Assessment of Learners
Types of Assessment

1. learner
2. teacher
3. programmatic
4. institutional
Q. What does the State of Pennsylvania require before allowing a 12-year-old to carry a loaded gun?
A. Passing a 50-item, multiple choice test.
What domains might you want to assess?
Hunting can be dangerous!
Domains

- Can they shoot their guns well?
  - targets
  - skeet
- Close order drill (safety)?
- Will they get lost?
To Consider

1. Design your evaluating system to assess what you think is important.

2. Ensure that your system accurately and consistently evaluates all learners.

3. Make sure your curriculum matches the evaluation and it is transparent.
Assessment can be criterion-based or normative
Criterion Reference

Inspection Stickers
Normative Reference
Note: assessment has dual roles
Most methods can be:
formative (feedback)
summative (pass judgement)
How good are our measurements?
Reliability and Validity

Reliability: Precision of measurement

Validity: The degree to which an instrument measures what it set out to assess
Reliability and Validity
Similar to Sensitivity and Specificity, you may need to compromise between reliability and validity.
Types of Reliability

- Internal consistency (alpha)
- Case specificity (knowledge holes)
- Test/Retest
- Inter/Intra rater (e.g. reading x rays)
Types of Validity

Face ("looks okay to me")

Construct (smart is smart, performances get better with more training)
- OSCE scores correlate with MCQ exam
- Residents do better than MS3

Predictive (Step 1, Step 2, Board Exam)

Criterion (gold standard)
So, let’s think about our learners –

MS³ clinical clerk: **Internal Medicine**
R⁴ resident: **Surgery**
MSOP Competencies
Medical Students

Altruism    Skillful
Knowledgeable    Dutiful
Clerkship Student

1. Cognitive Knowledge Base
2. Clinical Skills
3. Interpersonal Skills
4. Professionalism
# Student Assessment in U.S. Medical Schools

Mavis. TLM 2001

<table>
<thead>
<tr>
<th>Modality</th>
<th>Year 1</th>
<th>Year 2</th>
<th>Year 3</th>
<th>Year 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCQ local</td>
<td>98</td>
<td>98</td>
<td>91</td>
<td>23</td>
</tr>
<tr>
<td>Orals</td>
<td>6</td>
<td>11</td>
<td>38</td>
<td>16</td>
</tr>
<tr>
<td>Lab Practical</td>
<td>88</td>
<td>64</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Written Notes</td>
<td>14</td>
<td>41</td>
<td>92</td>
<td>79</td>
</tr>
<tr>
<td>Preceptor Ratings</td>
<td>61</td>
<td>78</td>
<td>98</td>
<td>98</td>
</tr>
<tr>
<td>OSCE</td>
<td>5</td>
<td>11</td>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td>Live Observe</td>
<td>39</td>
<td>64</td>
<td>89</td>
<td>88</td>
</tr>
<tr>
<td>Logs</td>
<td>4</td>
<td>10</td>
<td>52</td>
<td>24</td>
</tr>
<tr>
<td>CBT</td>
<td>10</td>
<td>10</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>NBME Subject</td>
<td>33</td>
<td>36</td>
<td>75</td>
<td>6</td>
</tr>
</tbody>
</table>

N = 87. indicates % use ≥ 3 times/yr.
ACGME Competencies
Residents

1. Patient Care
2. Medical Knowledge
3. Interpersonal and Communication Skills
4. Practice-Based Learning
5. Systems-Based Practice
6. Professionalism
Residents

1. Knowledge Base
2. Clinical Skills
3. Interpersonal Skills
4. Management Skills
5. Technical Skills
6. Systems Approach
ACGME lists “most desirable tools” by competency
eg. MCQ for knowledge or OSCE for communication

Surveyed 272 IM programs (70%)

Most use “most desirable” (64-98%) for each of 6 competencies
Traditional Assessment Measures

1. Multiple Choice Question (MCQ) Exam
2. Objective Structured Clinical Exam (OSCE)
3. Descriptive Evaluation
Why Use These?

1. NBME Subject exam:
   - reliability .8-.9
   - increasingly valid
   - nationally normed

2. USMLE

3. ITE
In-Training Exam
Garibaldi. Ann IM 2002

Created by: ACP, APM and APDIM

voluntary, self-evaluation for IM residents
formative intent: not summative
over 80% use it. NBME scores it
MCQ testing knowledge of H&P, Dx, Rx
analyzed over 12 years
ITE Result

Construct validity:  R3 > R2 > R1

Demographics:      Canadian > IMG > USMG

Confounding:       lack of high stakes
Why Use These?

2. OSCE: can assess directly
   History & Physical Exam Skills
   Counseling Skills
   Technical Skills
   Teaching Skills
Objective Structured Clinical Exam (OSCE)

Format: set tasks using standardized patients in fixed rotation

Pluses: simulate actual experiences (valid)

Minuses: cost, logistics, security, reliability, case specificity issues

(need 10 stations, 3-4 hours for $r > .8$)

Score: checklists v. global ratings
Why Use These?

3. Descriptive Evaluations
   can be done in 360 fashion
   assess numerous items
   can be synthetic (RIME)
   capture degrees of proficiency
Descriptive Evaluations

Most commonly used (94% in CDIM survey)
Reliability problems, bias

To improve:

- train evaluators annually
- specify educational objectives
  - increase number (need ≥ 7 per Carline)
- use behavioral descriptors
- use accepted framework (AAMC, ACGME)
Synthetic System: RIME
Pangaro

[O = Observer]
R = Reporter
I = Interpreter
M = Manager
E = Educator

Supporting: reliability > .8, predictive validity, generalizable
Utilizes formal evaluation systems
Improving Resident Performance Assessment
Littlefield. Acad Med 2005

Surgery Programs: NWU, UTSA

used complex adaptive systems theory – identify small group within system & internal behavior models. Then use for system improvements.

used: PD, individual faculty evaluators, infrastructure (e.g. evaluation training)

made QI project with goals: improve return rate, precision of evaluations

compared pre/post intervention
Improving Resident Performance
Assessment
Littlefield (cont.)

**Intervention:** PD – identified issues, synthesized data

- Evaluators – established importance, PD involved, showed examples

- Infrastructure – chair and administrators enlisted, revised forms, created databases

**Results:** return rate improved significantly
- mean scores unchanged (5/7)
- generalizability coefficients (per 10 raters)
  - baseline .6 to .8
Newer Assessment Measures

• Mini- CEX
• Triple Jump
• Computer-based Testing
• Portfolios
• Peer Reviews
Mini CEX in Medicine Clerkship
Kogan. Acad Med 2003

Mini CEX directly observes clinical skills;
   developed by ABIM.
assessed here: interviewing, physical exam,
   judgement, counseling, humanism, overall
   competence
feasible: mean time with feedback – 19 minutes
reliability (8 raters) = .77
correlations with exam (r = .22, p < .01),
descriptive (r = .35 & .43, p < .01)
Triple Jump and Multi-Step

1. Hypothesis generation and testing
   Paper scenario presented, discuss hypotheses to explain findings

2. Independent Study
   Read and prepare conclusions to diagnostic problem

3. Final problem formulation
   Report conclusions

Note: used at McMasters, Hawaii, USUHS
Computer-Based v. SP Testing
Edelstein. Acad Med 2000

CBX: uncued questions unfold
based on student responses
tests diagnostic reasoning, management
correlate to USMLE 2 = CBX .4, SP .3
correlate CBX & SP = .24
Portfolios: Cleveland Clinic
Dannefer. Acad Med 2007

Graduation contingent on demonstrating mastery of 9 competencies

Portfolio is depository for multiple assessments

- Formative: done with advisor, reflective writings, learning plans
- Summative: reviewed by promotion committee (e.g. research, exams, logs, clinical assessments)

Note: Clear expectations on timeline

Caveats: faculty development needed, labor intensive
Portfolios in Residencies
Culbert. Teach and Learn 2008

Types:
- learning (learner controls)
- structured (program/learner share)
- showcase (creators demonstrate best work, summative)

Uses:
- gauge resident competence
- teach complicated competencies (eg. practice based learning)
- encourage self-reflection and self-directed learning
Resident Portfolios
Culbert (cont.)

Can be summative or formative (don’t mix)

Portfolio scores correlate with exams, clinical evaluations (valid)

Caveats: learners must be trained, time consuming, little reliability testing
Peer Review in Residency
Thomas. JGIM 1999

J. Hopkins IM Program
Used 2 of 4 inpatient firms (control group)
Assessed: intern and resident of intern, pooled and done anonymously
Instrument: 10 items (similar to ABIM form)
9-point Likert scales
residents trained in giving FB, interns not
Results: Peer Review

Thomas (cont.)

Response rate: interns 35%, residents 70%
mean ratings: residents and interns all items
  > 8.0 faculty 7-8
reliability: alpha .96, inter-item correlation .73;
  faculty/resident correlations > .60
FAN: cognitive/technical skills and interpersonal skill
System rated (v. controls): 4.2–4.7 v. 3.6–4.3 (p< .03)
Concluded: reliable, feasible (+/-), new info, acceptable to learners
Levels of Evaluation

Change Outcomes or behaviors
Learn better than other system
They learn (pre/post)
They like it (feedback analysis)
Systematic Review (38 studies) and meta-analysis (19 studies)

Methods: tested heterogeneity, random or fixed effects models, generated correlation coefficients

Predictors: NBME 1 & 2, GPA, OSCE, Dean’s ranks, honor society

Outcomes: NBME, ITE, specialty board exams, PD ratings, practice performances
## BEME Review (continued)

**Supervisor ratings (n = 8)**

<table>
<thead>
<tr>
<th></th>
<th>r</th>
<th>CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>NBME 1</td>
<td>.22</td>
<td>(.13-.30)</td>
</tr>
<tr>
<td>NBME 2</td>
<td>.27</td>
<td>(.16-.38)</td>
</tr>
<tr>
<td>Clerkship GPA</td>
<td>.28</td>
<td>(.22-.35)</td>
</tr>
<tr>
<td>OSCE</td>
<td>.37</td>
<td>(.22-.50)</td>
</tr>
<tr>
<td>Dean’s Letter</td>
<td>.22</td>
<td>(.12-.31)</td>
</tr>
<tr>
<td>Pre-clinical GPA</td>
<td>.25</td>
<td>(.19-.31)</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>r</td>
</tr>
<tr>
<td>--------------------------</td>
<td>---</td>
<td>------</td>
</tr>
<tr>
<td>NBME 3</td>
<td>2</td>
<td>.59</td>
</tr>
<tr>
<td>NBME 1 &amp; 2</td>
<td></td>
<td>.72</td>
</tr>
<tr>
<td>NBME 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specialty Boards</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>NBME 1</td>
<td></td>
<td>.58</td>
</tr>
<tr>
<td>Intern Performance</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>med school GPA</td>
<td></td>
<td>.40</td>
</tr>
</tbody>
</table>
BEME Conclusion

No single measure suffices to predict competency. It is a multidimensional issue, necessitating the use of multiple assessment methods.
Difficult Areas
Epstein. NEJM 2007

- Professionalism
- Teamwork
- Communication
- Expertise