Towards a new paradigm for assessment?

Based on a paper by:

Cees van der Vleuten
School of Health Professions Education
Maastricht University
The Netherlands
A model for programmatic assessment fit for purpose


1Maastricht University, The Netherlands, 2Flinders Medical School, Australia, 3Leiden University Graduate School of Teaching, The Netherlands, 4Utrecht University of Applied Sciences, The Netherlands, 5Utrecht University, The Netherlands
Background

- « climbing » the pyramid
- A shift from an psychometric to an edumetric framework

Assessment methods

Programmatic assessment
Simple competence model


Knows

Shows how

Knows how

Does

Knows

Behaviour

Cognition

Professional authenticity
**Assessment formats used**

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

- **Knows how**
  - *Stimulus format:* (patient) scenario, simulation
  - *Response format:* menu, 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

---

**Stimulus vs Response Format**
Miller’s pyramid

Assessing Knows, knows how and shows how

**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
Assessment principle 1

- Competence is specific, not generic
## Competence is not generic

<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;3&lt;/sup&gt;</th>
<th>Oral Exam&lt;sup&gt;4&lt;/sup&gt;</th>
<th>Long Case&lt;sup&gt;5&lt;/sup&gt;</th>
<th>OSCE&lt;sup&gt;6&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>
<th>In-cognito SPs&lt;sup&gt;8&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>
<td>0.61</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>
<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>
<td>0.92</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>
<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
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
  - Be aware of (sizable) false positive and negative decisions
  - Build safeguards in examination regulations
Assessment principle 2

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

| Testing Time in Hours | MCQ\(^1\) | Case-Based Short Essay\(^2\) | PMP\(^1\) | Oral Exam\(^3\) | Long Case\(^4\) | OSCE\(^5\) | Mini CEX\(^6\) | Practice Video Assessment\(^7\) | In-cognito SPs\(^8\) |
|-----------------------|-----------|-----------------------------|----------|----------------|----------------|----------|-------------|--------------------------|----------------|-|
| 1                     | 0.62      | 0.68                        | 0.36     | 0.50           | 0.60           | 0.47     | 0.73        | 0.62                     | 0.61   |
| 2                     | 0.76      | 0.73                        | 0.53     | 0.69           | 0.75           | 0.64     | 0.84        | 0.76                     | 0.76   |
| 4                     | 0.93      | 0.84                        | 0.69     | 0.82           | 0.86           | 0.78     | 0.92        | 0.93                     | 0.92   |
| 8                     | 0.93      | 0.82                        | 0.82     | 0.90           | 0.90           | 0.88     | 0.96        | 0.93                     | 0.93   |

\(^1\)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
Reliability oral examination (Swanson, 1987)

<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>
Practical implications

- Objectivity is not the same as reliability
  - Don’t trivialize the assessment (and compromise on validity) with unnecessary objectification and standardization
  - Don’t be afraid of holistic judgment
  - Sample widely across sources of subjective influences (raters, examiners, patients)
Assessment principle 3

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

- Once reliable (meaning sufficient sampling) correlations across formats are huge
- Cognitive activities follow the task you pose in the stimulus format
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)
Assessment principle 5

- Assessment drives learning
An alternative view

Curriculum

Assessment

Teacher

Student
An alternative view

Assessment may drive learning through:

- Content
- Format
- Programming/scheduling
- Regulations
- Teacher

Assessment

Student
Empirical findings

- The relationship between assessment and learning is complex
- Summative assessment systems often drive in a negative way
- Formative feedback has dramatic impact on learning
- Learners want feedback (more than grades), but not getting it
Assessment drives learning

- For every evaluative action there is an educational reaction
- Verify and monitor the impact of assessment (evaluate the evaluation); many intended effects are not actually effective -> hidden curriculum
- No assessment without feedback!
- Use the assessment strategically to reinforce desirable learning behaviours
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 (including the ‘old’ assessment methods depending on its utility within the programme)
  - Compare assessment design with curriculum design
    - Responsible people/committee(s)
    - Use an overarching structure
    - Involve your stakeholders
    - Implement, monitor and change (assessment programmes ‘wear out’)
Assessing Does

Assessment principles

1. A feasible sample is required to achieve reliable inferences
2. Bias is an inherent characteristic of expert judgment
3. The validity lies in the users of the instruments, more than in the instruments
4. Formative and summative functions are typically combined
5. Qualitative, narrative information carries a lot of weight
6. Summative decisions can be rigorous by using non-psychometric qualitative research procedures