The Reliability and Validity of Simulation-Based Assessments

John (Jack) R. Boulet, Ph.D.

Foundation for Advancement of International Medical Education and Research

Simulation in Medical Education (SiME) Lecture, Stanford University School of Medicine, Stanford, CA, March 28, 2007

Purpose

- To provide
  - examples of where simulations are currently used for assessment and evaluation (e.g., certification/licensure)
    - Types of simulations
  - a synopsis of some of the unique challenges of using simulations for assessment and evaluation
    - Psychometrics (reliability, validity)
  - Future directions

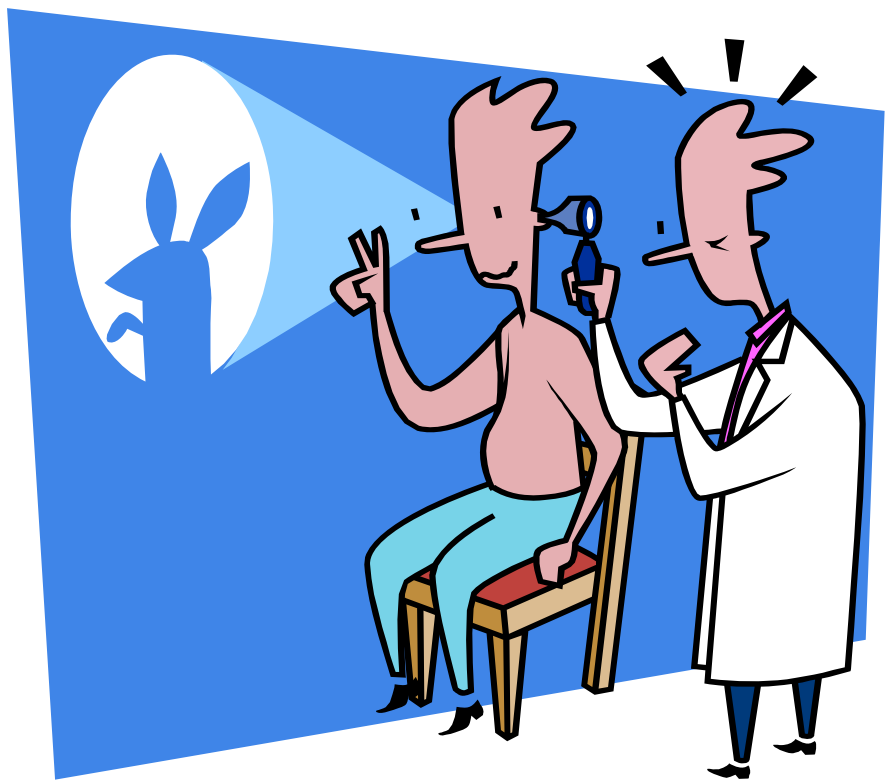
Assessment Simulations in Medicine

- United States Medical Licensing Examination (USMLE™)
  - Step 2 CS (ECFMG Clinical Skills Assessment)
  - Step 3
- National Board of Osteopathic Medical Examiners
- Medical Council of Canada
- General Medical Council (UK)
  - Professional Linguistics and Assessment Board
- Various specialty board examinations
  - American Board of Emergency Medicine
- etc.

- Dillon, G.F., Boulet, J.R., Hawkins, R.E. & Swanson, D.B. (2004). Simulations in the United States Medical Licensing Examination™(USMLE™). Quality & Safety in Health Care, 13(Suppl 1), i41-i45.

Types of Simulations

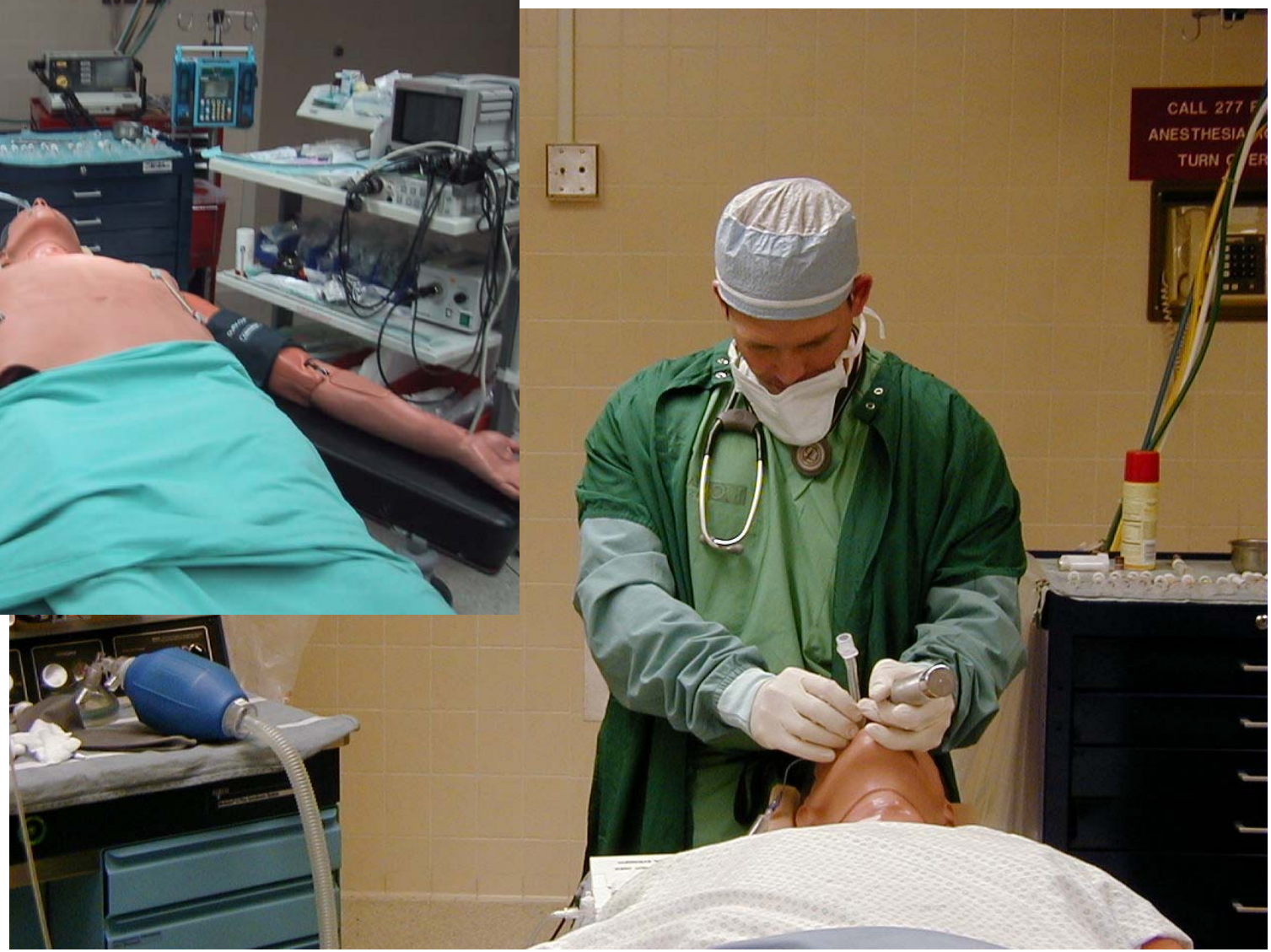
- Paper-based patient management problems
- Computer-based clinical scenarios
- Part-task trainers
- **Electromechanical mannequins**
- **Standardized (simulated) patients**



Electromechanical 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

Electromechanical Mannequins

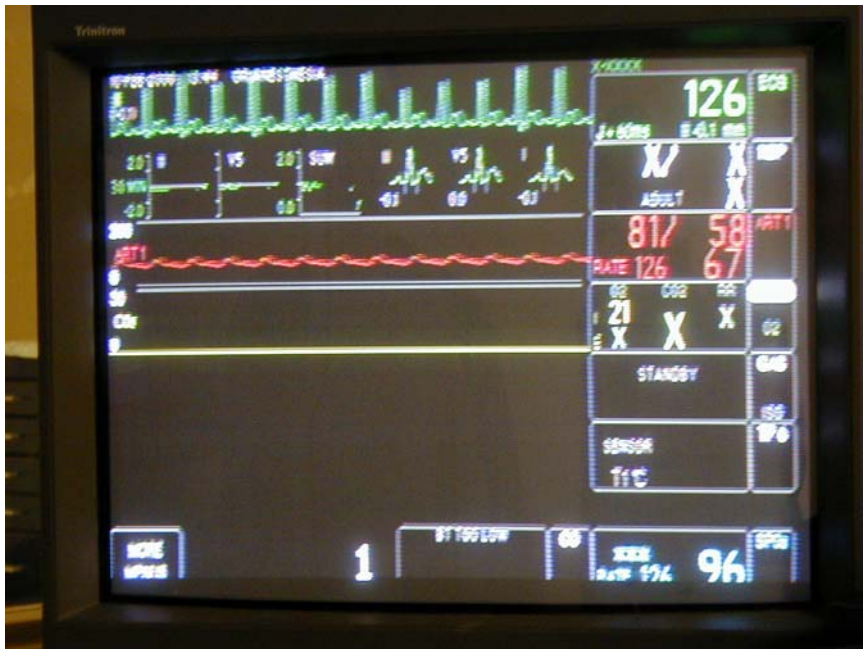


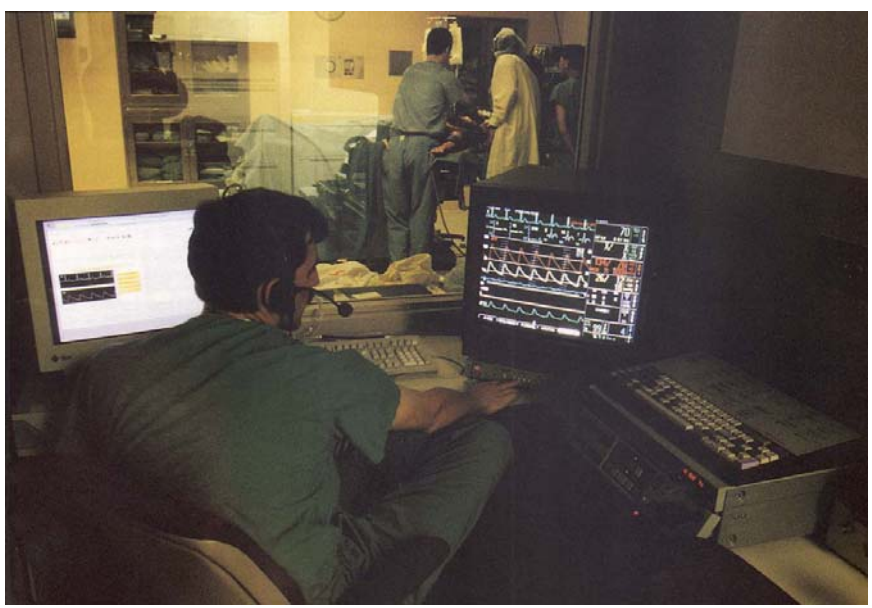
Standardized (Simulated) Patient Examinations

- Performance-based
- "Standardized"
  - Same conditions for all test takers
- Series of interactions in simulated encounters
- Numerous scoring options
  - Explicit process
  - Implicit process

Why Use Simulation for Assessment and Evaluation?

- Model “real” situations
- “standardization”
  - Variability of “real” patients
  - Introduce complexity in controlled way
- Many events are rare
- Errors not reversible
- Errors are expensive
- Patient safety



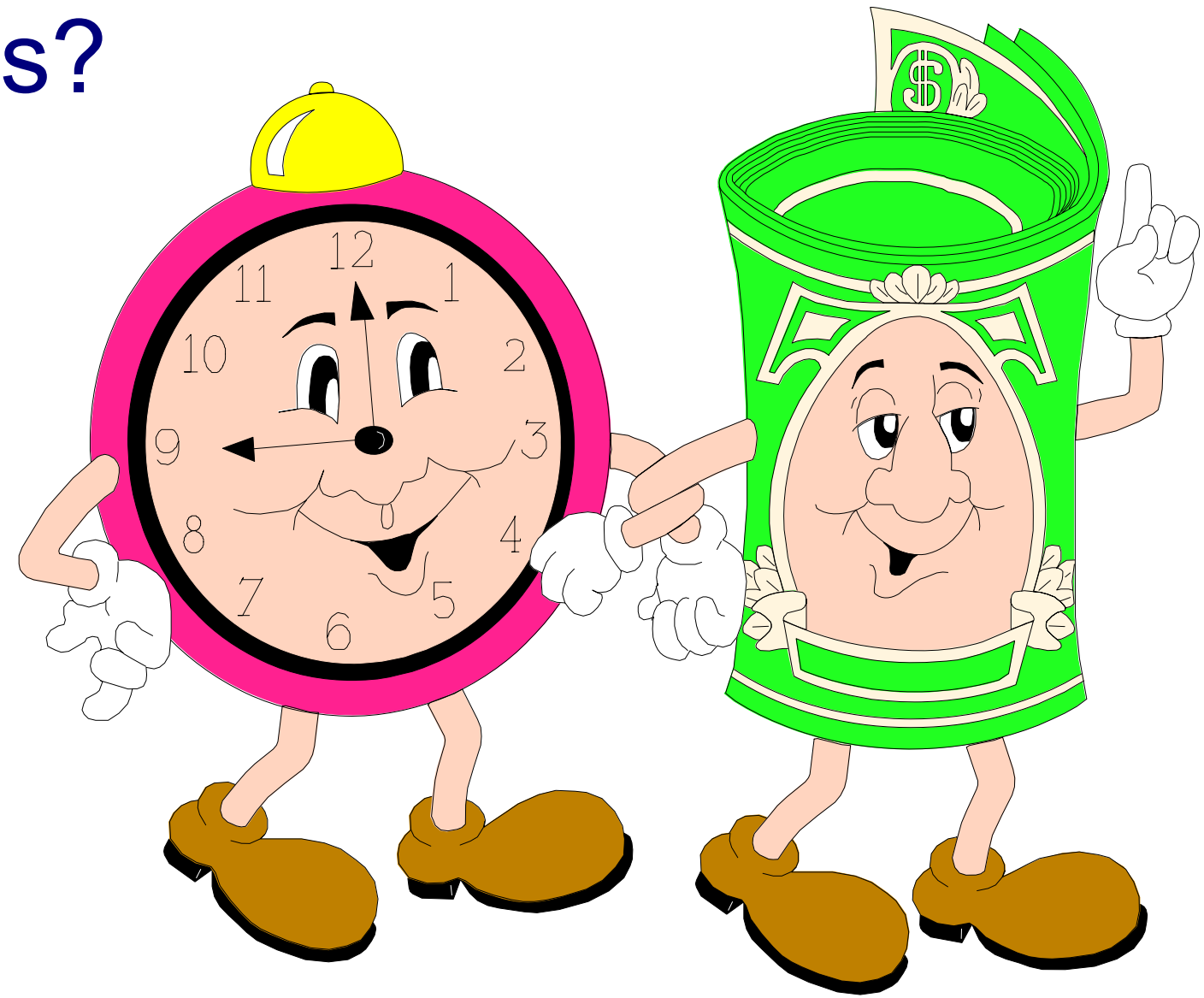


Patient Safety

Challenges

- Cost
- Logistics
- Interdisciplinary cooperation
- Integration
- Measurement Issues
  - Scoring
  - Generalizability/ Reliability
  - Validity

What is the Societal Cost of Having Providers with Inadequate Skills?

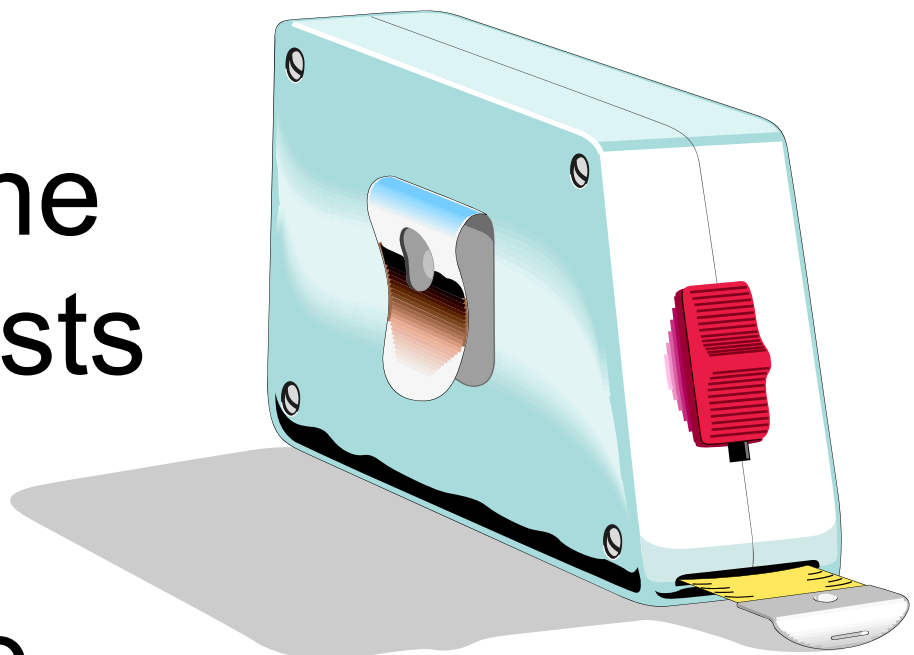


Logistics

Cooperation/ Teamwork

Key Measurement Issues

- Evaluation (scoring) of the performance
- Generalization of the results to similar tests
  - reliability
- Extrapolation of the assessment results
  - validity



Types of Scores

- Explicit Process
  - Checklists
  - Key actions
- Implicit Process
  - Rating scales
    - Timing
    - Sequencing

Intraoperative Asthma Episode

- Review Vital Signs
- (Key)Increase $FI0_2$ to 100 %
- Increase Anesthesia Depth after Increase Fi02
- Establish Lung Compliance is Increased by Hand Ventilation
- (Key) Auscultate Chest
- (Key)Diagnose Presence of Bilateral Wheezing
- Above Steps in Less than 60 seconds
- Pass Suction Catheter Through Endotracheal Tube
- (Key)Begin Nebulizer Therapy (Any B-agonist or Combined Atrovent)
- Corticosteroid IV
- Beta-Agonist IV
- Suggest Arterial Blood Gas
- Order Chest X-ray

Checklist for “Atypical Pneumonia” Standardized Patient Case

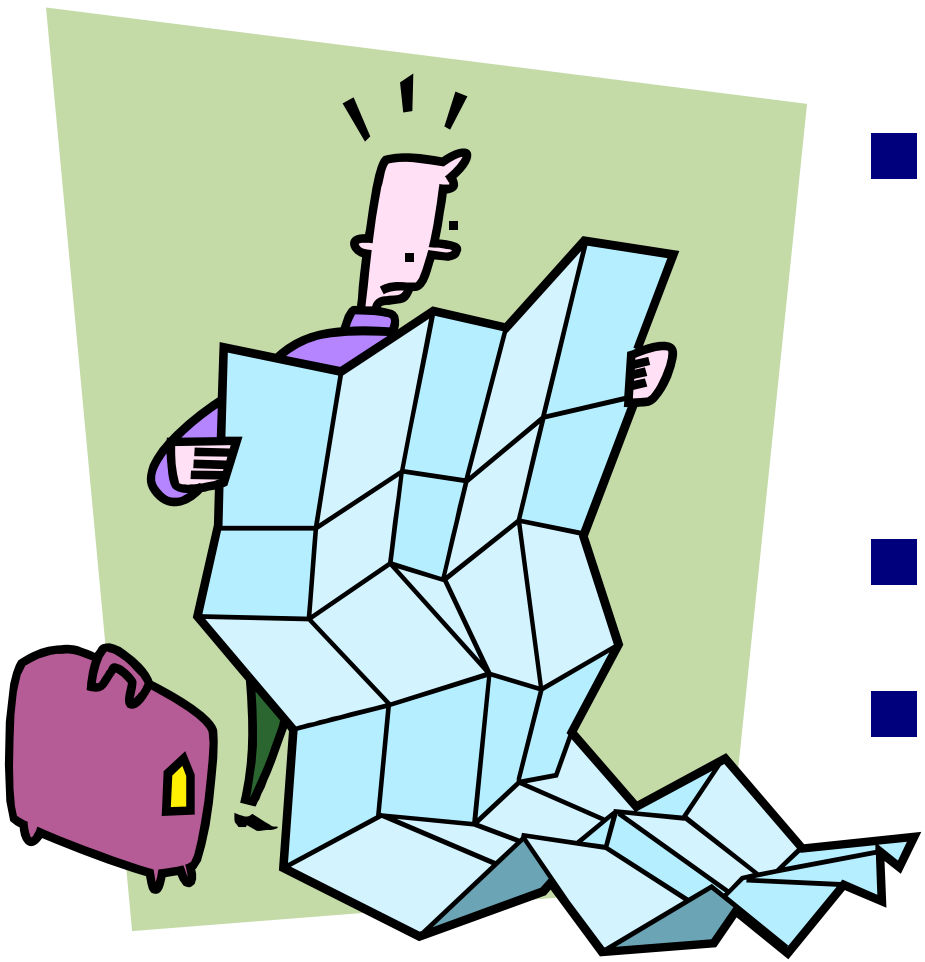
- Muscle or body aches
- Fever, chill, sweats
- Pain when taking a deep breath
- Medications
- Vomiting, diarrhea
- Etc.

- Examines throat
- Palpates for anterior cervical lymph nodes
- Flexes neck
- Etc.

Advantages and Disadvantages of Checklists

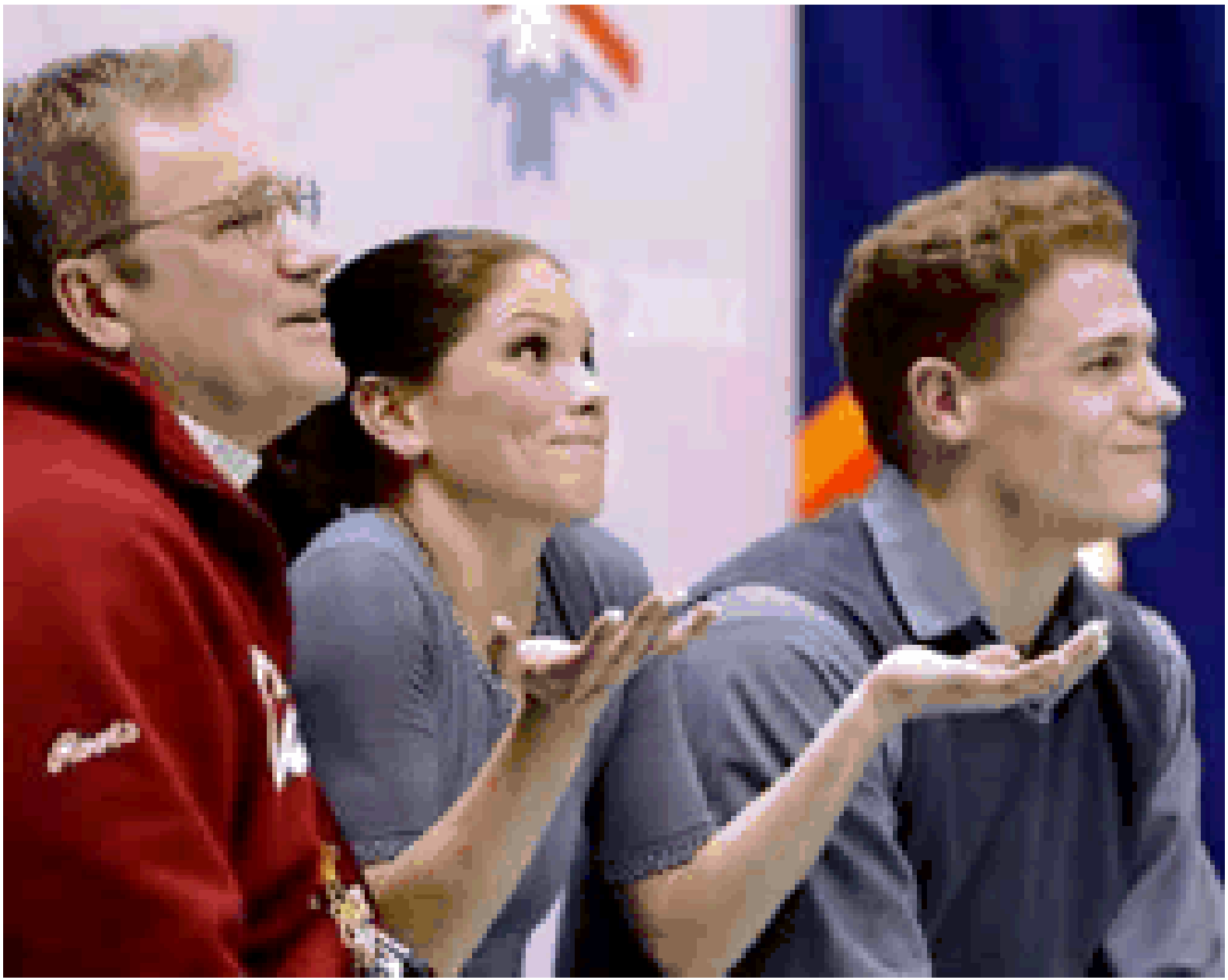
- Fairly easy to develop
- “Objective”
- Record of what was done (feedback)
- Can be used by non-physicians
- Students perceive that they are being evaluated by patients
- Difficult to assess complex skill sets

Checklists for Assessing Acute Care Skills



- Certain actions are much more important than others
- Sequence is important
- Timing is important

Holistic (“Expert”) Ratings



Patient Note Exercise

ECFMG CSA USMLE Step 2 CS

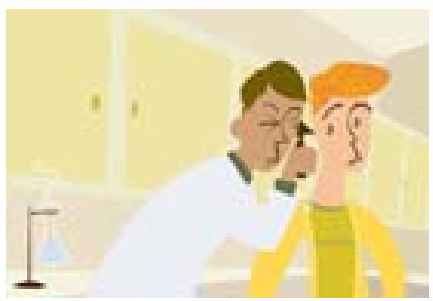
PN Number

Test Date | Session Number | Room Number
Candidate | Candidate Number
Rater ID | Encounter Number

□□□ □□□ □□□

HISTORY - Include significant positives and negatives from history of present illness, past medical history, review of system(s), social history and family history.

PHYSICAL EXAMINATION - Indicate only pertinent positive and negative findings related to patient's chief complaint.

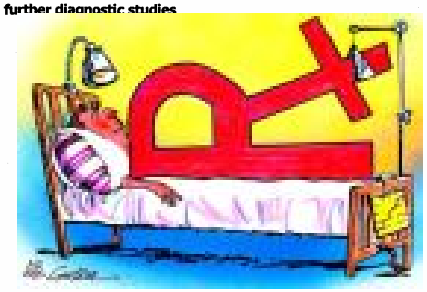


DIFFERENTIAL DIAGNOSIS - In order of likelihood write no more than 5 differential diagnoses for this patient's current problems.

1.
2.
3.
4.
5.

DIAGNOSTIC WORK UP - Immediate plans for no more than 5 further diagnostic studies

1.
2.
3.
4.
5.



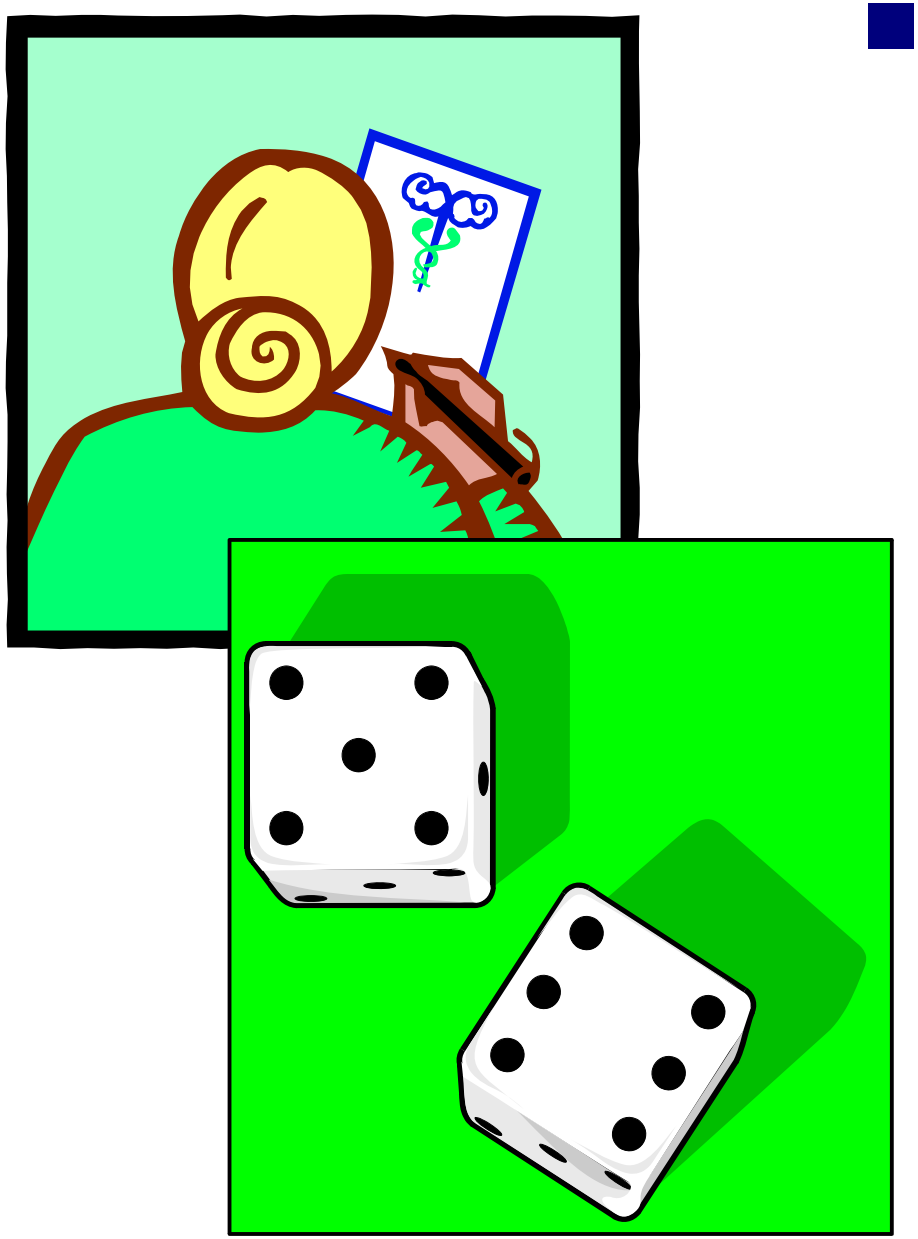
Advantages and Disadvantages of Global Ratings

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

Who Should Provide the Scores?

- Expert (physician, nurse, etc.) examiners
  - "face" validity
  - Expertise
    - Practice of medicine is complex
  - Perceived subjectivity
- Other 'observers'
  - Objective
  - First-hand understanding of skill being measured
  - Economical/ efficient

Reliability



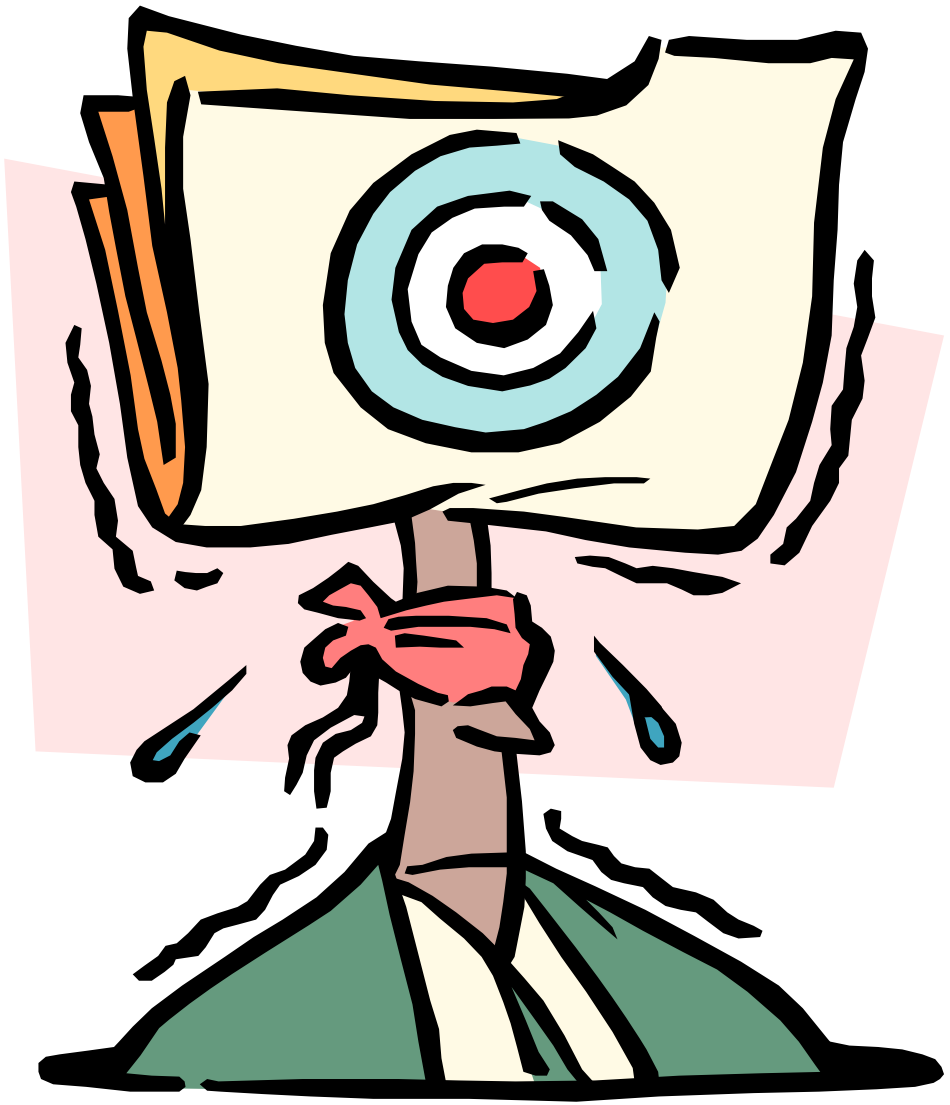
- **How consistent are the examinee/trainee scores?**
  - want to ensure that an examinee's observed score is a reasonable reflection of his/her "true" ability
  - minimize errors of measurement

Sampling Perspective

	Judge 1	Judge 2	Judge 3	…	Judge n
Case 1	A B C	A	A	A	A
Case 2	B	C			
Case 3	B		C		
…	B			C	
Case n	B				C

Enhancing Precision

- Choice and number of tasks
  - Task specificity
- Raters
- Settings
- Administration conditions
- etc.



Lessons from the Literature

- **There are problems with all scoring systems**
  - Find ways to minimize measurement errors
    - Training the raters
  - Generalizability studies to estimate error sources

Validity

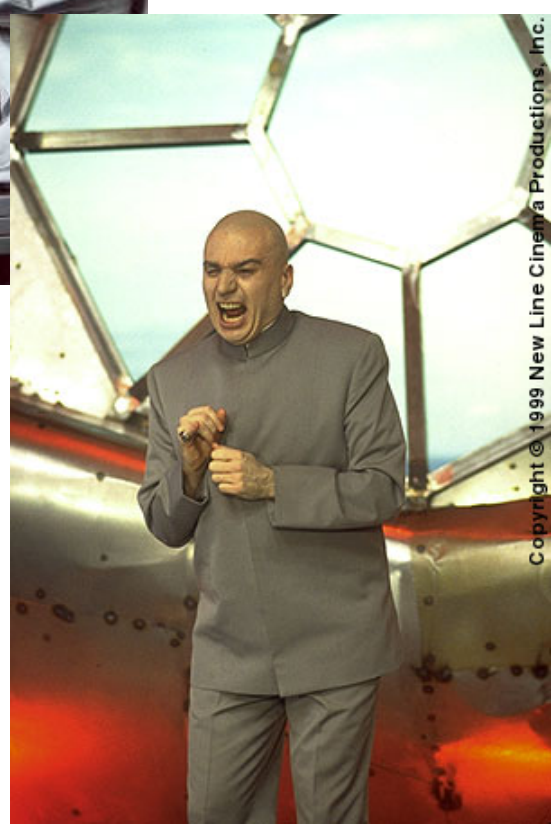
- Development of evidence providing "… a sound scientific basis for the proposed score interpretations"
- Does the assessment provide measure of what it is supposed to?

Assuring the Validity of the Assessment Scores

- Case/ simulation Development
  - Sampling
    - skills
    - content area
  - Scoring criteria
    - necessary tasks/questions to provide patient care

Validity Evidence

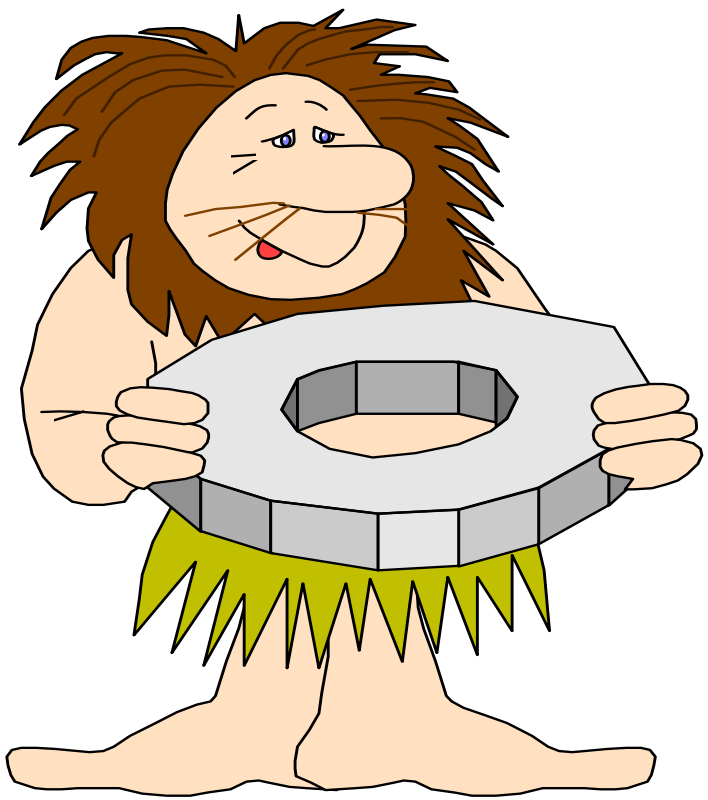
- Start process as evaluation is being developed
  - Content
  - Response processes
  - Internal structure
  - Relationships with criterion measures
  - Consequences
    - evaluation drives learning



Test Content for Simulations

- Skills
  - Data Gathering (History Taking & Physical Examination)
  - Doctor-patient communication
  - Written communication
- 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

Scenario Development Issues



- Cases (simulations) are "vehicles" to measure skills
  - Who are the "target" examinees?
  - Specificity
  - Difficulty
  - Essential maneuvers and questions?
  - Sampling from domain

Content Under-representation

- Some conditions cannot be simulated very well
  - Not as important for basic skills
- Programming for mannequins is imperfect

Response Processes

- Timing of scenarios
  - Evaluation is not “speeded”
- “Item” (scorer) issues
  - fatigue does not affect accuracy
  - candidate performance remains constant
- Feedback
  - realism, appropriateness of content, etc.
- Validation of standards
  - convergence of expert judgments and decisions based on scores

Internal Structure

- Internal correlations among assessment components
  - Biomedical skills - communication
- How scenarios function for identifiable subgroups (controlling for ability)
  - Interaction of rater-candidate characteristics does not appreciably impact assessment scores
  - Threats to validity

Relationship with Other Variables

- External associations with other medical test scores
- Survey of stakeholders
  - residents with successful performance judged to be “ready” for advanced training/ practice?
- Trainee characteristics
  - Individuals with more advanced training expected to perform better

Performance with “Real” Patients

- Difficult to establish “predictive” value … at least in the short term
  - Aviation simulation
  - Driver’s test

Consequences of Testing

- Candidate self-selection / preparation
  - want to learn
  - specialized preparation courses
- Changes in curriculum
- Development of training centers

Discussion

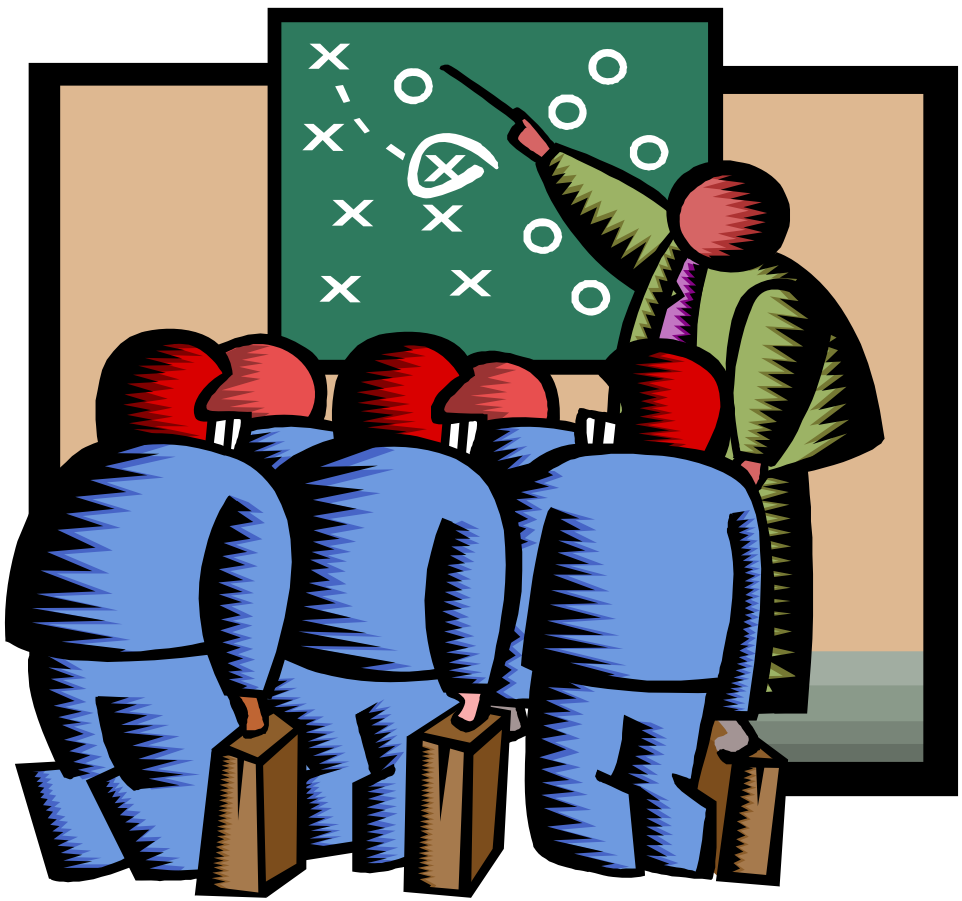
- Utility of simulations for assessment is dependent on
  - Content
    - Proper case development is the key
    - Expert judgment
    - Relevant to practice
  - Developing reliable and valid scoring systems
    - Psychometric analyses
    - Research

Going Forward

- Aligning the scoring system with the complexity of the task
- Generalization of results of the performance
- Research to support the validity of the scores

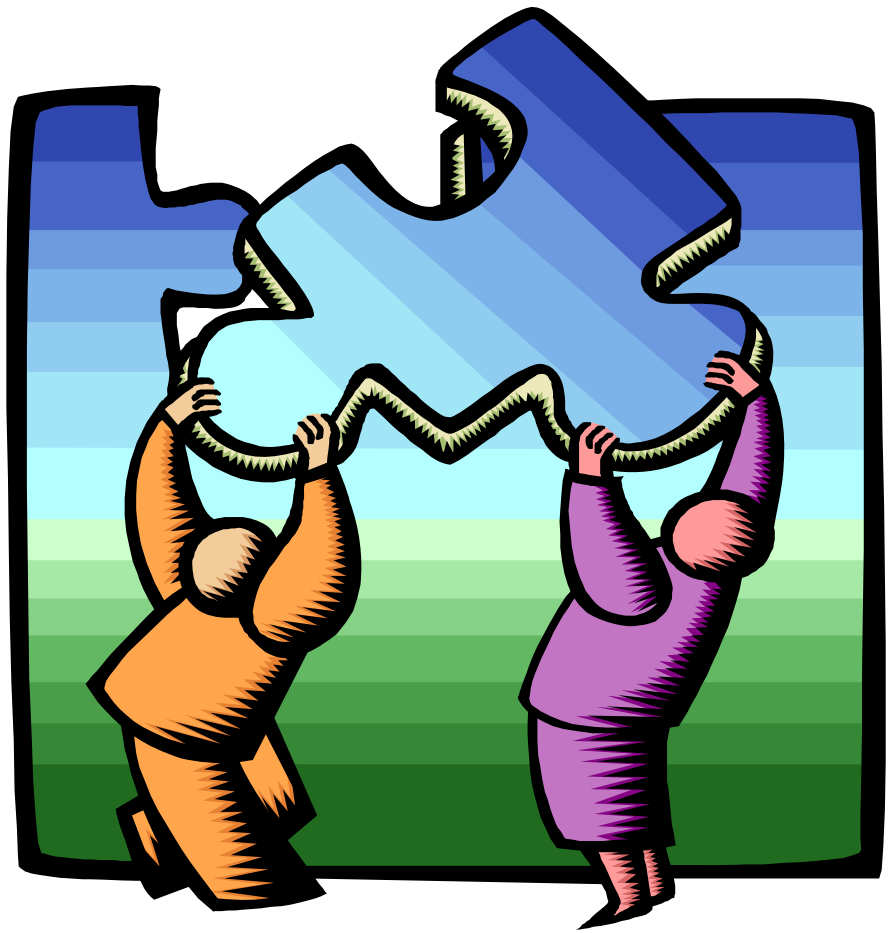
Scoring Systems

- Explicit outcome
  - Patient status
  - complications, etc.
- Combined criteria
  - Sequence
  - Timing



Generalization

- Selection of tasks
  - “evidence based”
- Structure/ length of the assessment
- Assessor training
- Familiarity with simulator



Research (Validity)

- Examinee responses
  - Effect of knowledge of scoring system
- Patient safety
  - Just-in-time training
- Other assessment domains
  - Teamwork
  - Integrated simulations