General principles in assessment of professional competence

European College of Veterinary Internal Medicine Congress 2012

Debbie Jaarsma (d.jaarsma@amc.nl)
Academic Medical Centre, University of Amsterdam
The Netherlands

With many thanks to prof. dr. Cees van der Vleuten, Maastricht University
SHORT BIOGRAPHY

- VETERINARIAN BY TRAINING
- SPECIALIST TRAINING PATHOLOGY AT FACULTY OF VETERINARY MEDICINE, UTRECHT UNIVERSITY (FVMU)
- TEACHER UNIVERSITY OF APPLIED ANIMAL SCIENCES AT DEN BOSCH
- SALES MANAGER FARMACEUTICAL INDUSTRY (JOHNSON & JOHNSON)
- PhD IN VETERINARY EDUCATION AT FVMU*
- ASSISTANT PROFESSOR ‘QUALITY IMPROVEMENT VETERINARY EDUCATION’ AT FVMU
- FULL PROFESSOR ‘EVIDENCE BASED EDUCATION’ AT ACADEMIC MEDICAL CENTRE, UNIVERSITY OF AMSTERDAM

Welcome everyone!

- Are you familiar with assessment literature?
  
  A. Not at all  
  B. To some extent 
  C. Pretty familiar 
  D. As my backyard
Overview of presentation

- General principles of assessment
- Implications for practice
- Criteria for good assessment
- Final note
Assessment:

“involves testing, measuring, collecting and combining information, and providing feedback”

- Drives and stimulates learning
- Provides information on educational efficacy institutions/teachers
- Protects patients and society
Simple competence model

- Does
- Shows how
- Knows how
- Knows

Simple competence model

- Knows
- Shows how
- Knows how
- Does

Under construction
Established technology

Assessment formats used

- **Knows**
  - *Stimulus format*: fact oriented
  - *Response format*: written, open, computer-based, oral

- **Knows how**
  - *Stimulus format*: hands-on (patient) standardized scenario or simulation
  - *Response format*: direct observation, checklists, rating scales

- **Shows how**
  - *Stimulus format*: (patient) scenario, simulation
  - *Response format*: written, open, oral, computer-based

- **Does**
  - *Stimulus format*: habitual practice performance
  - *Response format*: direct observation, checklists, rating scales, narratives
What do you think?

- Clinical reasoning, particularly with experts, is:

  A. More generic than context specific
  B. More context specific than generic
  C. Equally specific and generic
Assessment principle 1

- Competence is specific, not generic
Competence is not generic

<table>
<thead>
<tr>
<th>Testing Time in Hours</th>
<th>MCQ(^1)</th>
<th>Case-Based Short Essay(^2)</th>
<th>PMP(^1)</th>
<th>Oral Exam(^3)</th>
<th>Long Case(^4)</th>
<th>OSCE(^5)</th>
<th>Mini CEX(^6)</th>
<th>Practice Video Assessment(^7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.62</td>
<td>0.68</td>
<td>0.36</td>
<td>0.50</td>
<td>0.60</td>
<td>0.47</td>
<td>0.73</td>
<td>0.62</td>
</tr>
<tr>
<td>2</td>
<td>0.76</td>
<td>0.73</td>
<td>0.53</td>
<td>0.69</td>
<td>0.75</td>
<td>0.64</td>
<td>0.84</td>
<td>0.76</td>
</tr>
<tr>
<td>4</td>
<td>0.93</td>
<td>0.84</td>
<td>0.69</td>
<td>0.82</td>
<td>0.86</td>
<td>0.78</td>
<td>0.92</td>
<td>0.93</td>
</tr>
<tr>
<td>8</td>
<td>0.93</td>
<td>0.82</td>
<td>0.82</td>
<td>0.90</td>
<td>0.90</td>
<td>0.88</td>
<td>0.96</td>
<td>0.93</td>
</tr>
</tbody>
</table>

\(^1\)Norcini et al., 1985  
\(^2\)Stalenhoef-Halling et al., 1990  
\(^3\)Swanson, 1987  
\(^4\)Wass et al., 2001  
\(^5\)Petrusa, 2002  
\(^6\)Norcini et al., 1999  
\(^7\)Ram et al., 1999  
\(^8\)Gorter, 2002
Practical implications

- Competence is specific, not generic

- One measure is no measure
  - Increase sampling (across content, examiners, patients…) within measures
  - Combine information across measures and across time
What do you think?

- Multiple choice questions are objective and therefore more reliable.

A. True
B. False
Assessment principle 2

- Objectivity is not the same as reliability
# Objectivity is not the same as reliability

<table>
<thead>
<tr>
<th>Testing Time in Hours</th>
<th>MCQ&lt;sup&gt;1&lt;/sup&gt;</th>
<th>Case-Based Short Essay&lt;sup&gt;2&lt;/sup&gt;</th>
<th>PMP&lt;sup&gt;1&lt;/sup&gt;</th>
<th>Oral Exam&lt;sup&gt;3&lt;/sup&gt;</th>
<th>Long Case&lt;sup&gt;4&lt;/sup&gt;</th>
<th>OSCE&lt;sup&gt;5&lt;/sup&gt;</th>
<th>Mini CEX&lt;sup&gt;6&lt;/sup&gt;</th>
<th>Practice Video Assessment&lt;sup&gt;7&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.62</td>
<td>0.68</td>
<td>0.36</td>
<td>0.50</td>
<td>0.60</td>
<td>0.47</td>
<td>0.73</td>
<td>0.62</td>
</tr>
<tr>
<td>2</td>
<td>0.76</td>
<td>0.73</td>
<td>0.53</td>
<td>0.69</td>
<td>0.75</td>
<td>0.64</td>
<td>0.84</td>
<td>0.76</td>
</tr>
<tr>
<td>4</td>
<td>0.93</td>
<td>0.84</td>
<td>0.69</td>
<td>0.82</td>
<td>0.86</td>
<td>0.78</td>
<td>0.92</td>
<td>0.93</td>
</tr>
<tr>
<td>8</td>
<td>0.93</td>
<td>0.82</td>
<td>0.82</td>
<td>0.90</td>
<td>0.90</td>
<td>0.88</td>
<td>0.96</td>
<td>0.93</td>
</tr>
</tbody>
</table>

<sup>1</sup>Norcini et al., 1985  
<sup>2</sup>Stalenhoef-Halling et al., 1990  
<sup>3</sup>Swanson, 1987  
<sup>4</sup>Wass et al., 2001  
<sup>5</sup>Petrusa, 2002  
<sup>6</sup>Norcini et al., 1999  
<sup>7</sup>Ram et al., 1999  
<sup>8</sup>Gorter, 2002
Reliability oral examination

<table>
<thead>
<tr>
<th>Testing Time in Hours</th>
<th>Number of Cases</th>
<th>Same Examiner for All Cases</th>
<th>New Examiner for Each Case</th>
<th>Two New Examiners for Each Case</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>0.31</td>
<td>0.50</td>
<td>0.61</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>0.47</td>
<td>0.69</td>
<td>0.76</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>0.47</td>
<td>0.82</td>
<td>0.86</td>
</tr>
<tr>
<td>8</td>
<td>12</td>
<td>0.48</td>
<td>0.90</td>
<td>0.93</td>
</tr>
</tbody>
</table>

Swanson, 1987
Practical implications

- Objectivity is not the same as reliability
  - Don’t trivialize assessment (and compromise on validity) with unnecessary objectification and standardization
  - Don’t be afraid of holistic judgement
  - Sample widely across sources of subjective influences (raters, examiners, patients)
What do you think?

Which format measures ‘understanding’ best?

A. MCQs
B. Essay questions
C. Orals
D. All of the above
Assessment principle 3

- What is being measured is more determined by the stimulus format than by the response format.
Assessment formats used

- **Knows**
  - Stimulus format: fact oriented
  - Response format: written, open, computer-based, oral

- **Knows how**
  - Stimulus format: (patient) scenario, simulation
  - Response format: written, open, oral, computer-based

- **Shows how**
  - Stimulus format: hands-on (patient) standardized scenario or simulation
  - Response format: direct observation, checklists, rating scales

- **Does**
  - Stimulus format: habitual practice performance
  - Response format: direct observation, checklists, rating scales, narratives
Empirical findings

- Once reliable (= sufficient sampling): correlations across formats are huge

- Cognitive activities follow the task you pose in the stimulus format
Knows:
What is arterial blood gas analysis most likely to show in dogs with cardiogenic shock?

A. Hypoxemia with normal pH
B. Metabolic acidosis
C. Metabolic alkalosis
D. Respiratory acidosis
E. Respiratory alkalosis

A 7-year-old bitch is brought to the emergency department. She is restless and panting. On admission, her temperature is 37.7°C, pulse is 120/min, and resp are 40/min. During the next hour, she becomes increasingly stuporous, pulse increases to 140/min, and respirations increase to 60/min. Blood gas analysis is most likely to show:

A. Hypoxemia with normal pH
B. Metabolic acidosis
C. Metabolic alkalosis
D. Respiratory acidosis
E. Respiratory alkalosis

Practical implications

- What is being measured is more determined by the stimulus format than by the response format

  - Don’t be married to a format (e.g. essays)

  - Worry about improving the stimulus format

  - Make the stimulus as (clinically) authentic as possible (e.g. in MCQs, OSCEs)
What do you think?

- The best strategy for constructing good test material is:

  A. Training staff to write test material
  B. Peer review of test material
Assessment principle 4

- Validity can be ‘built-in’
Empirical findings

- Validity is a matter of good quality assurance around item construction (Verhoeven et al 1999)
- Generally, medical (and veterinary) schools can do a much better job (Jozewicz et al 2002)
Item review process

Pre-test review

item pool

review committee

item analyses
student comments

test administration

Info to users

item bank

Post-test review

anatomy
physiology
int medicine
surgery
psychology

Pre-test review
Practical implications

- Validity can be ‘built-in’
  - Outcomes and content need to be clear and known to the item constructors AND the learner
  - Assessment is as good as you are prepared to put into it
  - Develop quality assurance cycles around test development
  - Share (good) test material across institutions
What do you think?

What drives students’ learning most?

A. The teacher
B. The curriculum
C. The assessment
Assessment principle 5

- Assessment drives learning
An alternative view

Assessment may drive learning through:

- Content
- Format
- Programming/scheduling
- Regulations
- ............
Empirical findings

- The relationship between assessment and learning is complex
- Learning strategy is mediated by the perception of the students on the assessment
- Summative assessment systems often drive in a negative way
- Formative feedback has dramatic impact on learning
- Learners want feedback (more than grades)
Practical implications

Assessment drives learning

- For every evaluative action there is an educational reaction
- Verify and monitor the impact of assessment: many intended effects are not actually effective
- No assessment without feedback!
- Embed the assessment within the learning programme
- Use the assessment strategically to reinforce desirable learning behaviours
What do you think?

The best method of assessment is:

A. Vignette based MCQs
B. Orals
C. Portfolios
D. None of these
Assessment principle 6

- No single method can do it all
Empirical findings

- One measure is no measure
- All methods have limitations (no single superior method exists)
- Different methods may serve a different function
- In combination, information from various methods provide a richer picture and combines formative and summative functions
Practical implications

- No single method can do it all
  - Use a cocktail of methods across the competency pyramid
  - Arrange methods in a programme of assessment
  - Any method may have utility
  - Compare assessment design with curriculum design
    - Responsible people/committees
    - Use an overarching structure
    - Involve your stakeholders
    - Implement, monitor and change (assessment programmes ‘wear out’)

- Responsible people/committees
- Use an overarching structure
- Involve your stakeholders
- Implement, monitor and change (assessment programmes ‘wear out’)
Assessment principles

1. Competence is specific, not generic
2. Objectivity is not the same as reliability
3. What is being measured is more determined by the stimulus format than by the response format
4. Validity can be ‘built-in’
5. Assessment drives learning
6. No single method can do it all
Criteria for good assessment

- Validity or Coherence
- Reproducibility or Consistency
- Feasibility
- Educational effect
- Catalytic effect
- Acceptability

Theme group “Criteria for Good assessment”, Norcini et al.
Medical Teacher 2011
Assessment in medical education has a rich history of research and development with clear practical implications.

Veterinary education is catching up ..... 

Assessment is much more than psychometrics; it involves educational design. 

Lots of exciting developments lie still ahead of us!
THANK YOU SO MUCH!


Literature