

Simulation in High-Stakes Exams

Challenges and Problems

Acknowledgement of the major source of information for this presentation:

- Boulet, John R., PhD. Summative Assessment in Medicine: The Promise of Simulation for High-stakes Evaluation. Academic Emergency Medicine 2008; 15:1017-1024.

High-Stakes Exams in USA Using Simulation: USMLE

- USMLE (United States Medical Licensing Examination)
- Started using computer-based case simulations as part of Step 3 assessment in 1999
- Designed to assess whether a physician can apply his medical knowledge for the unsupervised practice of medicine by using computer-based simulations

USMLE: Step 2

- In 2004, USMLE introduced Step 2 Clinical Skills (CS) examination
- This consists of using Standardized Patients in a multi-station exam
- Uses simulated clinical encounters to assess history taking, physical exam skills, and oral communication skills with a patient and written communication skills with the health team

Other High-Stakes exams using simulated clinical encounters

- NBOME: National Board of Osteopathic Medical Examiners use the COMFLEX (Comprehensive Osteopathic Medical Licensing Examination)

Challenges involved in these exams:

- Modeling a scenario to measure defined skills
- Trying to create realism by modeling scenarios after actual clinical practice
- Standardizing what the SP (standardized patient) says and does in response to the examinee
- Standardizing the evaluators who score the clinical encounters

Challenges Continued

- Creating measuring tools that minimize subjectivity
- Creating measuring tools that have consistency from one examiner to the next
- Determining the psychometric properties of the measures
- Establishing the reliability of the scores
- Validating inferences one can make based on the scores
- Overall fairness of the scores and consistency from center to center

What is the Bottom Line of these Exams??

- If they pass the exam, are they competent in the real world, and if they fail the exam will they be incompetent in the real world????
- Is there any data on this????

Data on Exam Scoring to real world competency

- No hard data
- However, in multiple medical schools it has been noticed that there is a very strong correlation to how a student does on these exams and to how the student has generally performed in their clinical rotations
- Students that have performed poorly on the clinical rotations have had more difficulty passing the exam and their clinical performance can be used as a predictor for those who will tend to fail the exam

4 Areas that need addressed for High-stakes Exams

- Defining the Skills and Choosing the Appropriate Simulation Tasks
- Developing the appropriate metrics
- Assessing the reliability of test scores
- Providing evidence to support the validity of test score inferences

Defining Skills and Choosing the Appropriate Simulation Tasks

- Test must have clear purpose
- Well defined set of skills
- Well defined knowledge base needed for the exam
- Well defined target group for the exam with the expected skills and knowledge base

Difficulties with SP

- Difficult to impossible to find clinical abnormalities that are the same from one patient to another
- Despite training, there can be significant variability to answers given by the SP to the examinee's questions often giving misleading information. The directions and coaching the SP must be very clear and it takes practice and review to have consistency

Manikins

- Can reproduce certain clinical findings accurately and consistently
- However, cannot adequately reproduce many associated findings which can diminish or lessen their clinical accuracy (example: pneumothorax)
- Must understand their technical limitations and accuracy

Developing Appropriate Metrics

- Analytic
- Holistic or Global

Analytical scoring metrics

- The prevailing methodology for analytical scores involves the use of checklists
- Committees are employed to determine specific checklist content
- Checklists may include history taking questions, physical examination maneuvers, and management strategies

Analytic Scoring Metrics

- Usually based on checklists
- May or may not weight items

Experience with checklists

- Checklist have provided modestly reproducible scores
- The are dependent on the number of simulated scenarios to be accurate

Problems with checklists

- Although objective in terms of scoring they can be subjective in terms of construction
- There can be considerable debate as to which actions are important or necessary
- Without expert consensus, one could question the validity of the scenario scores
- Guidelines are specific for some clinical problems and non-specific and variable for others

Other problems with checklist

- Use of checklist may promote behavior such as:
- Employment of rapid-fire questioning techniques and or
- Performing as many physical examination maneuvers as possible in the time allotted to accrue more points (example is the Heart Code ACLS)

More problems with checklists

- Checklists may not be conducive to recording/scoring the timing or sequencing of tasks
- There may be several different pathways to handle a given scenario with each pathway requiring a different number of critical actions making the scoring more difficult to compare from pathway to pathway

Holistic or Global Scoring Metrics

- Holistic scoring is where the entire performance is rated as a whole
- Although this would appear to be more subjective, there is evidence that global rating scales are often adequate, especially in measuring complex and multidimensional tasks such as communication and teamwork

(Barker et al. The role of teamwork in the professional education of physicians: Current status and assessment recommendations. *Jt Comm J Qual Patient Saf.* 2005; 31:185-202) (Van Zanten et al. Assessing the communication and interpersonal skills of graduates of international medical schools as part of the United States medical licensing exam step 2 clinical skills exam. *Acad Med.* 2007;82(10 Suppl 1):S65-8.)

Problems with Global Scoring

- Rater training is important in global scoring
- Raters must be trained in using holistic rating scales
- Raters should have well defined, written, training protocols that include specific exercises for various quality assurance measures

Metrics for the Future

- With very sophisticated physiologic manikins or virtual reality simulation, could the future test scoring depend totally on the final outcome??
- For example, with the sophisticated endovascular simulations, could the test be simply to place a stent across a partially obstructed coronary artery without causing complications to the virtual patient and to do so within a time limitation??

Assessing the reliability of test scores

- Sources of error:
- Rater to rater variability
- The number of simulation encounters that are being measured, the more encounters that are measured, the better the reliability
- In studies that have been done, it has been shown that the biggest source of error is when only a few performance samples are being measured

Enhancing the reliability of the test scores

- To minimize rater effects, it is most effective to employ as many different raters as possible for any given examinee
- To minimize error in performance assessment, it is important to have a larger number of performance tasks to accomplish

Supporting the Validity of Test Score Inferences

- Does an “Expert” perform better on the test than a “Novice”?
- Have those who do not pass the test show evidence in other related activities to be having trouble or difficulties? (For example, for a student not passing the simulation test, has this same student had difficulty during his clinical rotations?)
- Do simulation test scores correlate with performance on knowledge testing exams?
- Do the test scores correlate with other data regarding the person’s real world performance when that performance is measured by some standard????

The Type of Format for Our Simulation Testing

- Critical Action Lists
- Start with 100 points and subtract points when a critical action is not done
- Expert consensus for the critical action lists
- Weighing each critical action so that actions omitted that are believed to be more important than others have a higher negative value if omitted

The Format for our Simulation Testing Continued

- Allowance in the scenarios to follow multiple pathways
- Use of the negative scoring (starting with 100 points and subtracting points) allows the scenario to take multiple pathways towards conclusion without penalties or rewards for taking different pathways since only the exclusion of a critical action counts

Explanation of why the negative scoring allows multiple pathways

- Example: Path A has 20 Critical actions associated with it and Path B has 40 critical actions associated with it: If both pathways are done perfectly, then the student would end up with 100 points with either pathway, despite one path having more critical actions than another
- If one pathway is felt to be better than another, this can then be accounted for by giving a set number of negative points for

Results of the Testing

- 15 Candidates
- 11 Passed
- 4 Did not pass
- 3 Examiners per candidate

Results of the Testing

- Scores among the 3 examiners within 10%
- Average score of those who passed was 15 points above the Angoff score
- Average score of those who did not pass was 8 points below the Angoff score
- A panel of experts set the minimum passing grade by a process called Angoffing

Angoff Scoring

- “What percentage of minimally competent physicians would answer this question correctly?”
- If 20% would answer correctly, that is an Angoff score of .2; if 80% answer correctly, that is an Angoff score of .8.
- A five question exam with Angoffs of .2, .8, .6, .7 and .8 would have a passing score of 3.1 in hard numbers or 62% in percentages. ($3.1/5 = 62\%$)
- A really hard exam might have a passing score of 40% and still be valid. A really easy exam might have a passing score of 80% and still be valid.
- A panel of experts meets together and goes over each critical action to decide the Angoff score for each action

Results of Testing Continued

- Of those who did not pass, the 4 candidates asked for a review of their performance
- All performances were videotaped and the tapes were sent to 3 independent experts in the field for scoring
- The 3 independent experts scored these candidates similarly to the original scoring and the scores were again below the Angoff score needed to pass

Now, a Mock Simulation

- We will present a mock scenario with a scenario presentation and critical action list
- All will have a chance to watch the scenario and be graders using the checklist provided
- At the end we will tally the checklists and give the scores to see how the checklists performed
- We will then discuss the strengths and weaknesses of this format for simulation testing