Accreditation Council for Graduate Medical Education

Programmatic Aspects of Resident Performance Assessment

ACGME

Disclosures

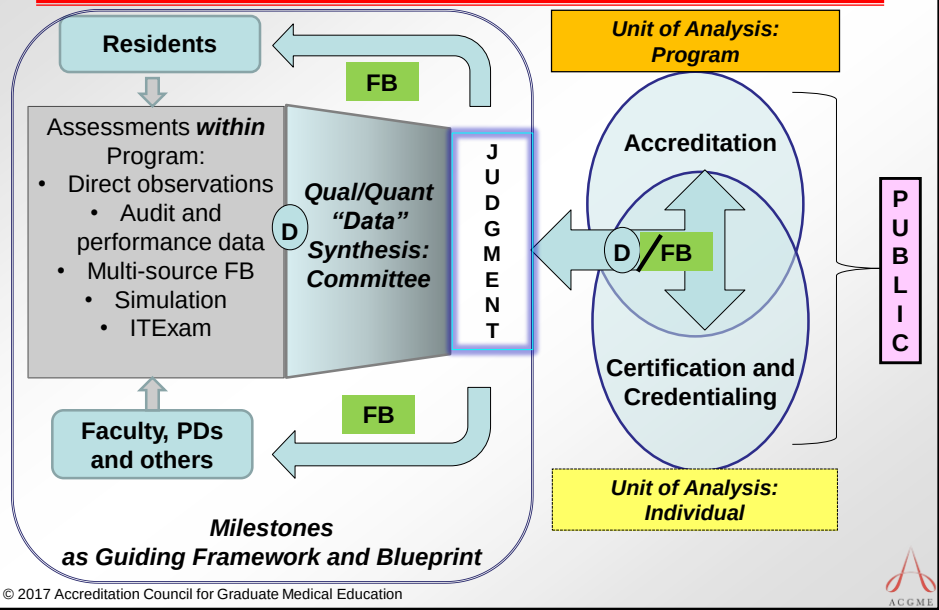
- Currently employed by the Accreditation Council for Graduate Medical Education (ACGME)
- I receive royalties from Elsevier for a textbook on assessment

ACGME

Outline

- Professional self-regulatory assessment system
- Definitions: program and systems
- Structure, process and outcomes
- Assessment structure and process
- Rethinking rating scales
- Group process in assessment

Self-Regulatory Assessment "System"



What's a Program?

- A group of related activities managed in a coordinated manner to obtain benefits and control NOT available from managing them individually.
- The activities have a common goal or success "vision" under integrated management.

What's a Program?

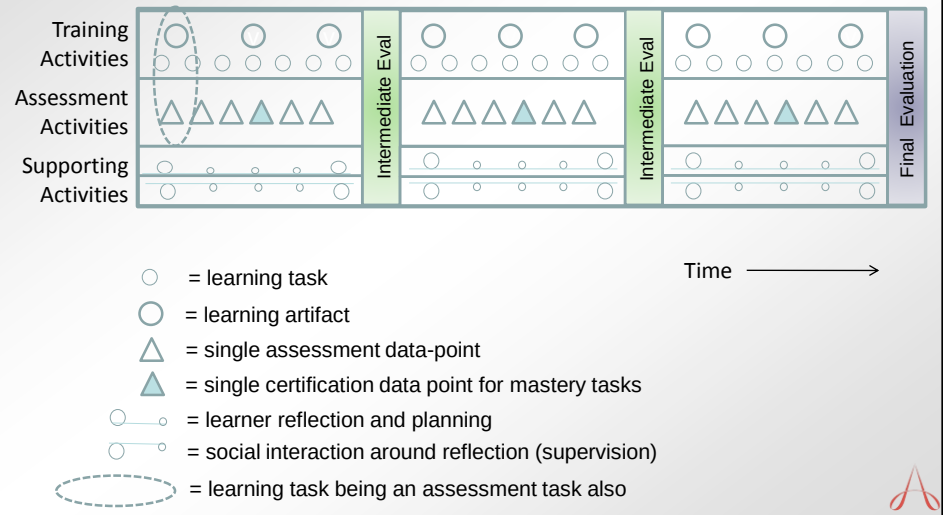
- These activities consist of:
  - People
  - Technology, and
  - Processes

  …aimed at implementing significant educational and clinical care delivery change.

  Assessment is an essential programmatic activity.

Model For Programmatic Assessment

(With permission from CPM van der Vleuten)



What is a "System?"

- Deming:
  - "Two or more interdependent parts that work together to accomplish a shared aim."
- Key concepts:
  - Working together, interactional and interdependent.
  - CBME as a system is not simply the sum or average of the curricular and assessment components, but the product of all the interactions among the components.

Complex Adaptive Systems

"A complex adaptive system is a collection of individual agents with freedom to act in ways that are not always totally predictable, and whose actions are interconnected so that one agent's actions changes the context of other agents."

Plsek P, Greenhalgh T. The challenge of complexity in health care. BMJ 15 Sept 2001; 323:625-628.

Complex Systems – Basic Rules

- ***System*** – specify what the system will provide
- ***Pathway*** – specify who will provide what to whom
- ***Connection*** – specify how those responsible for successive stages should communicate
- ***Work Activity*** – specify how work is accomplished
- ***Improvement*** – problems are immediately solved close to their occurrence

Adapted from E. Armstrong; HMI 2009.

Assessment Program as Subsystem

- An assessment program should function as a subsystem primarily as:
  - A group of people who work together on a regular basis to perform assessment and provide feedback to a population of trainees over a defined period of time
- The assessment program must ultimately produce a valid entrustment judgment

Assessment Program as Subsystem

- This group shares:
  - Educational goals and outcomes
  - Linked assessment processes
  - Information about trainee performance
  - A desire to produce a trainee truly competent (at a *minimum*) to enter practice or fellowship at the end of training

CAS Small Group Exercise

- Using the System Grid, fill-out the "basic rules" for current assessment programs

Assessment Program as Subsystem

- The assessment program has a *structure* to carry out assessment *processes* that produce an *outcome*

Measurement Model

Donabedian Model (adapted)

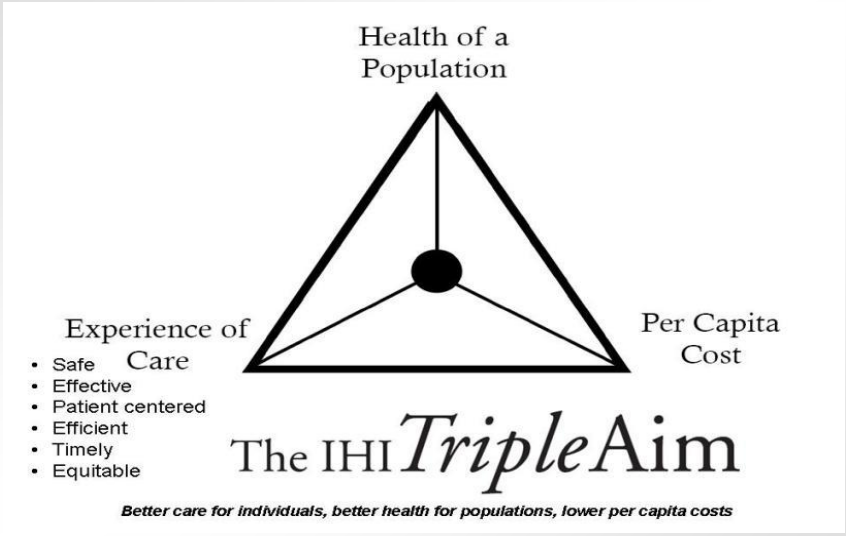
- Structure: the way a training program is set up and the conditions under which the program is administered
  - Organization, people, equipment and technology
- Process: the activities that result from the training program
- Outcomes: the changes (desired or undesired) in individuals or institutions that can be attributed to the training program

Donabedian: Producing Quality

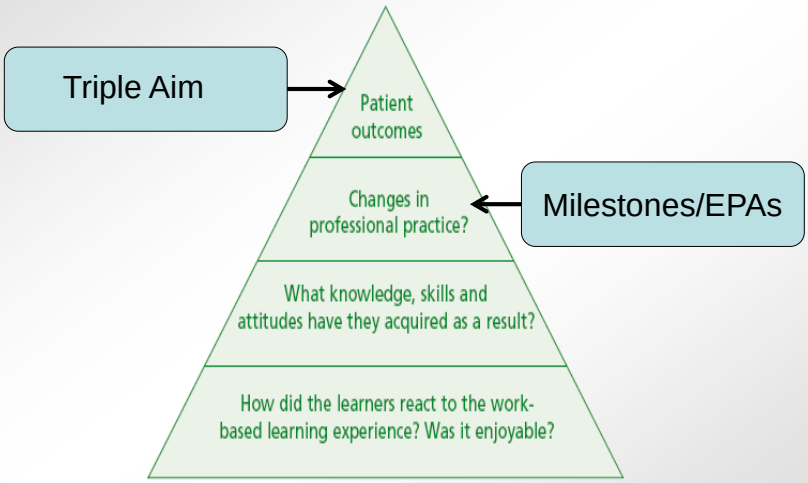
Structure X Process ≈ Outcomes

Institutional Environment & Performance

What Are The Outcomes?



Kirkpatrick Model: Medical Program Perspective



National Health Service – UK.
http://www.wipp.nhs.uk/tools_gpn/unit6_education.php

Needed Perspective

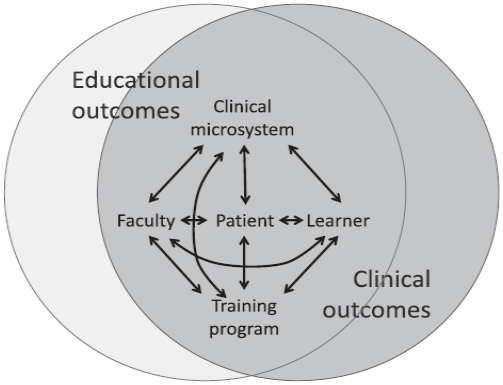
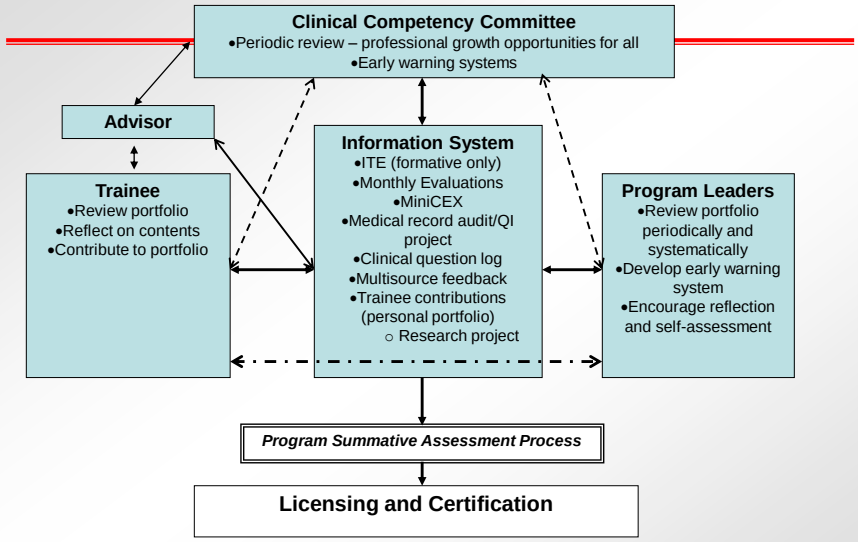


Figure 2 Schematic of the proposed framework for academic faculty perspective and educational design of graduate medical education training programs, where both educational and clinical outcomes are centered around the patient. This reorganization recognizes that (1) the dynamic interplay between the faculty, learner, training program, and clinical microsystem ultimately influences the quality of physician that emerges from the training program *and* the environment, and (2) patient outcomes relate to the quality of education and the success of clinical microsystems.

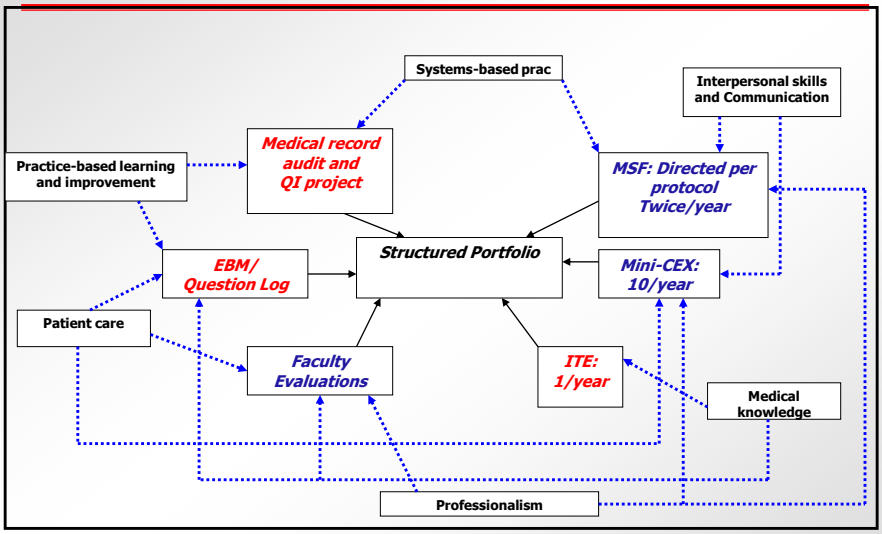
Producing Quality in GME

Assessment Structure

Assessment During Training: Structures



Multi-modal Assessment



■ Trainee-directed

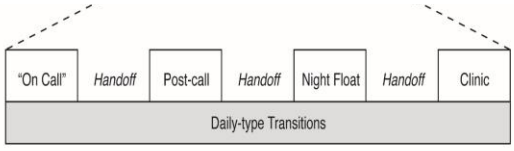
■ Direct observation

Multi-modal Assessment

- No single "tool" sufficient to evaluate all components of competence
  - Pick best combination that meets your needs in context of local resources
- Evaluation tools and faculty
  - Nothing ever works perfectly
  - Embed CQI into evaluation subsystem

Medical Education Architecture[1]

How does your curricular design affect the nature and quality of your assessment program?



[1]Holmboe E, Ginsburg S, Bernabeo E. The rotational approach to medical education: time to confront our assumptions. Med Educ. 2011; 45(1):69-80.

Producing Quality in GME

Assessment Process

Effective Assessment Processes…

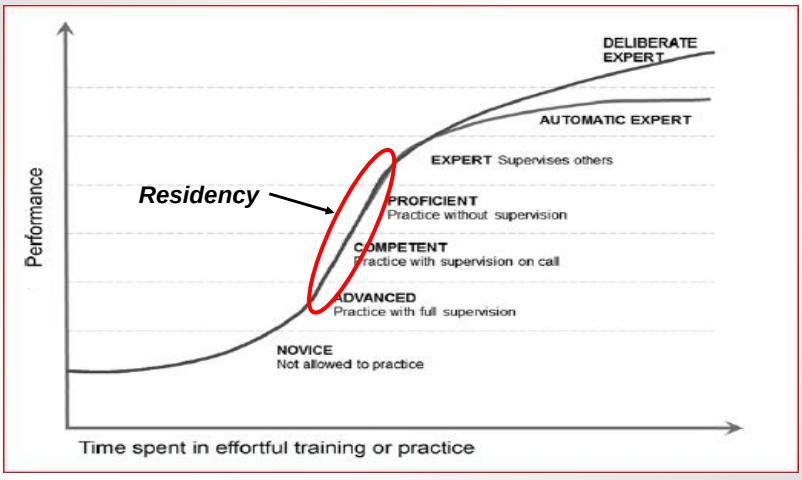
…must start with a shared mental model of the outcomes.

- That is a major purpose of the Milestones
  - Create developmental language to explain and describe the competencies

Milestones

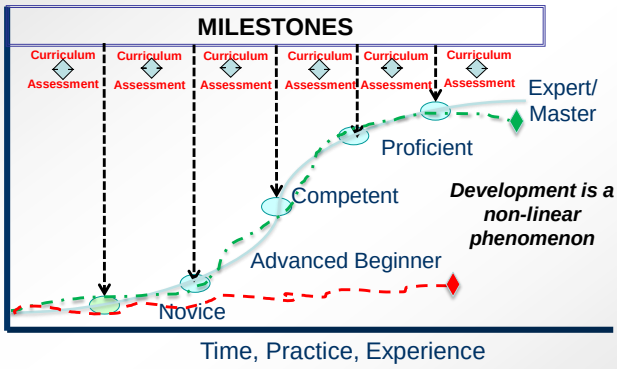
- By definition a milestone is a significant point in development.
- Milestones should enable the learner and training program to know an individual's trajectory of competency development.
- They serve as educational outcomes

Learning Curves



From Pusic, et. al. Acad Med. 2014

Professional Development: Dreyfus and the Learning Curve



Dreyfus SE and Dreyfus HL. 1980
Carraccio CL et al. Acad Med 2008;83:761-7

Small Group Exercise

- What have you done and/or are currently doing to help your faculty attain a shared mental model of professional development using competencies and milestones?

Choosing the Right Assessment Tools

"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?

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 Microsystems

Criteria for "Good" Assessment[1]

- **Validity or Coherence**
- **Reproducibility or Consistency**
- **Equivalence**
- **Feasibility**
- **Educational effect**
  - **Learning that occurs in preparation for an assessment (e.g. certification exam; MRCP)**
- **Catalytic effect**
  - **Assessment resulting in feedback that "drives future learning forward."**
- **Acceptability**

[1]Ottawa Conference Working Group 2010

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

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
- Inference
- Idiosyncrasy

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.

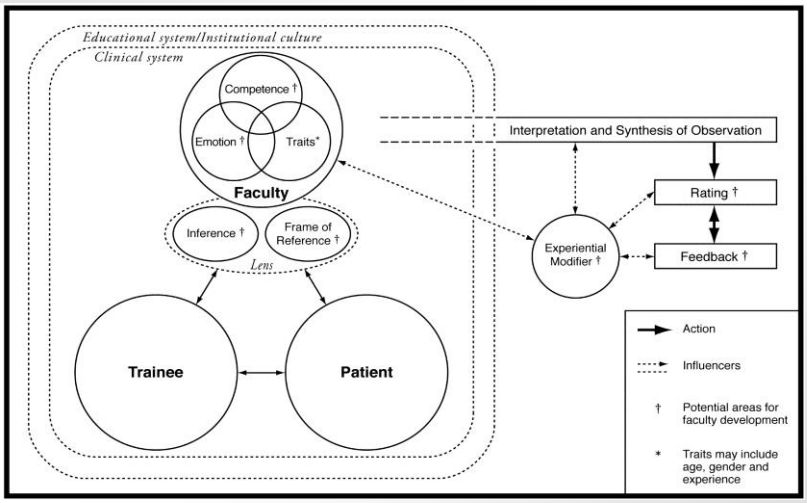
Faculty Clinical Skills - OSCE

Competency (N =44)	Mean (SD)	Range	Generaliz-ability
History Taking	65.5% (9.6%)	34% - 79%	0.80
Physical Exam	78.9% (13.6%)	36% - 100%	0.52
Counseling	77.1% (7.8%)	60% - 93%	0.33
Patient Satisfaction[1]	5.62 (0.48)	4.43 – 6.63	0.60

[1]On 7-point scale

Kogan JR, Hess BJ, Conforti LN, Holmboe ES. What Drives Faculty Ratings of Residents' Clinical Skills? The Impact of Faculty's Own Clinical Skills. Acad Med. 2010;85(10 Suppl):S25-8

Direct Observation: A Conceptual Model



Kogan JR, et al. Med Educ. 2011

Rating Scales: Types of Anchors

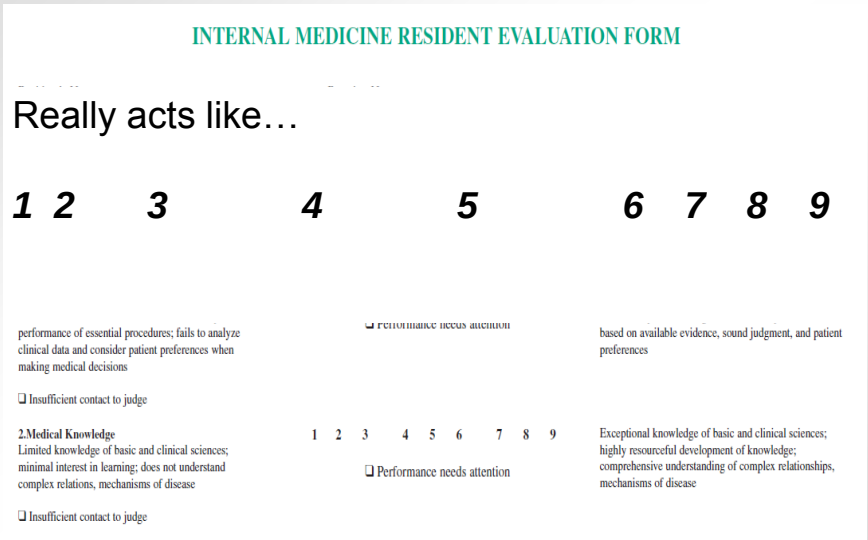
- Adjectival - performance "quality"
  - E.g. Unsat-satisfactory-superior
- Frequency
  - Rarely – always
- Normative
  - Level of comparative performance
- Developmental
- Entrustment/supervision
- Narrative

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

performance of essential procedures; fails to analyze clinical data and consider patient preferences when making medical decisions

❑ Insufficient contact to judge

❑ Performance needs attention

based on available evidence, sound judgment, and patient preferences

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

❑ Insufficient contact to judge

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

How Scales Actually Get Used

1	2	3	4	5	6	7	8	9
Not What I Do			Close to What I Do			What I Do (or better)		
SELF								
Below Expectation			At Expectation			Exceeds Expectation		
NORMATIVE								
????			????			????		
GESTALT								
Missing evidence based elements			Most evidence based elements			All evidence based elements		
BEST PRACTICE								

Kogan JR et al. Med Educ. 2011;45:1048-60

How Scales Actually Get Used

In GME context:

- Self was most common
- Normative and gestalt were also described by faculty
- Best practice (criterion-referenced) was ***rare***

Kogan JR et al. Med Educ. 2011;45:1048-60

Implications

Sources of Potential Error

Variable frames of reference
Assessors' clinical skills
Inference
idiosyncrasy

Unusable Assessments

Learners
Patients
Assessors

Rethinking Rating Scales

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.

Zwisch Scale

- Developed for surgery
- Form of a 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

ACGME

Revised mini-CEX

1	2	3	4	5
Learner can be present but only as observer (i.e. The learner cannot perform this skill. Learner can be present, but only as observer)	Learner can practice skill with direct supervision (supervisor in room) (i.e. I need to watch the learner perform the skill in real time)	Learner can practice skill with indirect supervision (supervision available within minutes) (i.e. I don't need to watch the learner in the room, but I am going to reassess the patient/confirm findings with the patient)	Unsupervised practice allowed (distant oversight) (i.e. I don't need to watch the learner but I am available if the learner comes for help or to provide feedback)	Learn can supervise junior learners in the skill (i.e The learner can supervise others)

ACGME

Small Group Exercise

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

Discuss in your groups what are the potential challenges with entrustment scales?

Safe Patient Care as Frame of Reference

- Importance of appropriate supervision
- Entrustment

Trainee performance* X
Appropriate level of supervision**
Must = Safe, effective patient-centered care

* a function of level of competence in context
**a function of attending competence in context

Entrustment in Surgery

Thresholds of principles and preference:

- Principles
  - Non-negotiable procedural approach
- Optional preferences
  - Idiosyncratic
  - Meaningfully different between 2 surgeons
    - Learner chooses the best path that "works for them" and its "OK"

Apramiam T, et. al. How do thresholds of principle and preference influence surgeon assessments of learner performance. Ann Surg. 2017; online first.

Entrustment in Surgery

Summary of article:

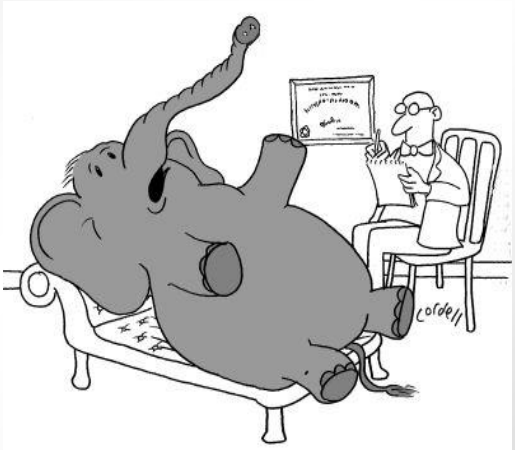
- Agreement on generic (non-negotiable) principles (11 attendings)
- Little agreement on ***personal*** application of principles
  - *"What is a principle for one surgeon may well be a preference for another"*
- Wide variability in what was deemed inconsequential preference

Apramiam T, et. al. How do thresholds of principle and preference influence surgeon assessments of learner performance. Ann Surg. 2017; online first.

© 2017 Accreditation Council for Graduate Medical Education

The Frame of Reference Problem

Several studies demonstrate that faculty heavily use self as the frame of reference in judging competence and entrustment. Assessment approaches assume faculty "self" is competent.



Useful Dictums

- Faculty, not assessment tools and scales, are the true assessment instrument
- Assessment tools are only as good as the individuals using them
- Assessment depends predominantly on observation

Therefore faculty need training in observation and assessment!

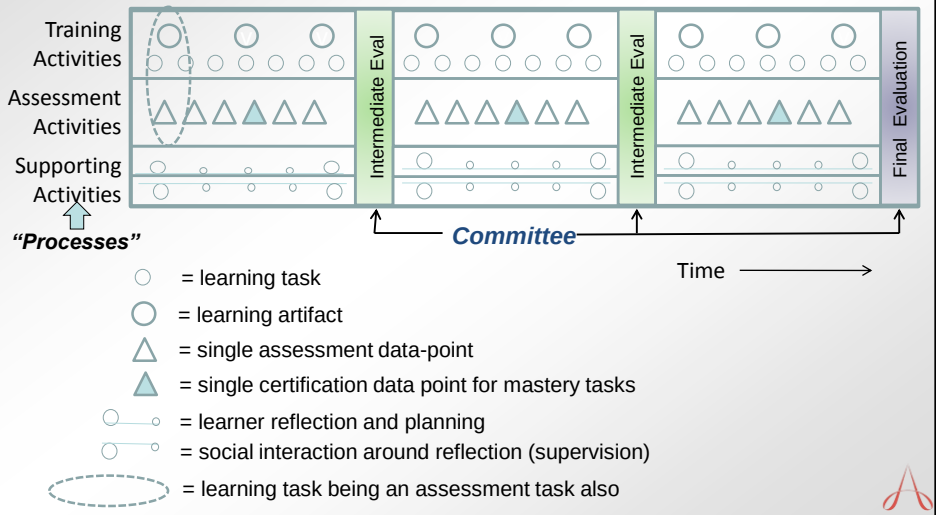
Small Group Exercise

How can you use the concept of shared mental models to more effectively prepare your faculty to use evaluation forms and rating scales?

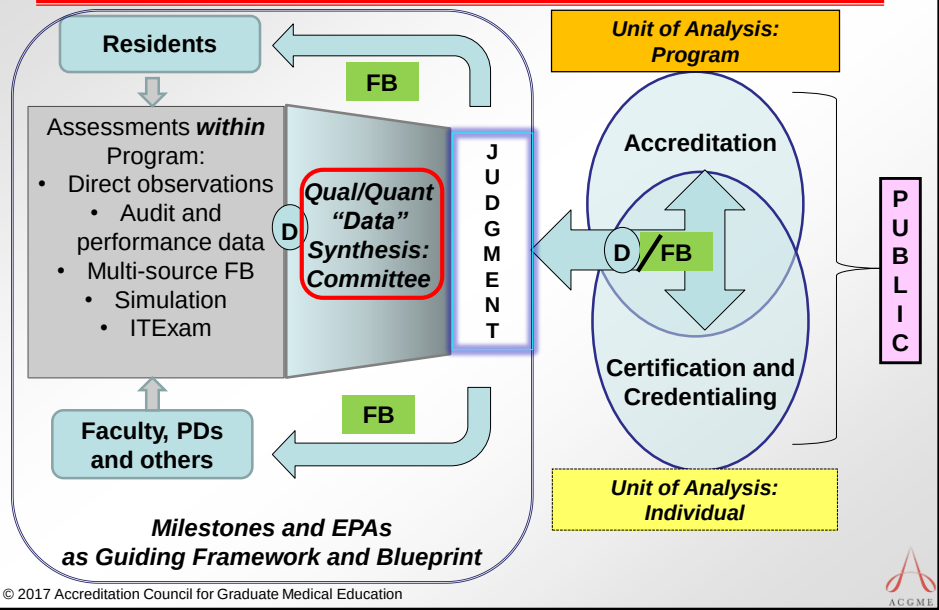
Group Process in Assessment

Model For Programmatic Assessment

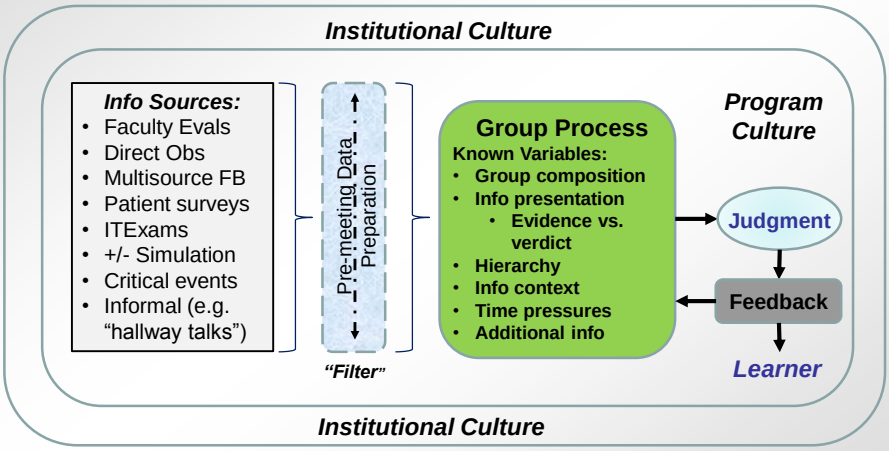
(With permission from CPM van der Vleuten)



Self-Regulatory Assessment "System"



Outline and Flow of CCC Processes

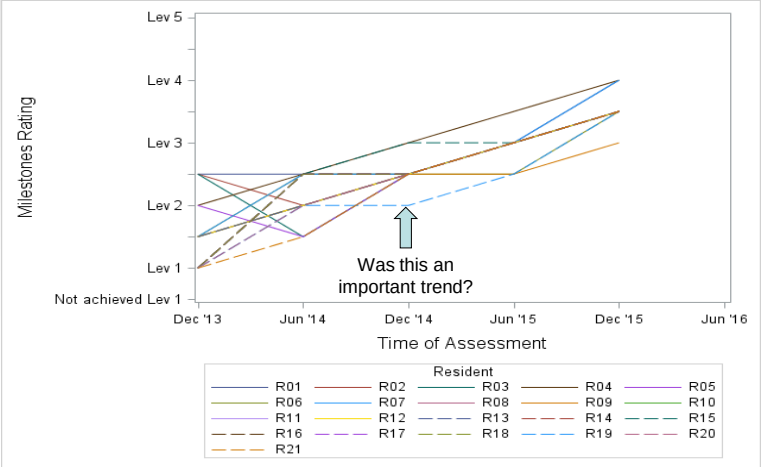


Small Group Exercise

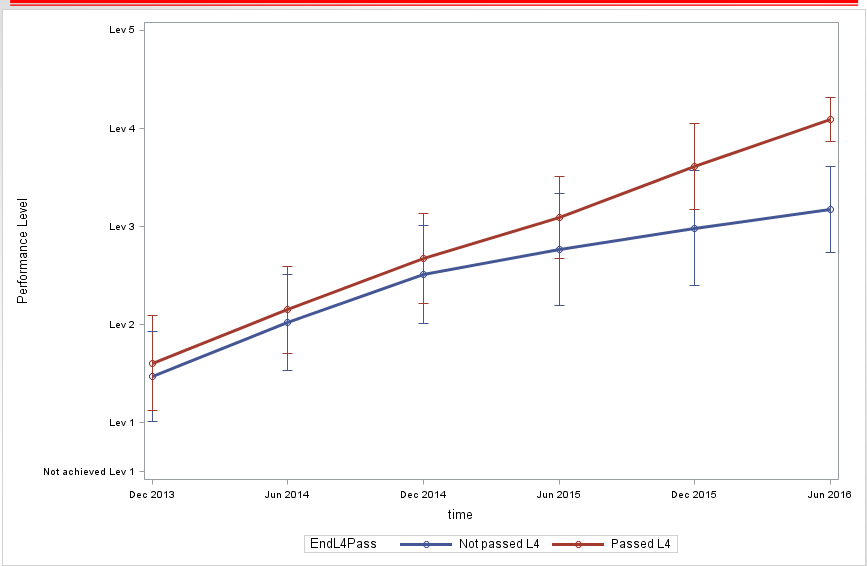
- How do you prepare the assessment information for the CCC?
- How is the data (information) turned into knowledge?
- Do you review the data longitudinally?

Getting Started: Track Progression Over Time

Within Program example - PC1 (Emergency Medicine)



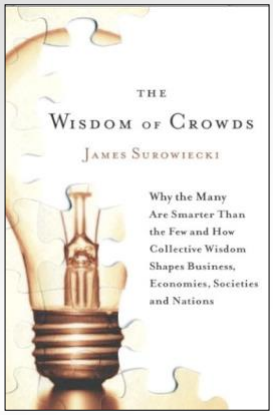
Mean Trajectory by L4 Passage



Group Decision Making

- Key Issues
  - What is the environment in which the committee performs its work?
    - What is the local culture?
      - Groups within groups
    - What is the medical culture of your institution?
  - What are the effects of hierarchy on group decision making?
    - Berg: Medicine one of the most hierarchical of all professions
  - Single variable of effectiveness: extent to which people are willing to say “positive” and “negative” comments and observations in a group

The Wisdom of Crowds



- The wisdom of many is often better than the wisdom of the few
- To maximize the probability of good judgments:
  - Sample
  - “Independence”
  - Diversity

  are important…

Basic Committee Principles

- Evidence-based versus verdict-based "jury"
  - Start and review all evidence *before* a decision
    - *Do not start* with a conclusion/decision
  - Confirmation bias
- Be careful not to emphasize consensus over dissent
  - Minority opinions, even if "wrong", still helpful
  - Be sure all voices are "heard" and watch carefully for negative effects of hierarchy

Theories Supporting Group Process

- Social decision scheme theory (Stasser)
  - Social decision schemes are the methods used by a group to combine individual responses into a single group decision
- Conversation theory (Pask; Pangaro)
  - Creating understanding and meaning through dialogue
- Paradox of Group Life (Berg)
  - Paradox an inherent part of group life

What Empowers a CCC?

- ***Shared Mental Model***
  - The most important aspect of preparing for a CCC meeting is to make sure the members develop a shared mental model of what resident/fellow performance looks like, and understand their roles and responsibilities on the committee, as well as how the CCC operates to judge resident/fellow performance.
  - Faculty members should reach a common understanding on the meaning of the narratives of each milestone in the context of their specialty. This may require "meeting before the meeting."
  - A shared mental model is facilitated by having a written description of the CCC process, and providing faculty development for committee members

Small Group Exercise

- Using the Hauer table:
  - Review the key aspects of effective group process and make some brief notes on how you think your CCC is performing
  - Discuss your observations with your colleagues
  - What, if any, changes will you consider making to your CCC process?

Summary: Creating Assessment Programs

- Use systems thinking
- 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

Questions