A Programmatic Approach to Assessment

ADEA/ADEE Meeting: Shaping the Future of Dental Education

8-9 May, London, UK

Cees van der Vleuten

Maastricht University

The Netherlands

www.maastrichtuniversity.nl/she

www.ceesvandervleuten.com

With gratitude to the support ofLiftUPP!


the personal journey

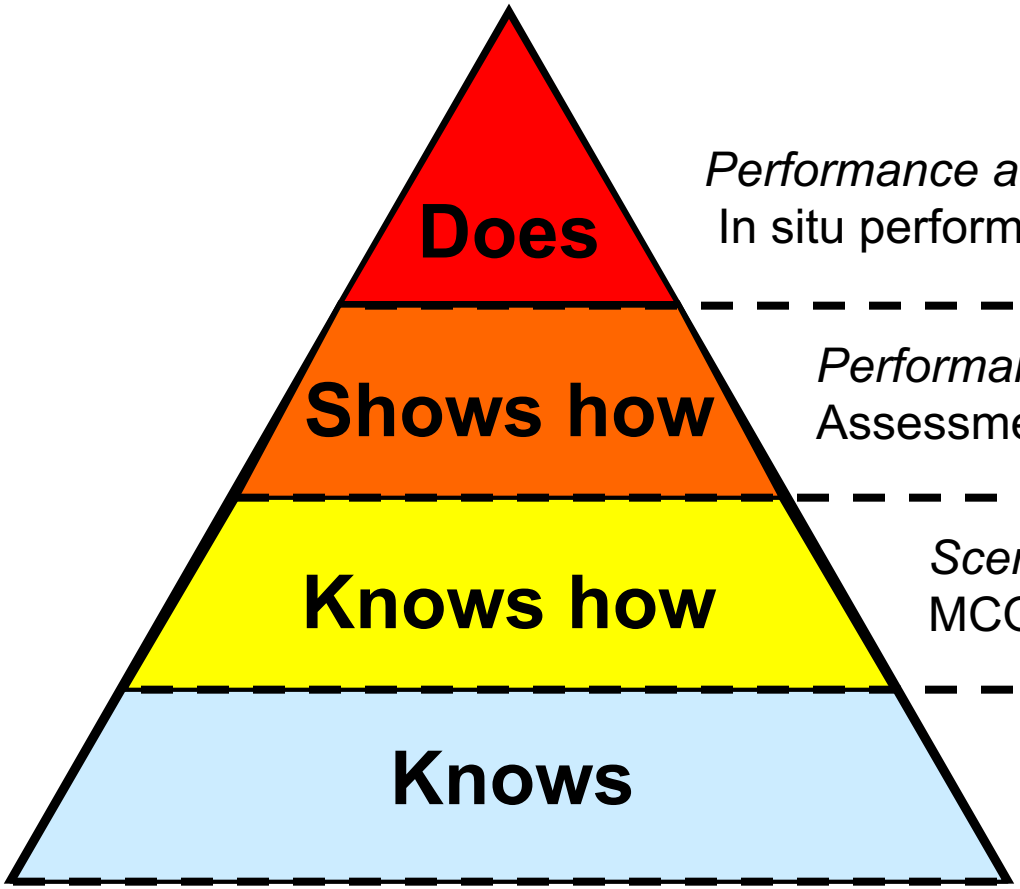
Overview

- From practice to research
- From research to theory
- From theory to practice
- Conclusions

The Toolbox

- MCQ, MEQ, OEQ, SIMP, Write-ins, Key Feature, Progress test, PMP, SCT, Viva, Long case, Short case, OSCE, OSPE, DOCEE, SP-based test, Video assessment, MSF, Mini-CEX, DOPS, assessment center, self-assessment, peer assessment, incognito SPs, portfolio…….

The way we climbed……



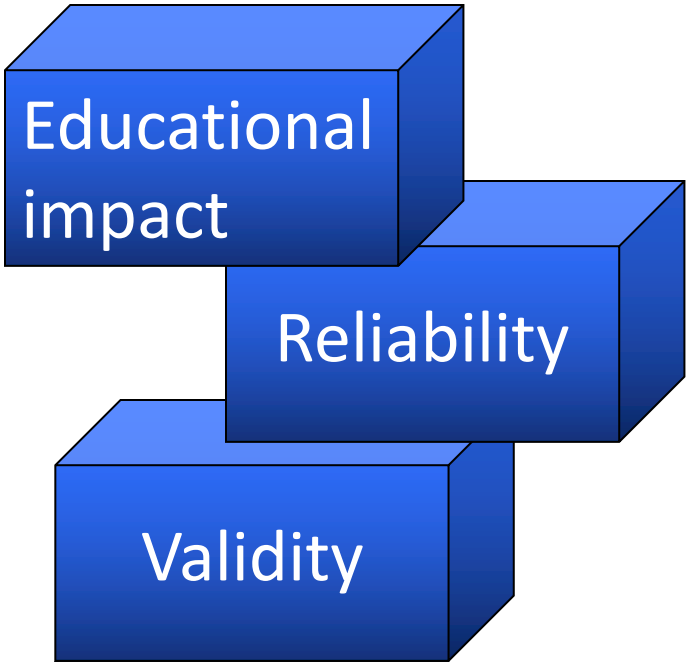
Performance assessment in vivo:
In situ performance assessment, 360°, Peer assesme

Performance assessment in vitro:
Assessment centers, OSCE…..

Scenario or case-based assessment:
MCQ, write-ins, oral…..

Fact-oriented assessment:
MCQ, write-ins, oral…..

Characteristics of instruments



Validity: what are we assessing?

- Curricula have changed from an input orientation to an output orientation
- We went from haphazard learning to integrated learning objectives, to end objectives, and now to (generic) competencies
- We went from teacher oriented programs to learning oriented, self-directed programs

Competency-frameworks

CanMeds

- Medical expert
- Communicator
- Collaborator
- Manager
- Health advocate
- Scholar
- Professional

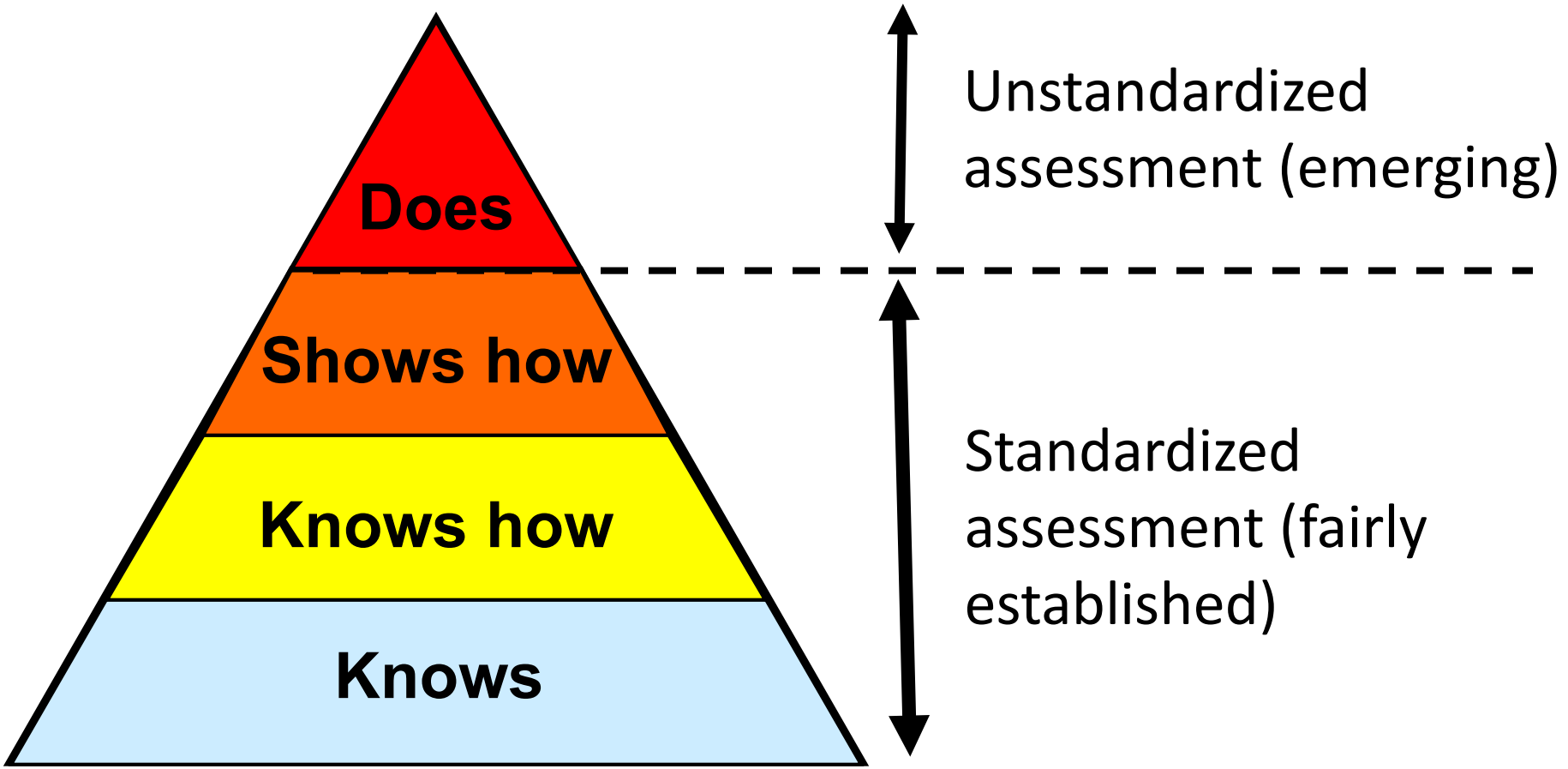
ACGME

- Medical knowledge
- Patient care
- Practice-based learning & improvement
- Interpersonal and communication skills
- Professionalism
- Systems-based practice

GMC

- Good clinical care
- Relationships with patients and families
- Working with colleagues
- Managing the workplace
- Social responsibility and accountability
- Professionalism

Validity: what are we assessing?



Messages from validity research

- There is no magic bullet; we need a mixture of methods to cover the competency pyramid
- We need BOTH standardized and non-standardized assessment methods
- For standardized assessment quality control around test development and administration is vital
- For unstandardized assessment the users (the people) are vital.

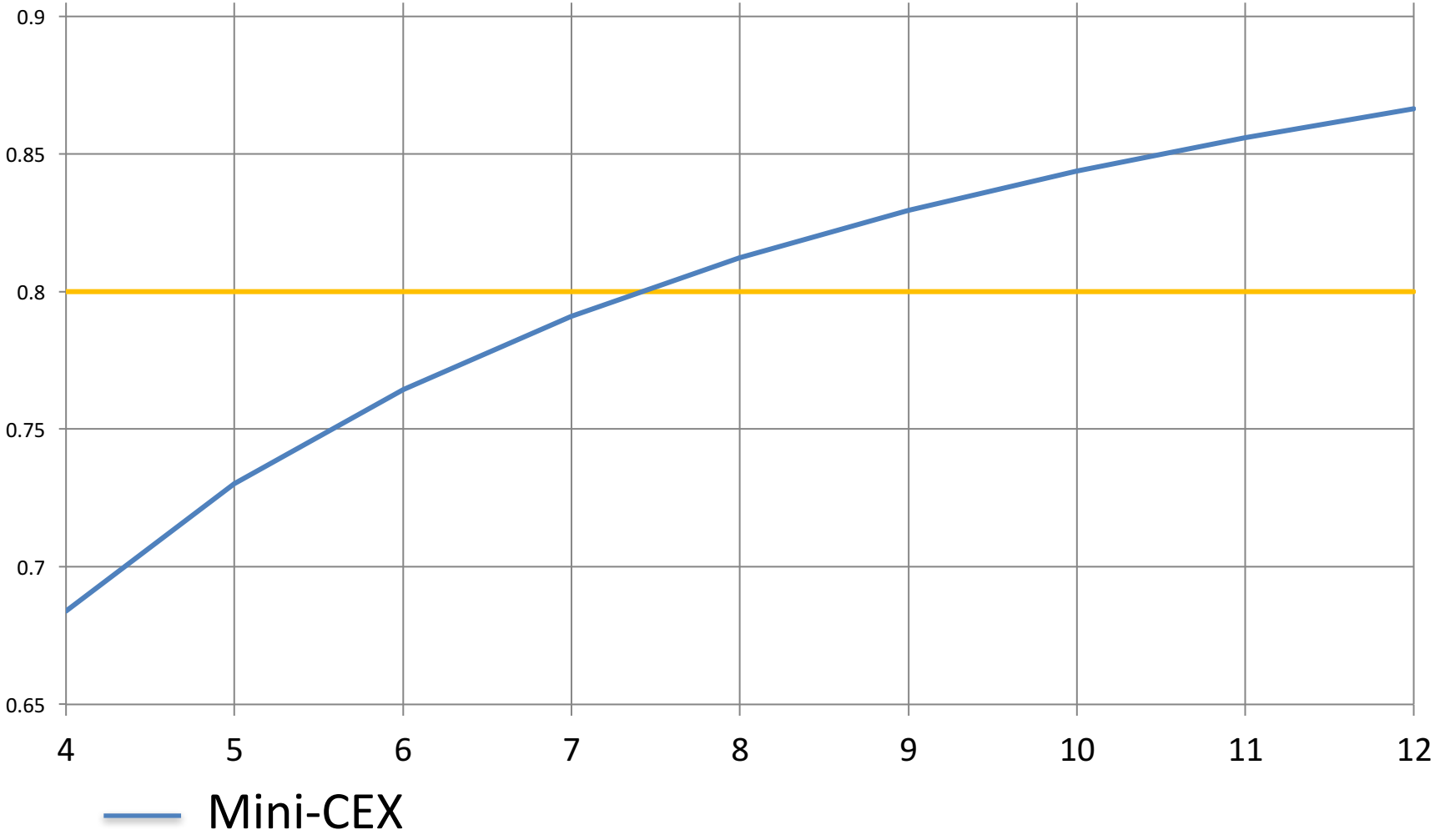
Method reliability as a function of testing time

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 Assess-ment[7]	In-cognito SPs[8]
1	0.62	0.68	0.36	0.50	0.60	0.54	0.73	0.62	0.61
2	0.77	0.81	0.53	0.67	0.75	0.70	0.84	0.77	0.76
4	0.87	0.89	0.69	0.80	0.86	0.82	0.92	0.87	0.86
8	0.93	0.94	0.82	0.89	0.92	0.90	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]Van der Vleuten, 1988
[6]Norcini et al., 1999
[7]Ram et al., 1999
[8]Gorter, 2002

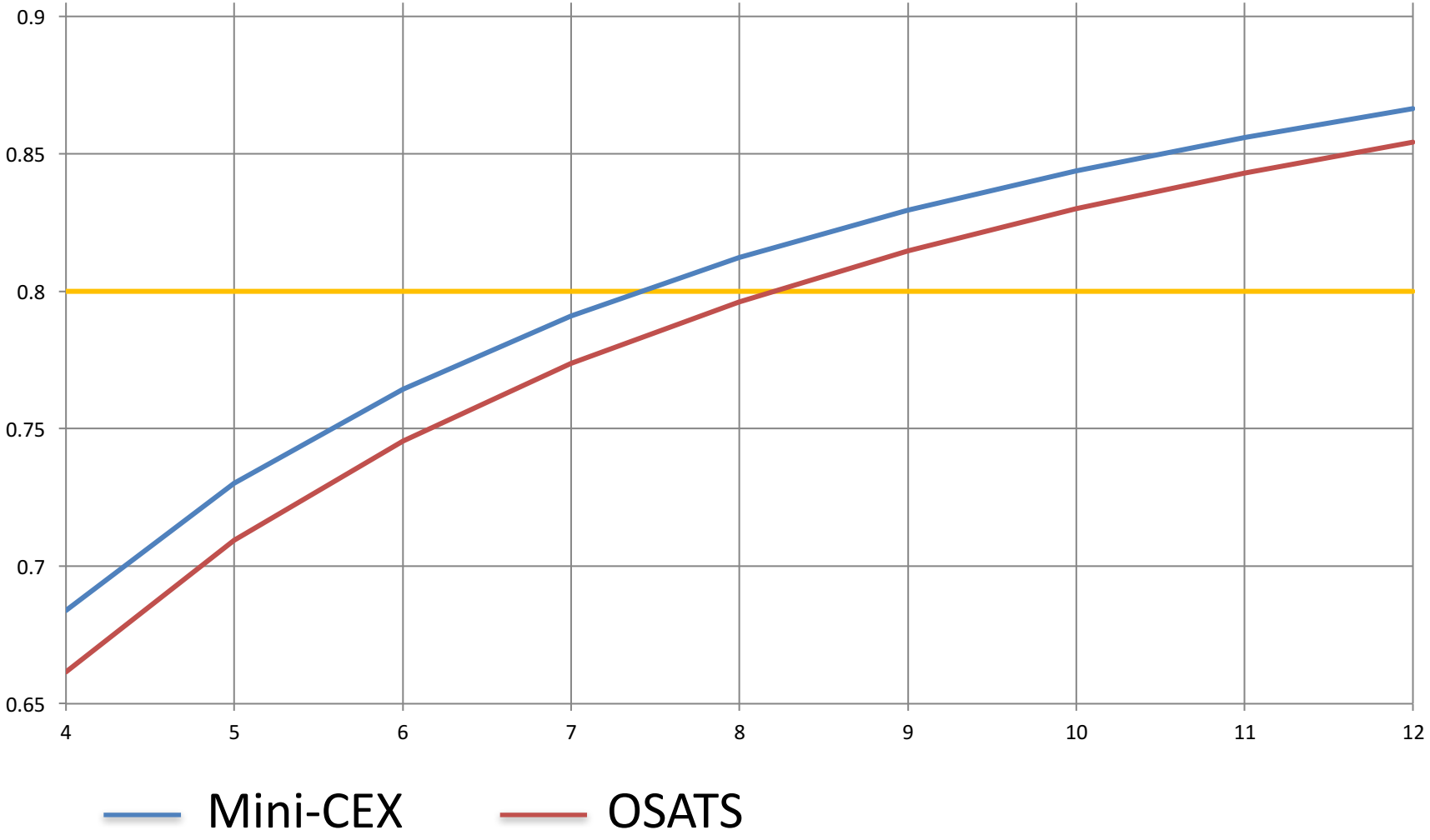
Reliability as a function of sample size

(Moonen et al., 2013)



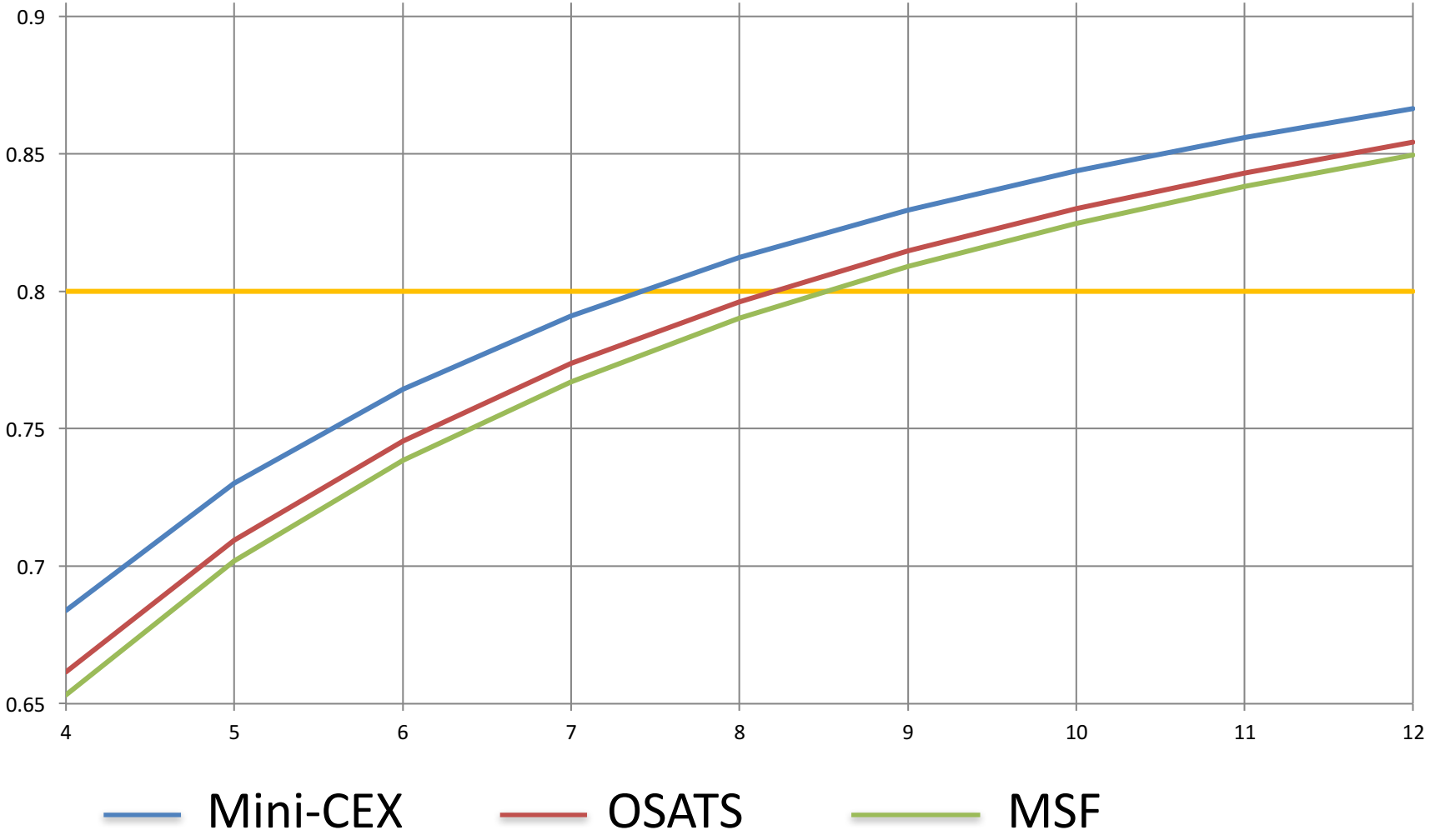
Reliability as a function of sample size

(Moonen et al., 2013)



Reliability as a function of sample size

(Moonen et al., 2013)



Effect of aggregation across methods

(Moonen et al., 2013)

Method	Sample needed when used as stand-alone	Sample needed when used as a composite
Mini-CEX	8	5
OSATS	9	6
MSF	9	2

Resultaten Betrouwbaarheid

Per instrument

Alle jaren:

8 KPB, 9 OSATS, 9 MSF

Eerste jaar:

6 KBP, 6 OSATS, 6 MSF

Gezamenlijk

Alle jaren:

7 KPB, 8 OSATS, 1 MSF of

5 KPB, 6 OSATS, 2 MSF

Eerste jaar:

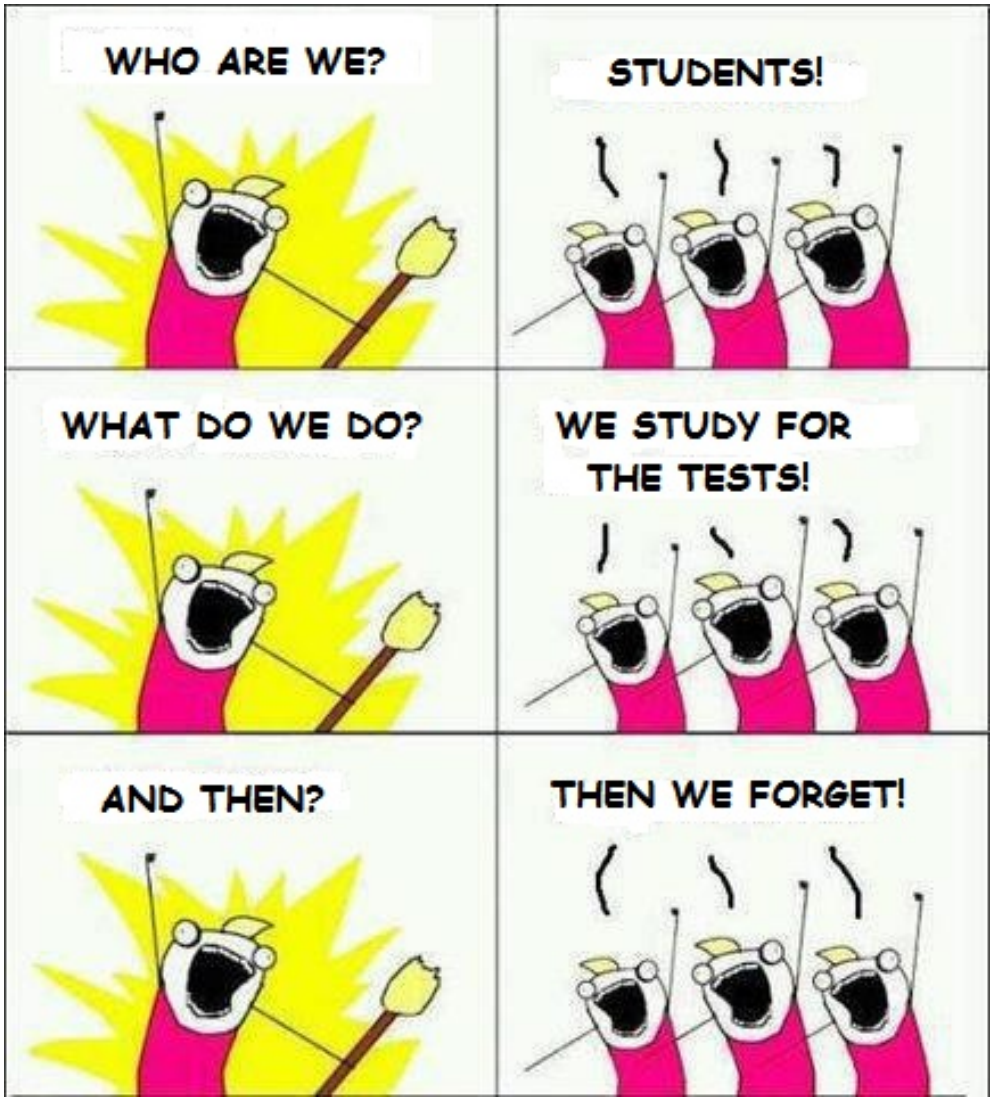
5 KBP, 6 OSATS, 1 MSF

Messages from reliability research

- Acceptable reliability is only achieved with large samples of test elements (contexts, cases) and assessors
- No method is inherently better than any other (that includes the new ones!)
- Objectivity is NOT equal to reliability
- Many subjective judgments are pretty reproducible/reliable.

Educational impact: How does assessment drive learning?

- Relationship is complex (cf. Cilliers, 2011, 2012)
- But impact is often very negative
  - Poor learning styles
  - Grade culture (grade hunting, competitiveness)
  - Grade inflation (e.g. in the workplace)
- A lot of REDUCTIONISM!
  - Little feedback (grade is poorest form of feedback one can get)
  - Non-alignment with curricular goals
  - Non-meaningful aggregation of assessment information
  - Few longitudinal elements
  - Tick-box exercises (OSCEs, logbooks, work-based assessment).


WHO ARE WE?
STUDENTS!
WHAT DO WE DO?
WE STUDY FOR THE TESTS!
AND THEN?
THEN WE FORGET!

- All learners construct knowledge from an inner scaffolding of their individual and social experiences, emotions, will, aptitudes, beliefs, values, self-awareness, purpose, and more . . . if you are learning ….., what you understand is determined by how you understand things, who you are, and what you already know.

Peter Senge, Director of the Center for Organizational Learning at MIT (as cited in van Ryn et al., 2014)

Messages learning impact research

- No assessment without (meaningful) feedback
- Narrative feedback has a lot more impact on complex skills than scores
- Provision of feedback is not enough (feedback is a dialogue)
- Longitudinal assessment is needed.

Overview

- From practice to research
- From research to theory
- From theory to practice
- Conclusions

Limitations of the single-method approach

- No single method can do it all
- Each individual method has (significant) limitations
- Each single method is a considerable *compromise* on reliability, validity, educational impact

Implications

- **Validity:** a multitude of methods needed
- **Reliability:** a lot of (combined) information is needed
- **Learning impact:** assessment should provide (longitudinal) meaningful information for learning

Programmatic assessment

Programmatic assessment

- A curriculum is a good metaphor;

  in a program of assessment:

  – Elements are planned, arranged, coordinated
  – Is systematically evaluated and reformed
- But how? (the literature provides extremely little support!)

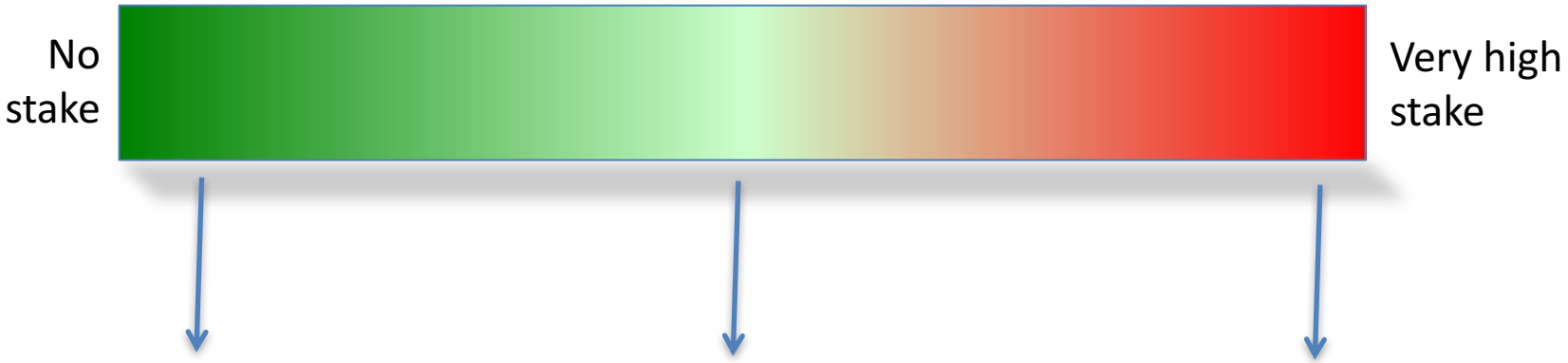
Programmatic assessment

- Dijkstra et al 2012: 73 generic guidelines
- To be done:
  - Further validation
  - A feasible (self-assessment) instrument
- ASPIRE assessment criteria

Building blocks for programmatic assessment 1

- Every assessment is but one data point (Δ)
- Every data point is optimized for learning
  - Information rich (quantitative, qualitative)
  - Meaningful
  - Variation in format
- Summative versus formative is replaced by continuum of stakes (stakes)
- N data points are proportionally related to the stakes of the decision to be taken.

Continuum of stakes, number of data point and their function



One Data point:

- Focused on information
- Feedback oriented
- Not decision oriented

Intermediate progress decisions:

- More data points needed
- Focus on diagnosis, remediation, prediction

Final decisions on promotion or selection:

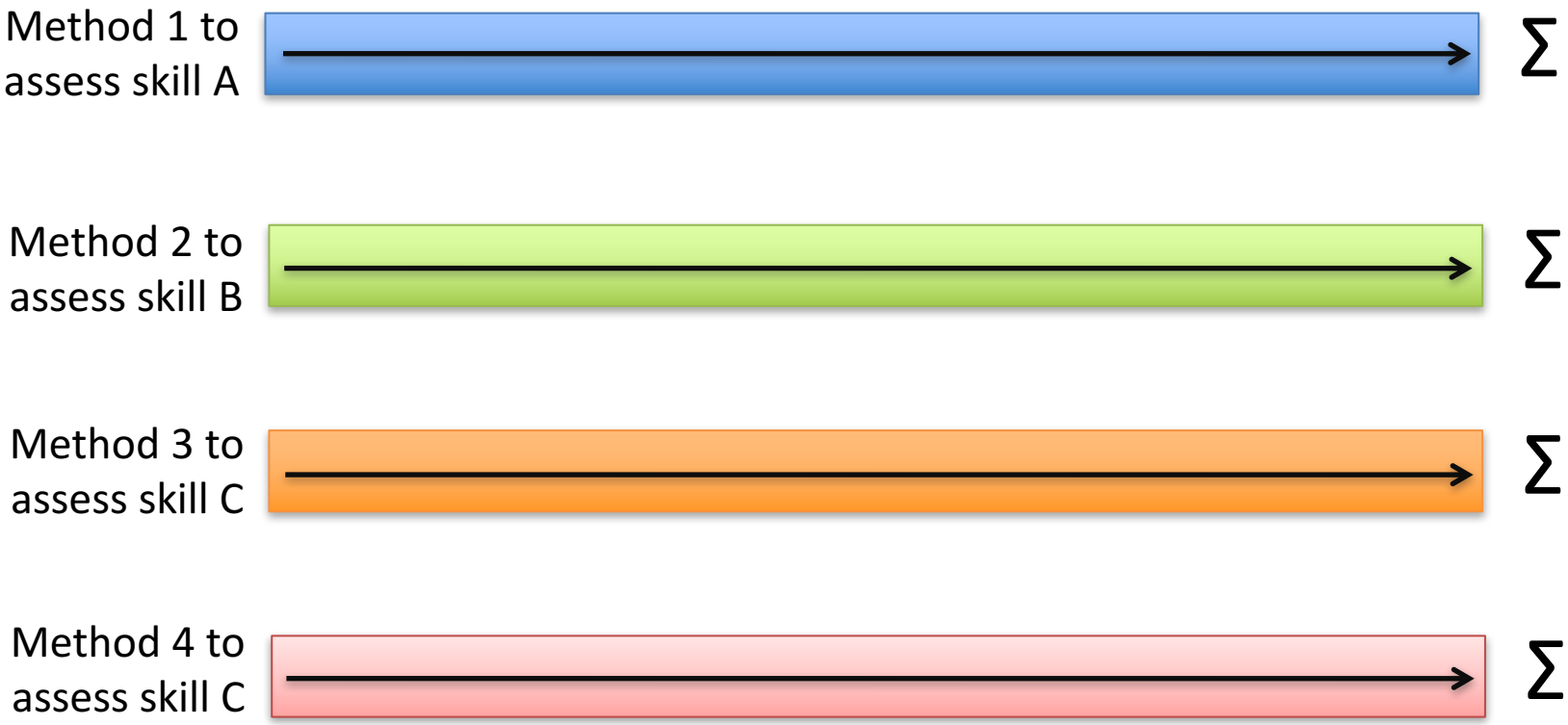
- Many data points needed
- Focused on a (non-surprising) heavy decision

Assessment information as pixels

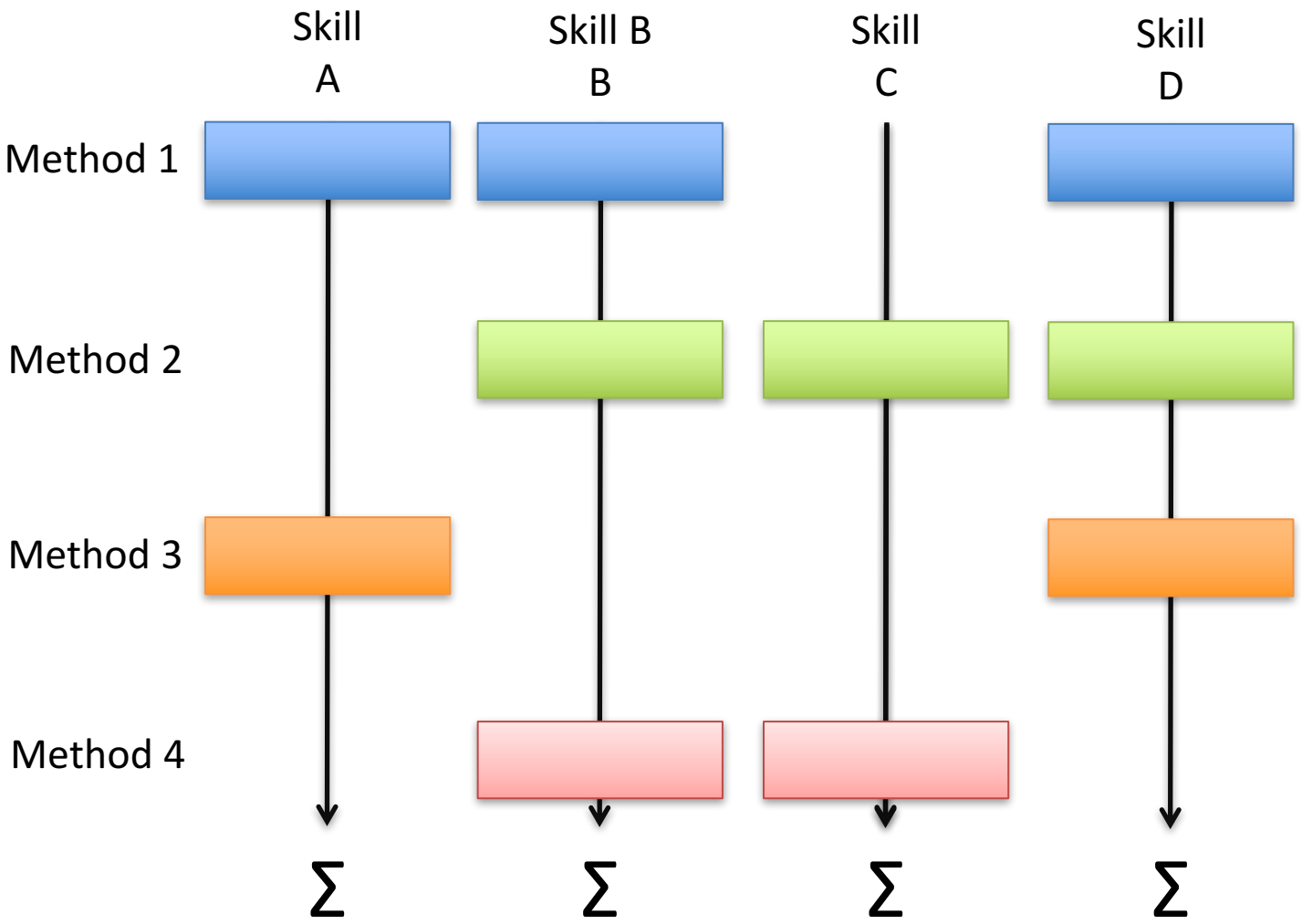




Classical approach to aggregation



More meaningful aggregation



2012; 34: 205–214 MEDICAL TEACHER

A model for programmatic assessment fit for purpose

C. P. M. VAN DER VLEUTEN[1], L. W. T. SCHUWIRTH[2], E. W. DRIESSEN[1], J. DIJKSTRA[1], D. TIGELAAR[3], L. K. J. BAARTMAN[4] & J. VAN TARTWIJK[5]

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

Abstract

We propose a model for programmatic assessment in action, which simultaneously optimises assessment for learning and assessment for decision making about learner progress. This model is based on a set of assessment principles that are interpreted from empirical research. It specifies cycles of training, assessment and learner support activities that are complemented by intermediate and final moments of evaluation on aggregated assessment data points. A key principle is that individual data points are maximised for learning and feedback value, whereas high-stake decisions are based on the aggregation of many data points. Expert judgement plays an important role in the programme. Fundamental is the notion of sampling and bias reduction to deal with the inevitable subjectivity of this type of judgement. Bias reduction is further sought in procedural assessment strategies derived from criteria for qualitative research. We discuss a number of challenges and opportunities around the proposed model. One of its prime virtues is that it enables assessment to move, beyond the dominant psychometric discourse with its focus on individual instruments, towards a systems approach to assessment design underpinned by empirically grounded theory.

2014, 1–6, Early Online MEDICAL TEACHER

TWELVE TIPS

12 Tips for programmatic assessment

C.P.M. VAN DER VLEUTEN[1], L.W.T. SCHUWIRTH[2], E.W. DRIESSEN[1], M.J.B. GOVAERTS[1] & S. HEENEMAN[1]

[1]Maastricht University, Maastricht, The Netherlands, [2]Flinders University, Adelaide, Australia

Abstract

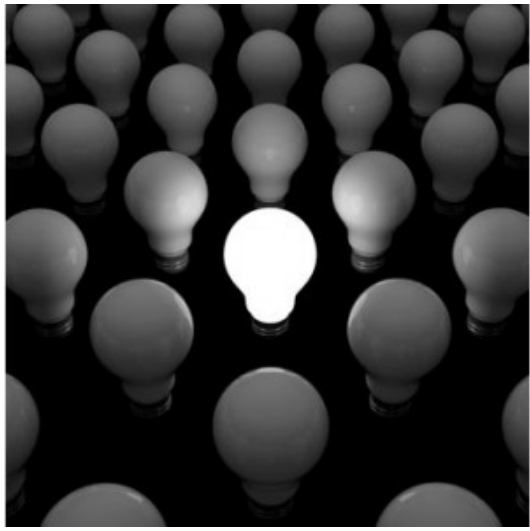
Programmatic assessment is an integral approach to the design of an assessment program with the intent to optimise its learning function, its decision-making function and its curriculum quality-assurance function. Individual methods of assessment, purposefully chosen for their alignment with the curriculum outcomes and their information value for the learner, the teacher and the organisation, are seen as individual data points. The information value of these individual data points is maximised by giving feedback to the learner. There is a decoupling of assessment moment and decision moment. Intermediate and high-stakes decisions are based on multiple data points after a meaningful aggregation of information and supported by rigorous organisational procedures to ensure their dependability. Self-regulation of learning, through analysis of the assessment information and the attainment of the ensuing learning goals, is scaffolded by a mentoring system. Programmatic assessment-for-learning can be applied to any part of the training continuum, provided that the underlying learning conception is constructivist. This paper provides concrete recommendations for implementation of programmatic assessment.

Overview

- From practice to research
- From research to theory
- From theory to practice
- Conclusions

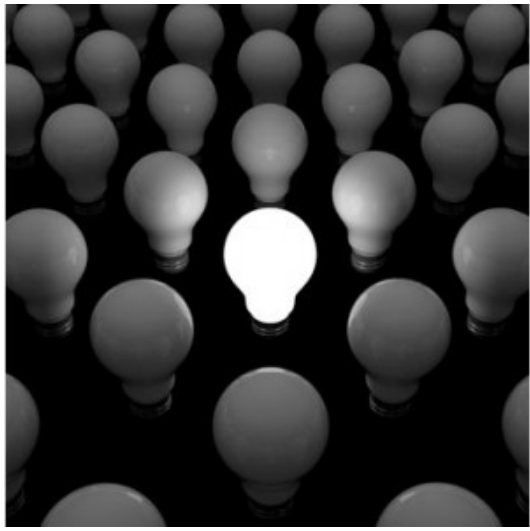
From theory back to practice

- Existing best practices:
  - Veterinary education Utrecht
  - Cleveland Learner Clinic, Cleveland, Ohio
  - Dutch specialty training in General Practice
  - McMaster Modular Assessment Program in Emergency Medicine
  - Graduate entry program Maastricht



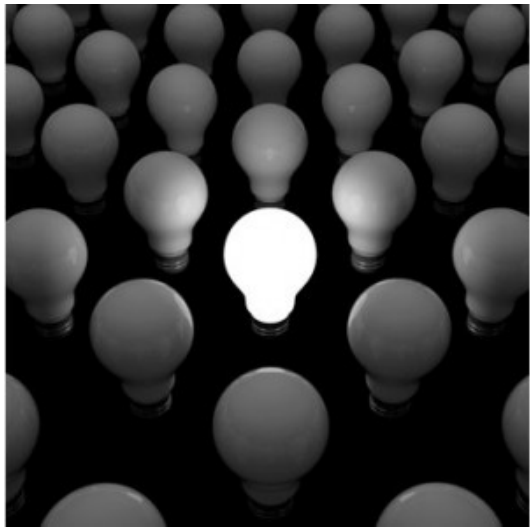
From theory back to practice

- Existing best practices:
  - Veterinary education Utrecht
  - Cleveland Learner Clinic, Cleveland, Ohio
  - Dutch specialty training in General Practice
  - McMaster Modular Assessment Program in Emergency Medicine
  - Graduate entry program Maastricht



From theory back to practice

- Existing best practices:
  - Veterinary education Utrecht
  - Cleveland Learner Clinic, Cleveland, Ohio
  - Dutch specialty training in General Practice
  - McMaster Modular Assessment Program in Emergency Medicine
  - Graduate entry program Maastricht

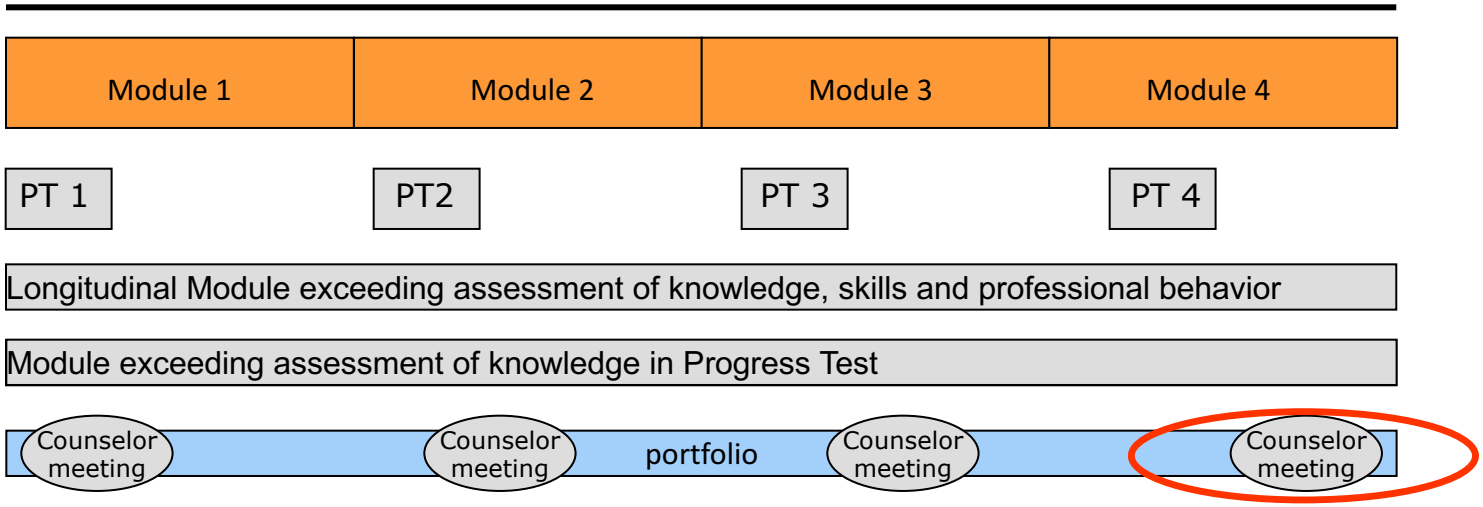


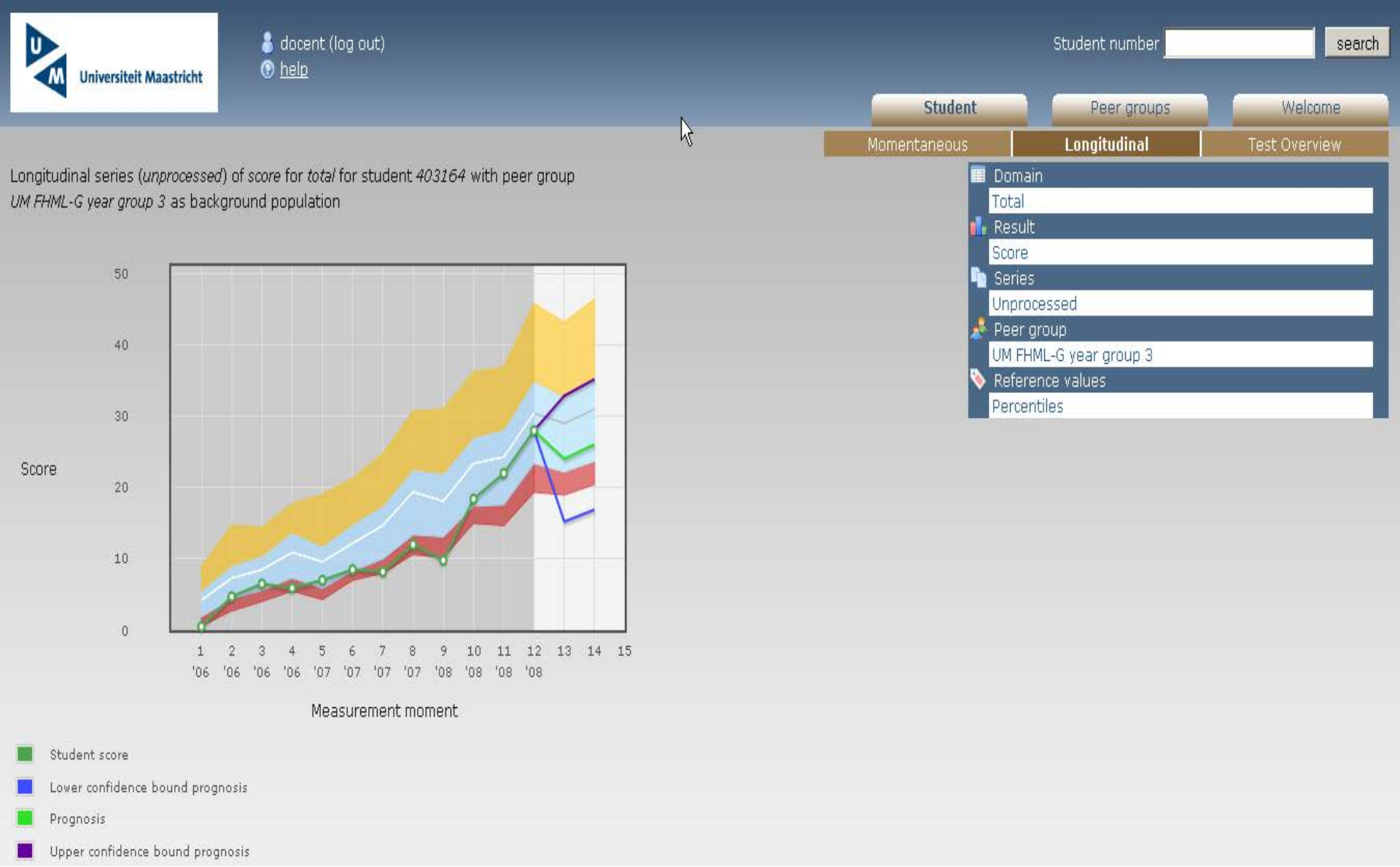
Physician-clinical investigator program

- 4 year graduate entry program
- Competency-based (Canmeds) with emphasis on research
- PBL program
  - Year 1: classic PBL
  - Year 2: real patient PBL
  - Year 3: clerkship rotations
  - Year 4: participation in research and health care
- High expectations of students: in terms of motivation, promotion of excellence, self-directedness

The assessment program

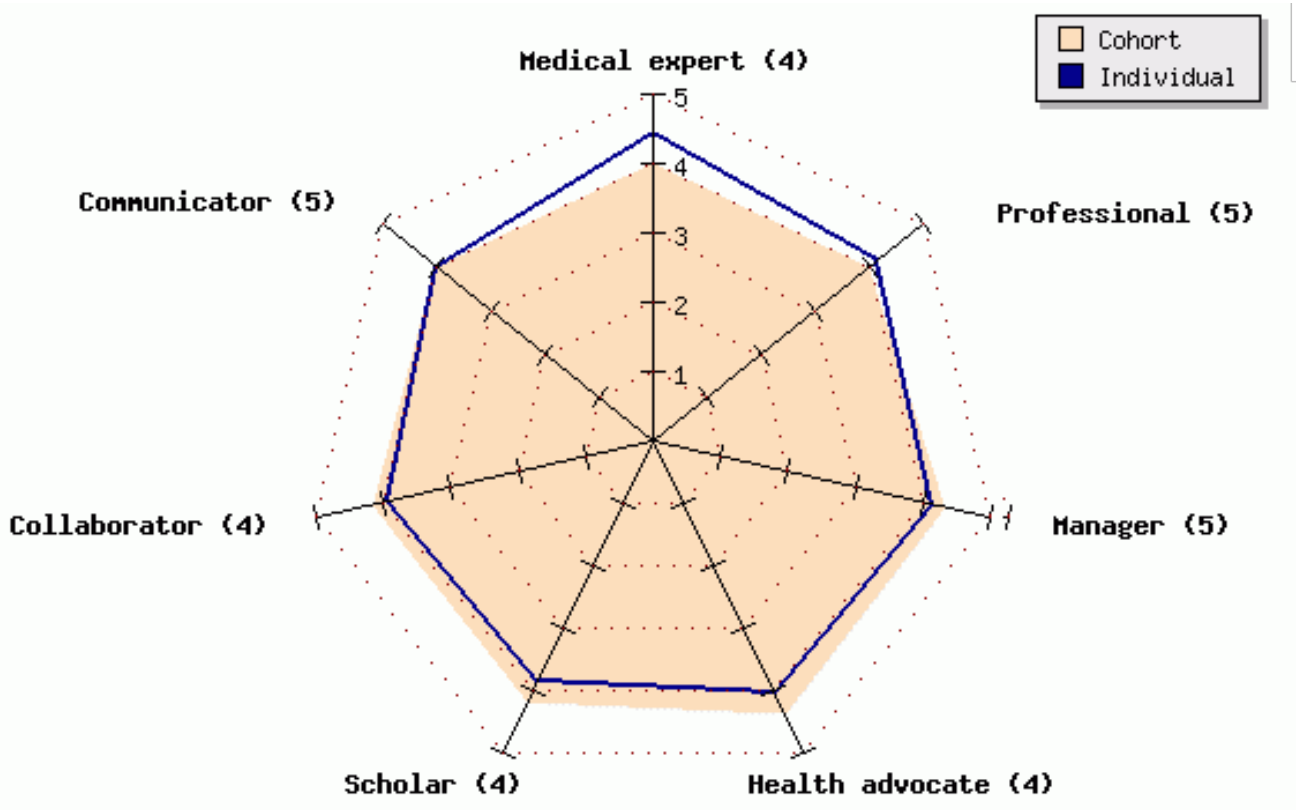
- Assessment in Modules: assignments, presentations, end-examination, etc.
- Longitudinal assessment: assignments, reviews, projects, progress tests, evaluation of professional behavior, etc.
- All assessment is informative and low stake formative
- The portfolio is central instrument





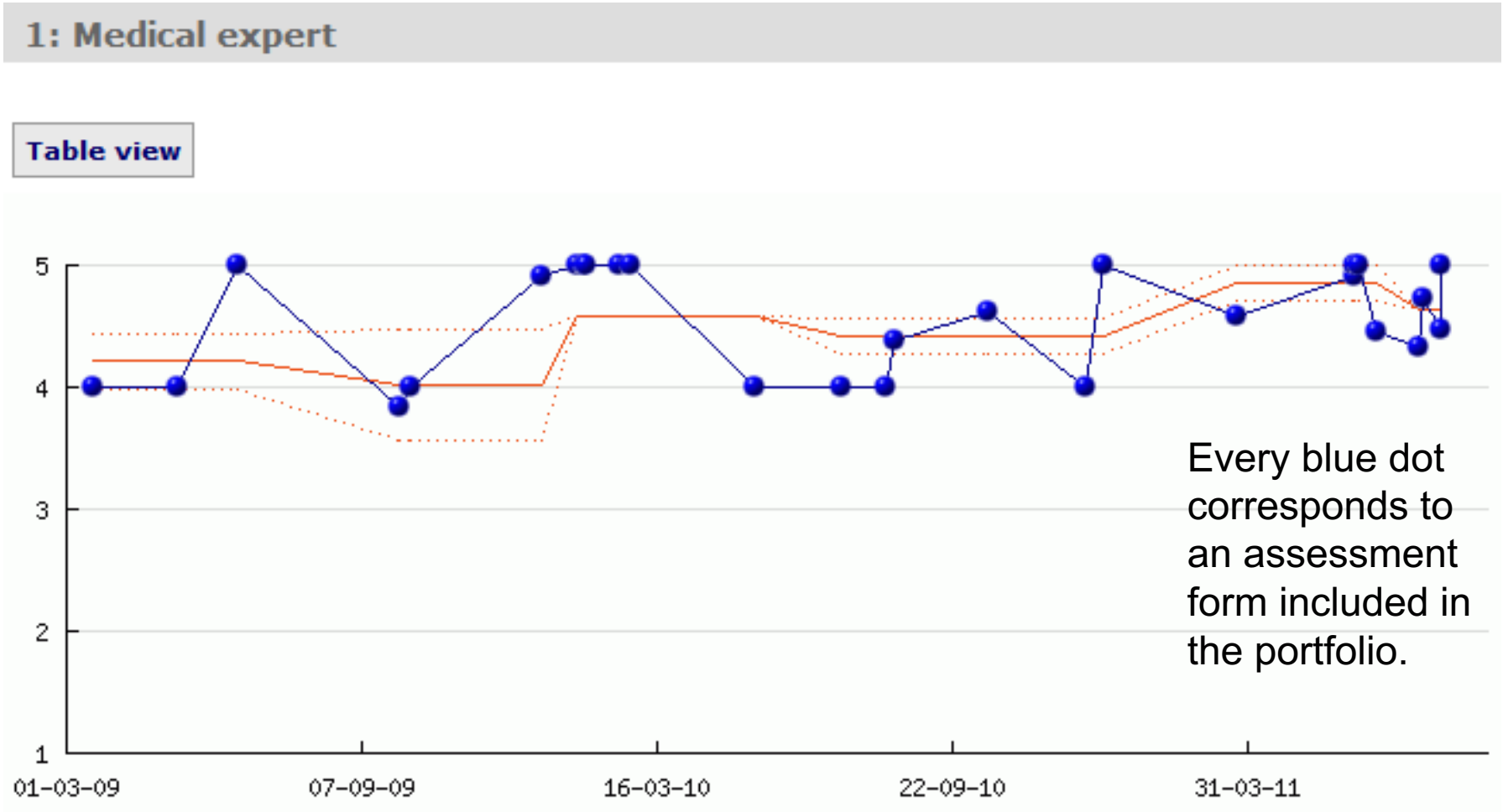
Longitudinal total test scores across 12 measurement moments and predicted future performance

Maastricht Electronic portfolio (ePass)



Comparison between the score of the student and the average score of his/her peers.

Maastricht Electronic portfolio (ePass)



Narrative feedback

Feedbacktype: all

Competency: all

Date	Feedbacktype	Competency	Narrative feedback	Form
06-11-2013	Improvement	General	don't repeat too much, no irrelevant details Conclusion: antenatal care in pregnancy may be done by a midwife and delivery will be done by a gynecologist, I revise	Mini-CEX-N
06-11-2013	Strength	General	included all information.	Mini-CEX-N
06-11-2013	Improvement	General	don't repeat too much, no irrelevant details. Conclusion: antenatal care in pregnancy may be done by a midwife, delivery will be done by a gynecologist, I revise.	Mini-CEX-N
06-11-2013	Strength	General	included all info.	Mini-CEX-N
18-10-2013	Improvement	General	more communication with the patient, in this case difficult because of language barrier more communication with supervisor	OSATS

Coaching by counselors

- Coaching is essential for successful use of reflective learning skills
- Counselor gives advice/comments (whether asked or not)
- He/she counsels if choices have to be made
- He/she guards and discusses study progress and development of competencies

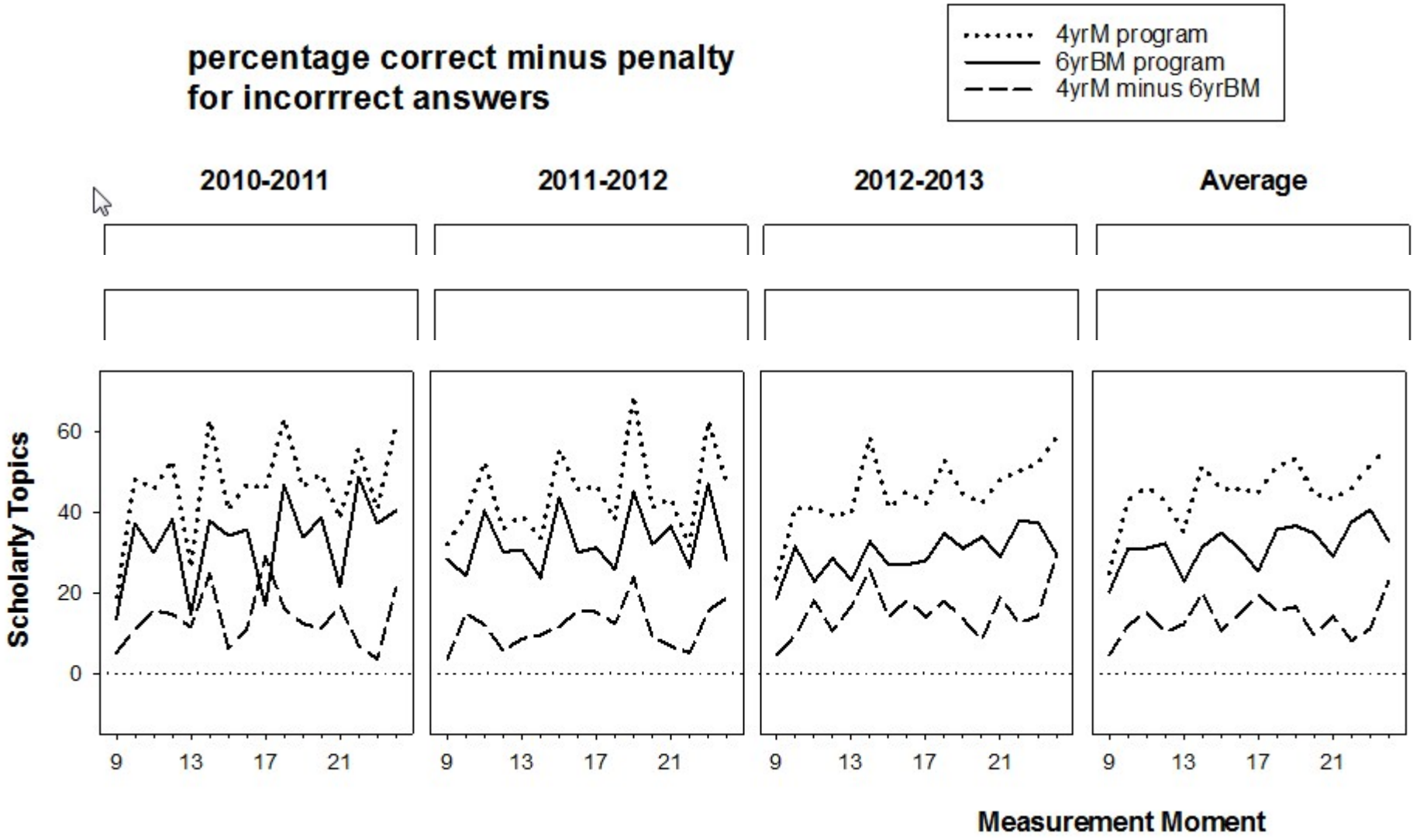
"'LUNCH'? WELL, YES--BUT WHAT ARE YOUR *LONG*-TERM GOALS?"

Decision-making by committee

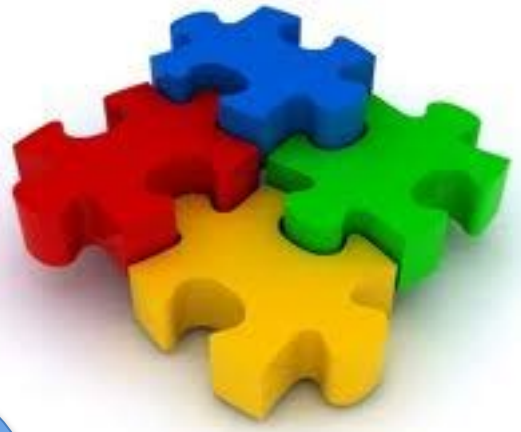
- Committee of counselors and externals
- Decision is based on portfolio information & counselor recommendation, competency standards
- Deliberation is proportional to clarity of information
- Decisions are justified when needed; remediation recommendation may be provided

Strategy to establish trustworthiness	Criteria	Potential Assessment Strategy (sample)
Credibility	Prolonged engagement	Training of examiners
	Triangulation	Tailored volume of expert judgment based on certainty of information
	Peer examination	Benchmarking examiners
	Member checking	Incorporate learner view
	Structural coherence	Scrutiny of committee inconsistencies
Transferability	Time sampling	Judgment based on broad sample of data points
	Thick description	Justify decisions
Dependability	Stepwise replication	Use multiple assessors who have credibility
Confirmability	Audit	Give learners the possibility to appeal to the assessment decision

Progress test embedded in programmatic assesssment – *use of information and feedback to selfdirect learning*



Overview



- From practice to research
- From research to theory
- From theory to practice
- Conclusions

Conclusions 1: The way forward

- We have to stop thinking in terms of individual assessment methods
- A systematic and programmatic approach is needed, longitudinally oriented
- Every method of assessment may be functional (old and new; standardized and unstandardized)
- Professional judgment is imperative (similar to clinical practice)
- Subjectivity is dealt with through sampling and procedural bias reduction methods (not with standardization or objectification).

Conclusions 2: The way forward

- The programmatic approach to assessment optimizes:
  - The learning function (through information richness)
  - The pass/fail decision function (through the combination of rich information)

Further reading:
www.ceesvandervleuten.com