Developing Skills at Making Observations

Lessons from Faculty Development and Rater Cognition Research

Eric S. Holmboe
Jennifer R. Kogan

Roadmap

1. Define workplace based assessment and the theories supporting direct observation
2. Identify barriers to high quality, frequent direct observation and solutions to overcome them

Workplace Based Assessment

- Assessment of day-to-day practice in the authentic clinical environment
- Assessment of what doctors actually do in practice


Miller’s Assessment Pyramid

- Expertise
- Authenticity

DOES
SHOWS HOW
KNOWS HOW
KNOWS

Direct Observation
Workplace Based Assessment
Standardized patients
Problem based learning
Multiple choice exam

Miller GE. Acad Med. 1990; 65:S63-7

Theories Supporting Direct Observation

- Importance and state of clinical skills
- Development of expertise
- Role in competency based medical education
- Necessity in supervision

Clinical Skills Matter

- History leads to diagnosis > 80% of the time
- Even in era of technology
- Required to avoid unnecessary testing
- Faulty data gathering common source of diagnostic errors

High Quality Care

- Timely
- Efficient
- Equitable
- Safe
- Effective
- Patient Centered

Outcomes of Patient Centered Care

- Improves communication
- Promotes patient involvement in care
- Increases patient knowledge and self-efficacy
- Creates positive relationships with the provider
- Improves adherence
- Improves well-being
- Improved outcomes
- Decreased costs

State of Clinical Skills

- Trainees
  - Wide variability in graduating students’ clinical skills measured as MS4s or starting internship
  - History taking
  - Exam

- Practicing physicians
  - Variability in physical exam skills
  - Missing elements of informed decision making

Why the Gap??

- Communication is a sophisticated procedure
- Needs to be taught and honed throughout one’s career
- Skills of patient-centered communication are rarely taught or practiced

Direct Observation to Assess Core Skills

- Legitimizes the subject
- Sends message skills are important
- Ensures assessment of essential skills

What Do They Have in Common?
Design and Sequencing of Training Activities

- Monitor students’ development
- Design and select training tasks for individual students

Dreyfus & Dreyfus Development Model

The Role of the Coach

- “They observe, they judge, and they guide”
- “That one twenty-minute discussion gave me more to consider and work on than I’d had in the past five years”
- “Medical practice is largely unseen by anyone who might raise one’s sights. I’d had no outside ears and eyes.”

Atul Gawande, New Yorker 10/3/2011

Miller’s Assessment Pyramid

In-Training Performance Assessment

- Assessment in authentic situations
  - Learners’ ability to combine knowledge, skills, judgments, attitudes in dealing with realistic problems of professional practice
- Assessment in day to day practice
  - Enables assessment of a range of essential competencies, some of which cannot be validly assessed otherwise

Observation and Safe Patient Care

Safe, effective patient-centered care

- Appropriate level of supervision**
  - a function of attending competence in context
- Trainee performance*
  - a function of level of competence in context
Utility Elements of Assessment

- Validity
- Reliability
- Educational impact
- Acceptability
- Cost effectiveness
- Identify gap between resident performance and desired outcome

Ensure high quality patient
Inform supervision

Trainee Learning & Assessment

Summary: Reasons for DO

Patient Safety
Assessment
Coaching

Problems with Performance Assessment

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

Multiple forms and types of error:

- Correlational errors
  - Halo effect
  - Horn effect
  - Ratings based mostly on perceived knowledge and personality
- Distributional errors
  - Leniency error ("Doves")
  - Severity error ("Hawks")
  - Central tendency
Key Issues: Human Limitations

- Limitation in working memory and mental processing
- Subconscious processes
  - Bias and stereotyping
- Cognitive Load

Yeates: Contrast effects


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

Key Issues: Individual Effects

- Inference
- Variability among faculty
  - Strengths and weaknesses
    - Clinical
    - Educational
    - Assessment
  - Variable frames of reference
- Idiosyncrasy
- Contextual factors
High Level Inference

The Problem with Inference

- Inferences are not recognized
- Inferences are rarely validated for accuracy
- Inferences can be wrong

Types of Inference about Residents

- Skills
  - Knowledge
  - Competence
  - Work-ethic

- Prior experiences
  - Familiarity with scenario

- Feelings
  - Comfort
  - Confidence
  - Intentions
  - Ownership

- Personality

- Culture

Frames of Reference

<table>
<thead>
<tr>
<th></th>
<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>Unsatisfactory</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Satisfactory</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Superior</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**ORDINAL**

- Below Expectation
- At Expectation
- Exceeds Expectation

**NORMATIVE**

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>????</td>
<td>????</td>
<td>????</td>
<td></td>
</tr>
</tbody>
</table>

**GESTALT**

Assessors' Own Clinical Skills

Variable and sometimes deficient

- History taking
- Physical exam
- Counseling/shared decision making
- Patient centered communication

1 2 3 4 5 6 7 8 9

Not What I Do Close to What I Do What I Do (or better)

**SELF**

Missing evidence based elements Most evidence based elements All evidence based elements

**BEST PRACTICE**

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

Paauw DS et al. JAMA 1995;274:1380-5
Vukanovick-Criley JM et al. Arch Intern Med. 2006;166:610-16
Levinson W. BMJ Qual Saf 2011;20:823-30
Faculty OSCE Clinical Skills

<table>
<thead>
<tr>
<th>Competency</th>
<th>Mean (SD)</th>
<th>Range</th>
<th>Generalizability</th>
</tr>
</thead>
<tbody>
<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 Satisfaction¹</td>
<td>5.62 (0.48)</td>
<td>4.43 – 6.63</td>
<td>0.60</td>
</tr>
</tbody>
</table>

¹On 7-point scale

Other Factors That May Impact Ratings

- Minimal impact of demographics
- Age, gender, clinical and teaching experience
- Faculty’s own clinical skills may matter
  - Faculty with higher history and patient satisfaction performance scores provide more stringent ratings.

Idiosyncrasy: What if…

Low inter-rater reliability was found to come from experts forming different and/or conflicting, yet equivalently relevant, interpretations?

Assessment as “Saturation”

- Narrative just as, or perhaps more, meaningful as ratings through numbers
- Sampling of multiple idiosyncratic, yet meaningful, observations can lead to robust information “saturation”
- Part of rationale for the inclusion of group process in the accreditation system

Contextual Factors

- Encounter complexity
- Resident characteristics
- Institutional culture

Direct Observation: A Conceptual Model
Solution: Faculty Development

- Performance dimension training
- Synthesis to final judgment

Performance Dimension Training

- Identify specific dimensions of a competency in behavioral terms
- Discuss the criteria and qualifications required for each dimension of that competency
- Develop a SHARED MENTAL MODEL
- Achieve evidence-based standardization and calibration

Benefits of Performance Dimension Training

- Direct observation
  - Standardized, broad, systematic approach
  - Attentiveness to interpersonal/communication skills
- Feedback
  - Breadth of skills discussed using more granular vocab
  - Self-efficacy giving specific, constructive feedback
  - Ability to deconstruct holistic assessments
- Faculty clinical skills
  - Acquisition of new knowledge

Synthesis to Judgment

- Goal: Improve the quality and accuracy of the educational “judgment” using a compare and contrast process

Steps: Synthesis to Judgment

- Review vignettes of different performance levels
- Judge using behaviorally-based frameworks (e.g. evidence based frame of reference)
- Trainer provides feedback on assessment accuracy
- Discuss discrepancies between scripted performance and participants’ assessments

Satisfactory Compared to What?

<table>
<thead>
<tr>
<th>Compared to</th>
<th>Frame of reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>What I do</td>
<td>Self</td>
</tr>
<tr>
<td>What resident at similar PGY level does</td>
<td>Normative</td>
</tr>
<tr>
<td>Readiness for independent practice</td>
<td>Criterion referenced</td>
</tr>
</tbody>
</table>

Kogan JR et al. JAMA. 2009;302:1316-26
Cox SB et al. JAMA. 2003;290:24-30
**What is Needed by the Patient**

![Graph showing the progression of patient skill levels from Novice to Advanced Beginner to Competent to Proficient to Expert/Master over time, practice, and experience.]

**Entrustment as Assessment Construct**
- Cognitively aligned scale resonates with raters’ experience
- Increases discrimination
- Reduces disagreement
- Reduces # assessments for good reliability (generalizability coefficient 0.7)
  - Mini-cex: 6->3
- Decreases assessor workload approx 50%

**Shifting to Entrustment as FoR**
- Define competency based assessment
- Define competence
- Performance dimension training
- Peer support/group discussions
- Social and cultural factors

**Lessons in Rater Cognition**
- Assessment (rater cognition) is a complex process
  - Training can help, but will not solve “all problems”
    - Clarity on outcomes
    - Shared mental models
    - Own clinical skills matter
  - Sampling remains essential
    - Multiple raters in multiple setting
  - Not all variation is bad, but not limitless
    - Variation is a bounded condition

**Questions**