



Naar een toekomst van programmatische toetsing

SYMPOSIUM LEREN VAN TOETSEN

Vlissingen

3 juni 2016

Cees van der Vleuten

Universiteit Maastricht

www.ceesvandervleuten.com

VAN EXAMENS KRIJG JE UITSLAG

Loesje

EXAMENS

**VIER JAAR ZWOEGEN
OM A**

B

C

OF D

IN TE VULLEN

Loesje

EXAMENS

Postbus 1045 6801 BA Arnhem www.loesje.nl

DE TIJD VAN KORTE ROKJES EN ANTWOORDEN OP JE BOVENBEEN

Loesje

HERKANSINGEN

Postbus 1045

6801 BA Arnhem

www.loesje.nl

**MAAR MENEER
KUNT U ONS
NIET GEWOON
EEN ZEVEN GEVEN
DAT SCHEELT ONS
BEIDE EEN HOOP WERK**

Loesje

**WAT IS
EEN EINDEXAMEN
ANDERS DAN**

**WETEN
ZWETEN
EN VERGETEN**

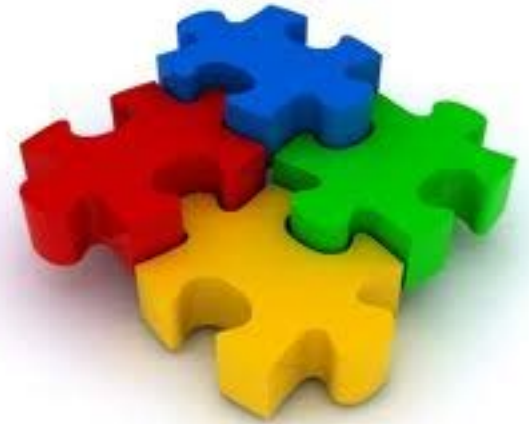
Loesje

POSTBUS 1045
6801 DA ARNHEM



Overzicht

- Van praktijk naar onderzoek
- Van onderzoek naar theorie
- Van theorie naar praktijk
- Conclusies



Disclaimer!

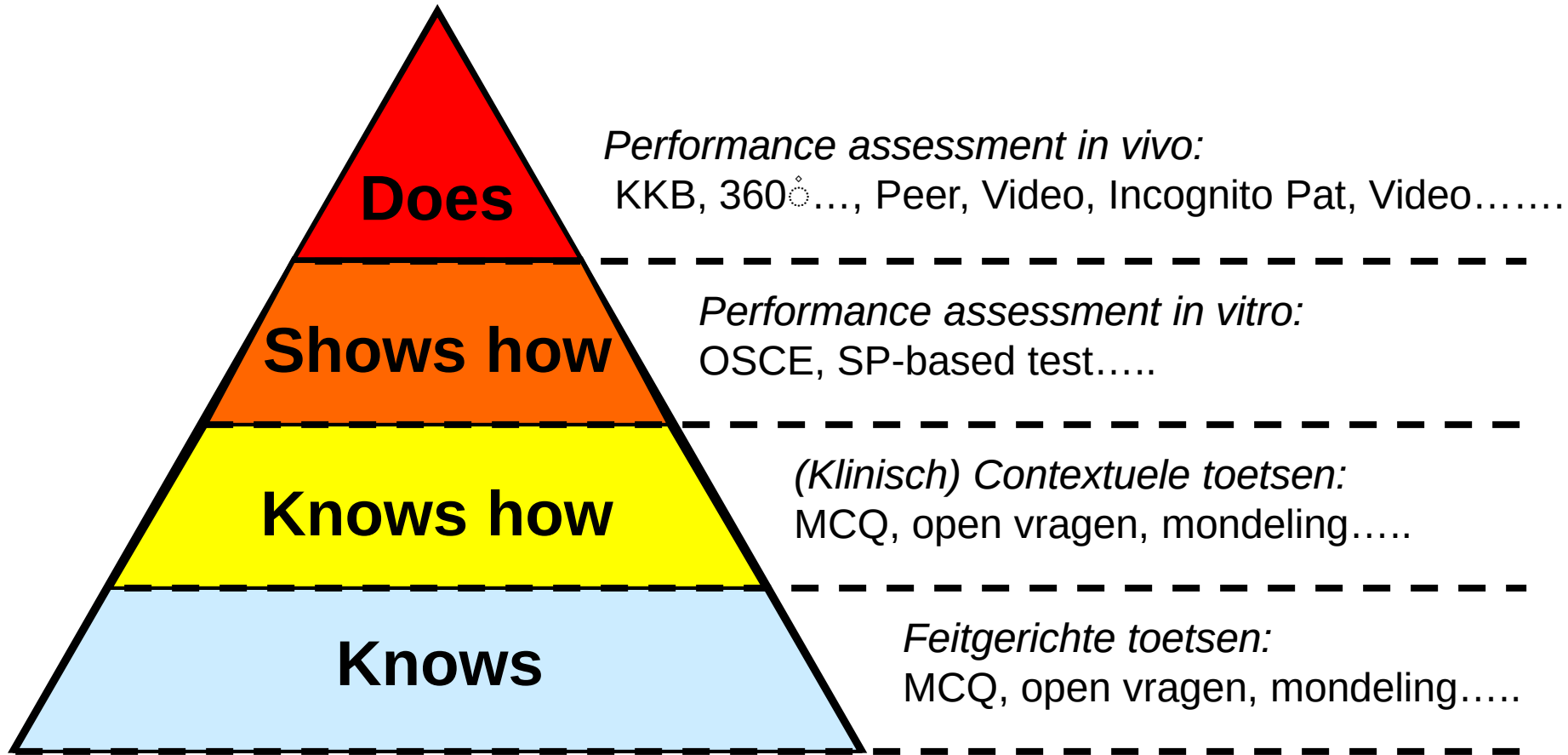


Toetsvormen

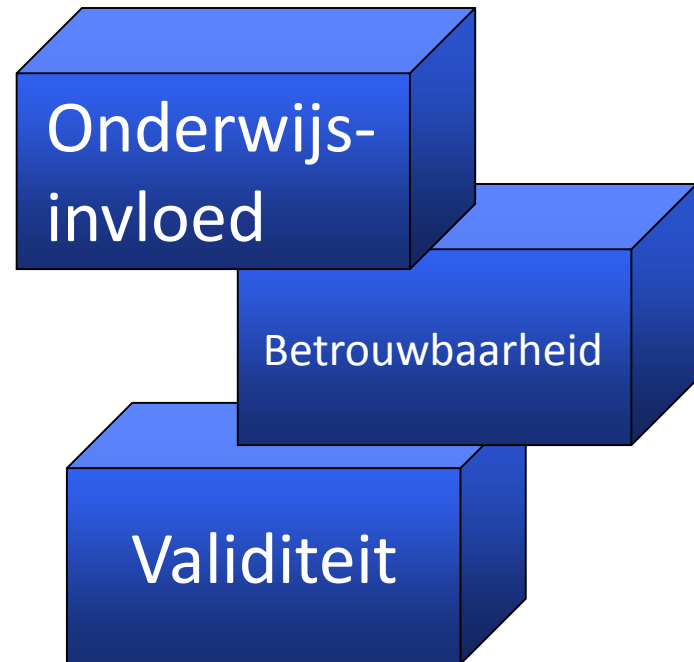
- MCQ, MEQ, OEQ, SIMP, Write-ins, Key Feature, PMP, SCT, Patiënt examen, OSCE, OSPE, DOCEE, SP-based test, Video assessment, MSF, Mini-CEX, DOPS, assessment center, self-assessment, peer assessment, incognito SPs.....



De beklimming.....



Karakteristieken van instrumenten



Validiteit: wat beoordelen we?

- Curricula zijn veranderd van input sturing naar output sturing
- Van disciplines naar leerdoelen, eindtermen, en nu competenties
- Van docentsturing naar zelfsturing

Competency-frameworks



CanMeds

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



ACGME

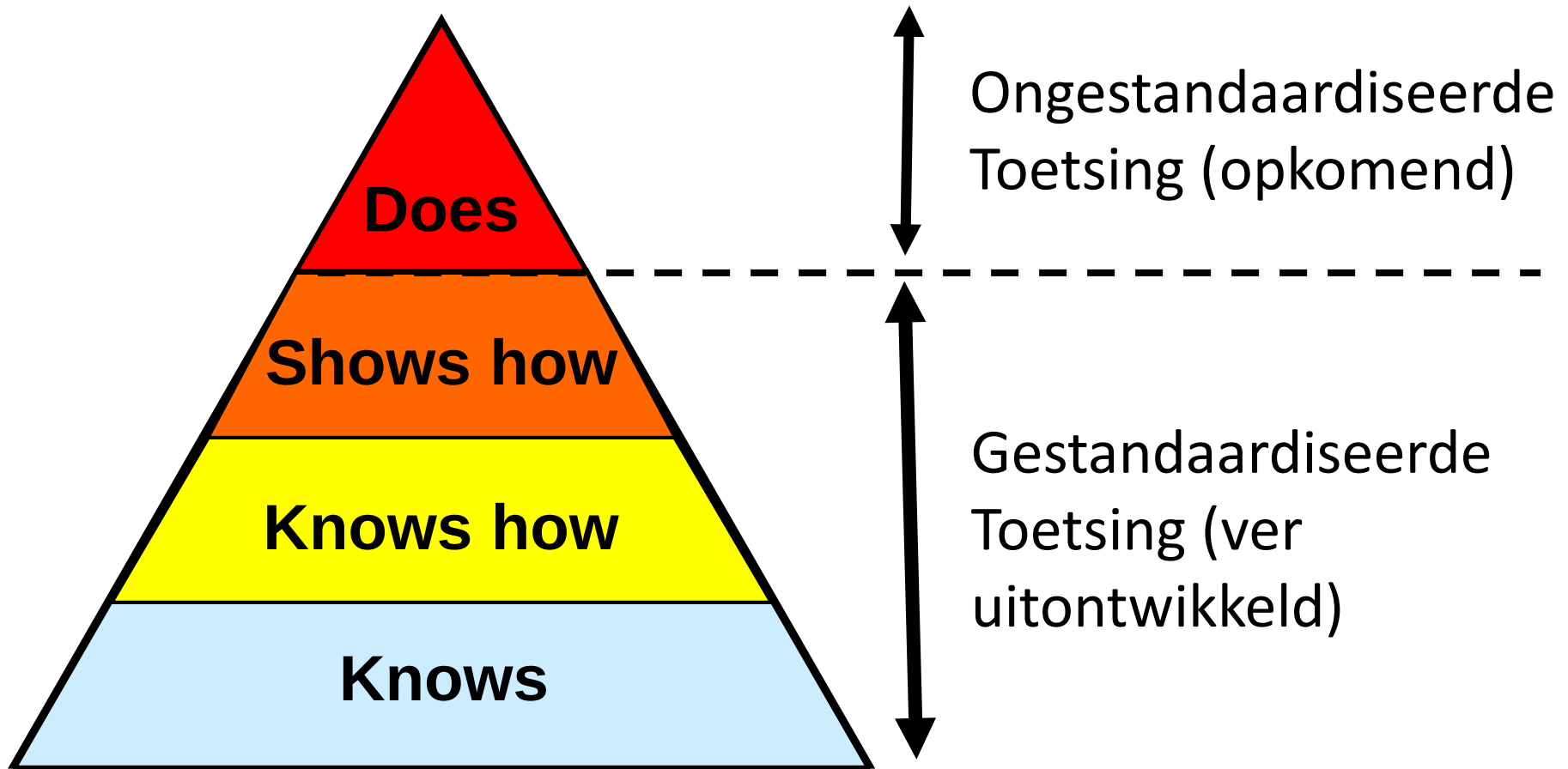
- Medical knowledge
- Patient care
- Practice-based learning & improvement
- Interpersonal and communication skills
- Professionalism
- Systems-based practice



GMC

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

Validiteit: wat beoordelen we?



Inzichten over validiteit

- We hebben een veelheid aan toetsmethoden nodig om de hele competentiepiramide te dekken
- Nodig zijn gestandaardiseerde en ongestandaardiseerde toetsmethodieken
- Kwaliteitscontrole van instrumenten is essentieel voor gestandaardiseerde beoordeling
- De gebruikers (de mensen) zijn essentieel bij ongestandaardiseerde beoordeling.



Betrouwbaarheid: hoe precies beoordelen we?

Toets-Tijd in Hours	MCQ ¹	Korte casus gerichte Essay ²	Simulaties ¹	Monde-ling ³	Patient examen ⁴	OSCE ⁵	KKB ⁶	Praktijk Video Observatie 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	0.76	0.73	0.53	0.69	0.75	0.69	0.84	0.76	0.76
4	0.93	0.84	0.69	0.82	0.86	0.82	0.92	0.93	0.82
8	0.93	0.82	0.82	0.90	0.90	0.90	0.96	0.93	0.86

¹Norcini et al., 1985

²Stalenhoef-Halling et al., 1990

³Swanson, 1987

⁴Wass et al., 2001

⁵Van der Vleuten, 1988

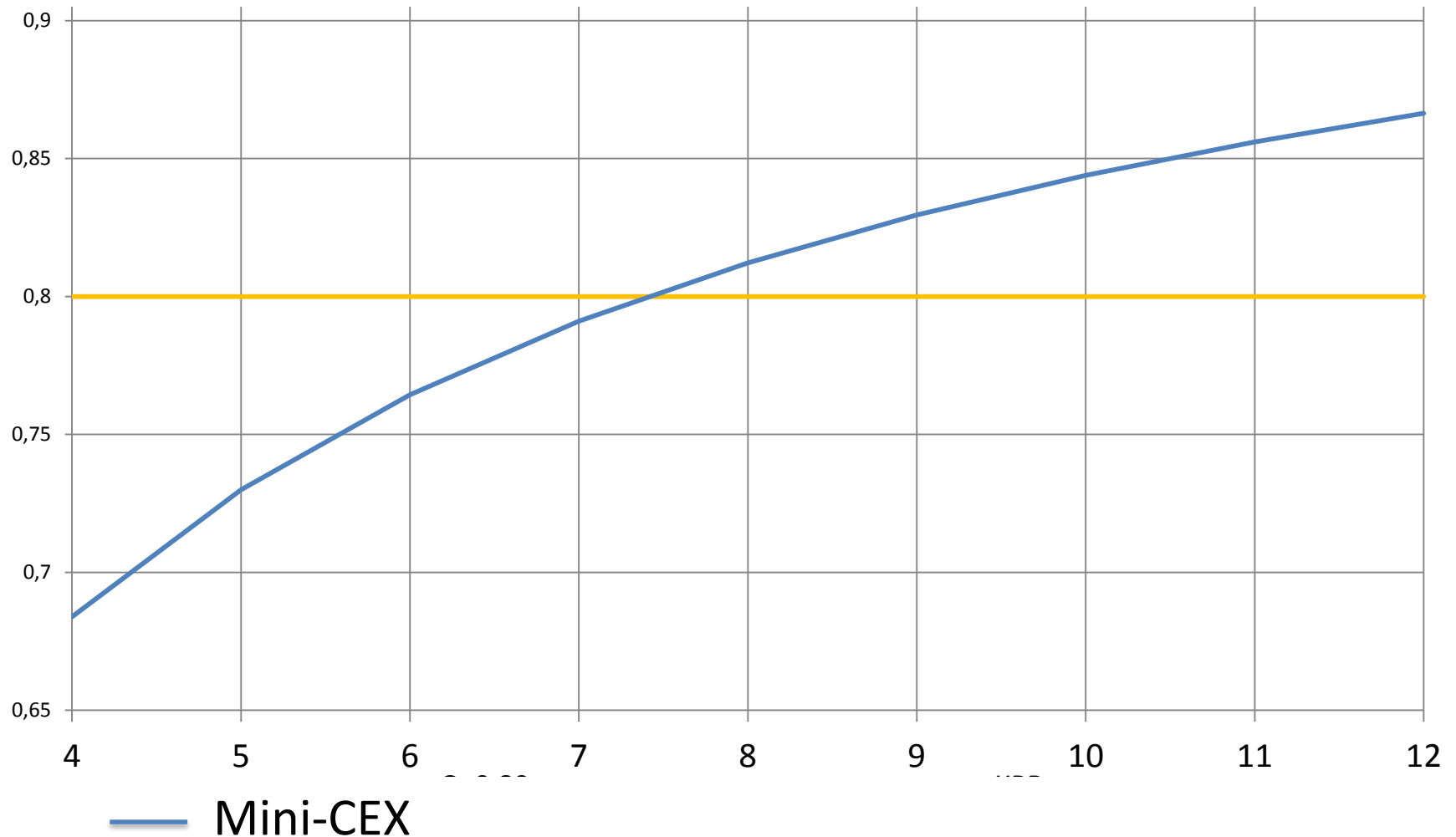
⁶Norcini et al., 1999

⁷Ram et al., 1999

⁸Gorter, 2002

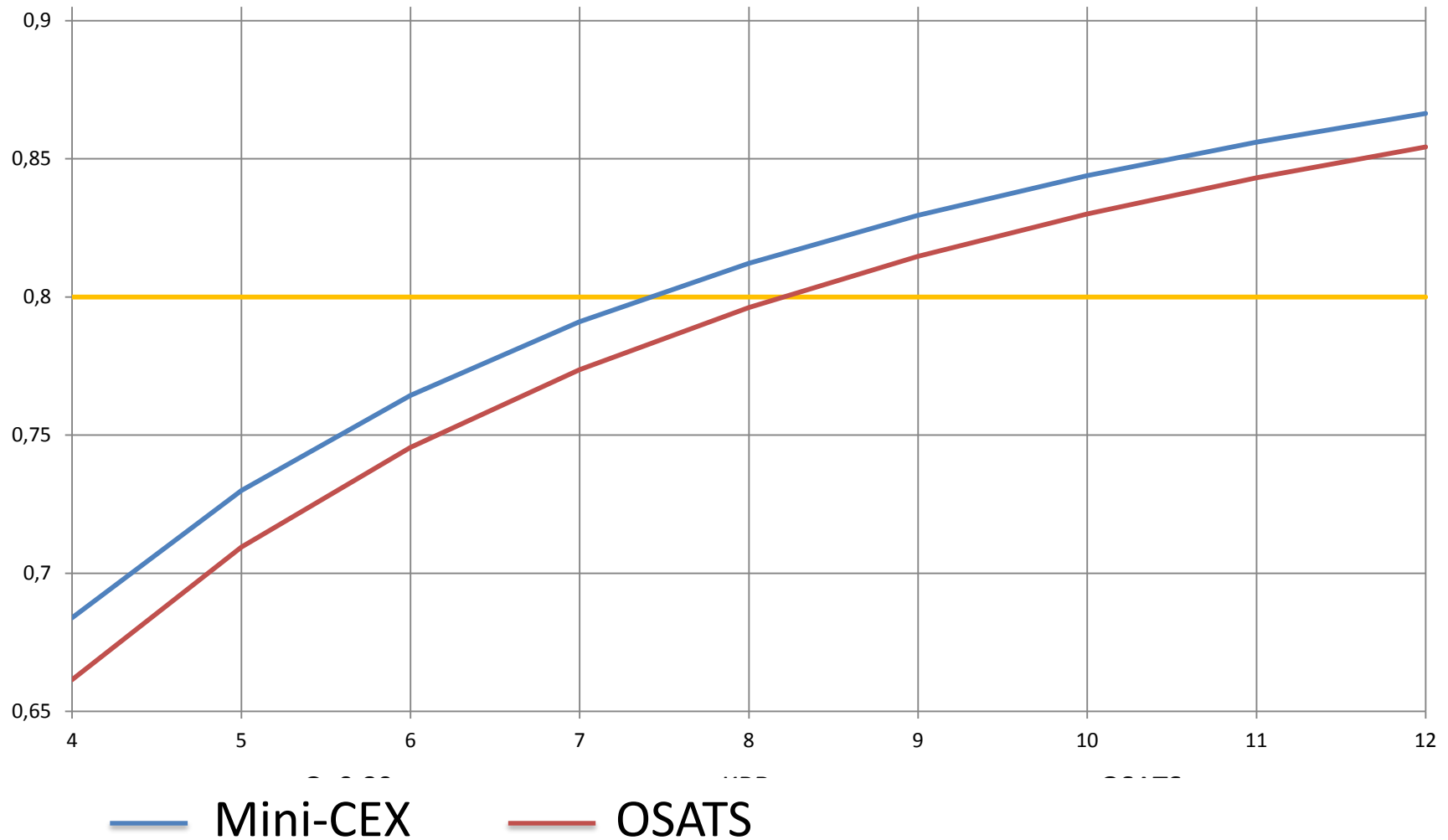
Reliability as a function of sample size

(Moonen et al., 2013)



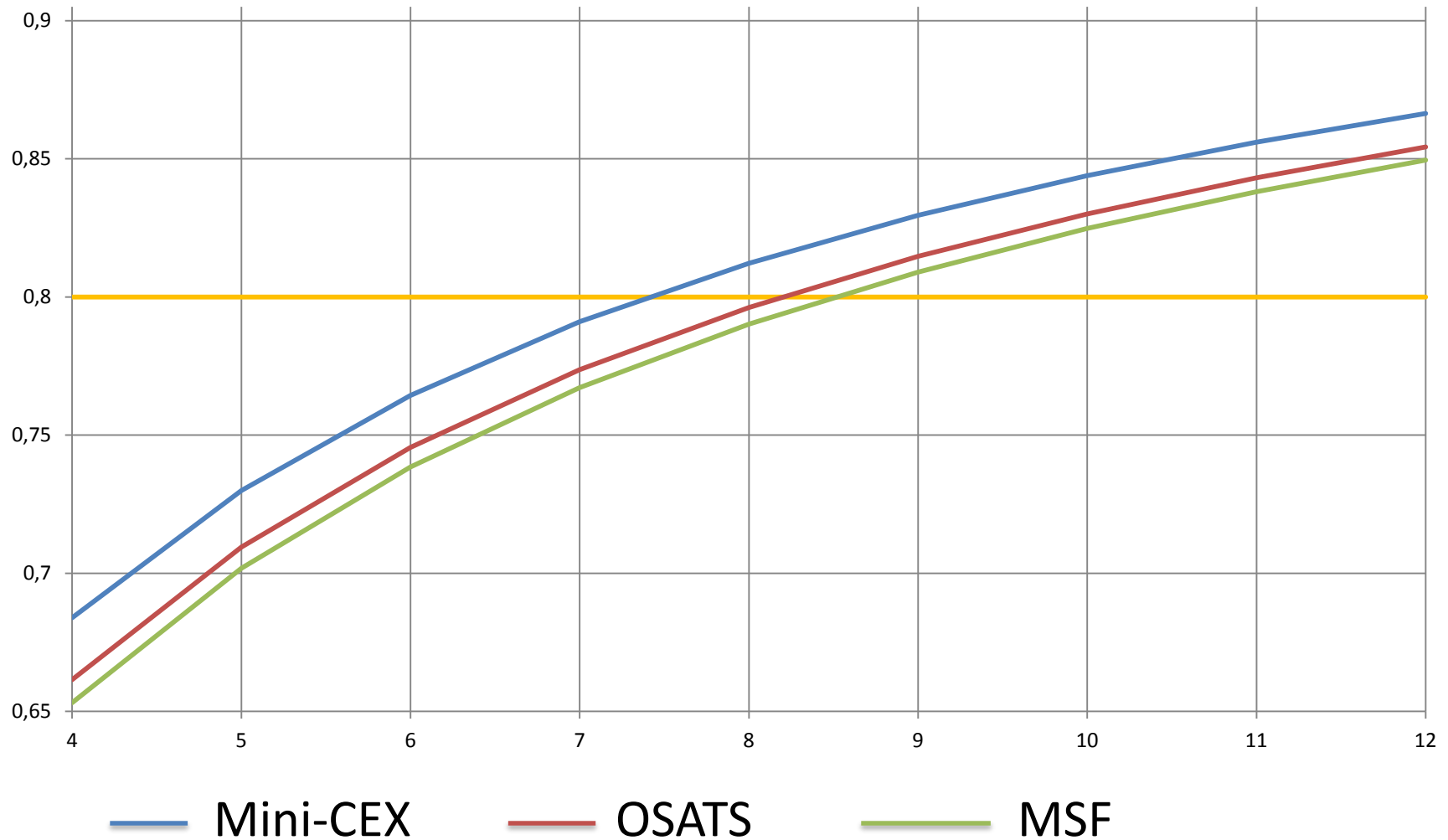
Reliability as a function of sample size

(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

Inzichten over betrouwbaarheid

- Acceptabele betrouwbaarheid wordt slechts bereikt bij een grote steekproef aan toetselementen en beoordelaars
- Geen enkele methode is inherent beter dan een andere (ook de nieuwere niet!)
- Objectiviteit is NIET gelijk aan betrouwbaarheid
- Vele onafhankelijke subjectieve oordelen maken een objectief oordeel.



Onderwijsinvloeden: Hoe stuurt toetsing het leren?

- Sturing is complex (zie Cilliers, 2011, 2012)
- Veel negatieve invloeden
 - Povere leerstijlen
 - Cijferproblematiek (zesjes cultuur, genade-zes, jagen op punten)
 - Hoge werkplekbeoordelingen
- Veel reductionisme in toetsing
 - Weinig feedback (cijfer is slechte vorm van feedback)
 - Toetsconcept niet aansluitend op onderwijsconcept
 - Aggregatie van informatie over niet betekenisvolle eenheden
 - Weinig longitudinale opvolging
 - Veel vinken, weinig vonken (OSCE, werkplekbeoordelingen).

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)

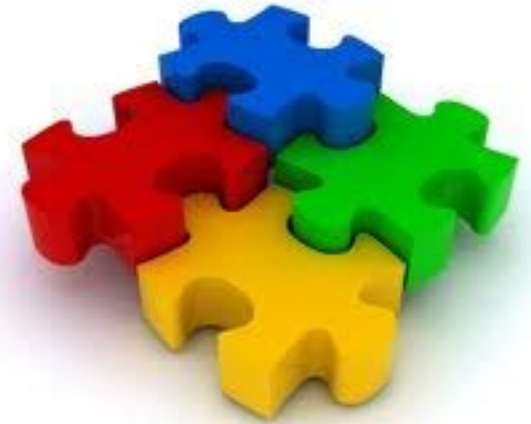


Inzichten over onderwijsgevolgen

- Geen enkele toets of beoordeling zonder betekenisvolle feedback
- Narratieve feedback heeft meer impact dan scores op complexe vaardigheden
- Feedback alleen is niet genoeg voor gebruik
- Meer longitudinale beoordeling is wenselijk.



Overzicht



- Van praktijk naar onderzoek
- **Van onderzoek naar theorie**
- Van theorie naar praktijk
- Conclusies

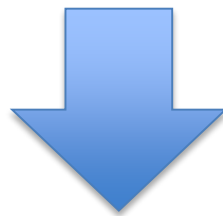
De beperkingen van de enkelvoudige toetsbenadering

- Geen enkele toets kan alles
- Elke toets heeft beperkingen
- Elke toets houdt een fors compromis in



Implicaties

- **Validiteit:** een veelheid aan toetsen nodig
- **Betrouwbaarheid:** veel (gecombineerde) informatie nodig
- **Onderwijsgevolgen:** toetsing moeten (longitudinaal) betekenisvol voor het leren zijn



Toetsprogramma's

Toetsprogramma's



- Curriculum programma is goede metafoor;
in een toetsprogramma zijn:
 - Onderdelen gepland, gearrangeerd, gecoördineerd
 - Systematisch geëvalueerd en bijgesteld
- Maar hoe doe je dat dan? (de literatuur biedt bitter weinig houvast!)

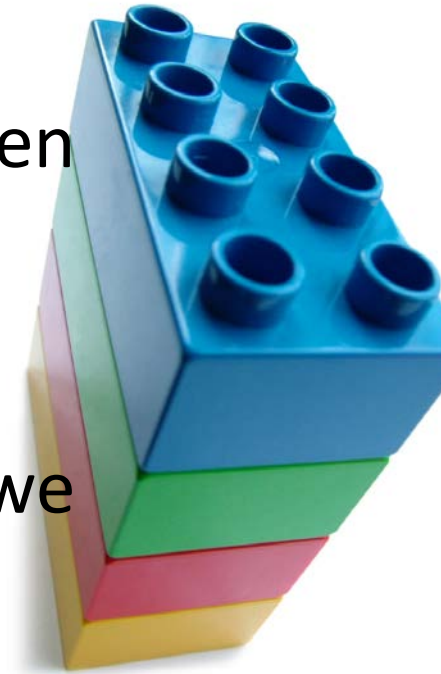
Toetsprogramma's

- Toetsprogramma geïntroduceerd in 2005 (Van der Vleuten & Schuwirth, 2005)
- Set van kwaliteitscriteria met self-assessment instrument voor beoordeling kwaliteit (Baartman et al, 2006, 2007)
- Model voor richtlijnen (Dijkstra et al, 2009) en richtlijnen zelf (Dijkstra et al, 2012)
- Nodig is een theoretisch model voor ontwerp van toetsprogramma's

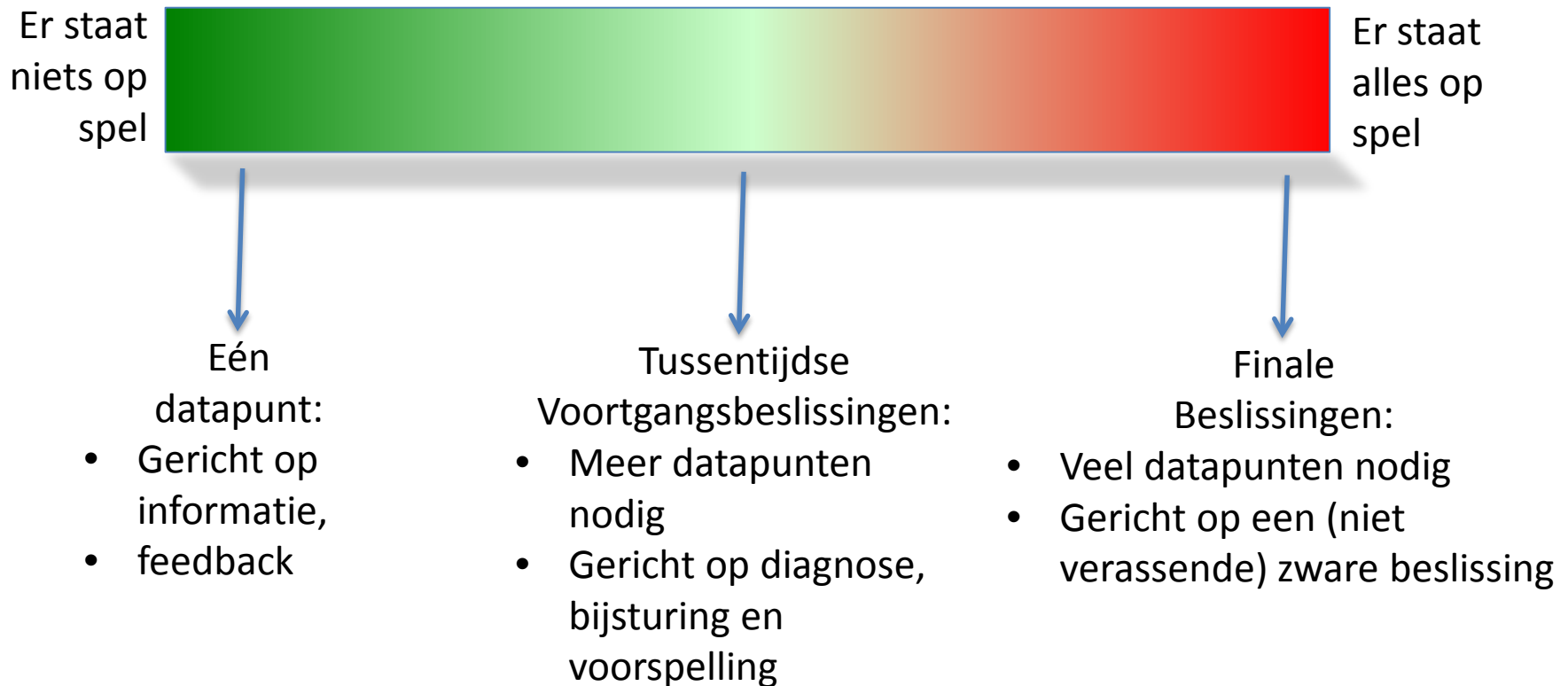


Bouwstenen Toetsprogramma's 1

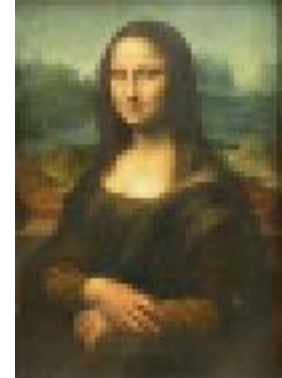
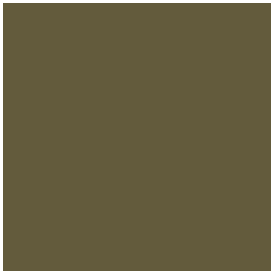
- Elke toets of beoordeling is één datapunt (Δ)
- Elk datapunt is geoptimaliseerd voor leren
 - Informatierijk (kwantitatief, kwalitatief)
 - Betekenisvol
 - Gevarieerd in vorm
- Summatief versus formatief vervangen we door een continuüm van wat er op het spel staat (stakes)
- N datapunten zijn gerelateerd aan aard van beslissing



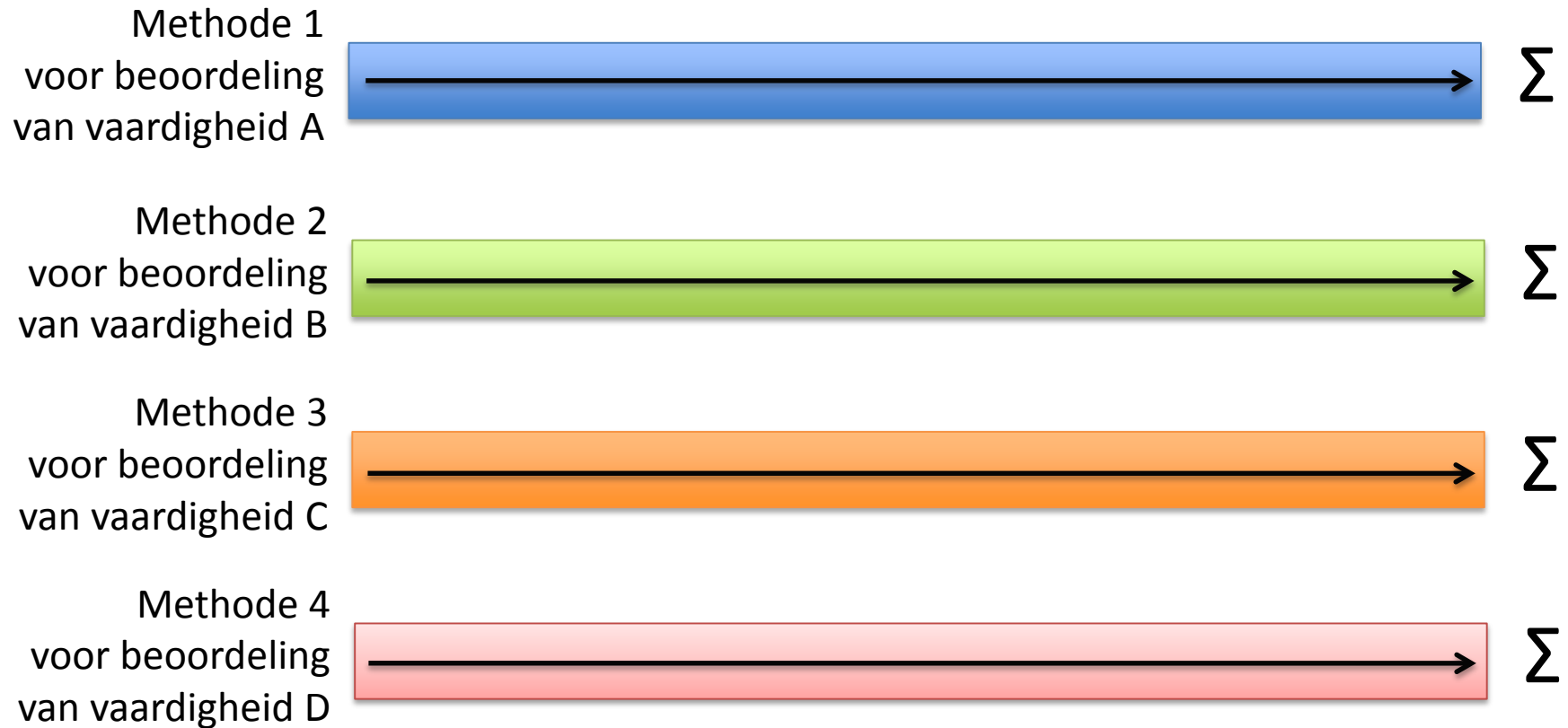
Continuüm van wat er op spel staat, relatie met datapunten en hun functie



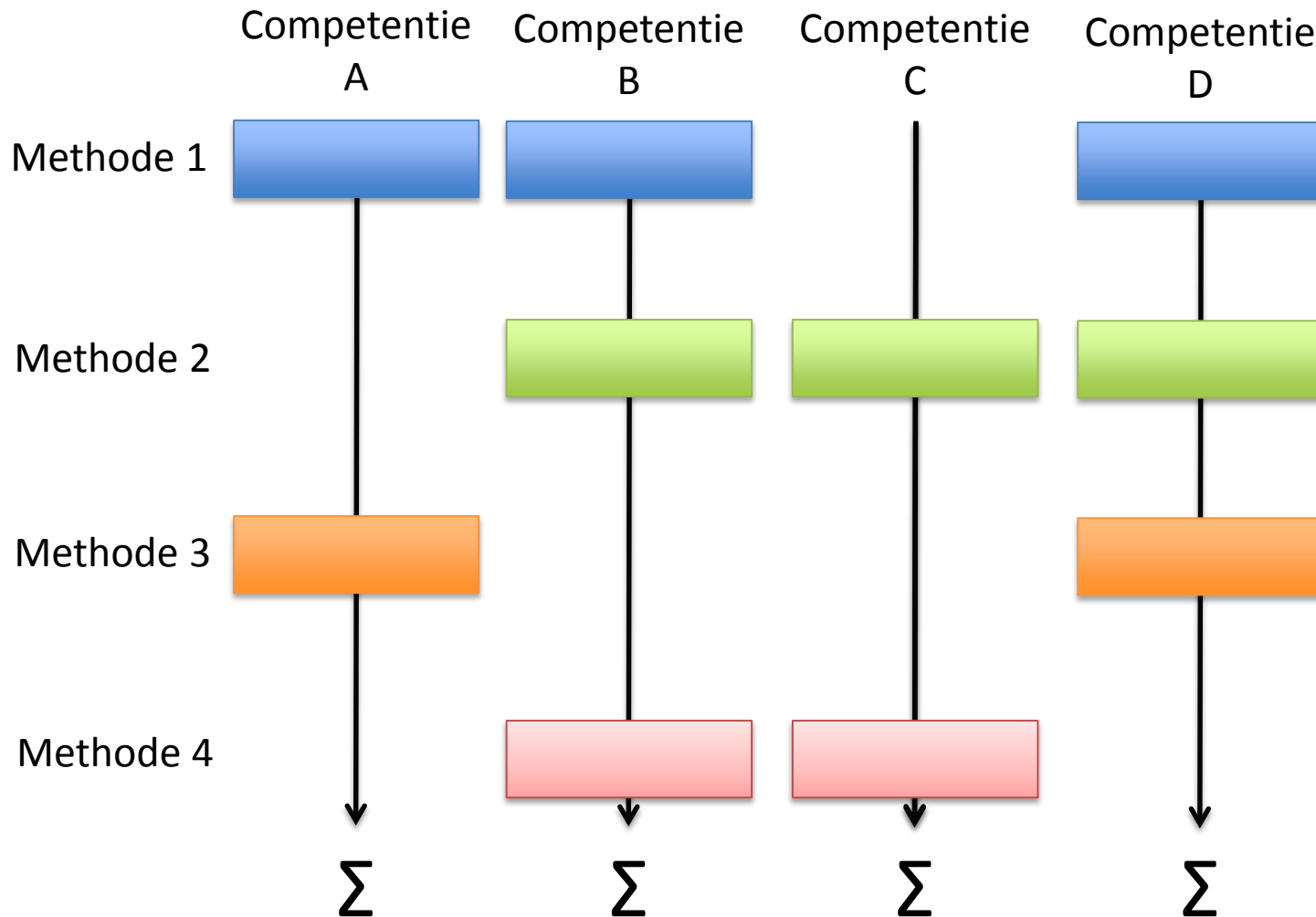
Toetsinformatie als pixels



Klassieke benadering van aggregatie



Betekenisvolle benadering van aggregatie



A model for programmatic assessment fit for purpose

C. P. M. VAN DER VLEUTEN¹, L. W. T. SCHUWIRTH², E. W. DRIESSEN¹, J. DIJKSTRA¹,
D. TIGELAAR³, L. K. J. BAARTMAN⁴ & J. VAN TARTWIJK⁵

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

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.

TWELVE TIPS

12 Tips for programmatic assessment

C.P.M. VAN DER VLEUTEN¹, L.W.T. SCHUWIRTH², E.W. DRIESSEN¹, M.J.B. GOVAERTS¹ & S. HEENEMAN¹

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

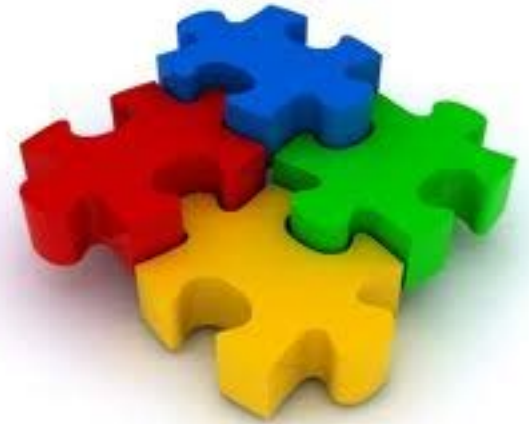
Abstract

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.



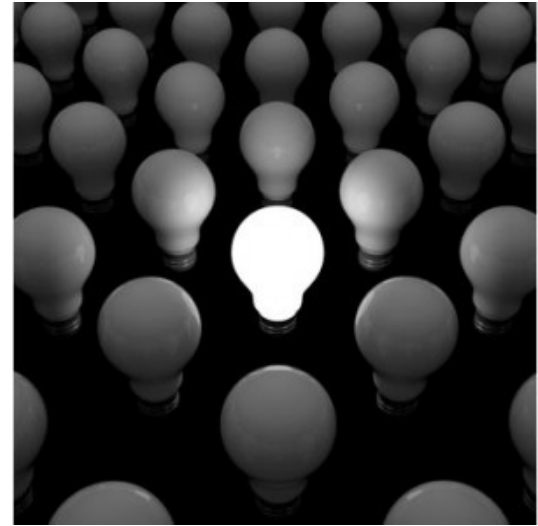
Overzicht

- Van praktijk naar onderzoek
- Van onderzoek naar theorie
- **Van theorie naar praktijk**
- Conclusies



Terug van theorie naar praktijk

- Bestaande toetspraktijken:
 - Master Diergeneeskunde Utrecht
 - Master geneeskunde Maastricht
 - Huisartsopleiding Nederland en Canada
 - Sommige specialistenopleidingen zijn hard op weg
 - Cleveland Learner Clinic, Cleveland, Ohio
 - Arts-Klinisch-Onderzoeker (AKO), graduate entry geneeskunde opleiding Maastricht



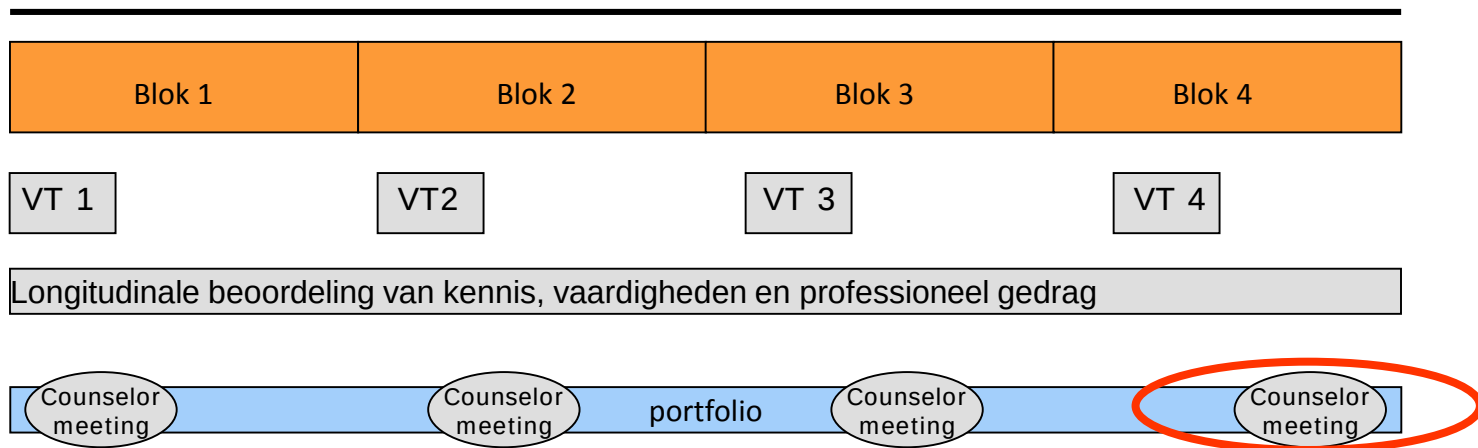
Arts-Klinisch Onderzoeker opleiding

- 4 jarig graduate entry programma
- Competentiegericht (Canmeds) met nadruk op research
- PGO programma
 - Jaar 1: Klassiek PGO
 - Jaar 2: PGO met echte patiënten
 - Jaar 3: stages
 - Jaar 4: participatie in gezondheidszorg en onderzoek
- Hoge verwachtingen van studenten mbt motivatie, excellentie, zelf-sturing.



Het toetsprogramma

- Blokgebonden toetsing: opdrachten, presentaties, tussen- en eindtoetsen, etc.
- Longitudinale toetsing: opdrachten, reviews en essays, projecten, voortgangstoetsen, professionele gedragsbeoordeling, etc.
- Alle toetsing is formatief en informatief
- Portfolio is het centrale instrument voor longitudinale monitoring door systeem van mentoring en voor besluitvorming over studievoortgang



Student

Peer groups

Welcome

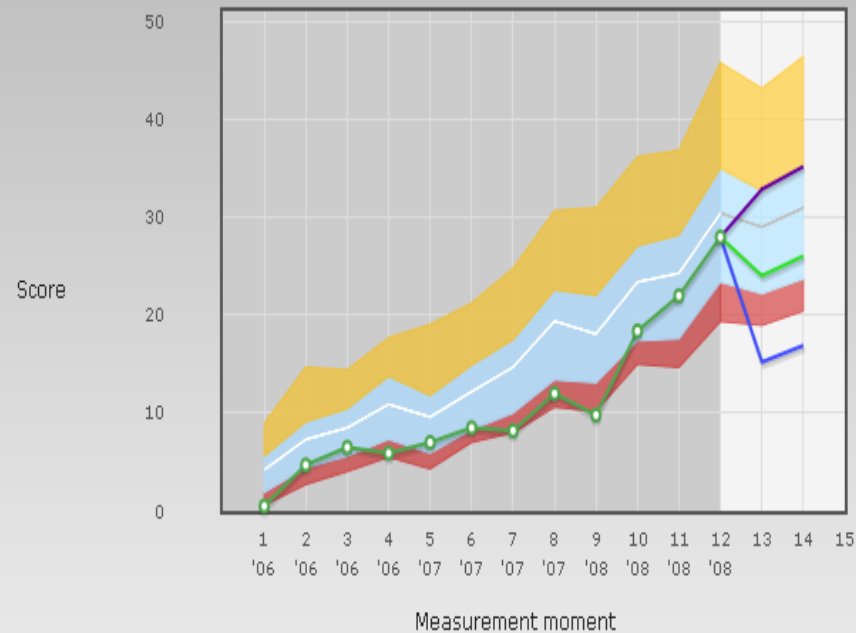
Momentaneous

Longitudinal

Test Overview

Domain	Total
Result	Score
Series	Unprocessed
Peer group	UM FHML-G year group 3
Reference values	Percentiles

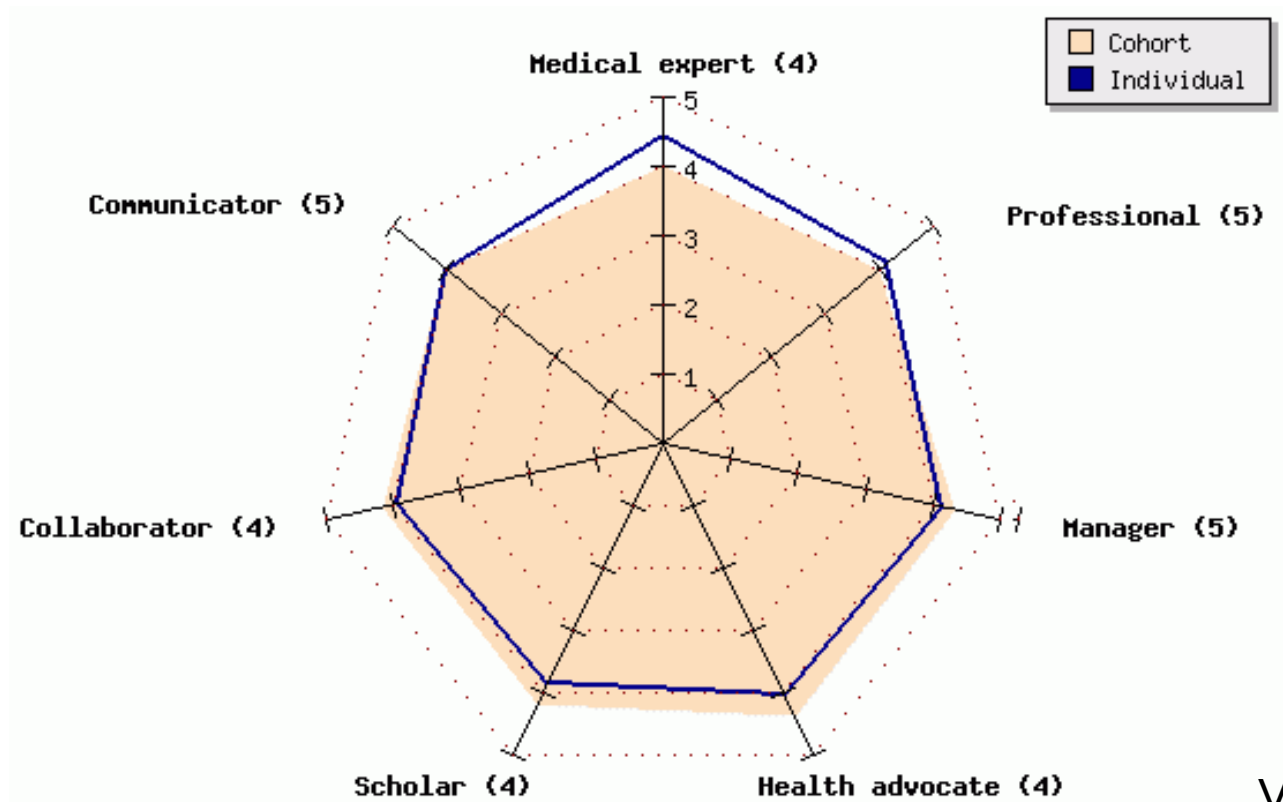
Longitudinal series (*unprocessed*) of score for *total* for student 403164 with peer group *UM FHML-G year group 3* as background population



- Student score
- Lower confidence bound prognosis
- Prognosis
- Upper confidence bound prognosis

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

Electronisch portfolio (ePass)

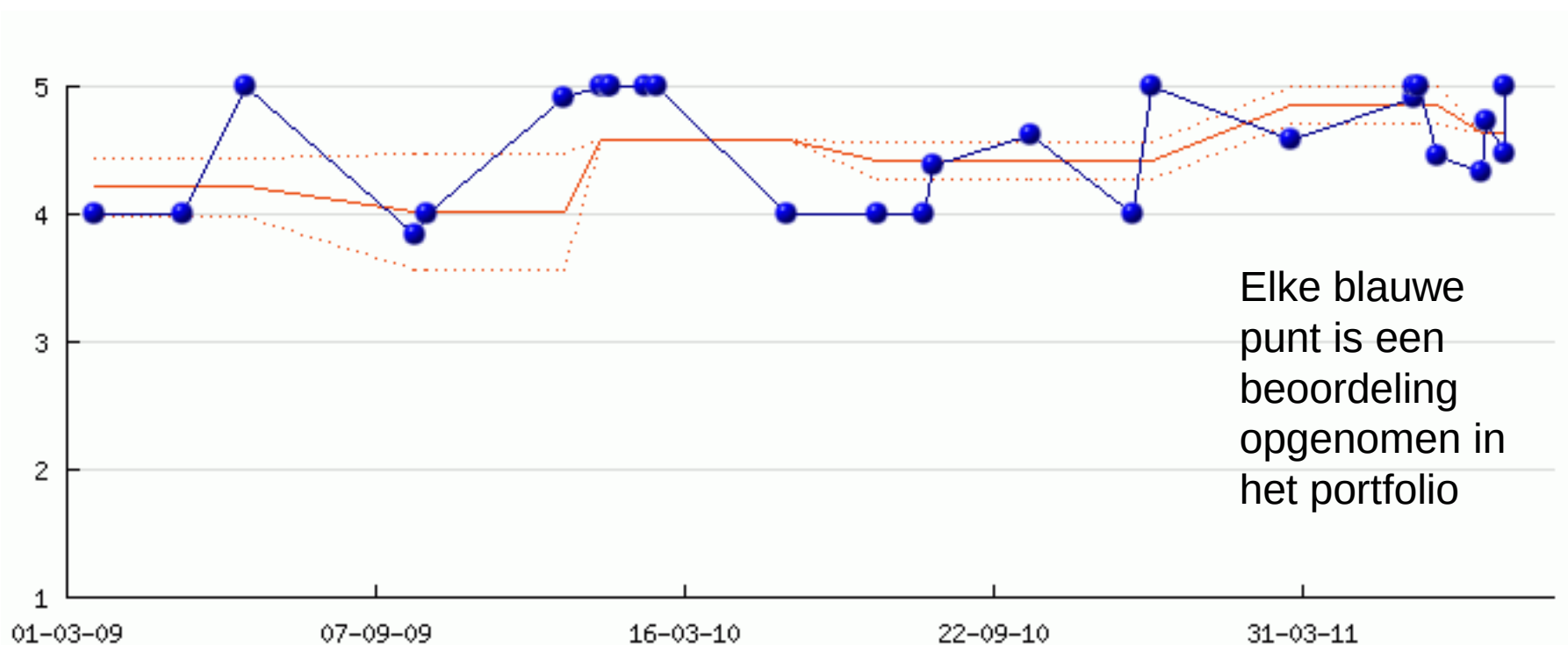


Vergelijking
tussen de score
van de student
vergeleken met
het gehele cohort

Maastricht Electronic portfolio (ePass)

1: Medical expert

Table view



Coaching door counselors

- Coaching is essentieel voor gebruik van feedback en planning
- Counselor geeft advies/commentaren (gevraagd en ongevraagd)
- Voortgang wordt bewaakt op alle competenties, voortgangsadviezen worden gegeven



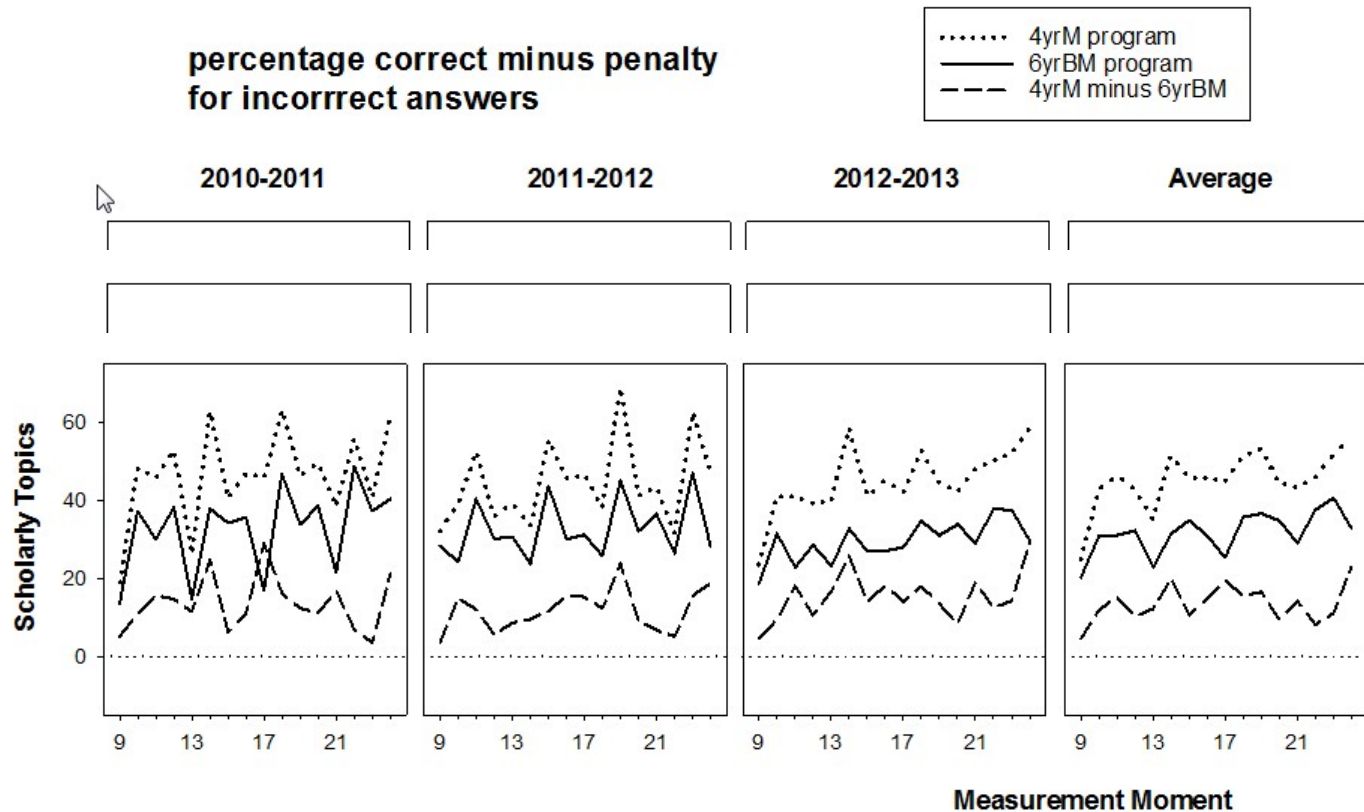
Besluitvorming door commissie

- Commissie van counselors én externen
- Beslissing obv portfolio informatie, counselor aanbeveling en competentie standaarden
- Deliberatie is proportioneel aan helderheid van informatie (efficiënt en tijdbesparend proces)
- Besluiten worden zo nodig toegelicht; remediëringsadviezen kunnen worden gegeven



Strategy to establish trustworthiness	Criteria	Potentiele Toepassingen in Toetsing
Credibility	Prolonged engagement	Trainen van beoordelaars
	Triangulation	Toenemende inschakeling van experts op basis van mate van zekerheid over het oordeel
	Peer examination	Benchmarken beoordelaars
	Member checking	Incorporeer een zelf-oordeel van de lerende
	Structural coherence	Controle op inconsistenties in besluitvorming
Transferability	Time sampling	Gebruik van veel datapunten
	Thick description	Gemotiveerde besluitvorming
Dependability	Stepwise replication	Grootte van de commissie bestaande uit gerespecteerde leden
Confirmability	Audit	Creëer beroepsmogelijkheid

4-jarige (AKO) en 6-jarige prestaties vergeleken op VGT



Heeneman, S., Schut, S., Donkers, J., Van der Vleuten, C., Muijtjens, A. (Under editorial review). Embedding of the progress test in an assessment program designed according to the principles of programmatic assessment.

Overzicht

- Van praktijk naar onderzoek
- Van onderzoek naar theorie
- Van theorie naar praktijk
- Conclusies



Conclusies 1



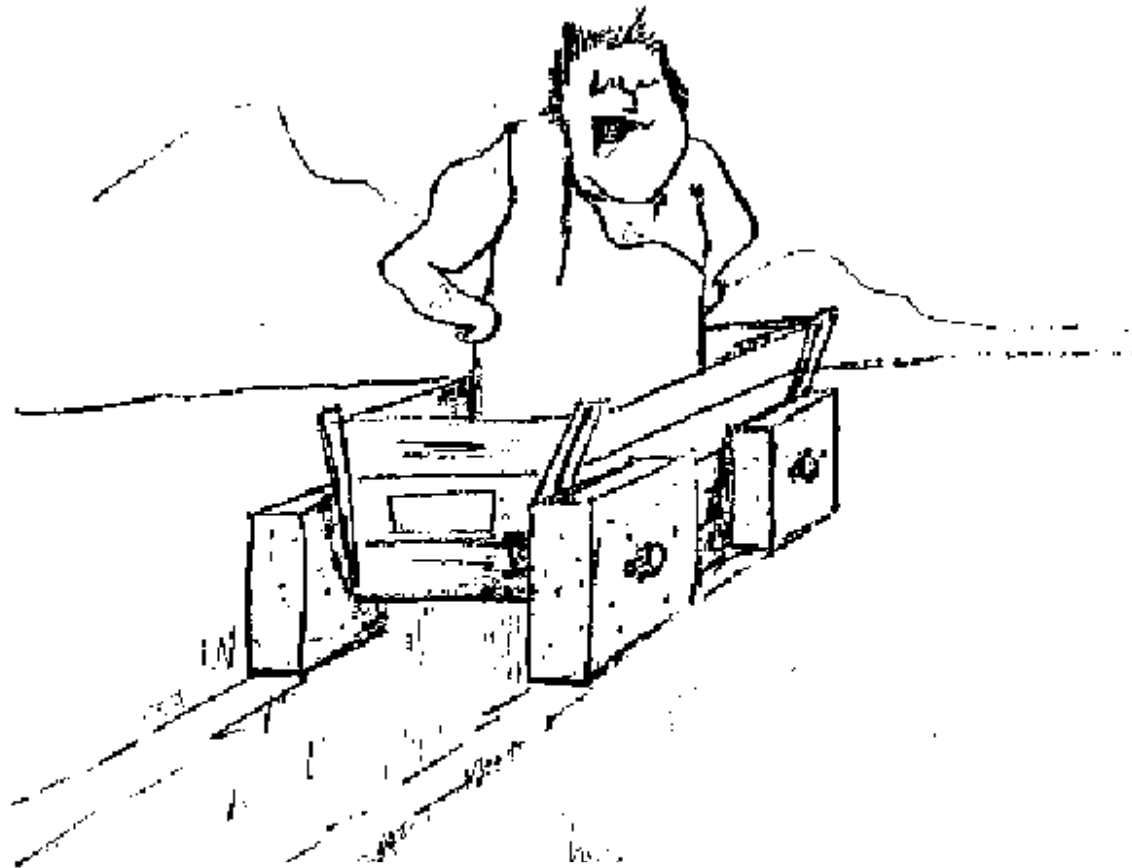
- We moeten af van het exclusief denken in individuele toetsmethoden
- Een systemische, programmatische aanpak is daarvoor noodzakelijk, longitudinaal gericht
- Elke toetsmethode kan hierin functioneel zijn (oud en nieuw; gestandaardiseerd en ongestandaardiseerd)
- Professionele oordeelsvorming is onontbeerlijk (vergelijkbaar aan de klinische praktijk)
- Subjectiviteit wordt gepareerd met sampling en met procedurele maatregelen (en niet met standaardisering of objectivering)

Conclusies 2

- Het toetsprogramma optimaliseert:
 - De leerfunctie (door informatierijkheid)
 - De beslisfunctie (door combinatie van informatie)







“Did you ever feel you’re on the verge of an incredible breakthrough?”