Objective Structured Clinical Examinations (OSCE) IV:

Training Physicians to Rate OSCE Patient Notes

NBME®

National Board of Medical Examiners
3750 Market Street
Philadelphia, PA 19104

Objective Structured Clinical Examinations (OSCE) IV: Training Physicians to Rate OSCE Patient Notes

Training Physicians to Rate OSCE Patient Notes

Lesson Objectives

By the end of this lesson, you will be able to:

- Describe three tools that help physicians rate objective structured clinical examination (OSCE) post-encounter tasks (PETs)
- List the steps needed to train physicians to effectively rate patient notes
- Identify four potential rater biases

Introduction

OSCEs involve a planned interaction between a learner and a standardized patient. An OSCE is an assessment method based on testing and direct observation of student performance during planned clinical encounters or test stations. To complete the examination, students rotate through a series of stations in which they are expected to perform specific clinical skills in a set period of time dependent on the task, normally between 10 and 20 minutes. Often, the learner is asked to engage in a PET. The PET may consist of specific questions relating to the encounter or may require the learner to perform specific tasks such as writing a history and physical or a note for the patient's chart.

When PETs are used for high-stakes purposes, such as contributing to a clinical clerkship grade or deciding learner advancement, it is important that the evaluation of the PET is accurate. Effective training of physicians who are responsible for rating PETs is a key step to ensuring the reliability and validity of the test. This lesson reviews key features of training physicians to rate history and physical or SOAP-style patient notes. SOAP is a format for documentation in patient charts. It refers to four components of the note: subjective, objective, assessment, and plan.

S = Subjective: This is the description of the patient's current condition or history of the presenting illness. It includes statements and feedback from the patient.

O = Objective: This is the description of the facts about the patient's status (eg, results of the physical examination, laboratory results, vital signs) and observations (eg, "the patient appeared short of breath").

A = Assessment: This is the analysis of the patient's status and may include a differential diagnosis.

P = Plan: This is the description of the care to be provided and/or whether a change in care is needed.

Physicians as Raters

OSCE PET scores should be as accurate as possible, especially if they are used in the summative evaluation of learners. The advantages of using physicians as raters include:

- Physicians have the background and experience to understand the task required of the learner.
- Physicians are familiar with the process of interacting with patients and the way in which patient responses dovetail with medical concepts to lead to effective diagnostic and treatment planning.
- When situations arise that are not fully explained in the training materials, physicians can use their clinical judgment to make appropriate decisions concerning rating.

Developing a PET

When developing an OSCE PET, it is important to identify which learner skills are going to be assessed. The following are examples of skills that could be assessed in an OSCE PET:

- Documentation of patient history
- Documentation of the physical examination
- Critical thinking about a differential diagnosis
- Development of a therapeutic plan

Depending on the purpose of the OSCE, you may wish to assess one or a combination of these skills.

Rating Tools

The rating tools used should address the unique aspects of each of the discrete tasks to be assessed. In other words, decide what you want to test and then build your OSCE scenario and rating tools around this. Raters need the right tools to give accurate ratings and to maximize inter-rater reliability. The following table offers a brief description of three tools.

Rating Tool	Description
Case-specific scoring key	Lists desired features that should be present in the learner's response
Benchmark notes	Provides an example of a learner's performance with scoring
Scoring rubric	Provides descriptive anchors across the spectrum of learner behavior to help raters arrive at an accurate assessment of a learner's performance

Case-Specific Scoring Keys

Case-specific scoring keys identify the features that should be present in a learner's response for a specific OSCE scenario. They define characteristics of possible learner responses that range from superior to poor performance. For example, if documentation of the physical exam is one of the goals of the assessment, then identifying specific exam maneuvers that candidates should include for the scenario being tested will help differentiate between outstanding, average, and poorly performing learners.

Example

Consider the following OSCE scenario: *Mr. Jones, a 57-year-old quadriplegic man, presenting with fever and increased malodorous drainage from his stage 3 pressure ulcer.*

There are a number of possible physical examination findings that could be included in a case-specific scoring key for this scenario. Examples of features you might choose to include are:

- Documentation of fever
- Documentation of size and location of wound
- Presence/absence of surrounding erythema
- Description of drainage
- Presence or absence of tracking, visible bone or muscle tissue

Based on the information provided, your key may look something like this:

Physical Examination:

- Temperature of 101.4°F
- Size (3 × 4 cm) and location (sacral) of wound
- 0.5-cm rim of surrounding erythema
- Yellow malodorous drainage
- Base of wound without tracking or visible bone/muscle

Benchmark Notes

Benchmark notes are example notes that illustrate different aspects of the training rubric or case-specific key.

In order to hone their rating skills, raters need to be exposed to a variety of learner performances. This can be accomplished by having raters rate notes from previously scored cases and review the ratings with either a trainer or other raters. Areas that require clarification can be identified. Raters should be provided with several examples of outstanding, average, and poor performance, along with descriptions of which PET characteristics resulted in the ratings. Such explanations need not be lengthy but should reference the tools that the raters are given so that raters can see how these tools are applied in each case.

Scoring Rubrics

There are two main types of **rubrics**:

- **Analytic or Rule-Based Rubrics**: In an analytic rubric, the rater scores individual elements of the performance, then sums the individual scores to obtain a total. Analytic rubrics yield a higher inter-rater reliability but tend to be more time-consuming to develop and learn.

- **Holistic Rubrics**: In a holistic rubric, the rater scores the overall performance as a whole, without judging each component separately. Holistic rubrics are easier to develop but don't have as much inter-rater reliability. However, because ratings based on holistic scorings correlate fairly well with analytic scoring, it is reasonable to employ holistic scoring if time and resources are limited.

There may be instances in which some aspects of both analytic and holistic rubrics are applied in a single evaluation.

How to Train Raters

Effective training of physician raters is essential to ensure the assessment is accurate. The following process is advocated:

1. **Introduce raters to the specific learning objectives of the assessment**: In order to accurately assess learner performance, raters should understand the specific skills that the OSCE is designed to assess. This will help them focus on features that impact these skills.

2. **Explain and demonstrate the rating process**:
   a. Read the note or a portion of the note.
   b. Identify the strengths and weaknesses of the note.
   c. Use the rubric and case-specific key to determine whether the note is poor, average, or outstanding, as well as to synthesize a score.

3. **Allow raters to practice:** Provide raters with annotated benchmark notes for the cases that they will be rating. These can be drawn from previous OSCEs. If previous notes are not available, benchmark notes can be developed using a subset of notes from the OSCE. As in all learning activities, repetition is key to success, so the more benchmarked notes the rater is exposed to during training, the more reliable he or she is likely to be when rating PETs.

Rater Bias

Rater training should also address common rater biases. Biases commonly impacting human raters include:

- **Severity or leniency**: The tendency to give all learners higher or lower ratings than the learner's performance justifies
- **Halo effect**: The failure to distinguish between more than one independent aspect of the learner's performance based on an outstanding or poor performance in a single area
- **Central tendency**: The tendency to use ratings around the midpoint of the scale and avoid ratings at the extremes of the scale
- **Restriction of range**: Diminished extent to which a rater's scores discriminate among learners' performances

Awareness of these common biases may help raters assign fairer PET ratings.

Putting it All Together I

Using the OSCE scenario presented earlier and the following case-specific scoring key, evaluate the post-encounter note provided using the key as a guide. Does the note reflect low or high examinee performance?

Scenario

Mr. Jones, a 57-year-old quadriplegic man, presenting with fever and increased malodorous drainage from his stage 3 pressure ulcer.

Case-Specific Scoring Key

History:

- Chief complaint of fever for 3 days
- History of quadriplegia
- Preexisting stage 3 pressure ulcer
- Increased yellow wound drainage for 3 days
- Foul odor noticed
- No other sources of infection

Physical Examination:

- Temperature of 101.4° F
- Size (3 × 4 cm) and location (sacral) of wound
- 0.5 cm rim of surrounding erythema
- Yellow malodorous drainage
- Base of wound without tracking or visible bone/muscle

Note

Mr. Jones is a 57-year-old man with a history of quadriplegia secondary to a MVC 10 years ago. About 1 week ago, he started feeling feverish and his sacral wound began to drain. The drainage is yellowish and foul-smelling. He has had this wound for about 1 year and it usually isn't painful or draining. He was told it is stage 3. He has not had recent antibiotics for the wound. He routinely goes to a wound care center for treatment. He has back pain but no other new complaints. He denies cough, shortness of breath, nausea/vomiting/diarrhea, burning with urination, foul-smelling or cloudy urine, or other skin wounds.

On Physical Exam:

Appears comfortable, in no acute distress.

Vital signs: Temp 101.4°F

Skin: Wound present over the sacrum. There is yellowish, malodorous drainage and some redness around the wound. The wound base is pink without eschar or exposed bone. It does not appear to track more deeply. There is no tenderness around the wound.

This example demonstrates a PET that falls into the high-performance category. Most key history items are included in this organized and focused note. There are some errors, however. In order to receive full credit, the examinee should have documented that onset of symptoms was 3 days ago, instead of 1 week ago. According to the note, the patient has had back pain, when in fact pain is not part of the history.

The physical examination is focused and is limited to items that are most pertinent to this presentation. It includes many of the required features that describe the sacral wound. In order to receive full credit, the examinee should be more specific by adding detail about the size of the wound and the extent of surrounding erythema.

Putting it All Together II

Using the OSCE scenario presented earlier and the following case-specific scoring key, evaluate the post-encounter note provided using the key as a guide. Does the note reflect low or high examinee performance?

Scenario

Mr. Jones, a 57-year-old quadriplegic man, presenting with fever and increased malodorous drainage from his stage 3 pressure ulcer.

Case-Specific Scoring Key

History:

- Chief complaint of fever for 3 days
- History of quadriplegia
- Preexisting stage 3 pressure ulcer
- Increased yellow wound drainage for 3 days
- Foul odor noticed
- No other sources of infection

Physical Examination:

- Temperature of 101.4°F

- Size (3 × 4 cm) and location (sacral) of wound
- 0.5 cm rim of surrounding erythema
- Yellow malodorous drainage
- Base of wound without tracking or visible bone/muscle

Note

CC: “Fever”
HPI: Mr. Jones is a 57-year-old man with a history of quadriplegia secondary to a car accident 10 years ago. He comes in today complaining of a new fever. He was diagnosed with a pressure ulcer about a year ago. He hasn’t had any previous infections in his wound. He sees a wound care specialist for his wound on a regular basis. He is turned regularly. Mr. Jones has no cough, no SOB, no heart palpitations, no change in urine, no diarrhea, no N/V, no muscle aches or pains, no hot or cold intolerance, and no change in his hearing or vision.

On Physical Exam:

NAD

CV: RRR no m/g/r
Lungs: clear to auscultation bilaterally
Abdomen: soft NT/ND
Extremities: no edema noted
Skin: 3 cm × 4 cm decubitus ulcer located directly over the sacrum. There is a yellow, foul-smelling drainage coming from the wound. No granulation tissue or eschar is noted. No other rashes are noted.

This is an example of a note that would receive a lower score than the previous example. In the history, the examinee has successfully documented the presence of a fever, the history of paraplegia, presence of a preexisting ulcer, and several elements that might point to other sources of infection. However, the examinee does not document the duration of the fever, the staging of the previous ulcer, or the presence of drainage or odor. The list of negatives included by the examinee seems more like a review of systems, rather than a targeted search for other sources of infection. This “shotgun” approach omits desirable pertinent negatives such as a description of changes in the patient’s urine, which might be an alternative source of infection in this case.

In the physical examination, the examinee has documented the examination of systems that do not add to the understanding of the source of infection. The examinee documents the size, location, and drainage of the wound successfully but omits the fever, surrounding rim of erythema, and pertinent negatives regarding tracking and visible bone/muscle.

Take-Home Messages

When PETs are used for high-stakes purposes, such as contributing to a clinical clerkship grade or deciding learner advancement, it is important that the evaluation of the PET is accurate. Effective training of physicians who are responsible for rating PETs is a key step to ensuring the reliability and validity of the test.

Resources

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

Huot, B. (1990). The Literature of Direct Writing Assessment: Major Concerns and Prevailing Trends. *Review of Educational Research*, 60, 237–63.

Preusche, I., Schmidts, M., & Wagner-Menghin, M. (2012). Twelve tips for designing and implementing a structured rater training in OSCEs. *Med Teach*, 34, 368–72.

Training Raters and Staff. (2009). In J. A. Penny, B. Gordon, & R. L. Johnson (Eds.). *Assessing Performance: Designing, Scoring, and Validating Performance Tasks*. New York: The Guilford Press.

We acknowledge the contribution of these content experts:

Jessica Salt, MD, MBE, FACP, Clinical Skills Evaluation Collaboration

Ellen Turner, MD, Clinical Skills Evaluation Collaboration