PRIME COLLABORTIVE

ACGME

What is Good Assessment?

Eric Holmboe, MD

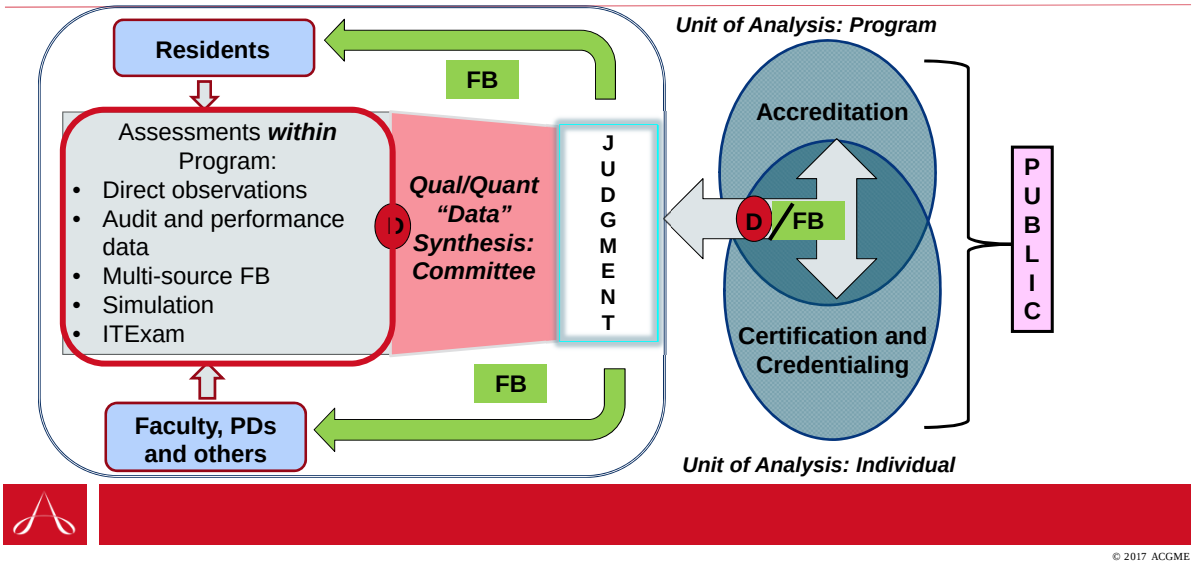
In Partnership with:

Geisinger Commonwealth School of Medicine

PennState College of Medicine

Jefferson HOME OF SIDNEY KIMMEL MEDICAL COLLEGE

The GME Assessment "System"



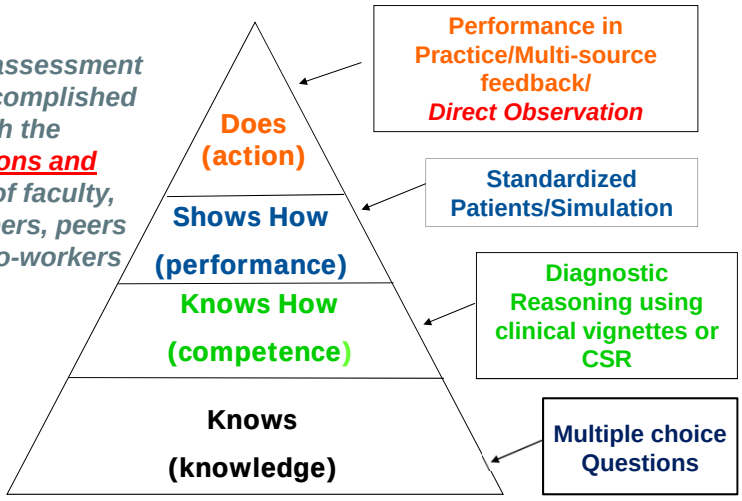
Process vs. Outcome Approach

	Educational Program	
Variable	**Structure/Process**	***Outcome-based***
Driving force: process	**Teacher**	***Learner***
Path of learning	**Hierarchical (Teacher→student)**	***Non-hierarchical (Teacher↔student)***
Responsibility: content	**Teacher**	***Student and Teacher***
Goal of educ. encounter	**Knowledge acquisition**	***Knowledge application***
Typical assessment tool	**Single competency measure**	***Multiple measures, multiple competencies***
Assessment tool	**Proxy**	***Authentic (real tasks of profession)***
Setting for evaluation	**Removed (gestalt)**	***Direct observation***
Evaluation	**Norm-referenced**	***Criterion-referenced***
Timing of assessment	**Emphasis on summative**	***Emphasis on formative***

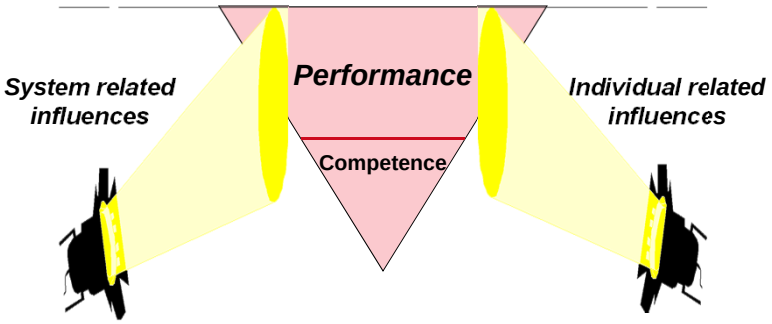
Adapted from Carracchio, et al. 2002.

Assessing for the Desired Outcome

*Work-based assessment is mostly accomplished through the **observations and questions** of faculty, team members, peers and other co-workers*



Cambridge Model: "Righting" the Pyramid



Work-based assessment has to be the primary focus of our assessment systems

Rethans, Norcini, et al, 2002

Common Assessment Methods

- Descriptive evaluation by teachers
- Records of clinical encounters
- External/internal evaluations
  - MCQ
  - Key features/script concordance
  - Short answer questions/essays
- Simulations
- OSCES
- Oral examinations
- Chart (record) reviews
- Standardized patients
- A-V reviews
- Educational prescription contracts
- Portfolios
- MSF (360) evaluations
- Patient logs
- Checklists
- Rating scales

How to Choose?

Nothing is Perfect

"Fit for Purpose"

- One of the most important decision points in choosing an assessment method and tool is whether it is "fit for purpose."
  - How will the method/tool help the program assess and provide feedback on professional development?
  - How does it fit within a program of assessment?

Utility: Choosing the Right Method

Cees van der Vleuten's utility index:

Utility = V x R x A x EI x CE/Context

Where:

V = validity

R = reliability

A = acceptability

E = educational impact

C = cost effectiveness

Reliability and Validity Simplified

Not reliable
Not valid

Reliable
Not valid

Reliable
Valid

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

From CPM Van der Vleuten; ACGME 2016

Validity: Messick

1. Content: relationship between the tool's content and the construct it intends to measure.
2. Response process: evidence showing raters have been properly trained (faculty development).
3. Internal structure (reliability): Internal consistency, test-retest reliability, agreement (inter-rater reliability), generalizability.
4. Relationship to other variables (concurrent, predictive validity): correlation of scores with other assessments or outcomes; differences in scores by learner subgroups.
5. Outcomes (educational outcomes): Consequences of assessment.

Criteria for "Good" Assessment[1]

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

[1]Norcini J et al. Med Teach 2011;33:206-14

Educational Impact

Educational Effect

"The assessment motivates those who take it to prepare in a fashion that has educational benefit."

Catalytic Effect

"The assessment provides results and feedback in a fashion that creates, enhances, and supports education; it drives future learning forward."

Norcini J et al. Med Teach 2011;33:206-14

Continuum of Stakes, Number of Data Point and Their Function

No stake

Very high stake

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

From CPM Van der Vleuten

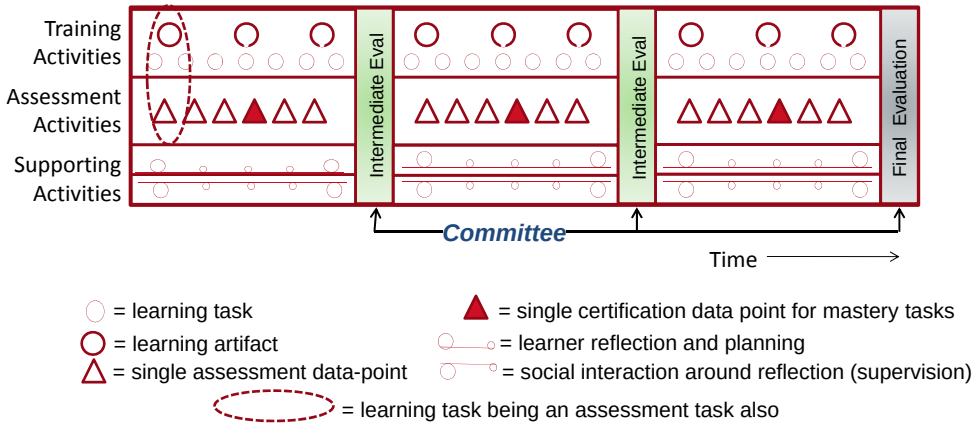
Creating Assessment Programs

- Competence is specific, not generic. Sample across contexts, assessors, time
- Use multiple assessment methods
- Quantitative not necessarily better than qualitative
- Move assessment back to workplace
- Use credible standards
- Validity resides in instrument user

Van der vleuten CPM et al Med Educ 39:309–17.
Van der vleuten CPM et al. Best Practice & Research Clinical Obstetrics and Gynaecology. 2010(24):703–19

Model For Programmatic Assessment

(With permission from CPM van der Vleuten)



Small Group Exercise – Part 1

- Choose an assessment tool your faculty use in your program
  - Is this assessment tool "fit for purpose"?
    - In other words:
      - Why this tool?
      - What competencies are addressed?
      - How does it help the learner? The program?

Small Group Exercise – Part 2

Rate the "utility effectiveness" of the tool in *your program*

What approaches have you used to improve the use of the tool?

What has been effective?

What barriers have you encountered that prevent the maximum utility being realized?

Small Group Exercise – Part 3

- Map this assessment tool to pertinent subcompetency/milestones in your specialty.
  - How well does this tool assess the pertinent subcompetencies/milestones?
  - How does this tool link to your curriculum?

PRIME COLLABORTIVE

ACGME

Issues in Faculty Assessments

Geisinger Commonwealth School of Medicine

PennState College of Medicine

Jefferson HOME OF SIDNEY KIMMEL MEDICAL COLLEGE

Key Issues: Psychometric

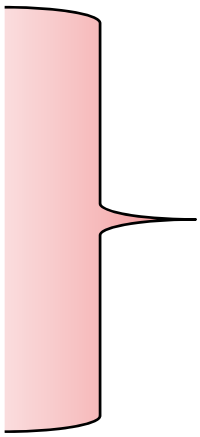
- Multiple studies demonstrating major issues in intra- and inter-rater reliability
  - Usual response – ***change the form or tool…***
- Limited evidence for validity
  - Modest correlations between high-stakes assessments and faculty ratings
- Lack of discrimination among domains of competence
  - The "factor analytic" problem

Key Issues: Errors

- <u>Correlational errors</u>
  - Halo effect
  - Horn effect
    - Ratings based mostly on *perceived* knowledge and personality
- <u>Distributional errors</u>
  - Leniency error ("Doves")
  - Severity error ("Hawks")
  - Central tendency

Key Issues: Individual Effects

- Variability among faculty
  - Strengths and weaknesses
    - Clinical
    - Educational
    - Assessment
- Idiosyncrasy
- Inference



Jen Kogan

Key Issues: Human Limitations

- Limitation in working memory and mental processing
  - The “7 +/- 2” rule for short term memory
- Subconscious processes
  - Bias and stereotyping
- Cognitive Load

PRIME
COLLABORTIVE

ACGME

Assessor and Human Limitations

These slides have been modified from work performed by Dr. Peter Yeates, Keele University, UK

Geisinger Commonwealth School of Medicine

PennState College of Medicine

Jefferson
HOME OF SIDNEY KIMMEL MEDICAL COLLEGE

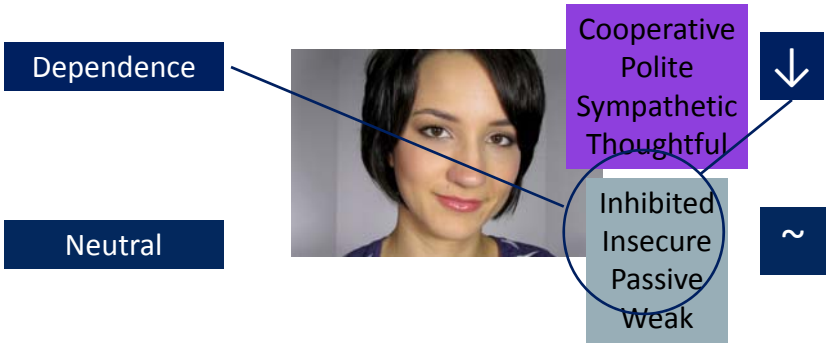
Human Working Memory

Limited working memory capacity

How we organise influences perception

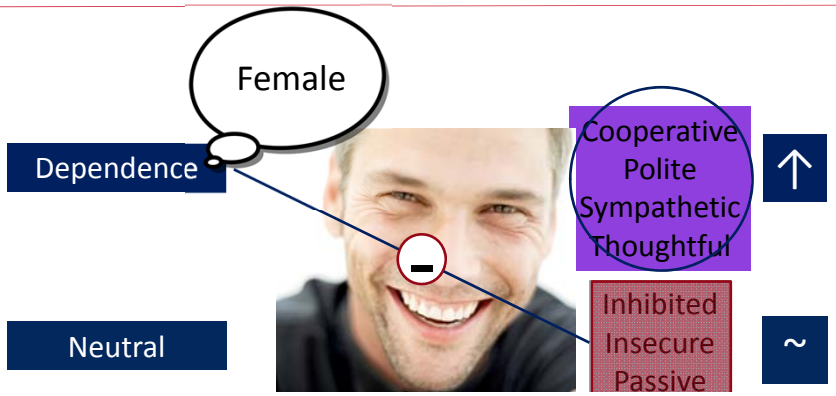
Baddeley, A.D., 1994. *Psychological Review*, 101(2), pp.353–356

An Example: Stereotypes



Banaji, M *et al.* 1993 Journal of Personality and Social Psychology 65(2): 272-281

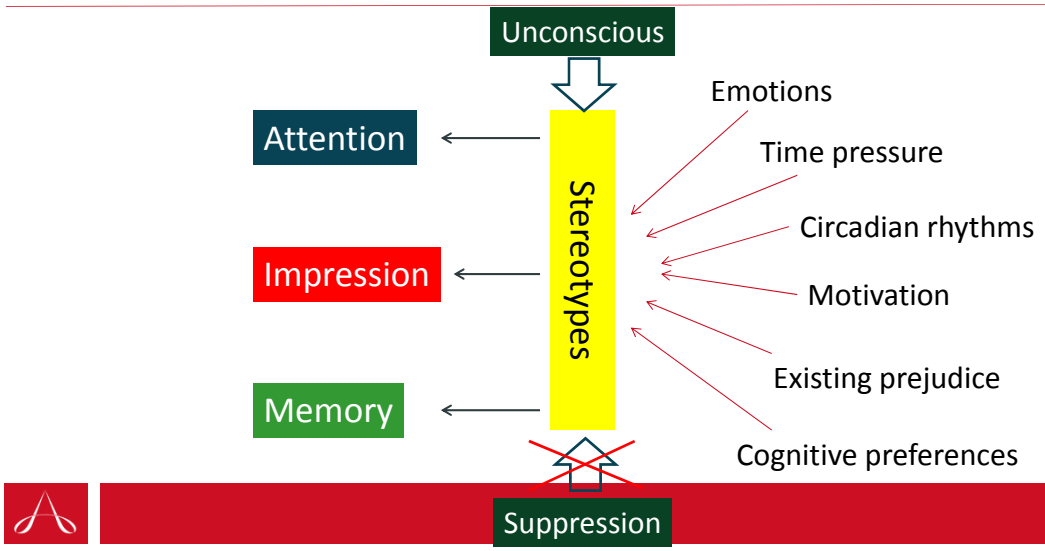
An Example: Stereotypes



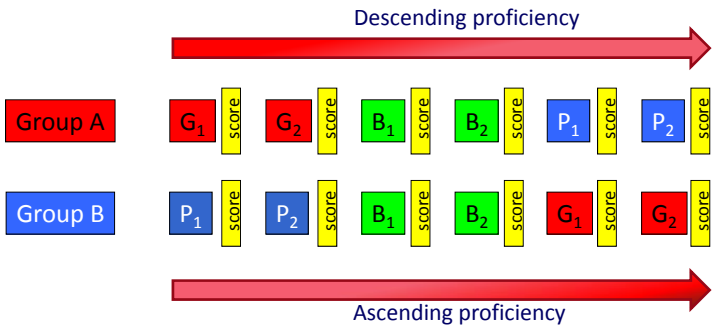
Senior doctors' have stereotypes of medical students

Woolf K. *et al.* (2008) *BMJ.* 337: a1220 doi:10.1136/bmj.a1220

Stereotypes

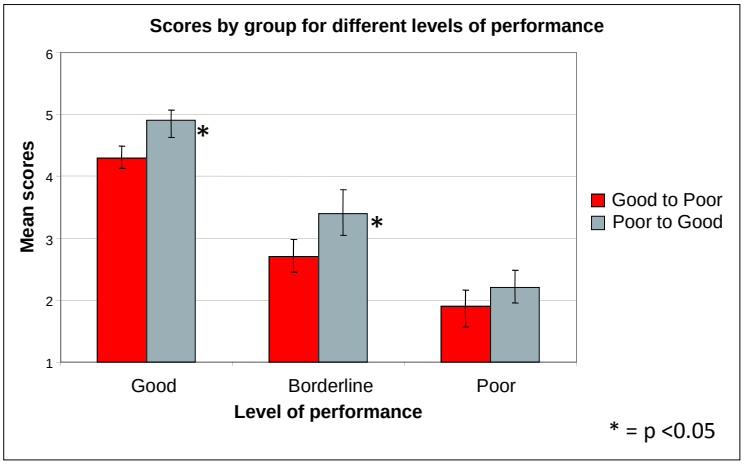


Contrast effects



Yeates, P. et al., *Medical education*, 2013, 47(9), pp.910–22.

Contrast Effect



Yeates, P. et al., *Medical education*, 2013, 47(9), pp.910–22.

Selective Attention

http://www.theinvisiblegorilla.com/videos.html

Cognitive Load

- There is a limit as to how much you can ask faculty to observe and capture
  - Clinical units: complex environment
  - Selective attention
- Byrne et. al. (Med Educ 2014)
  - Average cognitive load for faculty judging OSCE stations was higher than anesthesia trainees during induction for routine surgery
    - OSCE had 21-22 items in an 8 minute station

Cognitive Load

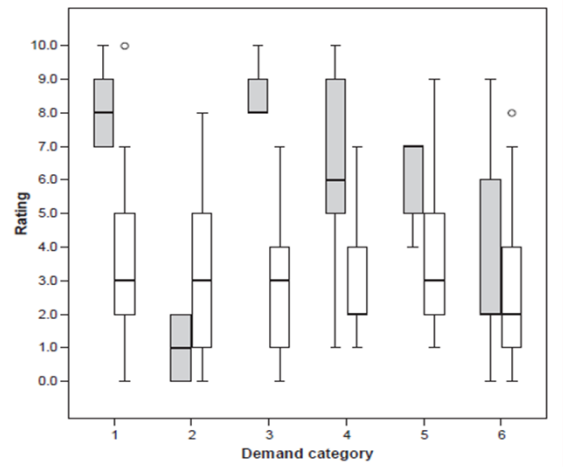


Figure 3 Comparison of NASA–Task Load Index (NASA-TLX) scores in the study subjects (grey boxes) and trainee anaesthetists (white boxes).

Demand categories:
1 = mental demand
2 = physical demand
3 = temporal demand
4 = performance/success
5 = effort
6 = frustration

Byrne A, Tweed N, Halligan C. A pilot study of the mental workload of OSCE examiners. Med Educ. 2014; 48: 262-67.

Useful Dictum

- The longer the form (or checklist) and the shorter the "exposure" or observation time, the more likely you are to get less useful ratings and information from the evaluation form
  - Long evaluation forms + short faculty rotations = trouble in assessment land

Implications

Training won't overcome these problems

More detailed frameworks / checklists unlikely to help
Could worsen if increase cognitive load

Tavares, W., & Eva, K. W. (2012). *Adv Health Sci Educ*. doi:10.1007/s10459-012-9370-3

Implications: Solutions

Increase sampling ?

- Cost
- Systematic biases

Cognitively-designed assessments ?

- Reduce Cog Load / Support memory
- Test assumptions first

Wood T. (2012) *Adv Health Sci Ed* DOI 10.1007/s10459-012-9396-6

Faculty Development

It is ***very*** unlikely you can create a successful assessment system without faculty training

Assessment tools are only as good as the individual using them

At the present time – ***still need faculty judgment***!

Need shared mental models and understanding of the competencies

PRIME
COLLABORTIVE

ACGME

Supervision and Entrustment to Guide Faculty Assessments

Geisinger Commonwealth School of Medicine

PennState College of Medicine

Jefferson
HOME OF SIDNEY KIMMEL MEDICAL COLLEGE

Video Exercise

What was this intern entrusted to do without direct supervision?

Entrustment

Trust:

"Trust involves the confident expectation that a person can be relied on to honour implied or established commitments to an individual and to protect [the individual's] interest. It renders the individual vulnerable to the extent (s)he cannot oversee or control the actions of the other, on whose expertise or integrity (s)he may depend."

From ten Cate, et al., Entrustment decision making in clinical training. Acad Med. 2016; 91: 191-98

Entrustment

Modes of trust:

- *Presumptive*
  - Based solely on credentials
- *Initial* (swift or thin trust)
  - Mostly based on first impressions
  - Prone to error
- *Grounded*
  - Based on essential and prolonged experience

From ten Cate, et al., Entrustment decision making in clinical training. Acad Med. 2016; 91: 191-98

Entrustment Decision Making

- Ad hoc
  - In the moment and usually based on a mix of estimated trustworthiness, risk of situation, urgency, suitability of task.
  - Does not necessarily set a precedent for future decisions
- Summative
  - Grounded in sufficient and robust assessment
  - Leads to supervision, licensing and certification decisions

From ten Cate, et al., Entrustment decision making in clinical training. Acad Med. 2016; 91: 191-98

Rating Scales: Types of Anchors

- Adjectival - performance "quality"
  - E.g. Unsat-satisfactory-superior
- Frequency
  - Rarely – always
- Normative
  - Level of comparative performance
- Developmental
- Entrustment/supervision
- Narrative

These can overlap depending on purpose

Rating Scales

- Rating scales are not *dimensional* data!
  - Equal intervals between anchors does not mean the data are truly dimensional
- Rating scales are almost always ***ordinal***

Adjectival Rating Form

INTERNAL MEDICINE RESIDENT EVALUATION FORM

Really acts like…

1 2 3 4 5 6 7 8 9

clinical data and consider patient preferences when making medical decisions

❑ Insufficient contact to judge

preferences

2.Medical Knowledge
Limited knowledge of basic and clinical sciences; minimal interest in learning; does not understand complex relations, mechanisms of disease

1 2 3 4 5 6 7 8 9

❑ Performance needs attention

Exceptional knowledge of basic and clinical sciences; highly resourceful development of knowledge; comprehensive understanding of complex relationships, mechanisms of disease

❑ Insufficient contact to judge

Construct Aligned Scales


Good questions, good answers: construct alignment improves the performance of workplace-based assessment scales

Jim Crossley,[1] Gavin Johnson,[2] Joe Booth[3] & Winnie Wade[3]

"Crossley and Jolly have suggested that effective assessment tools have construct alignment, which means that the tool reflects the expertise and priorities of the evaluator."

Crossley J, Johnson G, Booth J, Wade W. Good questions, good answers: construct alignment improves the performance of workplace-based assessment scales. *Medical Education* 2011; 45: 560–569

Entrustment Scales

- Per Rekman and colleagues, entrustability scales are a species of construct-aligned scales
- Entrustability scales are usually expressed by varying levels of supervision, oversight and/or actions of the attending

Rekman J, Gofton W, Dudek N, Gofton T, Hamstra SJ. Entrustability Scales: Outlining Their Usefulness for Competency-Based Clinical Assessment. Acad Med. 2016 Feb;91(2):186-90.

Entrustment Scale: O-SCORE

The Ottawa Surgical Competency Operating Room (O-SCORE) Scale[a]: An Entrustability-Aligned Anchor Scale

Level	Descriptor
1	"I had to do" (i.e., requires complete hands on guidance, did not do, or was not given the opportunity to do)
2	"I had to talk them through" (i.e., able to perform tasks but requires constant direction)
3	"I had to prompt them from time to time" (i.e., demonstrates some independence, but requires intermittent direction)
4	"I needed to be there in the room just in case" (i.e., independence but unaware of risks and still requires supervision for safe practice)
5	"I did not need to be there" (i.e., complete independence, understands risks and performs safely, practice ready)

[a]The authors adapted the scale from Gofton W, Dudek N, Wood T, Balaa F, Hamstra S. The Ottawa surgical competency operating room evaluation (O-SCORE): A tool to assess surgical competence. Acad Med. 2012;87:1401–407.

Alignment of Developmental Models

Milestone Level	Dreyfus Stage	Learner Behavior	Transition to Practitioner	Level of Supervision
1	Novice	Doing what is told, rule driven	Intro to clinical practice	Observation, no entrustment
2	Advanced beginner	Comprehension	Guided clinical practice	Act under direct supervision
3	Competent	Application to common practice	Early independence	Act under indirect supervision
4	Proficient	Application to uncommon practice	Full unsupervised practice	Clinical oversight
5	Expert	Experienced, up-to-date clinician	Aspirational growth	Supervise others

Small Group Exercise

- Entrustment scales look great for developmental assessments, so what could possibly go wrong using them?

Discuss in your groups the potential challenges with entrustment scales

Thank You

Questions and Discussion

eholmboe@acgme.org