spread is also similar.

Caveat: the relatively small numbers mean that differences are indicative rather than conclusive.

The elevated median for males with mixed-sex examiner pairs is likely an artefact of the small numbers for this combination.

Notes to the display:
- The solid diamonds show the median scores for the three examiner pair types: mixed sex, female-female, and male-male.
- The vertical lines show the interquartile range, that is, the range from the 25th percentile to the 75th percentile.
- The open diamonds show the equivalent median for the opposite sex candidate.
Figure 4B. Sex effect on scores (Display 4)
Figure 5. Item scoring (Display 5)
For each component (written and oral) total scores from the two examiners are averaged to give the component score.

We independently calculated this component score and examined the difference, if any, between the calculated component score and the recorded component score (i.e., the component score recorded for the candidate on the sheets supplied to us). We ignored differences that could be accounted for by rounding (e.g., 55.5 recorded as 55; 47.25 recorded as 47.2) or where the scores rounded to the nearest integer were not different (e.g., 57.3 recorded as 56.9; both would be rounded to 57; 72.05 recorded as 71.8, both would be rounded to 72).

Other differences we termed a 'discrepancy'. A discrepancy represents an error at some point in the calculation of the component score, or in the recording of the elements contributing to that score, that made a material difference to the integer recorded score (e.g., 74.675 recorded as 74.1; 76 recorded as 81).

Across the two components the discrepancy rate was 5.9%. None of the discrepancies appears to have affected a pass/fail decision in this dataset but the rate should be zero.

Some of these discrepancies were inconsistent rounding
- e.g., 80.5 rounded up to 81; 21.5 rounded down to 21 (rather than up to 22)
and rounding was inconsistent even when it was a minor difference
- e.g., 76.75 rounded up to 77; 82.75 rounded down to 82.5 (rather than up to 84).

While one apparent discrepancy was, in fact, the result of an intervention by the Chief Examiner, there appears to be no reason for the others.

There were also two cases where the discrepancy was 5 (recorded 81, calculated 76) and 10 (recorded 75.75, calculated 65.75). These discrepancies may be recording errors.

It is also possible that errors were introduced in the transcription of paper-based data to electronic files for the purposes of undertaking the analyses contained in this report.

Whatever the source of the discrepancies, however many of them there really are, and however large or small their effects, electronic data capture would reduce the clerical load on examiners.

Figure 6. Calculating component scores (Display 6)
GENERAL OBSERVATIONS

Of the issues, observations and uncertainties identified during this review and listed below, some have already been analysed while others are noted here for the record.

- There is nothing new under the sun: The general assessment literature is transferable to the context of the veterinary college examinations.
- The College is to be congratulated for its attention to detail and consideration of every possible contingency. Also, its policies and procedures are well documented.
- The Executive Officer and the Chief Examiner and any other person involved in designing this examination system have set up procedures or prepared strategies for dealing with the whole gamut of issues associated with tests and examinations, not necessarily using the associated technical jargon but with full knowledge of the problem to be dealt with.
- The examinations have high face validity and high levels of inter-rater agreement.
- The clerical finesse exhibited in conducting the examinations was impressive. The examinations were completed (as in final results available) in a comparatively short time (and paper-based at that).
- The algorithms applied for calculating final scores, although mathematically accurate, could be made simpler. This would reduce the role of hand-held calculators or cell phones in calculating averages and total scores.
- Two different grading methods were being used simultaneously – combining scores and making on-balance judgments. While it could be conjectured that this situation is related to some past notion of subjectivity and objectivity and examiners’ professional judgments, the situation is unnecessary.
- The ultimate decision to be made by the College is how to trade off between the ‘political’ imperative (face validity and perceived fairness) and the ‘economic’ imperative – there would be faster cheaper ways of coming to the same conclusions about whether candidates attain the standard for membership or fellowship of the College, with results that are just as reliable.

Psychometric considerations

- Giving candidates choice on the written paper
- Stating that all questions are of equal value
- Range and balance of questions on an examination
- Independence of judgments of examiners working in pairs or trios
- Inter-rater reliability (more formal capturing required for research purposes or in case of appeal)
- Parallel nature of written, oral, practical – being sure about the construct being assessed
- Acknowledgment of possible relativities in subject esteem – realising that the ‘bar’ might be ‘set higher’ for passing some subjects compared with others
- Place of higher-order thinking skills (vis à vis communicating knowledge)?
- Sub-group differences (gender, socio-economic background, geographical location etc.)?
- Interactions of gender, background, and examiner–candidate relationship?
- Authenticity of the assessment technique for orals (does it involve candidates in second guessing)?
- Marking backwards to the pass mark.

---

7 Acceptance of the apparent ability of a test or study to measure what it purports to measure. Face validity is an intuitive matter since it relies upon inspection and opinion rather than statistical methods.
RECOMMENDATIONS

Preamble to the recommendations

There might appear to be a large number of recommendations. It should be noted, however, that many of the recommendations are operational rather than conceptual and, as such, are merely suggestions for refinement of an already rigorous process.

It might be that some of the recommendations are judged to be in conflict with the philosophy of the College or are based on inaccurate perceptions of College procedures. Such recommendations should be ignored.

Some readers might interpret some recommendations as having negative overtones. This would not be a fair portrayal of the situation. All of the recommendations fall into one or other, even all, of the following categories:

- Making life easier for College, candidate, or examiner;
- Exploiting strengths already demonstrated by personnel in certain roles or expanding on procedures already in place and successful;
- Acknowledging the implications of certain procedures (without necessarily having to change such procedures); and/or
- Becoming even more sophisticated in assessment procedures, but at a financial cost.

There is no recommendation that is a grave reflection on current procedures.

The recommendations are numbered 1 to 13. Recommendations 1–3 are overarching recommendations. Recommendations 4–13 are operational.

Overarching recommendations

1. Maintain the oral/practical examination to complement the written examination but consider adapting the model as presently administered.

2. Settle on a conceptually simple model for marking; namely, holistic assessment by expert judges (the examiners having already demonstrated their expertise in this sort of assessment), and thus eliminate the necessity for examiners to perform unnecessary calculations with numerical marks.

3. Introduce electronic data capture for examiners in recording marks and assigning grades (results), and make explicit the current model that examiners use for making decisions about a candidate’s final grade (pass/fail/supplementary).

Operational recommendations

Setting

4. Consider payment of examiners in return for subjecting them to additional processes at setting and marking stages.

5. Expand the examination setting process with an editorial loop or service.

6. Expand the examination setting process to include production of marking schemes in standard format, agreed to by all examiners in a given subject.

7. Revisit the rationale for providing candidates with extensive choice of question in written examinations.
Questioning

8. Vary the style of questions so that the oral examination assesses a different set of skills and processes from the written examination (while acknowledging that there is sometimes the need for overlap).

9. In structuring the examinations to match the subject outcomes, balance the credit given to communication skills and higher-order thinking skills in the veterinary context.

10. Review written examination questions for variation in question styles such as open-ended questions to test breadth and depth of knowledge in subject, written communication skills, ability to select relevant information and so on, closed questions of the ‘analysis of data’ type where data are those that the candidate would encounter in the subject in practice and in refereed journal articles, and case studies.

11. Consider using videotape to record oral examinations for training purposes.

Marking and grading

12. Monitor the marking process in terms of agreement between examiners and of allocation of examiners to candidates.

13. Introduce more post-hoc analyses of data on the performance of candidates.
REFERENCES


EXECUTIVE SUMMARY

Terms of reference
1. Review procedures and written information for examinations.
2. Review processes for preparing examination papers and oral/practical material.
3. Review the conduct of examination and assessment procedures (including grading).
4. Appraise procedures and processes against College objectives and guidelines, and against accepted standards for professional assessment.
5. Recommend improvements or alternatives.

Period of the review
April to October 2007

Outline of the method
The review consisted of a real-time natural study located in the experiences of the College executive, examiners and candidates. The iterative, descriptive, action research was designed to produce well-grounded and informed opinion so that any procedural recommendations would be based on existing practice and theoretical notions of validity and reliability.

Frames of reference for analysis
- Research literature on assessment in general (especially format and orals);
- Research literature on assessment of professional competencies;
- Best practice in setting examination papers – technical and editorial aspects;
- Validity (Cronbach, 1989); Reliability (Moss, 1992).

Structure of the report
The report is in the form of a commentary and critique. It covers the preparation and conduct of examinations, the marking of candidate responses, and the grading of candidates for membership and fellowship of the Australian College of Veterinary Scientists. The review draws on the literature about the theory of educational assessment and the many facets of professional assessment in action.

The first major section addresses assessment in general, especially examination format and oral examinations. The second major section is about the assessment of professional competencies, including communication within a veterinary context. The third major section is about examination, the fourth about marking and grading, and the fifth about validity and reliability.
Following the main commentary and critique are sections with different emphases: formulation and testing of hypotheses; summary of conclusions; and presentation of three overarching recommendations and ten operational recommendations.

**General observations**

- There is nothing new under the sun: The general assessment literature is transferable to the context of the veterinary college examinations.
- The College is to be congratulated for its attention to detail and consideration of every possible contingency. Also, its policies and procedures are well documented.
- The Executive Officer and the Chief Examiner and any other person involved in designing this examination system have set up procedures or prepared strategies for dealing with the whole gamut of issues associated with tests and examinations, not necessarily using the associated technical jargon but with full knowledge of the issues at hand.
- The examinations have high face validity\(^8\) and high levels of inter-rater agreement.
- The clerical finesse exhibited in conducting the examinations is impressive. The examinations are completed (as in final results available) in a comparatively short time (and paper-based at that).
- The algorithms applied for calculating final scores, although mathematically accurate, could be made simpler. This would reduce the role of hand-held calculators or cell phones when calculating component and overall scores.
- Two different grading methods are being used simultaneously – combining scores and making on-balance judgments. While it could be conjectured that this situation is related to some past notion of subjectivity and objectivity concerning examiners’ professional judgments, this ‘double whammy’ is unnecessary. One of the overarching recommendations relates to this important finding.
- The ultimate decision to be made by the College is how to trade off between the ‘political’ imperative (face validity and perceived fairness) and the ‘economic’ imperative – there would be faster cheaper ways of coming to the same conclusions about whether candidates attain the standard for membership or fellowship of the College, with results that are just as reliable.

**Psychometric issues**

- Giving candidates choice on the written paper. Is this in their best interests?
- Stating that all questions are of equal value. Are they?
- Range and balance of questions on an examination. Could be enhanced.
- Independence of judgments of examiners working in pairs or trios. There appears to be a certain degree of inter-dependence.
- Inter-rater reliability. A more formal capturing of data would be useful for research purposes or in case of appeal.
- Parallel nature of written and oral/practical examinations. Examiners need to be sure about the construct being assessed in each of them and in the overlap of both.
- Acknowledgment of possible relativities in esteem for subjects offered for examination. Are the examinations comparable in difficulty between subjects within each standard (membership and fellowship)?

---

\(^8\) Acceptance of the apparent ability of a test or study to measure what it purports to measure. Face validity is an intuitive matter since it relies upon inspection and opinion rather than statistical methods.
• Place of higher-order thinking skills vis-à-vis communicating knowledge. What is the hierarchy of these dimensions in judging the quality of a candidate’s performance, written or oral? How often is the dimension, communication skills, taken into account?

• Sub-group differences (sex, socio-economic background, geographical location etc.). Sex differences turned out to be different empirically from what the reviewer had expected from observations of a sample of examinations. Other potential sources of bias need to be monitored.

• Interactions of sex, background, and examiner–candidate relationship. Requires further investigation even though numbers of candidates and examiners are statistically small.

• Authenticity of the assessment technique for orals. Could it possibly involve candidates in second-guessing? This is an important finding.

• Marking backwards to the pass mark. This observation was validated empirically. It has implications for the way the marking operation is carried out and results of candidates collected. The repercussions are not a problem at present given the level of agreement between examiners’ holistic judgments.

Empirically based findings

• Examiners first decide whether a candidate has passed or failed, and then use marks to support this decision.

• There is an assortment by sex in the pairing of examiners for the membership examinations.

• There may be some sex-related differences in scores.

• Examiners do not use a common scale for marking the oral examinations.

• The current method of arithmetic operations on marks for recording results and assigning grades is a potential source of discrepancies.

Preamble to the recommendations

There might appear to be a large number of recommendations. It should be noted, however, that many of the recommendations are operational rather than conceptual and, as such, are merely suggestions for refinement of an already rigorous process.

It might be that some of the recommendations are judged to be in conflict with the philosophy of the College or are based on inaccurate perceptions of College procedures. Such recommendations should be ignored.

Some readers might interpret some recommendations as having negative overtones. This would not be a fair portrayal of the situation. All of the recommendations fall into one or other, even all, of the following categories:

• Making life easier for College, candidate, or examiner;

• Exploiting strengths already demonstrated by personnel in certain roles or expanding on procedures already in place and successful;

• Acknowledging the implications of certain procedures (without necessarily having to change such procedures); and/or

• Becoming even more sophisticated in assessment procedures, but at a financial cost.

There is no recommendation that emanates from an identified serious deficiency in current procedures.

There are three overarching recommendations (1–3) and ten operational recommendations (4–13).
**Overarching recommendations**

1. Maintain the oral/practical examination to complement the written examination but consider adapting the model as presently administered.

2. Settle on a conceptually simple model for marking; namely, holistic assessment by expert judges (the examiners having already demonstrated their expertise in this sort of assessment), and thus eliminate the necessity for examiners to perform unnecessary calculations with numerical marks.

3. Introduce electronic data capture for examiners in recording marks and assigning grades (results), and make explicit the current model that examiners use for making decisions about a candidate’s final grade (pass/fail/supplementary).

**Operational recommendations**

**Setting**

4. Consider payment of examiners in return for subjecting them to additional processes at setting and marking stages.

5. Expand the examination setting process with an editorial loop or service.

6. Expand the examination setting process to include production of marking schemes in standard format, agreed to by all examiners in a given subject.

7. Revisit the rationale for providing candidates with extensive choice of question in written examinations.

**Questioning**

8. Vary the style of questions so that the oral examination assesses a different set of skills and processes from the written examination (while acknowledging that there is sometimes the need for overlap).

9. In structuring the examinations to match the subject outcomes, balance the credit given to communication skills and higher-order thinking skills in the veterinary context.

10. Review written examination questions for balance in question styles such as open-ended questions to test breadth and depth of knowledge in subject, written communication skills, ability to select relevant information and so on, closed questions of the ‘analysis of data’ type where data are those that the candidate would encounter in the subject in practice and in refereed journal articles, and case studies.

11. Consider using videotape to record oral examinations for training purposes.

**Marking and grading**

12. Monitor the marking process in terms of agreement between examiners and of allocation of examiners to candidates.

13. Introduce more *post-hoc* analyses of data on the performance of candidates.