

Objective Structured Clinical Examinations (OSCE) II:

Developing Rating Scales and Checklists for OSCEs

NBME®

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

Objective Structured Clinical Examinations (OSCE) II: Developing Rating Scales and Checklists for OSCEs

Developing Rating Scales and Checklists for OSCEs

Lesson Objectives

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

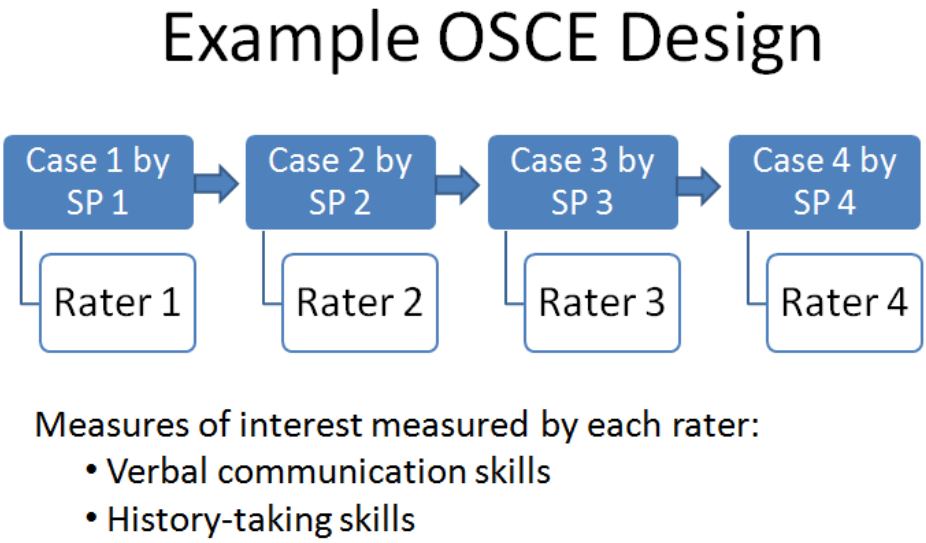
- Describe the differences between rating scales and checklists
- Identify scenarios in which rating scales and checklists might be used
- Explain the implications of scale choice on rater biases and other factors that impact scoring in an objective structured clinical examination (OSCE)
- Explain the implications of scale choice on the reliability and validity of an OSCE

Key Terms

A measurement scale is a scale used to quantify an activity, attitude, or ability. These scales may be used in surveys, educational assessments, psychological assessments, and many other venues. This lesson will focus on the use of different scales in the OSCE scenario.

A rubric is a guideline for scoring each element on a measurement scale.

The following example OSCE will be used throughout this lesson to illustrate concepts and computations: A clerkship would like to administer a standardized patient (SP)-based OSCE observed by trained raters. The goal is to have a single rater for each station assess each student on verbal communication skills and history-taking skills.



Verbal communication is defined here as the ability of the student to use the spoken word to gather information from and share information with the standardized patient, while demonstrating a professional manner.

History-taking is defined here as the skill of the student to obtain relevant clinical information (eg, reason for the office visit, medical background, family history) from the standardized patient via an interview.

Rating Scales

A rating scale is a measurement scale that has a defined rubric. A rating scale allows raters to quantify a variety of skills often assessed during OSCEs, including complex, interrelated, or noncognitive skills. While rating scales can be used to quantify subtle aspects of performance, the scale may be relatively objective or subjective depending on the scale design and training level of raters.

The first step in developing a rating scale is to determine the question the scale is meant to answer. The following are examples of questions on which we could base a rating scale related to verbal communication skills:

- Overall, how satisfactory are this student's verbal communication skills?
- What is the overall readiness of this student to progress to the next stage of education?
- Compared to other third-year students, how are this student's verbal communication skills?

The next step in developing a rating scale is to select the appropriate format for the scale, given what you want to measure. Two of the most common rating scale formats are:

- Likert-type scales
- Behaviorally anchored rating scales (BARS)

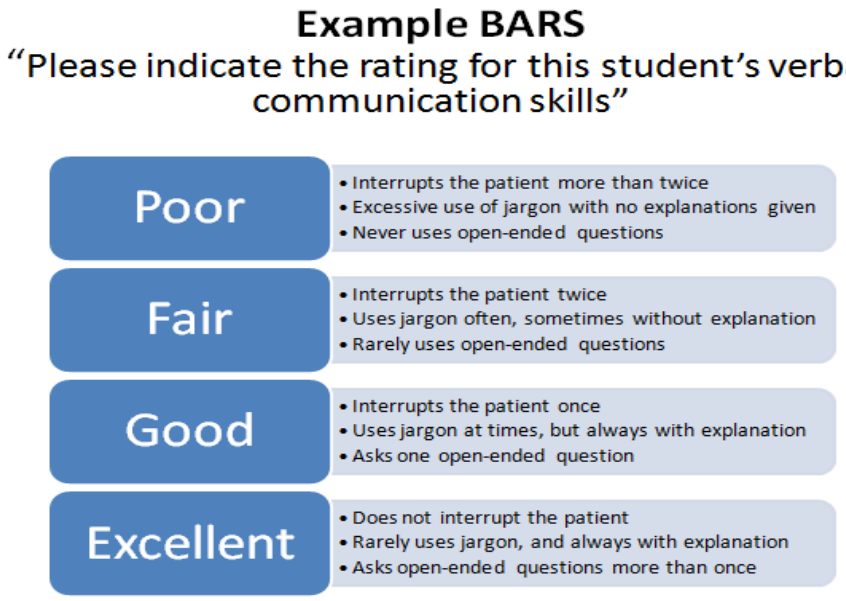
Likert-type scales are typically used for measuring attitudes or beliefs among raters, or for capturing global statements using a symmetric, bipolar rating continuum. Likert-type ratings make the assumption that all points are equidistant; eg, the distance between "strongly agree" and "agree" is the same as between "disagree" and "strongly disagree." An example follows:

Example Likert Scale

"This student demonstrates acceptable communication skills for a third-year student."

	Completely disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Completely agree
This student demonstrates acceptable communication skills for a third-year student	o	o	o	o	o

BARS are a type of Likert scale, but behavior anchors are provided for each scoring point. Thus, BARS items are suited for measuring traits among examinees. On an anchored scale, students are rated according to where their performance lies along a continuum that is defined, or anchored, by specific levels of behaviors or actions. An example follows:



Anchored scales require users to be trained to recognize behavioral distinctions along the scale. The selection and definition of these behaviors, and the training of the raters, can pose meaningful challenges for scale developers.

Likert-type vs. Anchored Scales

The choice of rating scale depends on what you want to measure and how that relates to the skill of interest. In the following table, the first and third scenarios are Likert-type scales. Both are global observations or impressions by the rater that are not anchored by student behavior.

The second scenario is an anchored scale, as a behavioral anchor is supplied as part of the scoring rubric. Although an anchor is shown here only for the highest possible score, additional behaviors are assumed to anchor the other scale points.

Scenario	Likert type	Behaviorally anchored
The faculty member indicates his or her impression of the student's introduction, using a 5-point scale that ranges from Unsatisfactory to Satisfactory.	X	
The faculty member indicates the appropriateness of the student's language for		X

communicating with the patient using a 7-point scale, where the highest value reflects a student who avoids jargon or provides additional explanations for medical terminology.		
The faculty member indicates his or her overall impression of the completeness of the closure, using a 9-point scale that ranges from Unacceptable to Excellent.	X	

Number of Scale Points

For both types of rating scales, there need to be enough scale points to capture meaningful differences in degrees of agreement or behavior between the students being rated, while not having so many that the distinctions become trivial or difficult for the rater to make. Most recommendations suggest no fewer than three points and no more than nine points for either Likert-type or anchored rating scales.

Checklists

The second type of scale addressed in this lesson is a checklist. Checklists can be useful for capturing whether a set of actions is performed correctly. In addition, checklists are often (but not always) more objective than rating scales because the scoring is often based directly on observed actions.

Using a checklist, raters can record what behaviors or skills students are demonstrating via categorical (usually dichotomous) responses for each behavior. One common way to score checklists is to sum the responses, so that students who perform more actions are judged to be better at the set of tasks. An alternate way of scoring is to provide weights for all checklist items that represent the varying importance of each item, with the weights applied prior to summing the items.

Following are some example items similar to those commonly used in history-taking checklists:

- The student asked when the patient first developed a fever.
- The student asked the patient about her family history of cancer.
- The student asked whether the patient's pain was relieved when the patient was lying down.
- The student asked whether the patient had ever used tobacco products.

Checklist items may be ordered or not ordered. If order is not made explicit, a student who asked all the appropriate questions in a confusing or illogical order might be awarded the same number of checklist points as the student who asked the questions in a more logical or expected order. One example of order in a set of checklist items is to require that the student begins with the most open-ended question and then follow up with more specific closed questions.

Checklists can be used to quantify noncognitive skills in addition to clinical skills.

Here are some example items similar to those used in communication skills checklists:

- The student used transitional statements.
- The student checked to make sure the patient understood his or her summary.
- The student allowed the patient to speak without interruption.

One challenge with the use of checklist items in assessing noncognitive skills is the need to clearly define what counts as "Done" or "Complete" for the rater, especially if the behaviors being observed are complex or may be done in a variety of ways. For example, the definition of "transitional statement" in the first checklist item above would need to be clarified for the raters.

Number of Checklist Items

Just as the number of scale points should be carefully determined for a rating scale, the number of checklist items should be carefully considered as well. This is especially true if raters are required to memorize the checklist or use a different checklist for every scenario. Vu et al. (1991) found the length of the checklist impacts the accuracy of standardized patients completing the checklist after an encounter (scoring from memory); they recommended 15 items as ideal or fewer than 30. There should be enough items to cover the construct of interest but not so many that unimportant or nonessential behaviors are listed.

Rater Training

To prevent bias in ratings, it is important to train and monitor raters throughout an OSCE administration. If raters emphasize different skills or are inconsistent in their ratings, the resulting scores may not be useful.

Raters should understand not only how a scale will be used but also why that scale was chosen. The development of a scale or checklist needs to fit with the purpose of the assessment. For example, a checklist about a physician's communication patterns is not useful if the physician is being assessed on his or her ability to explain the treatment plan to the patient.

Raters should also be made aware of common rater biases. A discussion of common sources of bias (error in ratings) will help raters to avoid these biases while rating.

Common Rater Biases

Common rater biases in a performance assessment are:

- Severity/leniency
- Halo effect
- Restriction of range
- Primacy effect

To a certain degree, rater biases are unavoidable. However, awareness of different biases may help raters to avoid them. Some methods to help prevent rater biases are:

- Providing experiences and exercises to help raters make reliable judgments
- Allowing the raters a chance to discuss and refine the rating criteria
- Providing raters with enough opportunities to observe sample performances and locate them along the scale correctly and consistently
- Providing raters with enough opportunities to observe student performance along the entire score scale

Severity/Leniency

Severity/leniency refers to the tendency of a rater to give low ratings (severity) or high ratings (leniency) to all students. This is not always considered a bias but rater training should be used to bring all raters to a shared idea of acceptable levels of severity (or leniency) in order to prevent extreme ratings.

Halo Effect

The halo effect occurs when raters rate participants on other (appealing or not-so-appealing) characteristics rather than the targeted ability or trait to be measured. For example, a rater may give a higher score to a checklist item related to nonverbal behavior because the rater thought the student was friendly or did well on the other checklist items but did not actually display excellent nonverbal behaviors. There may be other aspects of the student that are irrelevant to any construct.

Restriction of Range

Restriction of range, or central tendency, refers to the likelihood of raters to rate all performance in the middle of the scale, regardless of all the options available.

Primacy Effect

The primacy effect encourages raters to assume that the next performance is similar to the performances they have just seen, causing them to compare an individual with their group, rather than the scoring rubric.

Reliability

The type of scale that is selected can have an impact on the overall score reliability. For example, a rating scale that is too easy or too difficult for the group of students (eg, all of the students exhibit all of the behaviors to obtain the highest scores, or none of the students exhibit behaviors past the lowest scores) would reduce the reliability of the overall score. The following scenarios could also reduce the reliability of a score:

- A rating scale is used with too few points to represent the amount of student performance variability.
- An anchored scale is used in which the anchor points are not well-defined, so that raters are not always sure of what rating to assign.
- A checklist is constructed with too few items to cover the construct of interest.
- A single global rating scale is used where an extensive checklist would be more appropriate (eg, raters are asked to provide a holistic judgment about whether they think an interview is acceptable vs. observing a set of 10 behaviors that are all necessary for an interview to be considered acceptable).

Reliability can be thought of as the precision of a scale score. A reliability estimate for a scale score indicates how much of that score is a representation of the true ability of the student and how much is error. The more reliable the score, the more confidence one can have that the score represents a student's true ability.

Validity

Validity evidence for a scale score is the evidence that shows a scale is actually measuring what it is assumed to measure and supports the use of the score in making inferences, judgments, and decisions about students. Validity evidence should be collected along with reliability evidence to ensure that the ability is not only being measured reliably but is also the targeted ability of interest.

Many of the rater biases and scenarios that negatively impact reliability also have the potential to negatively impact the validity of the scale scores in making inferences. For example:

- If raters who are too lenient produce scores that are too high, students will appear to be more capable than they actually are.
- If a checklist is too long, rater fatigue could cause raters to make errors in observation or rating.
- If an anchored scale contains too many points, this may result in distinctions being made among student performances that are not actually meaningful from an educational or clinical standpoint.
- If raters allow irrelevant aspects of the student to impact the rating scale, this will result in scores that are measuring these aspects as well as the construct of interest.

Take-Home Messages

- Common scale types used in OSCEs are rating scales and checklists.
- Rating scales can be useful for gathering judgments of student performance that are global impressions or anchored by specific student behaviors.
- Checklists can be useful for gathering multiple categorical observations of student behaviors.
- The type of scale should be carefully linked to the desired measure.
- Scale development and rater training have the potential to impact the reliability and validity of the resulting scale.

Resources

De Champlain, A. F., Margolis, M. J., King, A., & Klass, D. J. (1997). Standardized patients' accuracy in recording examinees' behaviors using checklists. *Academic Medicine*, 72(10), S85–S87.

Feldman, M., Lazzara, E. H., Vanderbilt, A. A., & DiazGranados, D. (2012, Fall). Rater training to support high-stakes simulation-based assessments**.** *Journal of Continuing Education in the Health Professions*, 32(4), 279–86.

Gorter, S., Rethans, J. J., Scherpbier, A., van der Heijde, D., van der Vleuten, C., & van der Linden, S. (2000). Developing case-specific checklist for standardized-patient-based assessments in internal medicine: A review of the literature. *Academic Medicine*, 75(11), 1130–7.

Gray, J. D. (1996, January). Global rating scales in residency education. *Academic Medicine*, 71(1), S55–S63.

Hauenstein, N. M. A. (1998). Training raters to increase the accuracy of appraisals and the usefulness of feedback. In J. W. Smither (Ed.), *Performance appraisal*. San Francisco, CA: Jossey Bass.

Hawkins, R. E., & Boulet, J. R. (2008). Direct observation: Standardized patients. In E. S. Holmboe, & R. S. Hawkins (Eds.), *Practical guide to the evaluation of clinical competence*. Philadelphia, PA: Mosby Elsevier.

Hettinga, A. M., Denessen, E., & Postma, C. T. (2010, September). Checking the checklist: a content analysis of expert- and evidence-based case-specific checklist items. *Medical Education*, 44(9), 874–83.

Hodges, B., & McIlroy, J. H. (2003, November). Analytic global OSCE ratings are sensitive to level of training. *Medical Education*, 37(11), 1012–6.

Hodges, B., Regehr, G., McNaughton, N., Tiberius, R., & Hanson, M. (1999, October). OSCE checklists do not capture increasing levels of expertise. *Academic Medicine*, 74(10), 1129–34.

Iramaneerat, C., & Yudkowsky, R. (2007, September). Rater errors in a clinical skills assessment of medical students. *Evaluation in the Health Professions*, 30(3), 266–83.

McIlroy, J. H., Hodges, B., McNaughton, N., & Regehr, G. (2002, July). The effect of candidates' perceptions of the evaluation method on reliability of checklist and global rating scores in an objective structured clinical examination. *Academic Medicine*, 77(7), 725–8.

Pangaro, L., & Holmboe, E. S. (2008). Evaluation forms and global rating scales. In E. S. Holmboe, & R. S. Hawkins (Eds.), *Practical guide to the evaluation of clinical competence*. Philadelphia, PA: Mosby Elsevier.

Regehr, G., MacRae, H., Reznick, R. K., & Szalay, D. (1998, September). Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Academic Medicine*. 73(9), 993–7.

Vu, N. V., Marcy, M. L., Colliver, J. A., Verhulst, S. J., Travis, T. A., & Barrows, H. S. (1991). Standardized patients' accuracy in recording clinical performance checklist items. *Medical Education*, 21, 482–9.

Yudkowsky, R. (2009). Performance tests. In S. M. Downing, & R. Yudkowsky (Eds.), *Assessment in health professions education*. New York, NY: Routledge.

We acknowledge the contribution of these content experts:

Kimberly Swygert, PhD, National Board of Medical Examiners

Amanda Soto, EdD, National Board of Medical Examiners