Programmatic Aspects of Resident Performance Assessment

Disclosures

- Currently employed by the Accreditation Council for Graduate Medical Education (ACGME)

- I receive royalties from Elsevier for a textbook on assessment
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”

Assessments within Program:
- Direct observations
- Audit and performance data
- Multi-source FB
- Simulation
- ITExam

Residents

Faculty, PDs and others

Milestones as Guiding Framework and Blueprint

Qual/Quant “Data” Synthesis: Committee

JUDGMENT

D/FB

Unit of Analysis: Program

Accreditation

Certification and Credentialing

Unit of Analysis: Individual

© 2017 Accreditation Council for Graduate Medical Education
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.”


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 $\approx$ Outcomes

Institutional Environment & Performance

© 2017 Accreditation Council for Graduate Medical Education
What Are The Outcomes?

Health of a Population

Experience of Care
- Safe
- Effective
- Patient centered
- Efficient
- Timely
- Equitable

Per Capita Cost

The IHI Triple Aim

Better care for individuals, better health for populations, lower per capita costs

Kirkpatrick Model:
Medical Program Perspective

Triple Aim

Patient outcomes

Changes in professional practice?

What knowledge, skills and attitudes have they acquired as a result?

How did the learners react to the work-based learning experience? Was it enjoyable?

Milestones/EPAs

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

© 2017 Accreditation Council for Graduate Medical Education
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

Clinical Competency Committee
- Periodic review – professional growth opportunities for all
- Early warning systems

Information System
- ITE (formative only)
- Monthly Evaluations
- MiniCEX
- Medical record audit/QI project
- Clinical question log
- Multisource feedback
- Trainee contributions (personal portfolio)
  - Research project

Program Leaders
- Review portfolio periodically and systematically
- Develop early warning system
- Encourage reflection and self-assessment

Program Summative Assessment Process

Licensing and Certification

Multi-modal Assessment

Structured Portfolio
- Medical record audit and QI project
- EBM/Question Log
- Faculty Evaluations
- ITE: 1/year
- Mini-CEX: 10/year
- MSF: Directed per protocol Twice/year

Trainee-directed
- Practice-based learning and improvement
- Patient care

Direct observation
- Systems-based prac
- Interpersonal skills and Communication
- Medical knowledge
- Professionalism
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?

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

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

- 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

1Ottawa 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


Faculty Clinical Skills - OSCE

<table>
<thead>
<tr>
<th>Competency</th>
<th>Mean (SD)</th>
<th>Range</th>
<th>Generalizability</th>
</tr>
</thead>
<tbody>
<tr>
<td>N = 44</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>History Taking</td>
<td>65.5% (9.6%)</td>
<td>34% - 79%</td>
<td>0.80</td>
</tr>
<tr>
<td>Physical Exam</td>
<td>78.9% (13.6%)</td>
<td>36% - 100%</td>
<td>0.52</td>
</tr>
<tr>
<td>Counseling</td>
<td>77.1% (7.8%)</td>
<td>60% - 93%</td>
<td>0.33</td>
</tr>
<tr>
<td>Patient Satisfaction</td>
<td>5.62 (0.48)</td>
<td>4.43 – 6.63</td>
<td>0.60</td>
</tr>
</tbody>
</table>

1 On 7-point scale

Direct Observation: A Conceptual Model

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

Really acts like…

1 2 3 4 5 6 7 8 9
### How Scales Actually Get Used

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not What I Do</td>
<td>Close to What I Do</td>
<td>What I Do (or better)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SELF</strong></td>
<td>Below Expectation</td>
<td>At Expectation</td>
<td>Exceeds Expectation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>NORMATIVE</strong></td>
<td>????</td>
<td>????</td>
<td>????</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>GESTALT</strong></td>
<td>Missing evidence based elements</td>
<td>Most evidence based elements</td>
<td>All evidence based elements</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

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

Sources of Potential Error
Variable frames of reference
Assessors' clinical skills
Inference
Idiosyncrasy

Unusable Assessments
Learners
Assessors
Patients

Rethinking Rating Scales
Construct Aligned Scales

“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.”


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

**Entrustment Scale: O-SCORE**

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

<table>
<thead>
<tr>
<th>Level</th>
<th>Descriptor</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>“I had to do” (i.e., requires complete hands on guidance, did not do, or was not given the opportunity to do)</td>
</tr>
<tr>
<td>2</td>
<td>“I had to talk them through” (i.e., able to perform tasks but requires constant direction)</td>
</tr>
<tr>
<td>3</td>
<td>“I had to prompt them from time to time” (i.e., demonstrates some independence, but requires intermittent direction)</td>
</tr>
<tr>
<td>4</td>
<td>“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)</td>
</tr>
<tr>
<td>5</td>
<td>“I did not need to be there” (i.e., complete independence, understands risks and performs safely, practice ready)</td>
</tr>
</tbody>
</table>


---

**Zwisch Scale**

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

Zwisch Scale Examples

<table>
<thead>
<tr>
<th>Zwisch Stage of Supervision</th>
<th>Attending Behaviors</th>
<th>Resident Behaviors Commensurate with This Level of Supervision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Show and Tell</td>
<td>Does majority of key portions as the surgeon Narrates the case (i.e., thinks out loud) Demonstrates key concepts, anatomy, and skills</td>
<td>Opens and closes First assists and observes</td>
</tr>
<tr>
<td>Cues to advancement</td>
<td></td>
<td>When first assisting, begins to actively assist (i.e., anticipates surgeons’ needs)</td>
</tr>
<tr>
<td>Smart Help</td>
<td>Shuts 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</td>
<td>Can execute the majority steps of procedure with active assistance</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The above, plus: Shuts 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</td>
</tr>
<tr>
<td>Cues to advancement</td>
<td></td>
<td>Can transition between all steps with passive assist from faculty</td>
</tr>
<tr>
<td>Dumb Help</td>
<td>Assists and follows the lead of the resident (passive assist) Coaching regarding polishing and refinement of skills</td>
<td>Can work with inexperienced first assistant Can safely complete a case without faculty Can recover most errors</td>
</tr>
<tr>
<td></td>
<td>Follows the resident’s lead throughout the operation</td>
<td>Recognizes critical transition points</td>
</tr>
<tr>
<td>Cues to advancement</td>
<td></td>
<td>Learns how to select best technique (i.e. I don’t need to watch the learner but I am available if the learner comes for help or to provide feedback)</td>
</tr>
<tr>
<td>No Help</td>
<td>Largely provides no unsolicited advice Assisted by a junior resident or an attending acting like a junior resident</td>
<td>Learns how to select best technique (i.e. I don’t need to watch the learner but I am available if the learner comes for help or to provide feedback)</td>
</tr>
<tr>
<td></td>
<td>Monitors progress and patient safety*</td>
<td>Learns how to select best technique (i.e. The learner can supervise others)</td>
</tr>
</tbody>
</table>

Revised mini-CEX

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

© 2017 Accreditation Council for Graduate Medical Education
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**

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

Must = Safe, effective patient-centered care
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”


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

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.

"Whenever I walk in a room, everyone ignores me."

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)

```
Training Activities
Assessment Activities
Supporting Activities

“Processes”

○ = learning task
△ = learning artifact
△ = single assessment data-point
△ = single certification data point for mastery tasks
△ = learner reflection and planning
△ = social interaction around reflection (supervision)
△ = learning task being an assessment task also
```

Self-Regulatory Assessment “System”

- Assessments within Program:
  - Direct observations
  - Audit and performance data
  - Multi-source FB
  - Simulation
  - ITExam

© 2017 Accreditation Council for Graduate Medical Education
Outline and Flow of CCC Processes

Institutional Culture

**Info Sources:**
- Faculty Evals
- Direct Obs
- Multisource FB
- Patient surveys
- ITExams
- +/- Simulation
- Critical events
- Informal (e.g. "hallway talks")

Group Process

**Known Variables:**
- Group composition
- Info presentation
  - Evidence vs. verdict
  - Hierarchy
  - Info context
  - Time pressures
  - Additional info

Program Culture

Judgment

Feedback

Learner

Institutional Culture

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)

Was this an important trend?

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


Questions