Performance Assessments and the Health Professions: Planning Backwards to Move Forward

Health Professions Educational Research Symposium
NOVA Southeastern University

January 19, 2008
John R. Gimpel, D.O., M.Ed.
Vice President for Clinical Skills Testing
National Board of Osteopathic Medical Examiners (NBOME)

jgimpel@nbome.org
Objectives

- You will be able to discuss and reflect on elements of performance testing that are relevant to teaching and learning in your profession.
- You will be able to describe some of the unique challenges of using simulation for training and assessment in the health professions.
Objectives

- You will appreciate assessment as an integrating concept in curriculum development, and performance assessment as a method that promotes collaboration across departments and professions - and one that initiates curriculum planning.
Curriculum Design

The planned curriculum
- What is intended by the designers

The delivered curriculum
- What is organised by the administrators
- What is taught by the teachers

The experienced curriculum
- What is learned by the students

Prideaux, D. BMJ 2003;326:268-270
Planning a Curriculum

Where to Start?
From Theory to Practice

Kaufman, D. M BMJ 2003;326:213-216
Situational Model: Malcolm Skillbeck

Prideaux, D. BMJ 2003;326:268-270
CURRICULAR THEORY

Desired outcomes (students will be able to...)

Content

• Teaching
• Learning

Assessment

Evaluation

Prideaux, D. BMJ 2003;326:268-270

Outcomes-Based: Grant/Wiggins/McTighe

“Backwards” Design
The Role of Assessment in The Hidden and the Null Curriculum

- “Students value what you test, not just what you teach”
- “People pay attention to what you inspect, not what you expect”
- “Assessment Drives Curriculum”
## Standardized Patient Programs at Colleges of Osteopathic Medicine

<table>
<thead>
<tr>
<th>Program Status</th>
<th>2001 Results N=19</th>
<th>No. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active</td>
<td>12 (63)</td>
<td></td>
</tr>
<tr>
<td>In Development</td>
<td>2 (11)</td>
<td></td>
</tr>
<tr>
<td>Students used as Standardized Patients</td>
<td>1 (5)</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>4 (21)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Program Status</th>
<th>2005 Results N=23</th>
<th>No. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active</td>
<td>19 (83)</td>
<td></td>
</tr>
<tr>
<td>In Development</td>
<td>2 (9)</td>
<td></td>
</tr>
<tr>
<td>Students used as Standardized Patients</td>
<td>2 (9)</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>N/A</td>
<td></td>
</tr>
</tbody>
</table>
Standardized Patient Programs at Osteopathic Medical Schools

- Active
- In Development
- Used as SPs
- None

2001 (N=19)
2005 (N=23)
# Medical education: traditional vs. competency-based

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>TRADITIONAL STRUCTURE &amp; PROCESS-BASED</th>
<th>COMPETENCY-BASED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Driving force for curriculum &amp; goal of educational encounter</td>
<td>Content-knowledge acquisition</td>
<td>Outcome-knowledge application</td>
</tr>
<tr>
<td>Driving force for process</td>
<td>Teacher</td>
<td>Learner</td>
</tr>
</tbody>
</table>

# Medical education: traditional vs. competency-based

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>TRADITIONAL STRUCTURE &amp; PROCESS-BASED</th>
<th>VS</th>
<th>COMPETENCY-BASED</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ASSESSMENT TOOLS</strong></td>
<td>MCQ + (subjective) global rating</td>
<td>VS</td>
<td>Multiple objective measures</td>
</tr>
<tr>
<td></td>
<td>Proxy</td>
<td>VS</td>
<td>Authentic (mimics real tasks)</td>
</tr>
<tr>
<td></td>
<td>Norm-referenced</td>
<td>VS</td>
<td>Criterion-referenced</td>
</tr>
<tr>
<td><strong>Setting for evaluation</strong></td>
<td>Removed</td>
<td>VS</td>
<td>“In the trenches”</td>
</tr>
</tbody>
</table>

Performance Assessment

- “to attain competence, basic knowledge, skills and attitudes are required.

  Competence is the application of those basics.

  Performance is the translation of competence into action”

  George Miller, 1984
Performance Assessment

- The basics are- do they know it and know how?
- Competence is- can they do it?
- And performance is- do they do it?
Competency Assessment: Miller’s Pyramid

Performance assessment in vivo:
SPs, Video, Logs, Clinical Simulation, 360 degree, Faculty Ratings, Portfolios

Performance assessment in vitro:
OSCE, SP-based test, MJI/DITs

(Clinical) Context based tests:
MCQ, essay type, oral...

Factual tests:
MCQ, essay type, oral...
Steps for Developing A Performance Test

1. Decide What To Test
2. Decide The Assessment Context
3. Specify The Scoring Rubrics
4. Specify Testing Constraints
Deciding What To Test

- What kinds of essential tasks, achievements, or other competencies am I missing with present MCQ examinations?
Simulation-Based Medical Education
What is Simulation?

- Simulation is described as a **strategy** – not a **technology** – to mirror, anticipate, or amplify real situations with guided experiences in a fully interactive way.

- A simulator **replicates a task environment** with **sufficient realism** to serve a desired purpose.

Agency for Healthcare Research & Quality (AHRQ)
Types Of Clinical Simulations

- Paper-based or computer-based patient management scenarios
- Standardized (simulated) patients
- Partial-task trainers
  - Simple level
  - Higher level
- Mannequins
  - Low fidelity
  - High fidelity - Human Patient Simulators
- Virtual patients
- COMBINATIONS
  - Augmented sp encounters with technology
  - Crises management
Standardized Patient Assessments

- Performance-based
- “Standardized”
  - Same conditions for all test takers
- Series of interactions in simulated encounters
Gross Anatomy

Animal Models
  e.g. Suturing Practice
Patient simulators (manikins)
  Teamwork, procedures e.g. codes, ACLS

Procedure simulators
  Psychomotor skills, e.g.
  laproscopic surgery
Part-task / Part body trainers
Basic concepts
Psychomotor skills training
Mannequins

- Life-sized simulators with realistic airway and cardiovascular attributes
  - real-time responses to therapeutic interventions
  - can model rare events
  - errors are ‘reversible’
- Provide a standardized high-fidelity simulated environment to train and evaluate students/residents
Pelvic ExamSIM

Opening the mind’s eye...
Mannequins-Human Patient Simulators
Virtual Reality and Computer-Based Programs

- PC/Mac – Patient “in the computer”
- Haptic – Feel and touch
- 3D – Haptic plus virtual environment
Full-Immersion Virtual Reality

Diana – University of Florida
“Cardiology” Scenario

Students encounter a cardiology complaint (manikin) and discuss physiology / pharmacology issues with a science teacher.
“Death and dying” Scenario

Students encounter “dying patient” (manikin), then counsel “grieving family member” (SP)
“Suturing” Scenario

Students encounter practice suturing (p/task trainer) attached to a conscious patient (SP)
“Conscious - Comatose” Scenario

Students encounter a hospital patient (SP), then that same patient in a comatose state (robot)
“Pre-Encounter” Scenario

Students prepare for a sim encounter by meeting a web-patient (PC-VR), then meet the “actual patient” (manikin) in an ED setting
“Patient Management” Scenario

Students encounter a patient (SP), then that same patient in an acute state (manikin), then manage the patient’s treatment post-discharge (PC-VR)
“Simulator-Audience Response” Program

Students encounter a patient in an acute state (manikin), and through a live DV feed, an audience participates via an audience response system.
Simulation Connectivity System

- Digital video for easy storage
- Local and remote access through the web
- SP training / quality assurance
- Precepting – locally and remotely
Simulation Connectivity System

- Digital data collection, e.g. by SPs at a PC
- Data analysis / scoring
- Score reporting
- Longitudinal studies of competency acquisition
Faculty Development

What do they need?

- Performance training and assessment
- Clinical skills blueprint
- Cases, scenarios, checklists
- Test committee
- Formative and summative assessment
- Debriefing / feedback
- Logistics
Simulation and Adult Learning

Simulation enhances learner motivation or “need to know through experiential learning"
Simulation and Adult Learning

- “The adult learner enters the training environment with a deep need to be self-directing”
- High fidelity team simulation combined with reflective debriefing teaches learners to monitor and question their mental models & practice behaviors
- Vivid experiences in simulation stimulate the “need to know” that motivates adult learners

Simulation and Adult Learning

- Teaches how to manage performance in high stress environments
- Enhances adult learners’ motivation or “need to know” through experiential learning
- Involves reflective learning
- Enhances depth of memory encoding
- Builds on social aspects of adult learning
Cannon-Bowers and Salas have studied high performance teams under stress in the navy. High performance teams employ different coordination strategies than low performing teams in high workload and high stress environments. Simulation gives the opportunity for teams to practice under these stressful conditions & hone coordination strategies.
State-Depending Learning is a phenomenon of learning and recalling that is base upon the physiologic and mental state of the organism.

“When sad it is easier to remember sad events than happy ones”
Factors Affecting State-Dependent

- Environment
- Intoxication
- Emotional State
- Sensory Modality

“What is learned in one state may have little influence on the performance exhibit in a different state”
Why do students forget?

- Failure to understand
- No relationship to experiences
- No perceived utility
- No relationship to interest
- Contrary to bias or values

Klatsky, 1975; Curry, 1993; Whitman, 1995
What are the advantages of clinical simulation?

- Immediate Feedback
- Repetitive Practice
- Curriculum Integration
- Individualized Learning
- Reflective Learning
- Opportunity to “Go Wrong” without risk.
What are the advantages of clinical simulation?

- Standardizes clinical exposures (“educational equivalency”)
- Enhanced Patient Autonomy and “An ethical imperative” (Ziv, et al)
- Evaluate effectiveness of teaching and the curriculum
Reflective Learning

“Reflection appears to be the engine that shifts surface learning to deep learning, and transforms knowing in action to knowledge in action”

Challenges In Using Simulation in Teaching, Learning & Assessment
Sign Hanging In Einstein’s Office at Princeton

“Not everything that counts can be counted, and not everything that can be counted counts.”
"People who thinks something cannot be done should not interrupt people doing it."

Chinese fortune cookie
Challenges

- Validity
- Reliability
- Standard-Setting
- Other
Design, Content, and Scoring Issues

- Need to develop an assessment that yields **valid** and **reliable** scores
  - How?
Validity

Does the assessment provide measure of what it is supposed to?
What is Validity?

- Refers to the inferences & conclusions one draws from test scores
- 1 single study isn’t sufficient
- Similarly, validity can be compromised in many ways
- Multifaceted
Thinking About the Validity of Clinical Skills Examinations

- What are our inferences?
- Look at process from beginning to end
- Does the assessment measure what it’s supposed to?
- What are the potential problem points that would weaken our inferences?
- Focus on these areas
The Test Development Process

- Content Validity
- How are the cases developed?
- Does the test design allow for sufficient reliability?
- The blueprint reflects the types on situations candidates will encounter in PGY1 and how are the measurement instruments & scoring rubrics developed and refined?
- Passing this exam will assure they are ready to practice safe and effective medicine.
Ways to Ensure Validity in the Test Development Process

- Content validity
  - NAMCS, NHAMCS (survey use)
  - Expert judgment
- Committee
  - Collect Vitas for case development committee
  - Describe composition of committee
- Periodic case review for changes in medical content
- Key validation of cases after pre-testing
Validity Evidence

- Start process as exam is being developed
  - Content
  - Response Processes
  - Internal structure
  - Relationships with criterion measures
  - Consequences
    - evaluation drives learning
Ways to Establish Defensibility of SP Portrayal Accuracy

- Performance Fidelity checklist for training
- Standard videos
- Video monitoring of patient portrayal consistency
- Sign off
  - Subsequent to last training
  - Accountability
  - Multiple measures to increase reliability of sign off decision
- Periodic, unannounced measures (e.g. quizzes & video) to ensure consistency of portrayal
- Physician Trainer Review
Ways to Establish SP Recording Accuracy

- Verification and sign off
- Quality control
- Multiple measures AND multiple time points
- Understand what is causing checklist errors, not just quantifying them (qualitative approach)
- SP file: qualifications, sign off, quizzes, notes
- Track SPs over time (drift)
Assuring the Consistency of the Test Administration

- Candidate ID
- Information dissemination
  - All candidates receive the same information during orientation
  - All candidates hear the same responses to questions
- The test center provides the same testing experience to all candidates
- Quality assurance for doorway info, SOAP notes in correct folders, etc
- Policy for std administration during irregularities
Assuring the Integrity of the Test Center Experience

- License ID and photos taken
- Consistent orientation
- Standard responses to questions are developed and given
- Protocol the same for all candidates, all test dates
- Eye on students during breaks
- Document any problems that would jeopardize the standard or overall score (policies)
Assuring the Accuracy of Candidate Scores

- What is the scoring process?
  - How are subcomponents weighted?
  - When are holistic scores used?

- What quality control methods are used?

- A candidate’s score is a good estimate of their true clinical skills ability.

- How might the selection and assignment of raters influence a candidate’s score?

- Is the rater training and protocol sufficient?

- What scores are reported?
Ways to Ensure Accuracy of Candidate Scores

- Build quality control into scoring process
- Provide documentation for all decision regarding weighting of score components
- Ensure that all notes are scanned properly
- Verify candidate ID is correct on Note and OMT video
- Ensure Rater training and consistency
- Report score information appropriate to the reliability of the exam
Assuring Rater Consistency

- How are the raters trained?
  - How long are they trained?
- What is the process?
- Who is doing the ratings?
  - What are their qualifications?
  - Have they been involved in the test process before?
- What sign off process is used?
- How is consistency in ratings assured?

Inference: Raters can accurately and consistently score student behavior.
Test Content for Simulations

- **Skills**
  - Data Gathering (History Taking & Physical Examination)
  - Doctor-Patient Communication
  - Procedural maneuvers
  - Clinical reasoning
  - Defined by subject-matter experts
  - Curriculum

- **Content (Clinical Scenarios)**
  - Can be simulated
  - Important
  - Prevalence of ‘reasons for visit’ in health care settings

- **Case content determined by local/national needs**
  - Technology
Case Scenario Development Issues

- Who are the “target” examinees?
- Specificity
- Difficulty
- Essential maneuvers and questions?
- Sampling from domain
Skills

- Cases are “vehicles” for measuring skills
  - Level of difficulty
  - Complexity
    - Data Gathering
    - Doctor-patient Communication
    - OMM/OMT
Checklist

☐ Behaviors required for “adequate” patient care

- Relevant history questions, physical examination maneuvers
- <=20 items, if possible
Checklist Review

☐ Expect differences of opinion!!!

- 1 or 2 items will not have that big an impact
Reliability

- How consistent are the examinee scores?
  - Want to ensure that an examinee’s observed score is a reasonable reflection of his/her “true” ability
  - Minimize errors of measurement
More on Reliability

- “Reproducibility”
  - Want scores to be consistent over occasions and equivalent test versions
  - Standard Error of Measurement (SEM)
    - confidence interval for “true” score

- Generalizability analysis vs Cronbach’s alpha
Assuring the Reliability of Clinical Skills Examination Scores

- Case development
  - Avoid double barreled items
  - Targeted, objective, trainable leaving little room for variation
- Define scoring rubrics
- Eliminate extraneous sources of error
  - training
  - quality assurance
  - benchmark videos
Enhancing Precision

- Choice (and number) of tasks
- Raters
- Settings
- Administration conditions
- etc.
Scoring Simulated Encounters
Types of Scores

- **Analytic**
  - Checklists
  - Key actions
    - Timing
    - Sequencing

- **Holistic/ Global**
  - Rating scales
Advantages and Disadvantages of Checklists

- Fairly easy to develop
- "Objective"
- Record of what was done (feedback)
- Can be used by non-clinicians
- Students perceive that they are being evaluated by patients
- Difficult to assess complex skill sets
Advantages and Disadvantages of Global Ratings

- Rely on expert judgment
- Can consider many factors related to performance
  - Egregious actions
- Medical students prefer to be evaluated by their peers
- Need “experts”
- Some evaluators may not be objective
Other Scoring Modalities “Anaphylaxis”

- Time/ Key action
  - Time to diagnosis of anaphylaxis
  - Time to treatment regime for suspected anaphylaxis
  - Time to dose of Epinephrine

- Overall assessment
  - Organized approach to diagnosis and treatment, precision and clarity of inquiry, etc
Who Should Provide the Scores?

- Physician examiners
  - “face” validity
  - Expertise
    - Practice of medicine is complex
  - Perceived subjectivity
- Standardized patients or other observers
  - Objective
  - First-hand understanding of certain skill sets (e.g., communication)
  - Economical/efficient
How to Collect the Data?
Important Scoring Issues

- Choice of method
- Choice of person to score
  - Cost, efficiency, availability
- Number of ‘raters’ to score
- What to score
  - Discreet actions
  - Speed
  - Key items
- Accuracy
- When to score
  - Real-time, afterwards
Assuring a Defensible Standard
Standard-Setting

- No widely accepted and validated standard-setting methods for use with performance assessments
Background

- Adaptations of MCQ-based methods are not completely satisfactory
- Many “new” techniques look promising but additional research is necessary
- Examinee-centered methods clearly becoming the “standard”
Standard Setting

- Process used to arrive at a passing score
- Competent to practice Credential
- Master/ non-master
- etc.
Standard Setting Issues

- Performance standard setting is a judgmental process
  - There is *no* “gold standard”
  - Sound and defensible procedures can be adopted

- Performance standards are method dependent
  - Selection of method must be clearly defended
  - Mounting literature
Assuring a Defensible Passing Standard

- Select knowledgeable and diverse committee
- Document method with literature
- Utilize consultants where necessary
- Educate/Calibrate Raters
- External committee sets std with info from surveys and std setting meeting: triangulation
- G-Theory and other decision reproducibility studies
Standard Setting Issues

- Performance standards should be periodically revisited / evaluated
  - Cross-validation across judges and students is needed
Questions to be Asked

- Who should set the standard?
  - Experience + legitimate interest in the outcome
  - Knowledge of the content being assessed
  - Diversity of judges
  - Balance of academicians, practitioners, program directors, licensing board representatives
Triangulation- NBOME’s “Standard”

- Expert Panels
- Surveys of Stakeholders
- NBOME Executive Committee
Standard Setting Frameworks

- **Criterion-Referenced**
  - Standard defined with regard to an acceptable specific measure of performance
  - Test-centered (inspection of items of test components)
  - Examinee-centered (inspection of examinee performances)
  - Leniency/severity of judges will affect content-based performance standards

- **Norm-Referenced**
  - Standard set based on performance of some selected group
  - Standard will change as the ability of the normative sample changes
  - Unknown false-positive and false-negative rates
Standards for OSCEs & SP Exams

- “For performance tests and simulations in which examinees are asked to complete a few, relatively long tasks, the examinee-centered methods seem particularly appropriate …”
  - Kane et al. (1999)
Examinee-Centered

- Direct judgment about status of persons on the construct of interest
Examinee-Centered

- **Contrasting Group**
  - Form 2 (or more) groups of people (e.g., masters, non-masters)
  - Look at the 2 score distributions from the test and pick the point that maximizes the probability of correct decisions
  - NBOME, ECFMG

- **Borderline Group**
  - Identify sample of examinees who could be classified as “borderline”
  - Select a point in this distribution (e.g., median) to signify borderline performance
  - MCC Part II
Application for OSCEs & SP Exams

- Intuitively appealing
  - Can judge actual performance
- May be difficult to categorize performances for some examinees
- Requires numerous "expert" judgments
Standards Validation

- Do decisions based on passing scores achieve the purposes of the organization while avoiding any serious negative consequences?
  - Note - Performance of candidate just above passing score not likely to be much different from one just below
Standards Validation (Consequences)?

- Protect public
- Not unduly restrict access to profession

- Research on passing candidates
- Criterion measure?
Self-assessment of Competence Lacking

- Physicians have a limited ability to accurately self-assess their own competency - with those with the worst accuracy in self-assessment being those who are least skilled and those who are most confident

*Davis, et.al., JAMA 2006*
Methods

- Adequately defining the performance standard
  - Level of mastery needed?
  - Surveys
  - Focus groups
  - Analysis of practice patterns
  - Behavioral descriptors
    - Operational definitions of the skills needed
Methods (cont.)

- Protocol
  - Training panelists
    - Take exam under exam-like conditions
    - Meaningful feedback
  - Selection of performance samples
  - Identification of aberrant panelists
  - Robustness of various techniques
Methods (cont.)

- Statistical techniques
  - Logistic regression, cluster analysis, ROCs, etc.

- Error estimation
  - Re-sampling, monte-carlo methods, generalizability theory, bootstrapping
Some Continuing Challenges For SBME

- Integrating various simulation modalities into the curriculum
- Faculty development
- Content under-representation
- Fidelity issues
- Accurately measuring specific abilities (teamwork, professionalism)
Some Continuing Challenges For SBME

- Scoring / feedback
- Stability/Durability
- Establishing the validity of the simulation training/ evaluation
  - Research
CURRICULAR THEORY

Prideaux, D. BMJ 2003;326:268-270

Outcomes-Based: Grant/Wiggins/McTighe

“Backwards” Design
Summary

- **Start with assessment**: How will you show the public and the learners that they have acquired the knowledge/skills/values/competencies?
- Consider assessing performance (do they do it?) in your programs
- “Planning Backwards” will help us all move forward in facilitating better teaching and learning
☐ Thank you!