Accreditation Council for Graduate Medical Education

Building and Implementing Assessments Fit for Purpose

ACGME

Disclosures

- Eric Holmboe works for the ACGME and receives royalties from Mosby-Elsevier for a textbook.

ACGME

Process vs .Outcome Approach

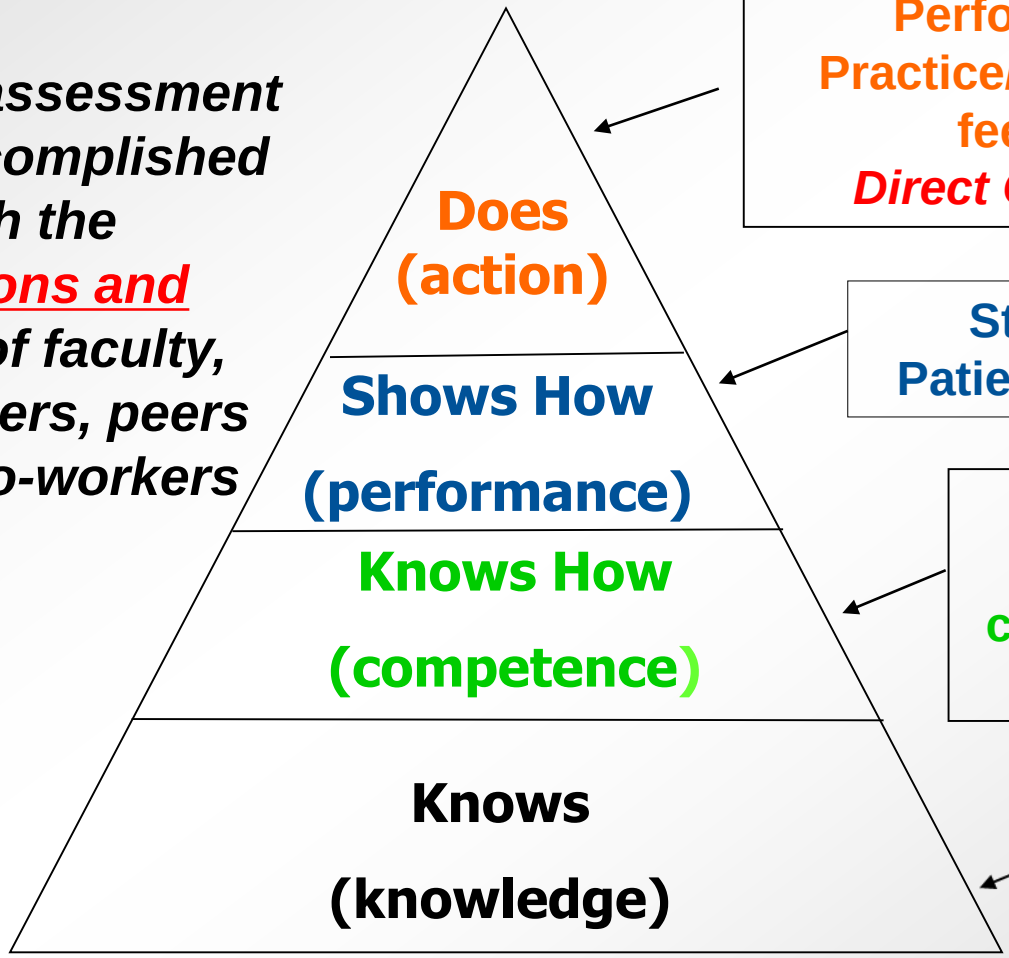
	Educational Program	
Variable	**Structure/Process**	***Competency-based***
Driving force: process	**Teacher**	***Learner***
Path of learning	**Hierarchical (Teacher→student)**	***Non-hierarchical (Teacher↔student)***
Responsibility: content	**Teacher**	***Student and Teacher***
Goal of educ. encounter	**Knowledge acquisition**	***Knowledge application***
Typical assessment tool	**Single subject measure**	***Multiple objective measures***
Assessment tool	**Proxy**	***Authentic (mimics 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***
Program completion	**Fixed time**	***Variable time***

Carracchio, et al. 2002.

ACGME

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



Performance in Practice//Multi-source feedback/ *Direct Observation*

Standardized Patients/Simulation

Diagnostic Reasoning using clinical vignettes or CSR

Multiple choice Questions

ACGME

Work-based Assessment Methods

- Performance reviews using quality measures
- Patient surveys
- Chart stimulated recall
- Direct observation
- Video reviews
- Portfolios
- Multisource feedback (can include self assessment)
- Procedural logs
- Faculty evaluations

Small Group Exercise

- List out all the work-based assessment tools you are currently using in your program
- Where do you use these tools?
- How do you prepare the assessor to use the tool?

Small Group Exercise

- Discuss first with a partner which of these assessment methods you are currently using, then ask…
  - How well do these methods help you as an educator know where your learners are in the professional development?
  - How well do these methods provide feedback to the learners to help guide them in their development?

How to Choose?

ACGME

Nothing is Perfect

Measurement Tools: Criteria

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

Context = ∑ Clinical environments where learners train

Criteria for “Good” Assessment[1]

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

[1]Ottawa Conference Working Group 2010

ACGME

Reliability and Validity

Not reliable

Not valid

Reliable

Not valid

Reliable
Valid

ACGME

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

ACGME

Educational Impact

Educational Effect

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

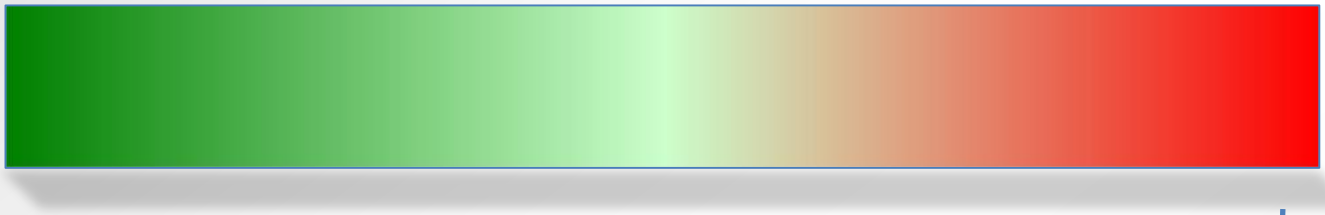
Catalytic Effect

"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

ACGME

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

Small Group Exercise

- Choose an assessment tool all of your faculty use in your program
  - 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?

Quick Break

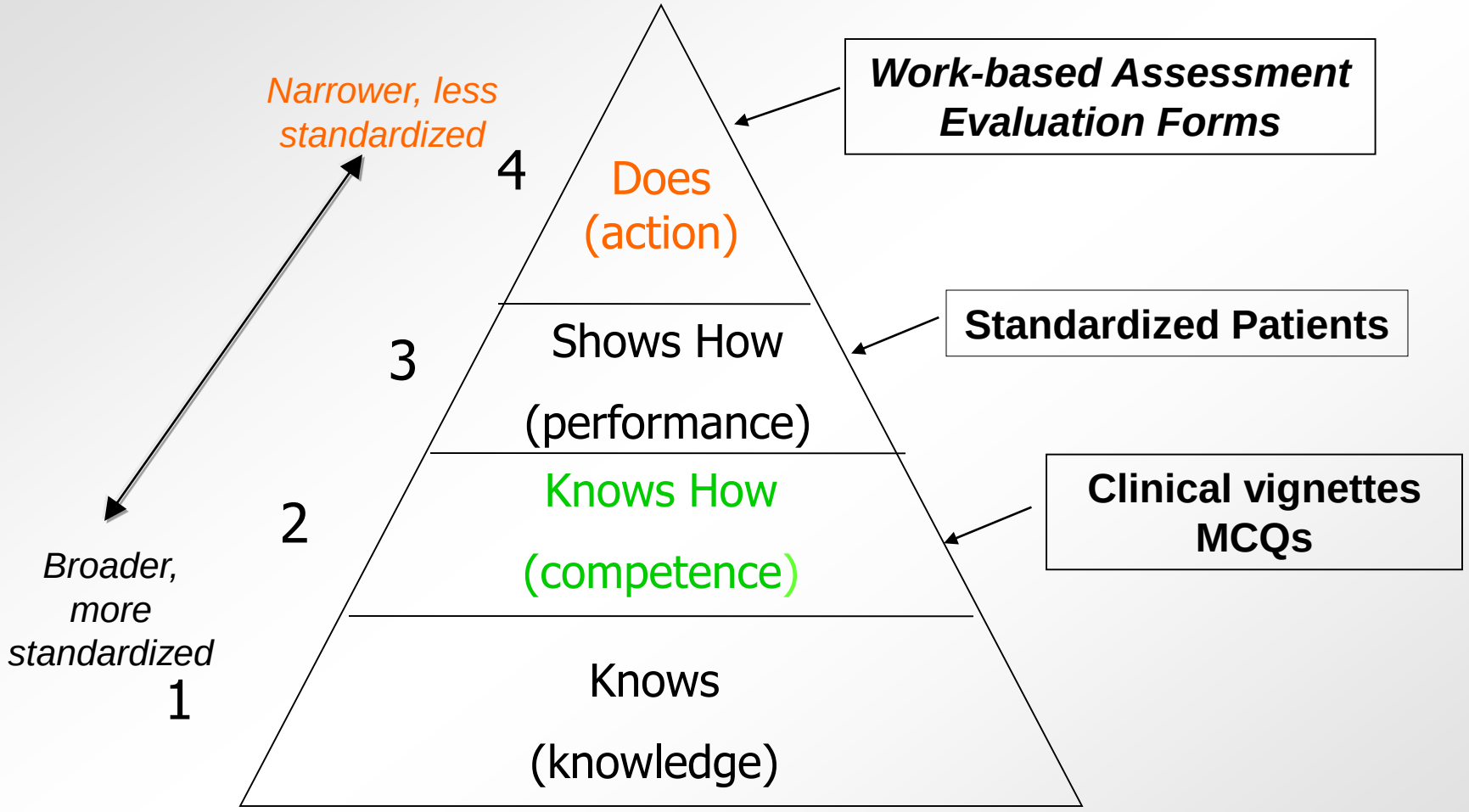
ACGME

Accreditation Council for Graduate Medical Education

Getting the Most out of Evaluation Forms

ACGME

Miller's Framework for Clinical Assessment (1990)



ACGME

Factors Influencing Faculty Ratings

- Own competencies
- Different frameworks for judgments/ratings
  - Self-as-reference (predominant)
  - Trainee level, absolute standard, practicing MD
- Contextual factors
  - Encounter complexity, resident characteristic and institutional culture
- Emotions surrounding constructive feedback
- Inference

Kogan JR, et al. Med Educ. 2011. 45(10):1048-60
Yeates P et al. Adv in Heath Sci Educ. In Press
Govaerts Adv Health Sci Educ. 2007.12(2):239-60.

Rating Scales: Frameworks

- Lack of faculty consensus on what they are evaluating
  - Lack of clear understanding of program goals and objectives
  - No standardized criteria for assessment processes
- Need for common framework for evaluation
  - Changes to evaluation form will account for less than 10% of the variance[1]

[1] Williams R. TLM. 2003

Rating Scales: Types of Anchors

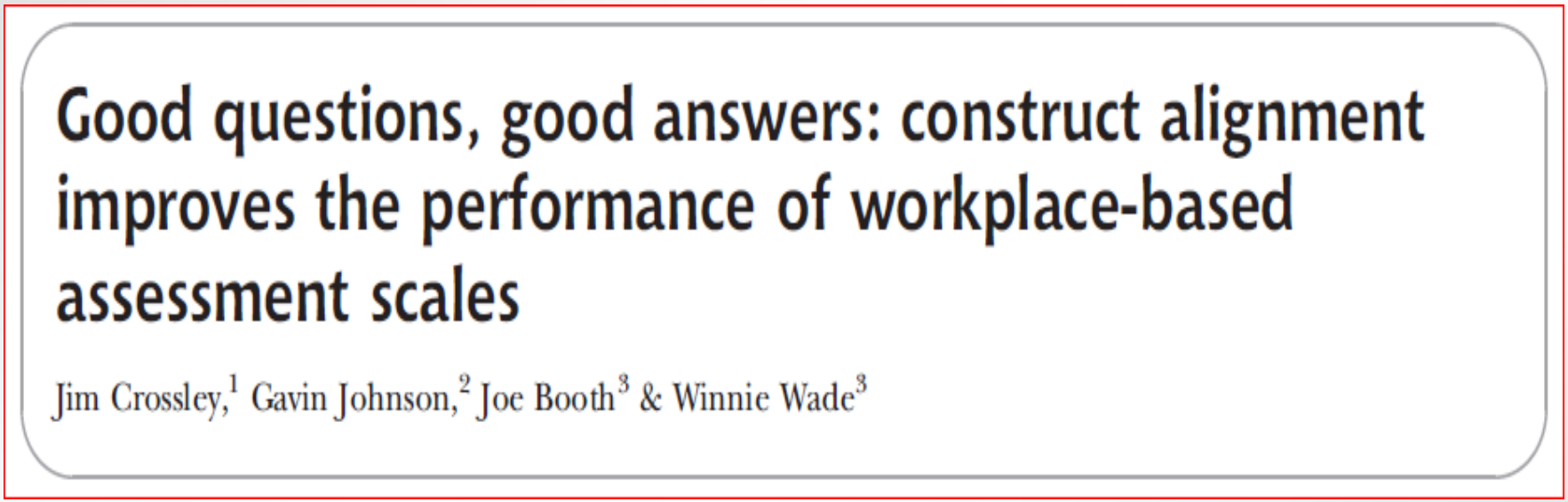
- Performance “quality”
  - E.g. Unsat-satisfactory-superior
- Frequency
  - Rarely – always
- Normative
  - Level of comparative performance
- Developmental
- Entrustment/supervision
- Narrative

These can overlap depending on purpose

Small Group Exercise

- What type of scale is used on your form?
- What has been your experience with this form?

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 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

Hybrid Example: UK MiniCEX

Rating	MiniCEX Description
Performed below level expected during Foundation Programme	Demonstrates basic consultation skills, resulting in incomplete history and/or examination findings. Shows limited clinical judgement following encounter
Performed at the level expected on completion of Foundation Programme/early Core Training	Demonstrates sound consultation skills, resulting in adequate history and/or examination findings. Shows basic clinical judgement following encounter
Performed at the level expected on completion of Core Training/early higher training	Demonstrates good consultation skills, resulting in a sound history and/or examination findings. Shows solid clinical judgement following encounter consistent with early higher training

Normative, developmental

ACGME

Zwisch Scale

- Developed for surgery
- Form of an “entrustment”/developmental scale:
  - Show and Tell
  - Active Help (“smart help”)
  - Passive (‘dumb help”)
  - Supervision only (“no help”)

DaRosa DA, Zwischenberger JB, Meyerson SL, George BC, Teitelbaum EN, Soper NJ, Fryer JP. A theory-based model for teaching and assessing residents in the operating room. J Surg Educ. 2013 Jan-Feb;70(1):24-30

Zwisch Scale Examples

TABLE 1. Zwisch Proposed Model for Teaching and Assessment in the Operating Room (Level Designated Based on Supervision Provided for the Majority of the Key Portions of the Case)

Zwisch Stage of Supervision	Attending Behaviors	Resident Behaviors Commensurate with This Level of Supervision
Show and Tell	Does majority of key portions as the surgeon Narrates the case (i.e., thinks out loud) Demonstrates key concepts, anatomy, and skills	Opens and closes First assists and observes
Cues to advancement		When first assisting, begins to actively assist (i.e., anticipates surgeons' needs)
Smart Help	Shifts between surgeon and first assist roles When first assisting, leads the resident in surgeon role (active assist) Optimizes the field/exposure Demonstrates the plane or structure Coaches for specific technical skills Coaches regarding the next steps Continues to identify anatomical landmarks for the resident	The above, plus: Shifts between surgeon and first assist roles Knows all the component technical skills Demonstrates an increasing ability to perform different key parts of the operation with attending assistance
Cues to advancement		Can execute the majority steps of procedure with active assistance
Dumb Help	Assists and follows the lead of the resident (passive assist) Coaching regarding polishing and refinement of skills Follows the resident's lead throughout the operation	The above, plus: Can "set up" and accomplish the next step for the entire case with increasing efficiency Recognizes critical transition point issues
Cues to advancement		Can transition between all steps with passive assist from faculty
No Help	Largely provides no unsolicited advice Assisted by a junior resident or an attending acting like a junior resident Monitors progress and patient safety*	The above plus: Can work with inexperienced first assistant Can safely complete a case without faculty Can recover most errors Recognizes when to seek help/advice

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

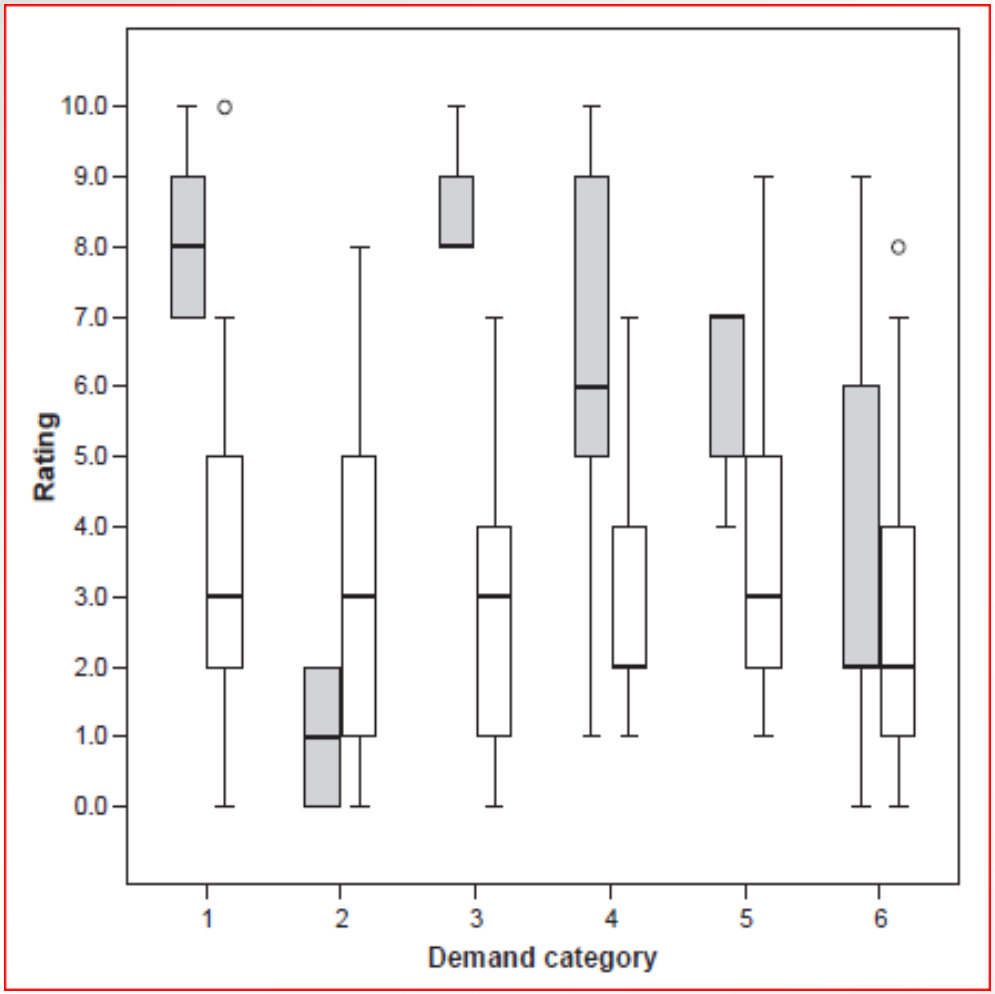


Figure 3 Comparison of NASA–Task Load Index (NASA-TLX) scores in the study subjects (grey boxes) and trainee anaesthetists (white boxes).

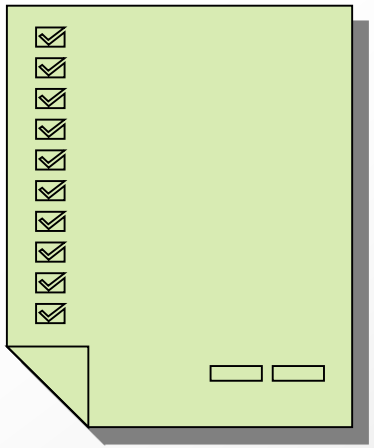
Demand categories:
1 = mental demand
2 = physical demand
3 = temporal demand
4 = performance/success
5 = effort
6 = frustration

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

ACGME

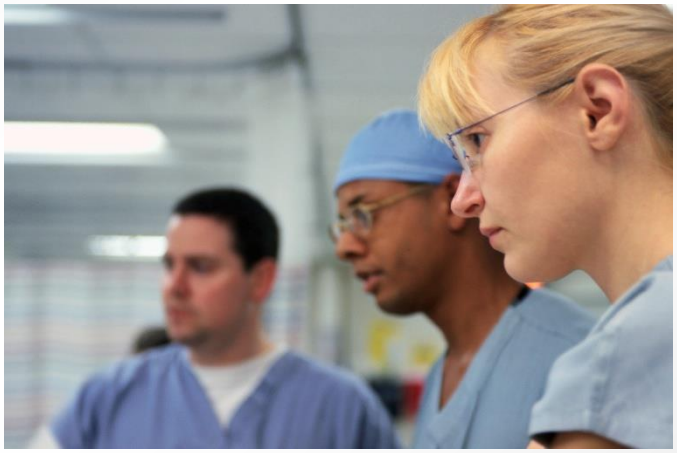
Which Assessment Forms Should We Use?

- Forms make only a small difference in the quality of assessment
  - Faculty and the encounters make a big difference
  - Forms should comport with what is to be assessed
    - Forms do not need to be long
    - Wording and scaling have only modest impact
  - Shared item pools would be very useful



From J. Norcini; AMEE 2013; FAIMER

Norcini: How do we train faculty?



Faculty development

- Methods of assessment will need to be based largely on observation
  - Faculty are the measurement instrument and they need training to develop shared mental models
- Milestones make training easier but they are not a substitute for it
  - 2-4 hour training exercise with periodic follow-up important (deliberate practice)

From J. Norcini; AMEE 2013; FAIMER

Small Group Exercise

How can you more effectively prepare your faculty to use evaluation forms and rating scales?

Thank You

Questions and Discussion

eholmboe@acgme.org

ACGME