Standard Setting for Performance-Based Assessments

John (Jack) Boulet, Ph.D.

Foundation for Advancement of International Medical Education and Research (FAIMER)

Jornadas de Educación Médica 2010
“Educación por Competencias: Cambio de Paradigma”

Universidad Nacional Autónoma de México
Facultad de Medicina UNAM
February 2010
Purpose

- To provide an overview of some standard setting methods that can be used for performance assessments, including simulation-based activities in the healthcare professions
  - Brief overview of standard setting for MCQ examinations
- Perform a standard setting exercise for an assessment exercise that utilizes simulated patients
Learning Objectives

- Be able to choose an appropriate standard setting methodology
- Design a basic standard setting study
- Understand and evaluate the standard setting process
  - Experience as a panelist
Background: Simulation-Based Assessment

- Performance-based
- Series of interactions in simulated encounters (usually)
- Numerous scoring options
  - Analytic
  - Holistic
  - Combinations
Background: Standard Setting

- No widely accepted and validated standard-setting methods for use with performance assessments
  - adaptations of MCQ-based methods are not completely satisfactory
  - many “new” techniques look promising but additional research is necessary
Standard Setting

- Process used to arrive at a passing score
  - competent to practice, credential, master, etc.
Standard Setting Issues

- All standard setting is judgmental
  - Empirical methods facilitate judgments
  - Need procedures that allow judges to make use of their expertise
- Judging minimally acceptable performance on complex cognitive tasks can be difficult
Why Set Standards?

- Is there a compelling reason to set a passing score?
  - What method?
- Are inferences implied by the application of a cut-score warranted?
  - Validity
Questions to be Asked

- Who should set the standard?
  - experience + legitimate interest in the outcome
  - knowledge of the content being assessed
- How many judges should be involved?
  - Errors due to sampling of raters may be larger than errors associated with case sampling
- How can the judgments be evaluated?
  - variability in ratings
  - Consequences
- Compensatory or non-compensatory standards?
Performance Standard vs Cut-Score

- Performance standard
  - Qualitative description of an acceptable level of performance and knowledge required in practice
  - “Conceptual” definition of competence
  - Performance standard is a construct

- Passing score (cut-score)
  - Selected point on the score scale that corresponds to this standard
  - “Operational” definition of competence
  - Cut-score is a number
Scoring Models

- Conjunctive
  - Candidate must obtain a passing score on each component/dimension of the assessment

- Compensatory
  - Candidate can compensate for poor performance in one area with excellent performance in another
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 performance)
  - Leniency/severity of judges will affect content-based performance standards

- **Norm-Referenced**
  - Standard set based on performance of some selected group
    - e.g., 70% pass
  - Standard will change as the ability of the normative sample changes
## 10 Steps for Setting Standards

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Select a large and representative panel</td>
</tr>
<tr>
<td>2.</td>
<td>Choose standard setting method, prep materials</td>
</tr>
<tr>
<td>3.</td>
<td>Prepare descriptions of performance categories</td>
</tr>
<tr>
<td>4.</td>
<td>Train participants to use the method</td>
</tr>
<tr>
<td>5.</td>
<td>Compile item ratings/judgments from participants</td>
</tr>
<tr>
<td>6.</td>
<td>Facilitate discussion among participants</td>
</tr>
<tr>
<td>7.</td>
<td>Provide opportunity for second round of ratings</td>
</tr>
<tr>
<td>8.</td>
<td>Provide final opportunity to review information</td>
</tr>
<tr>
<td>9.</td>
<td>Conduct evaluation of standard setting process</td>
</tr>
<tr>
<td>10.</td>
<td>Assemble documentation of standard setting process</td>
</tr>
</tbody>
</table>
Test-Centered Methods

- Primarily used for MCQ tests
  - Nedelsky (1954)
    - Review items and identify options a minimally competent examinee would rule out as incorrect
  - Ebel (1972)
    - Make judgments about items based on difficulty and relevance
  - **Angoff (1971)**, modified/extended
    - Most common method for credentialing examinations
Angoff Procedure

- **Task**
  - MCQ - estimate the proportion of minimally acceptable examinees who would get item correct

- **Cut-score**
  - For each judge, sum the proportions/points across items
  - Compute mean or median across judges
Angoff Methods (Brief Exercise)

- 4 Items
- Exam targeted at graduating medical student
- Of 100 minimally competent examinees, how many would get the item correct?
An 8-year old boy is stung by a bee. Within 5 minutes, he develops a 2-cm, raised, red, swollen lesion at the site of the injury. Which of the following findings will be predominant in tissue from the lesion?

A. Foreign body reaction  
B. Hemorrhage  
C. Lymphocytic infiltration  
D. Neurophilic migration  
E. Vasodilation
A six-month old boy bruises easily and has bleeding gums on several occasions for 2 months. A maternal uncle has a bleeding disorder. Examination shows several small bruises on the legs. Partial thromboplastin time is prolonged, and prothrombin time is normal. Which of the following is the most likely coagulation factor deficiency?

A. Factor III
B. Factor VII
C. **Factor VIII**
D. Factor X
E. Factor XIII
Serum samples from a normal woman with a history of regular 28-day menstrual cycles show a peak in the serum concentration of 17β-estradiol over the past 12 hours. No progesterone is detectable. Within 3 days, which of the following is expected to occur?

A. Cessation of menstruation
B. Decreased basal body temperature
C. Ovulation
D. Onset of menstruation
E. Regression of the corpus luteum
Item 4

Large amounts of the artificial sweetener aspartame should be avoided in children who have which of the following metabolic disorders?

A. Diabetes mellitus
B. Phenylketonuria
C. Hereditary fructose intolerance
D. Lactose intolerance
E. Maple syrup urine disease
Examinee-Centered

- Direct judgment about status of persons on the construct of interest
Methods

- **Contrasting Group**
  - Form 2 (or more) groups of performances (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
  - ECFMG CSA

- **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 Simulation-Based Examinations

- Intuitively appealing
  - can judge actual performance
- May be difficult to categorize performances for some examinees
- Requires numerous “expert” judgments
Process

- Define the performance standard
- Level of mastery needed?
  - Surveys
  - Focus groups
  - Analysis of practice patterns
  - Behavioural descriptors
    - Operational definitions of the skills needed
Process (cont.)

- Training panelists
  - Have them take exam under exam-like conditions
- Selection of performance samples
- Identification of aberrant panelist ratings
Process (cont.)

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

- Figure out where panelists agree and disagree
Standards Validation

- Compare results using different methods
- Create parallel panels
  - Equivalent standards?
- Gather procedural validity evidence
Exercise
Define Performance Standard

- Who are we setting standards for?
  - Level of performance
- Description adequate/inadequate performance
  - May be specific to particular patient management task
Selection of Simulation Scenario and Performance Samples

- Represents practice domain
- Valid scoring system
- Samples available from performance continuum
Description of Performance Task

- Clinical skills assessment
  - Interview and evaluate the patient
    - History taking
    - Physical examination
  - Communication
    - Written notes
- 15 minutes
- 11 stations
Performance Criteria

- Communication skills
  - Interviewing and collecting information
  - Counseling and delivering information
  - Rapport/ personal manner
  - Humanistic qualities
  - Professionalism
CASE 1 - Dementia

- 65-year-old female brought to the doctor by her daughter, who thinks that the patient is becoming forgetful
- Numerous incidents of forgetfulness
- No problems with ADL
- Hypertensive*
- Short-term memory impairment, not oriented to time
- Daughter and patient disagree about plans for patient’s living situation
CASE 2 - Fatigue

- 47-year-old male patient who complains of two months of fatigue, constipation, and weight loss
- Sleeps approximately 4 hours per night
- Lack of interest in daily activities
- Works two jobs
- Otherwise healthy
- Family supportive
- Recently witnessed an armed robbery
- Smoker
CASE 3 - Fall

- 80-year-old female patient who complains of pain in the right hip after a fall yesterday evening
- Patient is not sure exactly how it happened
- Immediately after the fall she felt fine
- Now experiences weakness and pain, abnormal gait, bruising
- Hypertensive
- Lives alone
- Patient requests hospital admittance
Panelist Task

- Watch video
- Rate performance (adequate, inadequate) based on
  A) Your expertise
  B) Expected level of practitioner
- Repeat
## Panelist Summary

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Candidate</th>
<th>Inadequate</th>
<th>Adequate</th>
<th>% adequate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Standards Validation

- Do decisions based on passing scores achieve the purposes of the organization while avoiding any serious negative consequences?
  - Compare results to past exercises
  - Provide consequential data
  - External validity?
Final Thoughts

- Standards are not perfect!
  - Generalizability related to choice of materials/ methods/ panelists
- Standards may need to be adjusted
  - False positive/ false negative rates
  - Decision consistency
- Standard setting process MUST be evaluated
  - Technical documentation