Moving Beyond
Surveys
External Assessments for Education & Research

Editors, Journal of Graduate Medical Education
ACGME Annual Meeting, 2015
Disclosures

- We are *JGME* editorial board members
- We hold memberships in other editorial boards & national medical & educational groups
- We will not be discussing commercial use of products
Background

Gail Sullivan, MD, MPH
Editor-in-Chief, JGME
Professor, UConn Center on Aging, University of Connecticut School of Medicine
Why this session?

- Evaluations required for faculty, trainees, experiences
- Ongoing changes to curriculum, programs – what works?
- Test new ideas
- For quantitative research, to compare interventions
- Disseminate scholarship
Objectives

• Expand ‘toolbox’ of assessments, for educational evaluation or research

• Avoid self-assessment, unless uniquely suited

• Match assessment to purpose & intervention
External vs. Self-assessments

Studies show:

• Those most confident are least competent

• Subjects may not ‘know what they don’t know’

• MD self-assessment does not correlate w/ external assessment

• *Exceptions*: stress/burn-out, acceptability of new intervention or teaching modality, opinions
Surveys

Best:
• Abstract ideas, concepts
• Opinions, beliefs, attitudes
• Behaviors, if can’t observe behavior & self-report likely to be accurate

Problematic:
• Behaviors & actual workplace performance
• Procedures
• Skills
• Retrospective: recall of trivial or large amounts of data
• Clinical outcomes
Match Outcomes to Purpose

• Improving knowledge              Assess knowledge

• Improving behaviors, skills      Assess performance, patient outcomes

• Teamwork, communication         Assess performance
Kirkpatrick’s Levels of Learning

4. Results  Change in patients, or organization’s practices
3. Behaviors  Change in behaviors or practice
2. Learning  Change in attitudes, knowledge, or skill
1. Reaction  Satisfaction, willing to participate
Miller’s Pyramid

- **Does**
  - Patient clinical performance assessment

- **Shows How**
  - Procedural assessment: OSCE, Simulation, SP

- **Knows How**
  - Clinical case synthesis: tests, essays

- **Knows**
  - MCQ & factual tests
Agenda

• Mini-talks by panel
• Q & A
• Stations: Tests
  OSCE, SPs
  Direct observation
  Multi-source feedback
  Chart reviews, audit & feedback
• Back to large group: share best practices
Tests

Anthony R Artino, Jr, PhD
Deputy Editor, JGME
Associate Professor & Deputy Director for Graduate Programs in Health Professions Education, Uniformed Services University of the Health Sciences
What is a Test?

• “Something (as a series of questions or exercises) for measuring the skill, knowledge, intelligence, capacities, or aptitudes of an individual or group” - Merriam-Webster

• “A systematic method for collecting a sample of an individual’s behavior in response to a set of structured tasks representing a construct domain” - Crocker, 2006
  • Includes objective tests (eg, multiple-choice tests), essay tests, performance assessments, self-reports, & observations surveys
Best for Assessing...

Kirkpatrick

- Reaction
- Learning
- Behavior
- Results

Miller’s Pyramid

- Knows
- Shows How
- Knows How
- Does

Tests
(Essays, MCQ & Oral Exams)

Behavior

Cognition
Considerations

• 2 most important issues in testing are **Reliability** and **Validity**

• **Reliability** -- Are the scores on the test consistent/reproducible?
  - Reliability is a necessary but not sufficient condition (prerequisite) for validity

• **Validity** -- Does the test measure what you say it measures?
  - Does evidence/theory support intended uses/interpretations of test scores?
  - Relates to aligning study outcome (ie, the test) to study purpose (ie, the RQs)
  - Lots of different sources of validity evidence, but for tests, **content** is critical
    - Benefit -- can sample more broadly than performance-based assessments
Limitations

• To know is not enough
  o “Tests of knowledge are surely important, but they are also incomplete tools... if we really believe there is more to the practice of medicine than knowing”
    - Miller, 1990

• Artificial ("non-authentic" assessments)
  o Most tests are artificial because they seldom directly measure instructional goals or job performance
    - Thorndike, 2005

• Feasibility
  o Can be difficult to develop a “valid” teacher/researcher-made test
  o Obtaining adequate score reliability requires many test items
  o Board scores (e.g., NBME, USMLE) are generally high-quality outcomes
  o Open-ended, essay-type tests are time-consuming and resource-intensive to administer and score
Objective Structured Clinical Examination (OSCE)

Denise M Dupras, MD, PhD
Associate Editor, JGME
Associate Professor of Medicine, Mayo Clinic
Miller’s Pyramid

- **Does**
- **Shows How**
- **Knows How**
- **Knows**

OSCE
- First described by Harden in 1975
- “Gold standard for clinical assessment”
  2002 - Norman
OSCE – Good For

- Physical Examination
- History taking
- Diagnosis
- Communication
- Management
  - Hand offs
  - Prescription writing
  - ....
- Procedures
- Problem solving
- “Synthesis”

- Undergraduate
- Graduate Medical Education
  - Internal Medicine
  - Pediatrics
  - Surgery
  - Family Medicine
  - Psychiatry
  - Physical Medicine
  - Ob – gyn
- Nursing
- Pharmacy
OSCE - Considerations

- Reliability
- Validity
- Feasibility
- Costs
- Formative or summative
- How many stations
- Duration of station
- Training of raters
- Training of “patients”
Direct Observation

Deborah Simpson, PhD
Deputy Editor, JGME
Aurora UW Medical Group

Kathyrn Andolsek, MD
Associate Editor, JGME
Duke University
Direct Observation
Does = Performance

Performance Integrated Into Practice
- eg through direct observation, workplace based assessment

Demonstration of Learning
- eg via simulations, OSCEs

Interpretation/Application
- eg through case presentations, essays, extended matching type MCQs

Fact Gathering
- eg traditional true/false MCQs

Based on work by Miller GE, The Assessment of Clinical Skills/Competence/Performance; Acad. Med. 1990; 65(9); 63-67
Adapted by Drs. R. Mehay & R. Burns, UK (Jan 2009)
Ex #1: Assessment

- Assessing learners in natural settings offers the opportunity to see beyond what they know and into what they actually do, which is fundamentally essential to training qualified physicians.

H Barrett Fromme, et al 2009

Ex #2 Discovery

- On the day-long follows that I used to do with mothers and their offspring - these chimp families that I knew so well - there was hardly a day when I didn't learn something new about them.

Jane Goodall
**BEST**

- Performance Based
  - Watch in typical environment
    - Settings: Overt? Covert?
  - Focus Narrow or Broad
    - All milestones or focused
    - Individual → Teams → Culture
  - Purpose: Formative or Summative

- Tools = Performance Focus
  - Descriptive “Snapshot”
  - Global Rating Forms (mini CEX)
  - Targeted Checklists (Communicate, Professionalism, Procedure)

- Trained Observers

- Data:
  - Quantitative & Qualitative
  - Combine other measures

**WORST**

- Ambiguous Focus*
- Susceptible to error/bias
  - Observers
  - Observed “Stress”, change performance
  - Tools

- Requires Resources
  - Validated Tools (Adopt, Adapt)
  - Rater training
  - Data Analysis

- Culture Doesn’t Value
  - Participants ≠ observation
  - Learners (“stressed”)
  - Raters “halo” checks

* Assessment vs. “Discovery”
Considerations

- **Purpose**
  - Assessment (formative, summative)
  - Discovery (research)

- **High Stakes?**
  - Culture
  - Resources to Minimize Bias/Error
    - Reliability
    - Validity
    - Feasibility
    - Costs

Leadership Support
Faculty Champions
Clinical/Hospital

- Evidence-based tools
- Rater Training
- Data Analysis
- Time /ROI
Multisource Feedback

Lalena Yarris, MD, MCR
Deputy Editor, JGME
Oregon Health & Science University
Multisource Feedback

• Recognized approach for assessing professional competence, behaviors, & attitudes in workplace performance

• MSF shown to have high reliability, validity, & feasibility
MSF – Best for

- Real-time performance
- 5 key domains:
  - Clinical competency
  - Professionalism
  - Communication
  - Manager
  - Interpersonal relationships
- Obtains more complete picture of learner performance than a single source
MSF – Limitations

• MSF accuracy is context/specialty-specific

• Forms are survey variants, prone to multiple sources of error

• Rely on observed performance; limited information about medical knowledge or cognitive processes

• May be subject to “halo” or “millstone” effect

• May be time-consuming & resource-intensive to administer
Chart-Stimulated Recall (CSR) and Medical Records Audits

Ingrid Philibert, PhD, MBA
Executive Managing Editor, JGME
Sr. Vice President, Field Activities
Accreditation Council for Graduate Medical Education
Chart-Stimulated Recall (CSR) and Medical Records Audits

Used to assess clinical reasoning and decision-making

Competencies assessed: patient care, medical knowledge (and its application in context), practice-based learning and improvement, systems-based practice

Type of assessment: Summative

Validity: High

Reliability: Medium; High with trained assessors
Chart-stimulated recall (CSR)

- Records for residents’ real patients are assessed in a standardized oral examination
- Examiner questions the resident about the care provided, the reasons behind the workup, diagnoses, interpretation, and treatment plans
- Assessor selects chart, reviews it, and develops open-ended and semi-structured questions
- Resident reviews chart before discussion; interview usually begins with open-ended question regarding case management.

Medical Records Audit

- Trained staff review medical records and abstract information such as medications, tests ordered, procedures performed, preventive care provided
- Records are summarized and compared to accepted patient care standards (standards exist for ~1600 diseases)
Importance of CSR and Medical Records Audits

• Can assess competencies that are difficult to measure
• Can assess integration of several competencies in real clinical performance
• Diagnostic reasoning, clinical decision-making, including use of resources, are important but internal processes
  o Usually inferred, not directly measured
  o Need to externalize trainees’ thought process to measure it
## Advantages and Disadvantages

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Availability</td>
<td>• Quality of the Documentation may be poor, even if care was “good”</td>
</tr>
<tr>
<td>• Feedback</td>
<td>• “Reimbursement focus” of much of the documentation in the electronic health</td>
</tr>
<tr>
<td>• Changing Clinical Behavior</td>
<td>record (EHR)</td>
</tr>
<tr>
<td>• Practicality</td>
<td>• Process versus Outcomes</td>
</tr>
<tr>
<td>• Evaluation of Clinical Reasoning</td>
<td>• Implicit Review</td>
</tr>
<tr>
<td>• Reliability and Validity</td>
<td>• Time and Cost</td>
</tr>
<tr>
<td>• Learning and Evaluating by Doing</td>
<td></td>
</tr>
<tr>
<td>• Self-Assessment and Reflection</td>
<td></td>
</tr>
</tbody>
</table>
## Overcoming Potential Disadvantages

<table>
<thead>
<tr>
<th>Limitation</th>
<th>Possible Solutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inferring clinical judgment</td>
<td>Combine MRA with chart-stimulated recall</td>
</tr>
<tr>
<td>Implicit review</td>
<td>Provide evaluator/auditor training</td>
</tr>
<tr>
<td></td>
<td>Provide framework and do not rely solely on the judgment of the reviewer</td>
</tr>
<tr>
<td></td>
<td>Use explicit criteria and templates whenever possible</td>
</tr>
<tr>
<td>Quality of documentation</td>
<td>Use problem lists and flowcharts for common conditions</td>
</tr>
<tr>
<td></td>
<td>Provide templates for medical history and physical examination</td>
</tr>
<tr>
<td></td>
<td>Combine MRA with direct observation</td>
</tr>
<tr>
<td>Cost/Time</td>
<td>Have trainees audit their own or peers’ charts</td>
</tr>
<tr>
<td></td>
<td>Use existing reports when available (from QI department, EHR)</td>
</tr>
</tbody>
</table>

Adapted from: ES Holmboe, Practice Audit, Medical Record Review, and Chart-Stimulated Recall
Faculty Development and Interface with Other Tools/Approaches

Faculty development:
• Create a framework for rating/scoring of responses
• Faculty should observe collectively several CSR) discussions and rate them.
• Faculty should discuss results to develop common rating anchors through group consensus.
• Teach use of relevant, non-judgmental, open-ended questions and semi-structured questions for each case

Interface with other tools:
• CSR can be combined with Medical Record Audit to assess educational needs specific to a medical condition or practice guideline.
• CSR can be used for post-testing after an educational intervention.
• Use CSR after direct observation to probe deeper into clinical decision-making and to assess documentation of visit.
Q & A
Stations – choose 3!

- Tests
- OSCE
- Direct observation
- Multi-source feedback
- Chart review, audit & feedback
What Did We Learn?

• Thank you for attending this session