# CHAPTER 7.3

# Assessment in Healthcare Simulation

*Wendy Anson, PhD, CHSE*

## ABOUT THE AUTHOR

**WENDY ANSON** is a Research Analyst for educational psychology and technology at MedStar Health. She is also a member of the SSH Accreditation Committee and has presented on assessment, tool rating, and competencies in healthcare simulation. Dr. Anson has served as a consultant for the National Center for Research on Evaluation, Standards & Student Testing, as well as other university organizations. She received the Annenberg Multimedia Scholar fellowship for a communication tool and was awarded a US patent for an online assessment tool that she developed.

**Acknowledgments:** The author would like to express tremendous gratitude to Rachel Yudkowsky, MD, MHPE, Associate Professor in the Department of Medical Education at the University of Illinois at Chicago College of Medicine and Director of the Dr. Allen L. and Mary L. Graham Clinical Performance Center, who graciously gave her time and expertise to look over multiple drafts and provide vital, timely feedback and input. The chapter could not have been done without her. The author would also like to thank USC professor Harold F. O'Neil, Jr., PhD for his teachings, Dr. Elizabeth Sinz who contributed helpful suggestions to the manuscript, and editor Dr. Janice C. Palaganas who consistently offered thoughtful, pertinent, judicious, and invaluable feedback throughout on content and organization.

## ABSTRACT

Peer-reviewed studies using healthcare simulators have shown reliable, reproducible data for assessing students within varied specialties and learning levels. Simulation as a modality has become a focus in the assessment of procedural, clinical decision making, behavioral, and communication skills of health professionals and teams. Simulation is being used more and more to demonstrate that learning has occurred in an environment wherein emphasis has moved to observed evidence of "competencies." This chapter outlines technical, trainee, trainer, and tool components of simulation-based assessment.

## CASE EXAMPLE

Sara, a Simulation Educator, works in a multilevel, multispecialty simulation center used by a variety of medical schools and hospitals for training of their medical students, residents, fellows, and RNs. A program director in a surgical specialty new to her center asked her for a central venous catheter (CVC) placement simulation assessment tool for a simulation exercise. The program director wants reportable outcomes. Understanding the complexity of assessment, Sara seeks out research help and has found that there are no psychometricians at her institution.

## INTRODUCTION

Outcomes-based education has become a focus in healthcare profession education, and there has been an increasing need to provide evidence that learning has occurred (Scalese & Issenberg, 2008). Healthcare simulation (HCS), integrated into the larger healthcare education curriculum, is currently being used to provide this evidence, specifically with a focus on observed evidence of competencies. According to McGaghie et al. (1978), "competence includes a broad range of knowledge, attitudes and observable patterns of behavior which together account for the ability to deliver a specified professional service" (p. 19).

Within the context of observing behaviors, there are two general types of assessment: "formative" (assessing learning during the teaching process or path, e.g., quizzes, question and answer, or in class discussion) and "summative" (assessment or testing of learning at the end of course/program, e.g., high-stakes testing or pass/fail grading; Scalese &

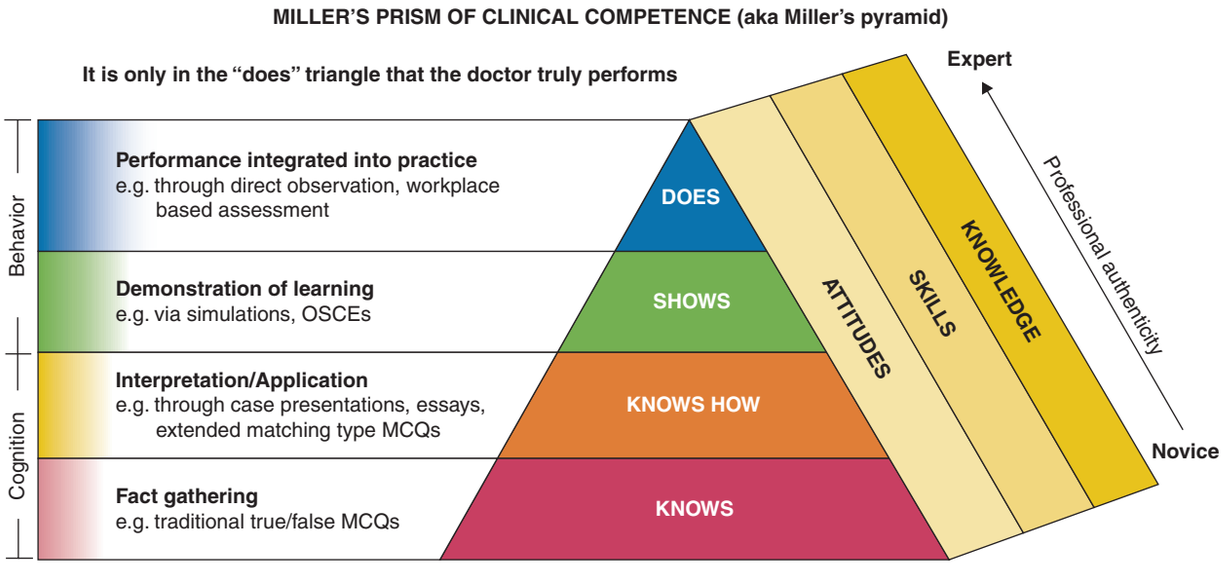


**FIGURE 7.3.1** Miller's Pyramid of Competence. (Based on work by Miller, G. E. [1990]. The assessment of clinical skills/competence/performance. *Academic Medicine*, *65*(9 Suppl), S63–S67. Adapted by Drs. R. Mehay & R. Burns, UK, January 2009.)

Issenberg, 2008). Learners are observed, and then given feedback on their performances. There are two varieties of simulation feedback—formative and summative. Performance feedback during the simulation itself and discussions during postsimulation debriefing often serve as formative assessment. Formative assessment in a medical simulation setting is frequently understood as the four-step model presented by Rudolph et al. (2008): (1) note salient performance gaps related to predetermined objectives, (2) provide feedback describing the gap, (3) investigate the basis for the gap by exploring the frames and emotions contributing to the current performance level, and (4) help close the performance gap through discussion or targeted instruction about principles and skills relevant to performance. Alternatively, the purpose of "summative assessment" is to collect, analyze, and summarize data that is then provided to decision makers in the organization or institution so that they can decide on the competence of the learners, for example pass or fail (Smith & Ragan, 1999). While formative assessment serves to inform the learner and the educator how to achieve learning for that learner or learner group, summative assessment serves to inform the educator whether or not the learner is competent to pass a course or level of competence. It is often referred to as "high-stakes" testing because the stakes of passing a summative assessment are often high, for example passing or failing a course, completing or failing a program, losing employment, or not being hired. This chapter focuses on the summative assessment of competencies using simulation and elucidates the use of simulation in providing valid and reliable outcomes on learner performance in a given domain or multiple domains.

Summative assessment often follows a set of determined criteria that, if met, demonstrates a level of competence. The use of simulation allows assessment of higher levels of competence. In Miller's Pyramid of Competence (Figure 7.3.1), the lowest "knows" stage level of the pyramid (e.g., "what is a central venous line?") can be assessed using simple knowledge tests, for example multiple-choice questions (MCQs). The "knows how" stage (e.g., "this is what needs to be done to insert a central venous catheter") can be assessed using MCQs, patient management problems, or essay questions. The "shows" or "does" competency level is problematic in the clinical environment—not only for ethical reasons of having learners demonstrate on real people for competency evaluation, but because assessment in that context would not be reliable due to multiple variables and challenges. For example, the number and types of real-world cases that occur in the clinic are unpredictable, and so it cannot be assumed that each student would be presented with the same number and types of cases within their set clinical rotation. For this reason, clinical rotations or time spent in a clinical environment are not reliable measures of competence. HCS can be structured to assess an individual or team at all levels of competence, including the "shows" and "does" levels.

Outcomes related to education have focused on competencies or skills, and more recently on robust credentialing (e.g., milestones, accreditation, or certification standards). Credentialed healthcare providers are also increasingly being required to demonstrate that they meet acceptable standards for continued practice through recredentialing or regular assessment of competencies (e.g., Maintenance of Certification [MOC]). More and more certification and MOC programs are mandating simulation assessment as part of the process (e.g., MOCA, NRP, FLS). Peer-reviewed studies using healthcare simulators have shown the modality to provide reliable, reproducible data (Boulet & Murray, 2010). Simulation as an assessment modality can also provide reliability through consistent testing environment, a selection of tasks that align with students' expected level of expertise, an environment focused on evaluation of student performance

rather than patient needs, and technology to represent, archive, and often compute performance outcomes with increased accuracy and objectivity. At the same time, a variety of considerations influence simulation assessment potential to embody a rigorous, viable methodology for the testing of healthcare providers and students. This chapter will explore the following five areas of simulation-based assessment:

1. Learner aspects (e.g., pre- and postlicensure)
2. Assessor aspects (e.g., simulationists and raters)
3. Programmatic aspects (e.g., program needs and resources)
4. Simulation modalities (e.g., standardized patients [SPs] or mannequin-based simulation [MBS])
5. Assessment tools

The last two areas (numbers 4 and 5) are combined into one section. In this section, considerations and examples of appropriate assessment tool use are described within each simulation platform category. It is noted that the purpose of the chapter is to increase awareness of the complexity of simulation-based assessment; speaking to comprehensive local programmatic and psychometric considerations within a full-scale, multilayered healthcare assessment program is beyond the current scope.

## LEARNERS

Individuals being assessed can be delineated as either prelicensure (e.g., undergraduate or graduate student at university) or postlicensure (e.g., licensed practicing provider). Most prelicensure assessment activities coincide with programmatic curriculum necessary to pass to be granted a degree by the program, while most postlicensure assessment activities correspond to hospital or unit practice competencies, policies, or procedures. The stakes are typically high in summative assessment despite the learner level—for example, the prelicensure student may not pass a course or graduate the program if (s)he fails an assessment and the postlicensure provider may lose his/her job or a portion of the job if (s)he fails an assessment. Because of the high-stakes nature of summative assessment, individuals involved with assessment are obligated to develop a fair simulation with every possible opportunity for fair assessment (Joint Committee on Fair Testing Practices in Education, 2004). This requires involvement with educators, researchers, and assessors who have a deep understanding of the learner group being assessed.



### EXPERT'S CORNER

### A CLEAR MESSAGE EMERGES FROM THE NLN HIGH-STAKES ASSESSMENT PROJECT

Mary Anne Rizzolo, EdD, RN, FAAN, ANEF
*Board of Directors, Society for Simulation in Healthcare*

Fair evaluation of student performance in the clinical area is a very weak link in the assessment of nursing students. Creating a fair testing environment with patients whose conditions can change at any moment in an often chaotic healthcare setting is a near-impossible task. Furthermore, many faculty are underprepared for their evaluator role. Not all faculty have been educated in assessment methods, and unintentional biases emerge when switching from the role of teacher to evaluator (Stroud et al., 2011). These are just some of the factors that led me to propose a project to explore the feasibility of using high-technology simulation to assess prelicensure students in schools of nursing. The project was funded by Laerdal Medical in 2010. It was designed to be a feasibility study and the basic questions were these: Can high-technology simulations provide a fair, valid, and reliable method for summative assessment? How hard is it to do? What are the biggest challenges? The project is ongoing with expected completion in the fall of 2014. An overview of the project and some findings to date provide the foundation and context for the concluding message.

#### OVERVIEW OF PROJECT

The project began with a Think Tank of esteemed individuals* who shared their expertise and wisdom and recommended scenarios to assess students at the end of their program. Dr. Pamela Jeffries then assembled a team of expert simulation authors to design the simulations while the evaluation team, comprising Drs. Marilyn Oermann and Suzan Kardong-Edgren, examined evaluation tools and planned the training of raters. The scenarios were piloted and refined, and then schools of nursing across the country ran the scenarios and sent video recordings of student performances. Raters, chosen on the basis of their expertise in both simulation and evaluation, scored the videos at two different points in time, and inter- and intrarater reliability statistics were generated. In the final phase, now in progress, the authors of the scenarios are scoring the video recordings, and then discussing the rationales for their scores to come to consensus on scoring criteria. The final scoring criteria will be given to the raters who will once again score the videos twice, and inter- and intrarater reliability scores will be calculated.

#### DESIGN AND IMPLEMENTATION ISSUES

We found that substantial time was required to design, develop, and pilot the scenarios, each with three parallel forms, but the task is achievable. The differences in designing scenarios for teaching versus assessment, and other lessons learned from this aspect of the process have been reported by the authors who created the simulations (Willhaus et al., 2014). Standardizing the implementation of the scenarios by different individuals at multiple sites was also a challenge, but can be handled by standardizing the testing environment and implementing rigorous training sessions for facilitators. Other suggestions on design and implementation issues are included in my chapter in *Clinical Simulations in Nursing Education: Advanced Concepts, Trends, and Opportunities* (Rizzolo, 2014).

#### TOOLS

Testing existing tools or developing new ones to use for scoring a simulation is a much bigger challenge and will require

*(continued)*

## EXPERT'S CORNER

### A CLEAR MESSAGE EMERGES FROM THE NLN HIGH-STAKES ASSESSMENT PROJECT *(continued)*

considerable research. Different tools or a combination of tools are required to measure different competencies of students at various points in the curriculum. Tools intended to evaluate higher-level learning, such as clinical judgment and reasoning, are extremely difficult to norm. Our study worked with one tool and found that the biggest challenge was gaining consensus on the criteria to use for scoring. For example, if the tool item is "collects appropriate assessment data" and there are four pieces of data that should be obtained, how do you score a student who only completes three assessments when the tool's scoring mechanism only permits a met or not met option? Resolving issues like this would be the challenge for any tool selected. A manuscript on the norming process used in this study is in development.

**THE CLEAR MESSAGE**

Watching simulation authors and raters all with a passion for simulation and fair evaluation practices, struggle with the norming process has provided me with new respect for the difficulty of this task. More importantly, it has opened my eyes to the varied expectations of faculty regarding clinical performance of students. For me, the clear message that has emerged from this study is that there is an urgent need for faculty to engage in deliberative conversations to clarify behaviors/expectations of students at the end of each course and at the end of the program. An exercise in creating a simulation exam can be a great stimulus to push for these conversations to occur. If it does, we will make giant strides, not only in improving our assessment practices, but also in resolving the huge discrepancy between faculty and employers regarding essential competencies of new graduates (Berkow et al., 2008).

Our current methods of evaluating the clinical practice competencies of students are woefully inadequate. While many challenges remain for using simulation as a method of summative assessment, I applaud the path taken by brave faculty (Wolf et al., 2011) and encourage others to follow their lead and help us fill in the potholes in the long, difficult road toward more fair, reliable, and valid assessment practices.

*S. Barry Issenberg, MD, Pamela R. Jeffries, PhD, RN, FAAN, ANEF, Kathie Lasater, EdD, RN, ANEF, Carrie B. Lenburg, EdD, RN, FAAN, ANEF, M. Bridget Nettleton, PhD, RN, Marilyn H. Oermann, PhD, RN, FAAN, ANEF, Mary Anne Rizzolo, EdD, RN, FAAN, ANEF, Theresa M. (Terry) Valiga, EdD, RN, FAAN, ANEF, Linda Wilson, PhD, RN, CPAN, CAPA, BC, CNE.

**REFERENCES**

Berkow, S., Virkstis, K., Stewart, J., & Conway, L. (2008). Assessing new graduate nurse performance. *Journal of Nursing Administration, 38*(11), 468–474.

Rizzolo, M. A. (2014). Developing and using simulation for high stakes assessment. In P. R. Jeffries (Ed.), *Clinical simulations in nursing education: Advanced concepts, trends, and opportunities* (Chap. 9). Washington, DC: National League for Nursing.

Stroud, L., Herold, J., Tomlinson, G., & Cavalcanti, R. B. (2011). Who you know or what you know? Effect of examiner familiarity with residents on OSCE scores. *Academic Medicine*, 86(10 Suppl.), S8–S11.

Willhaus, J., Burleson, G., Palaganas, J., & Jeffries, P. (2014). Authoring simulations for high-stakes student evaluation. *Clinical Simulation in Nursing*, 10(4), e177–e182.

Wolf, L., Dion, K., Lamoureaux, E., Kenny, C., Curnin, M., Hogan, M. A., . . . Cunningham, H. (2011). Using simulated clinical scenarios to evaluate student performance. *Nurse Educator*, 36(3), 128–134.

Acceptable passing standards for summative evaluation differ according to the learner level and may be constructive or absolute. Prelicensure student assessment is often constructive. For example, a finishing resident is judged to be competent or not via "static or pass/fail dichotomous ratings of competence," which are modified for novice residents early in training and at each level toward full competency in that skill (Holmboe & Hawkins, 2008). Program directors become responsible for framing accreditation competencies into the program and throughout the levels, often following a mastery learning model. In mastery learning, testing is used to gauge unit completion at a preset minimum passing standard (Kulik et al., 1990). Advancement to the next educational unit depends on a given measured achievement at or above the mastery standard (or there is continued practice until the mastery standard is reached), to ensure achievement of all educational objectives with the least amount of outcome variation (McGaghie et al., 2010). McGaghie et al. (2010) explain that mastery learning has seven complementary features:

1. Baseline (i.e., diagnostic) testing.
2. Clear learning objectives, sequenced as units ordered by increasing difficulty.
3. Engagement in educational activities (e.g., skills practice, data interpretation, reading) that are focused on reaching the objectives.
4. Establishment of a minimum passing standard (e.g., test score, checklist score) for each educational unit.
5. Formative testing to gauge unit completion at a preset minimum passing mastery standard.
6. Advancement to the next educational unit given measured achievement at or above the mastery standard.
7. Continued practice or study on an educational unit until the mastery standard is reached.

Because the goal of many postlicensure assessment programs is to bring all clinicians to an acceptable level of performance and not to rank-order them, only absolute performance standards are appropriate. Therefore, it is necessary to set reference standards for a defined performance measure (Lammers et al., 2008).

With respect to learner level, the process of standard setting should ensure that cut score setting (determination of competency score) is not arbitrary, but is reasonable, defensible, and fair. Absolute standards reflect a certain level of mastery. Most assessments set out to confirm that a domain of knowledge or skill has been mastered. A passing score should be determined through a systematic, reproducible, and unbiased process by a group of content experts for that learner group and level (Yudkowsky et al., 2009).

Summative assessment simulations often require more resources (e.g., faculty time, simulation staff, equipment) than written examination or educational simulations. When choosing to use HCS, educators must first determine whether or not HCS is the most appropriate assessment tool for the competency level (e.g., Miller's Pyramid) being assessed. Simulations should be developed and criteria chosen appropriate to the level of the learner. Dry runs (practice run-throughs) of the simulation with learners from the target group are necessary to determine appropriateness and fairness of the simulation-based assessment.

## CONSIDER THIS

### BIAS FOUND IN HEALTHCARE SIMULATION

Janice C. Palaganas, PhD, RN, NP
*Author, Researcher, Assessor, and Assessor Trainer for the National League for Nursing High-Stakes Assessment Multi-Site Research Study*

When developing or analyzing assessment activities, biases described in research, assessment, and education literature are often considered, and measures are taken to screen for such biases. There are different sources of bias that are described in research literature (e.g., threats to internal, external, construct, and statistical conclusion validity). A goal in constructing an assessment test is to create a test in which inferences made by assessors of a learner's competence or skill level is accurate (or valid) and the assessment platform is seemingly as real as possible to the event or skill being assessed. This accuracy can be undermined by bias. Some common research, assessment, and education biases and a description of each are listed briefly in Table 1. Like any assessment activity, these biases occur often in simulation assessment. There is extensive research behind each of these biases that are beyond the focus of this textbox and chapter. In addition to research and education bias, there is also simulation bias that is often overlooked in simulation-based assessment. This chapter will focus briefly here on these types of simulation biases commonly observed in simulation-based assessment.

#### FAMILIARITY EFFECTS

Learners previously exposed to simulations have a familiarity with simulation that can allow them to anticipate or miss many assessment items, causing a familiarity effect. Learners who are familiar with simulation may be more comfortable in the assessment and may perform better than those learners who are not. Paradoxically, learners familiar with simulation may also demonstrate simulation habits in the assessment that may harm their scores. Some common familiarity effects include the following:

- **Modality Familiarity Effect**. Learners influenced by this bias will perform physical assessments on the modality within its limitations. For example, if a mannequin was being used, the learner may know where to feel a pulse and easily distinguish mannequin breath sounds through the mechanical overlay in comparison with another learner not influenced by this bias. Similarly, if a standardized patient was used, a learner familiar with this modality may more readily physically assess the patient, or if a virtual simulator was used, a learner familiar with this modality may be more facile with the haptic equipment. Another presentation of the

**TABLE 1**

**Common Biases Seen in Simulation-Based Assessment**

| Bias in Literature | How It May Appear in Simulation | Additional Resources |
|---|---|---|
| Design bias | When the simulation was not structured or screened to control for internal (does not allow for observation of the item) and external (the item being assessed does not fit the simulation or learner group) validity. | Cook and Campbell (1979) |
| Assessor's bias | Untrained assessor rates lower or higher, may know some learners and grade differently, may not have the adequate experience or knowledge to assess the skill. This is also known as "experimentor bias" or "investigator bias." Also presents in the trained assessor as a "rater drift," where the assessor over time drifts toward their expectation despite adjustments made during rater training. | Fernandez-Ballesteros (2003); see "Assessors" section |
| Procedural bias | Learners feel psychologically unsafe in being assessed or in the simulation environment. | Cook and Campbell (1979) |
| Halo effect | When the assessor is positively influenced by his/her impression of a learner (physical appearance, reputation, personality). Reverse-halo effect is when the assessor is negatively influenced by learner factors. | Nisbett and Wilson (1977) |
| Inference distortion | When one group of learners (from different profession, learned from a different professor, different school) learned something different, used a different learning method, or did not learn the item being assessed. | Popham (2012) |
| Demographic bias | Gender, racial, socioeconomic, profession bias. | Popham (2012) |
| **Simulation Bias** | **How It May Appear in Simulation** | **Additional Resources** |
| Familiarity effect | Includes modality familiarity effect, simulated resource familiarity effect, trained simulation habit effect, informed effect, and hypothesis-guessed simulation | See Familiarity Effects |
| New-to-simulation effect | Includes pauses, delays, laughing, and other awkward behavior of a learner new to simulation during a simulation. | See New-to-Simulation Effects |
| Simulation external validity | When two or more simulations assessing a skill or set of skills differs. This may appear through the facilitation skills of the ESP or scenario director. | See Simulation External Validity |

*(continued)*

**CONSIDER THIS**

## BIAS FOUND IN HEALTHCARE SIMULATION *(continued)*

modality familiarity effect is a willing dismissal of any technological glitches, including the simulation monitor.

- **Simulated Resource Familiarity Effect**. The learner familiar with simulated resources may access embedded simulated providers more readily, interviewing them extensively throughout the simulation, as well as immediately accessing other resources such as code team, other team members, charge nurse, or attending physician.
- **Trained Simulation Habit Effect**. Learners develop simulation habits that stem from undergoing many simulations at a previous simulation program. These habits may positively or negatively bias assessment activities. Trained simulation habits are often difficult to assess because some learners may have learned the habit in clinical practice, whereas for others they may be natural habits. Some habits often noted as positive during assessment include washing hands immediately before and after patient contact, use of gloves, and communicating out loud so that the scenario facilitators are able to hear the learner's thoughts and needs. Other trained simulation habits include making inquiries to the scenario facilitators behind the camera or window (e.g., attempting to engage with a facilitator via microphones in the room who might answer their questions via speakers, also known as "voice of God").
- **Informed Effect**. As an assessment simulation repeats with learners over time, there is a possibility that a previous learner may have informed a current learner on the cases being simulated or any other aspect of the simulation-based assessment that may allow the current learner to anticipate events or skills to perform. Learner ethics and code of conduct may be questioned in this case. Frequently, because the simulation interacts with learner actions and the informed learner may not have the same actions as the informant learner, the scenario may progress differently than what was previously informed.
- **Hypothesis-Guessed Simulation**. Some learners are familiar with the flexibility of simulation, often psychologically engrossed with a previous simulation case, and may steer the simulation in the direction of a diagnosis or finding that they self-select during a simulation. In research, the bias "hypothesis guessing" occurs when a subject's performance is influenced by their knowledge of findings that they expect to gather (Cook & Campbell, 1979). For example, a learner may state out loud that there is a foreign object in the airway when there is not and progress through the simulation as though the case was an airway obstruction. This learner may have done this because of something he saw in the airway, likely attributed to the simulator, or because he was anticipating this scenario. Regardless of the reason, simulation is a flexible method. Learners are familiar with this flexibility and do not feel the limitations of the case, steering the simulation down a self-determined path. This bias may be redirected through the facilitation of an embedded simulated provider.

### NEW-TO-SIMULATION EFFECTS

Learners new to simulation are often distracted by the unfamiliar environment, equipment, process, or modality (e.g., mannequin, standardized patient, virtual simulator, etc.). This may create pauses or delays in performance. The unfamiliar simulation may impede on a learner's fiction contract (see Experts' Corner: Helping Learners "Buy in" to Simulation), allowing awkward behaviors such as laughing or appearing frozen. The learner not familiar with a modality may be reluctant to physically touch the modality or do what (s)he would do in a real-life setting. This may create false-negative results when being assessed in a simulated environment.

### THE PROBLEM WITH SIMULATION EXTERNAL VALIDITY

External validity in the setting of simulation is the consistency of a simulation that could be generalized to other simulation programs. If a learner were to be assessed on the same skill or set of skills at different sites or different days with different simulation facilitators, the learner should, in fair assessment, achieve the same assessment scores. Simulation programs organically form as a result of needs and resources, especially human resources. The differing nature of simulation programs is its own threat to simulation external validity. Because simulation programs differ from site to site, including those within organizations, it is extremely difficult to standardize a simulation. Cues may be given at one site or in one simulation, where it is not given (or at the same degree) as another. There may also be **Facilitator Bias** where a simulation is facilitated more smoothly by experienced embedded simulated participants or by the scenario director who can anticipate events and has the experience to interact more immediately with the learners' actions. Equipment may also differ, with one or more potentially subtracting or adding to the realism of the case. Assessment scores may differ from one site to another site or one simulation with certain simulation staff to another simulation with other simulation staff. One way to help overcome these problems with external validity and replication is to provide example videos of how the simulation should look and be conducted.

### REFERENCES

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings.* Chicago, IL: Rand McNally.

Fernandez-Ballesteros, R. (Ed.). (2003). *Encyclopedia of psychological assessment.* Thousand Oaks, CA: SAGE.

Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology*, 35(4), 250–256.

Popham, W. J. (2012). *Assessment bias: How to banish it.* Boston, MA: Pearson.

# ASSESSORS

The selection of appropriately qualified and fair assessors (or people who perform the assessment through observation and an assessment tool, also known as "raters") and the training of assessors is a critical component of fair and consistent assessment. A process should be developed to match the characteristics of the assessment with the qualifications of assessors by virtue of their experience and education. Once qualified assessors are selected, a systematic training program must be implemented (Feldman et al. 2012; see Table 7.3.2, below). Dry runs (practice run-through of the simulation prior to the exercise) are key in the process of training assessors. Reference copies of the relevant assessment tool, the

scenario progression, policies regarding the activity, the assessor evaluation rating tool, and fellow assessor files should be made available for all assessors at the appropriate times.

## How to Train Assessors

Holmboe and Hawkins describe two types of accuracy measures in rating: the first kind are decisions whether the behavior did or did not occur (checklists). The other type involves "judgmental measures," where the assessor must apply a judgment involving accuracy when providing a rating: accuracy in whether a learner has attained a level of performance (criterion accuracy), accuracy in distinguishing among learners (normative accuracy), and accuracy in discriminating between a performance or competence dimension (stereotype accuracy; Holmboe & Hawkins, 2008). Observational assessment of learners in clinical settings, and, by extension, to simulation settings, can fall prey to subjectivity, false impressions, ageism, racism, sexism, and misinterpretation (McGaghie et al., 2009). In Performance Dimension Training (PDT), raters receive expected performance standards for each level of performance. Assessors receive "Frame of Reference" training where they define "satisfactory" performance (the anchor point) and practice evaluating. Assessors need to reach agreement on common nomenclature for the desired expectations of interest, and agree on the relative importance of the different components of the behavior being assessed (Holmboe & Hawkins, 2008).

Assessor errors include halo, leniency, severity, central tendency, and idiosyncratic rating and are described in Table 7.3.1 (Downing, 2003). In addition, Feldman et al. (2012) list and define rater training components in Table 7.3.2.

Once the assessors are trained, Feldman et al. (2012) explain how to assess the reliability of raters' scores. They describe two primary methods for assessing the reliability of the Simulation-Based Training ratings as a whole:

1. Agreement is the most simple and most common. Agreement tells you whether or not your raters picked exactly the same score for a particular behavior.
2. Correlations tell you whether or not your raters followed similar patterns, but not whether they agreed exactly.

Assessor training should occur with use of the rating method with which the assessor will rate learners—whether it is a live simulation or video. The mean of the passing score for each item is determined, with the passing score for the case being the simple average of passing scores for all items. The researcher or educator directing the assessment program should seek to understand any differences in ratings of the same case and find ways to create consistency between assessors; this may include elimination of a rater who consistently scores too high or too low relative to the other assessors. Assessor performance should be routinely (annually at minimum) evaluated to ensure inter-rater reliability (consistency across assessors) and to evaluate individual assessor competence.

# PROGRAMMATIC ASPECTS

Assessment requires as optimally controlled an environment as possible to ensure consistent cases for learners and assessors, and requires the facilities, simulation staffing, and technology to support this level of standardization. The facilities and technology should also be appropriate to the individual or teams being assessed.

Dry runs are recommended to highlight potential inconsistencies between simulations. The assessment should begin with a standardized orientation of the learner to the environment and assessment tools, followed by implementation of the standardized simulation.

Assessment does not end after the activity. Adequate technical and research support is needed for the appropriate analysis of data. Just as the assessors should be appropriately qualified, so should the human factors, psychometric, and statistical support. Most institutions (educational and hospital) have resources and departments accessible for support. Because of the high-stakes implications in summative assessment, a process and secure plan must be developed, implemented, and maintained to ensure confidentiality, data, and test security.



**TABLE 7.3.1**

**Assessor Bias Errors**

| Rater Error | Description | Consequence |
|---|---|---|
| Central tendency | Avoiding extreme positive or negative ratings. | Reduces ability to discriminate performance. |
| Halo error | All ratings based on one very positive or negative observation. | Positively or negatively skewed ratings. |
| Primacy/recency effect | All ratings based on observations made early or late in the scenario. | Positively or negatively skewed ratings. |
| Contrast effect | Ratings are made relative to performance of previous group. | Positively skewed ratings when prior group performed very poorly, negatively skewed ratings when prior group performed very well. |
| Potential effect | Ratings based on perceptions of future potential. | Usually positively skewed ratings. |
| Similar-to-me effect | Ratings based on degree of similarity to the rater. | Tendency to rate people who resemble themselves higher. |
| Stereotype effect | Ratings based on group inclusion rather than individual differences. | Positively or negatively biased ratings for some groups. |

**TABLE 7.3.2**

## Feldman et al.'s How to Train a Rater Checklist

Feldman et al.'s "How-to" Rater Training Curriculum Components

- *How* to Assess Reliability
- **Over Time**
  Same observer scores the same video-taped scenario at two different points in time
- **Across Multiple Raters**
  Two different observers score the same scenario
  In real time
  Video
- **Percent Agreement**
  Agreements/(agreements + disagreements)
  *Number of times observers agree*
  Number of opportunities to agree
  *Tasks:*

1. Pair up with a buddy.
2. Use the measurement tool that both you and your buddy used during previous exercise.
3. Compute inter-rater agreement using the ratings you did in the previous exercise (p. 9).
4. Debrief to entire group.
   a. What teamwork dimensions did you use for this exercise?
   b. What was your inter-rater agreement?
   c. What were the challenges?

- Courtesy of Rosen, Types of Rater Training
- **Rater Error Training (RET)**
  Focuses on *avoiding* rater errors and biases.
- **Performance Dimension Training (PDT)**
  Focuses on *defining* skill dimension.
- **Frame of Reference Training (FOR)**
  Focuses on *discriminating* between skill levels.
- Rater Error Training
- Reduces occurrence of rating errors.
- Refine and Realign
- Raters discuss where and why they disagreed.
- Goal is to come to consensus on how to rate behavior using observed examples.
- Rating "rules" can be generated to help in addition to the rating guide.
- Rating skills are refined and realigned to common criterion.
- Normative Reference Strategy
- Rating standards and rules are developed through a consensus process between a group of raters.
- Reliability and validity are assessed by comparing ratings across raters and dimensions.
- Accuracy, sensitivity, and reliability are relative to the group of raters.
- Raters refine and realign until adequate level of reliability is reached.
- Criterion Reference Strategy
- Rating standards and rules are developed using an expert set of raters and set of examples.
- Reliability and validity are assessed by comparing raters to expert ratings using standard set of example scenarios.
- Accuracy, sensitivity, and reliability are relative to an expert set of "gold standard" ratings.
- Raters refine and realign until adequate reliability is found with expert ratings.
- Comparing Strategies
- Reliability and Rater Turnover
- Quality Monitoring and Improvement
- QMI should be a continuous process.
- Has implications for effectiveness of training, planning, and performance improvement.
- Key PointsRaters should be knowledgeable about the task being rated, but experts are more resistant to training to external criterion.
- Raters with similar backgrounds rate more similarly.
- Expert and novice raters have been shown to rate with adequate levels of reliability with effective rater training.
- Consider external factors such as availability, buy-in, and potential biases.

Feldman, M., Lazzara, E., Vanderbilt, A., & DiazGranados, D. (2012). Rater training to support high-stakes simulation-based assessments. *Journal of Continuing Education in the Health Professions, 32*(4), 279–286.

## ASSESSMENT TOOLS

There are generally two types of assessment tools with which learners are observed and graded: checklists and global rating scales. These are used to assess procedural skills, critical decision making, team skills, and communication. Assessment tools are typically scored in the following way:

- **Checklists:** items are scored dichotomously (e.g., done or not done).

- **Weighted Checklists:** items judged to be more critical to a successful procedure are assigned higher point values.
- **Rating Scales:** each item is assigned points on a scale.

Assessment tool development is highly complex. For example, when developing an assessment test or tool, the assessment group or tool developers need to determine what the passing score is, also known as the "test cut score" (Yudkowsky, 2009). This cut score and standard setting is determined by content experts. In a contrasting group procedure for standard setting, five or more assessors with content expertise are selected to divide the scores into categories such as pass/fail. The assessors must have a full understanding and be in agreement about the scope and dimension of all behaviors expected in each performance category for that learner level. The cut point demarcates the boundaries between those performance categories on the exam score distributions (Yudkowsky et al., 2009). In the "Angoff Standard-Setting Procedure," judgments are carried out at the item level. On a scale of 0 to 1.0, each assessor answers independently, "what is the probability that one borderline score will accomplish this item correctly?" (Yudkowsky et al., 2009, p. 133). Divergent ratings might then be discussed and assessors have the opportunity to modify their judgments if they wish.

A process should be developed for the selection of appropriate assessment tools, specifically to ensure that the tool selected is appropriate, reliable, valid, and fits well with the testing conditions and requirements. The creation of a tool that is valid and reliable often can take years of expert psychometric and statistical support. For this reason, educators and researchers with time constraints for their simulation education delivery are encouraged to reference existing valid and reliable tools published in the literature. These tools are typically developed and vetted for a specific research purpose at a specific institution, so the challenge faced by adopting or adapting published tools might be that while the tool is a good fit for its developed purpose, it may not be an adequate fit for another assessment focus at another institution. Challenging factors include the following:

- The tool was tested on a specific group of learners in a specific circumstance that may not be similar or apply to another assessor's learners or circumstances.
- The tool was designed for healthcare providers who are coming into the simulation with a different level or amount of knowledge and expertise than those students of another assessor.
- The tool was designed to "test out" credentialed healthcare providers, not teach healthcare students.

Many programs are not familiar with the best practices requirement for validity and reliability and choose homegrown tools without any history of testing for validity (the tool is accurately measuring the intended variable) and reliability (the items are clear enough to establish a general consistency with each item across assessors). Some programs that are familiar with the need for valid and reliable tools often adapt published tools and change items on the adopted published tool to better fit their assessment need. However, when tools are tested for validity and reliability, the test comprises a holistic analysis, meaning that any change made to an existing tool threatens and discredits the validity and reliability of that tool.

The process of selecting a site-specific institutional simulation assessment can be guided with the Assessment Tool Selection Template presented in this section (Table 7.3.3). This Assessment Tool Selection Template addresses the necessary considerations mentioned throughout this chapter, linking the specific objectives, learners, learning goals, and learning needs to the particular institutional performance goal. As an example, the opening case study will be revisited and addressed using this template.

Steps to using this framework as a guide:

1. First, together with your faculty, fill out who, what, which, and why. This determines the learner levels and specialties, learning goals, and the institutional and/or curricular need.
2. Now, fill in your ideal performances and which measures would indicate passing of that measure.
3. From this, you can decide the appropriate simulation environment and modality.
4. Finally, perform a peer-reviewed literature search, and select the appropriate assessment tool to use or plan to develop.

Once an assessment tool is chosen, assessors must be trained (see "Assessors" section above). This training establishes a level of inter-rater reliability (concordance in scores between assessors). During assessor training, assessors may adjust expectations of items to be in concordance with other assessors. Over time, this adjustment may "drift" toward their initial expectations. Because of this, repeated reevaluation of inter-rater reliability is needed.

## SIMULATION MODALITIES AND ASSESSMENT TOOLS

Assessment is traditionally done in domains of measurable skills from history taking, physical examination, procedural performance, clinical decision making, and patient management to such areas as teamwork, cultural competence, and professionalism. The ACGME Toolbox of Assessment Methods suggests that HCSs are most appropriate for evaluations of those outcomes that require trainees to "show how" they are competent to perform (Scalese & Issenberg, 2008).

One attractive benefit in using HCS for assessment is that it can provide a consistent platform for rater observation of learner performance. Instruments for simulation evaluation are increasingly reliable with some promising validity (Kardong-Edgren et al., 2009).

Learners are tested with these checklists or rating scales in a variety of simulation "scenarios" in one or another of

**TABLE 7.3.3**

**Assessment Tool Selection Template**

| Consideration | Example | Fill in as Appropriate |
|---|---|---|
| **WHO gets the simulation?** | | |
| Professions and levels | **Surgical PG Y-1 to PGY-2**<br>**Surgical Physician Assistant Students**<br>**Critical Care Nurse Practitioner Students** | |
| **WHAT is the learning goal?** | **CVC Placement** | |
| **WHICH CVC placement will you teach?**<br>…because? | **US-guided IJ CVC placement**<br>because: JCAHO recommendation | |
| **WHY (the need in your environment is)?**<br>**-Institutional**<br>**-Curricular**<br>**-Both** | **US-guided is safest for patient population:** previous complications, underlying vascular anomalies, limited access sites for attempts, difficult to identify surface landmarks, inexperienced operators<br>**Both** Pine Hospital reports mechanical errors; pneumothorax, artery punctures; learner levels have been assessed "not up to standard" during live IJ CVC placement | |
| **Ideal Performances** | **Perform US-guided IJ CVC insertion with no mechanical errors in under 15 minutes** | |
| **By which Passing Measures?** | a. Learners must make fewer than three needle passes<br>b. Learners must demonstrate successful location and cannulation of vein on first try—no artery attempts | |
| **Which Simulation Modality(s) are the most appropriate for these performance measures?** | **Partial Task Trainer (PTT)**<br>1. Blue Phantom anthropomorphic simulation US training model, Advanced Medical Technologies, LLC<br>2. central venous access head neck and upper torso simulation model, Advanced Medical Technologies, LLC | |
| **Choose Assessment tool type and tool** | **Checklist**<br>**Pine Hospital ICU CVC Simulation Placement Checklist (Author, Year)** | |

"Example column" based in part on Evans et al. (2010).

the simulation modalities. Testing of team, communication, cognitive, or psychomotor skills best takes place in the simulation modality appropriate to that skill domain. A key principle of Simulation-Based Education (SBE) is that educational goals must dictate decisions about the acquisition and use of simulation technology (McGaghie et al., 2010). In HCS-based assessment, assessments might be grouped into five major modalities: partial task trainers, SPs, hybrid, mannequin-based, and virtual reality (VR). These modalities are described below with example assessment tools appropriate or common to each modality.

## Partial Task Trainers

Static simulators that reproduce anatomic regions and/or clinical task events provide education and assessment in basic procedural skills such as intravenous line insertion, suturing, intubation, and lumbar puncture (Scalese & Issenberg, 2008). Partial task trainers mimic body parts or regions (e.g., the arms, pelvis, or torso). These trainers are scaled from infant to adult sizes. They range technically from simulated skin with "blood"-filled veins to moving organs with appropriate vasculature viewable on ultrasound machines. These trainers also differ by specialty (e.g., for surgical skills, multilayered pads are used with cutdown procedures, cyst removal, subcuticular suturing, and knot tying; for anesthesia, there are airway trainers; for obstetrics and urology, there are birthing trainers, and simulators of the pelvis/perineum). There are also computerized task trainers such as a cardiopulmonary patient simulator with blood pressure, arterial, venous, and precordial pulses as well as heart and lung sounds synchronized to simulate 30 different cardiac conditions (Scalese & Issenberg, 2008).

## Simulation Assessment Tools for Partial Task Trainers

There are two common assessment tools used for partial task trainers: the checklist and the rating scale.

1. **Technical dichotomous checklist, individual**
   Checklist items are statements or questions that reflect concrete, observable behaviors that can be scored dichotomously as "done" or "not done." An example of this is the "Objective Structured Assessment for Technical Skills" (OSATS) by Martin et al. (1997). Specifically, in Figure 7.3.2, the *Control of Hemorrhage Checklist below*, each OSATS technical dichotomous checklist comprises a list of items to be checked "done" or "not done".
2. **Global rating scale**
   The rating scale is used to demonstrate how well the individual performed a procedure or action in

RESIDENT STICKER

**STATION 4**

**CONTROL OF HAEMORRHAGE**

**INSTRUCTIONS TO CANDIDATES**

You have just identified a stab wound to the inferior vena cava. Control the haemorrhage and repair the vessel.

**Start Time:**

**CHECKLIST**

| | ITEM | Not Done/ Done Incorrectly | Done Correctly |
|---|---|---|---|
| | CONTROL OF HEMORRHAGE | | |
| 1. | Applies pressure to stop bleeding first | 0 | 1 |
| 2. | Asks assistant to suction field | 0 | 1 |
| 3. | Inspects injury by carefully releasing the IVC | 0 | 1 |
| 4. | Ensures all equipment needed for repair is at hand before starting | 0 | 1 |
| 5. | Control of bleeding point (use deBakey forceps /Satinsky clamp or prox/distal pressure) | 0 | 1 |
| | REPAIR | | |
| 6. | Select appropriate suture (4.0/5.0/6.0 polypropylene) | 0 | 1 |
| 7. | Select appropriate needle driver (vascular) | 0 | 1 |
| 8. | Select appropriate forceps (de Bakey) | 0 | 1 |
| 9. | Needle loaded 1/2–2/3 from tip 90% of time | 0 | 1 |

**FIGURE 7.3.2** Example of technical dichotomous checklist rating of individual, *OSATS Control of Hemorrhage Checklist.* (Reprinted with permission from Martin, J., Regehr, G., Reznick, R., Macrae, H., Murnaghan, J., Hutchison, C., & Brown, M. [1997]. Objective structured assessment of technical skill (OSATS) for surgical residents. *British Journal of Surgery*, *84*(2), 273–278.)



a static setting (e.g., how well the physical therapist provided information on cardiac rehabilitation or how well respiratory therapist handled ventilator). Rating scales rate the item along a spectrum of response options. The items on this type of tool often presents as numeral points, assessing aspects of a skill commonly measured on a 3-, 4-, 5-, or 7-point scale (also known as Likert scale after the psychologist Likert who studied the use of numeral point items to create a simple sum) (Figure 7.3.3).

3. **Technical Behaviorally Anchored Rating Scale (BARS), individual**
Because scales can lead to subjective judgment, behaviorally anchored items are often provided for the rating options to theoretically improve consensus among raters. This is called a "Behaviorally Anchored Rating Scale." For example, the OSATS "Technical Domain-Specific rating scale with behavioral anchors" (Martin et al., 1997) includes a number of items all assessing aspects of operative skill and the anchoring of points 1, 3, and 5 on the 5-point scale by behavioral descriptors (Figure 7.3.4). The OSATS technical rating scale is used to demonstrate how well the individual performed a procedure or action in a static setting, for example operatively controlled and repaired a hemorrhage in terms of "respect for tissue," "time and motion," "instrument handling." Rating scales rate the item along a spectrum of response options, usually no more than 7 (Yudowsky, Downing, & Tekian, 2009). Because scales can lead to subjective judgment, behaviorally anchored items are provided for the rating options and therefore improve consensus among raters. Kim et al. (2009) note that rating scales, in indicating

| Directions: Circle the number that most closely corresponds to your observations of the employee. Do this on at least two different occasions. Scale: 1=Poor 2=Below average 3=Average 4=Above average 5=Excellent | | | | | |
|---|---|---|---|---|---|
| Knowledge of work: | 1 | 2 | 3 | 4 | 5 |
| Clinical skills: | 1 | 2 | 3 | 4 | 5 |
| Communication skills: | 1 | 2 | 3 | 4 | 5 |
| Concern for patient safety: | 1 | 2 | 3 | 4 | 5 |
| Concern for team: | 1 | 2 | 3 | 4 | 5 |
| Overall quality of work: | 1 | 2 | 3 | 4 | 5 |
| Total: | | | | | |

**FIGURE 7.3.3** Example of nontechnical Global Rating Scale, rating of individual, *End-of-Year Global Rating Employee Assessment Scale.*

how well the action was performed and not simply that it was "done or not done," is best for reporting levels of expertise (see Figure 7.3.4). (Note: Use of this tool for partial task training assumes that there will be an assessment of fine-grained skills with tissue, e.g., surgical suturing with a beef tongue.)

## Standardized Patients

An SP is

> a person who has been carefully coached to simulate an actual patient so accurately that the simulation cannot be detected by a skilled clinician. In performing the simulation, the simulated patient presents the "gestalt" of the patient being simulated; not just the history, but the body language, the physical findings and the emotional and personality characteristics as well. (Barrows & Abrahamson, 1964; see chapter 3.3)

SPs are trained to provide a consistent account of their condition, answer a range of questions about themselves, and can portray people from a variety of cultures, ethnicities, and/or those with communication problems and/or specific mental or physical conditions. The SP is trained to provide standardized answers and behaviors so that a group of students can be reliably tested. SPs are to be distinguished from "Embedded Simulated Persons" (ESPs) in mannequin simulation. ESPs are embedded acting participants leveraged to facilitate the learners through objectives in the simulation (see chapter 3.3). SPs are also trained to evaluate the healthcare students' skills in history taking, physical exam, clinical reasoning, and ability to take part in a challenging conversation, as well as general communication skills on the basis of a checklist of items. The SP proceeds through the interaction with the student and then scores the student on the basis of their observations (McLaughlin et al., 2006). SP simulations are used exclusively (without task trainers or other simulation modalities) in many medical, graduate nursing, and physician assistant schools as Objective Structured Clinical Examinations (OSCE). OSCEs are also used in board exams such as the USMLE in the US and MCC in Canada.

## Simulation Assessment Tools for SPs

Below are examples of scales used in SP simulations that are designed to be rated by the SPs themselves, as well as scales that are designed to be rated by faculty (Figure 7.3.5, Tables 7.3.4 and 7.3.5).

## Hybrid Simulation (Mixed Modality)

"Hybrid" simulations use multiple modalities to achieve the objective of the assessment. For example, an SP can

**Please rate the candidate's performance on the following scale:**

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| **Respect for tissue** | Frequently used unnecessary force on tissue or caused damage by inappropriate use of instruments | | Careful handling of tissue but occasionally caused inadvertent damage. | | Consistently handled tissues appropriately with minimal damage. |
| **Time and motion** | Many unnecessary moves | | Efficient time/motion but some unnecessary moves. | | Economy of movement and maximum efficiency |
| **Instrument handling** | Repeatedly makes tentative or awkward moves with instruments. | | Competent use of instruments although occasionally appeared stiff or awkward. | | Fluid moves with instruments and no awkwardness. |
| **Knowledge of instruments** | Frequently asked for the wrong instrument or used on inappropriate instrument. | | Knew the names of most instruments and used appropriate instrument for the task. | | Obviously familiar with the instruments required and their names. |
| **Use of assistants** | Consistently placed assistants poorly or failed to use accidents. | | Good use of assistants most of the time. | | Strategically used assistant to the best advantage at all times. |
| **Flow of operation and forward planning** | Frequently stopped operating or needed to discuss next move. | | Demonstrated ability for forward planning with steady progression of operative procedure. | | Obviously planned course of operation with effortless flow from one move to the next. |
| **Knowledge of specific procedure** | Deficient knowledge. Needed specific instruction at most operative steps. | | Knew all important aspects of the operation. | | Demonstrated familiarity with all aspects of the operation. |

**Overall, on this task, should this candidate:** ☐ **Pass** ☐ **Fail?**

**FIGURE 7.3.4** Example of technical, Behaviorally Anchored Rating Scale, rating of individual, *OSATS*. (Reprinted with permission from Martin, J., Regehr, G., Reznick, R., Macrae, H., Murnaghan, J., Hutchison, C., & Brown, M. [1997]. Objective structured assessment of technical skill (OSATS) for surgical residents. *British Journal of Surgery*, *84*(2), 273–278.)

| | Strongly agree 1 | Agree 2 | Neutral 3 | Disagree 4 | Disagree strongly 5 |
|---|---|---|---|---|---|
| a. Rapport and relationship building—an overarching skill<br>Rating ____ | | | | | |
| b. Opens the discussion<br>Rating ____ | | | | | |
| c. Gathers information<br>Rating ____ | | | | | |
| d. Understands patient's perspective<br>Rating ____ | | | | | |
| e. Shares information<br>Rating ____ | | | | | |
| f. Reaches agreement on problems and plans<br>Rating ____ | | | | | |
| g. Provides closure<br>Rating ____ | | | | | |
| h. Addresses family interviewing skills<br>Rating ____ | | | | | |

**FIGURE 7.3.5** Example of nontechnical Likert scale, rating of individual, individual to be rated by faculty, *Kalamazoo Communication Rating Scale*. (Reprinted with permission from Makoul, G. [2001]. Essential elements of communication in medical encounters: The Kalamazoo consensus statement. *Academic Medicine*, *76*(4), 390–393.)

**TABLE 7.3.4**

**Example of Nontechnical Likert Scale, Rating of Individual; Individual to Be Rated by SP: UIC CIS Scale, 2006**

| Please Rate Your Agreement with Each Item | Rating |
|---|---|
| I felt you greeted me warmly upon entering the room. | ( ) Strongly disagree ( ) Disagree ( ) Neutral ( ) Agree ( ) Strongly agree ( ) Not applicable |
| I felt you were friendly throughout the encounter. You were never crabby or rude to me. | ( ) Strongly disagree ( ) Disagree ( ) Neutral ( ) Agree ( ) Strongly agree ( ) Not applicable |
| I felt that you treated me like we were on the same level. You never "talked down" to me or treated me like a child. | ( ) Strongly disagree ( ) Disagree ( ) Neutral ( ) Agree ( ) Strongly agree ( ) Not applicable |
| I felt you let me tell my story and were careful to not interrupt me while I was speaking. | ( ) Strongly disagree ( ) Disagree ( ) Neutral ( ) Agree ( ) Strongly agree ( ) Not applicable |
| I felt you showed interest in me as a "person." You never acted bored or ignored what I had to say. | ( ) Strongly disagree ( ) Disagree ( ) Neutral ( ) Agree ( ) Strongly agree ( ) Not applicable |
| I felt you were patient when I asked questions. | ( ) Strongly disagree ( ) Disagree ( ) Neutral ( ) Agree ( ) Strongly agree ( ) Not applicable |
| I felt the resident displayed a positive attitude during the verbal feedback session. | ( ) Strongly disagree ( ) Disagree ( ) Neutral ( ) Agree ( ) Strongly agree ( ) Not applicable |

Reprinted with permission from Yudkowsky, R., Downing, S. M., & Sandlow, L. J. (2006). Developing an institution-based assessment of resident communication and interpersonal skills. *Academic Medicine*, *81*, 1115–1122.

**TABLE 7.3.5**

**Example of Nontechnical Rating BARS; Individual to Be Rated by SP, RUCIS Scale, 2009**

**UIC CIS 2009 (RUCIS)**

Please choose the option that best describes how you feel toward the resident's communication skills. Some items also have a "not applicable" option. Select this option when the context of the case does not allow you to observe that aspect of the resident's performance.

1. **Friendly communication**
   - ( ) You did not greet me, or greeted me perfunctorily, or communicated with me rudely during the encounter.
   - ( ) Your greeting and/or behavior during the encounter was generally polite but impersonal or distant.
   - ( ) You greeted me warmly and communicated with me in a friendly, personal manner throughout the encounter.
   - ( ) Your greeting and overall communication were friendly and compassionate. Overall, you created an exceptionally warm and friendly environment that made me feel comfortable to tell you all of my problems.
2. **Respectful treatment**
   - ( ) You showed an obvious sign of disrespect during the encounter. For example: You treated me as an inferior.
   - ( ) You did not show disrespect to me. However, I observed some signs of condescending behavior. Although I believe it was unintentional, it made me feel that I was not at the same level with you.
   - ( ) You gave several indications of respecting me. If there was a physical exam, this includes draping me appropriately.
   - ( ) You were exceptionally respectful throughout the encounter. Your verbal and nonverbal communication showed respect for my privacy, my opinions, my rights, and/or my socioeconomic status, etc.
3. **Listening to my story**
   - ( ) You rarely gave me any opportunity to tell my story and/or frequently interrupted me while I was talking, not allowing me to finish what I said. Sometimes I felt you were not paying attention (e.g., you asked for information that I already provided).
   - ( ) You let me tell my story without interruption, or only interrupted appropriately and respectfully. You seemed to pay attention to my story and responded to what I said appropriately.
   - ( ) You allowed me to tell my story without inappropriate interruption, responded appropriately to what I said, and asked thoughtful questions to encourage me to tell more of my story.
   - ( ) You were an exceptional listener. You encouraged me to tell my story and checked your understanding by restating important points.
4. **Honest communication**
   - ( ) You did not seem truthful and frank. I felt that there might be something that you were trying to hide from me.
   - ( ) You did not seem to hide any critical information from me.
   - ( ) You explained the facts of the situation without trivializing negative information or possibilities (e.g., side effects, complications, failure rates).
   - ( ) You were exceptionally frank and honest. You fully explained the positive and negative aspects of my condition. You openly acknowledged your own lack of knowledge or uncertainty, and things you would have to consult with others. When appropriate, you also suggested I seek a second opinion.
   - ( ) **Not applicable**. There was no information for the clinician to provide.
5. **Interest in me as a person**
   - ( ) You never showed interest in me as a person. You only focused on the disease or medical issue.
   - ( ) In addition to talking about my medical issue, you spent some time getting to know me as a person.
   - ( ) You spent some time exploring how my medical issue affects my personal or social life.

Reprinted with permission from Iramaneerat, C., Myford, C. M., Yudkowsky, R., & Lowenstein, T. (2009). Evaluating the effectiveness of rating instruments for a communication skills assessment of medical residents. *Advances in Health Sciences Education*, *14*, 575–594.

be used in conjunction with partial task trainers and mannequins (Kneebone et al., 2005).

*Nontechnical scale with Likert evaluation.* Kneebone et al.'s (2006) "Integrated Procedural Performance Instrument (IPPI)" is used, where the assessment combines SPs with partial task trainers and medical equipment (Figure 7.3.6).

## Mannequin-Based Simulation

Complex clinical events such as team responses to simulated hospital emergencies require use of lifelike full-body mannequins that have computer-driven physiological features (e.g., heart rate, blood pressure) and ability for the performance of procedures that are invasive or traumatic (e.g., IV insertion, chest compressions). With staff facilitation or programming, these mannequins can respond to physical interventions, react appropriately, and record clinical events in real time (McGaghie et al., 2010).

Mannequin simulators comprise the functionality and programming to represent a wide range of pathophysiology and to respond dynamically to user actions. Many MBS are used in crisis management skills because they can be programmed with a wide range of responses and can adapt to emergencies. For example, an MBS may simulate blood pressure, multiple peripheral arterial pulses, breath and heart sounds, muscle twitches from nerve simulation, papillary reflexes, salivation, lacrimation, and bleeding from several anatomic sites (Scalese & Issenberg, 2008). In MBS, vital signs can be displayed in real time; it can respond to the administration of multiple medications and procedures, including intubation and ventilation, chest compressions, and defibrillation; needle or tube thoracostomy and arterial and venous cannulation. Many mannequin simulators have built-in preprogrammed patient profiles and can simulate scenarios involving those patients, and can also be customized. Owing to its peripheral pulses with oxygen saturation and electrocardiographic and other monitoring capabilities for evaluation of advanced cardiac and moulage, it can teach and assess trauma skills. Some mannequins can assess some adult and neonate obstetric skills including delivery and postpartum care. MBS can be

NB F2 completion refers to the end of the second year of the Foundation Programme.

Assessor: candidate:

| A. | | Below expectations for F2 completion | | Borderline for F2 completion | Meets expectations of F2 completion | Above expectations for F2 completion | | Unable to comment |
|---|---|---|---|---|---|---|---|---|
| 1 | Introduction/establish rapport | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 2 | Explanation of intervention including patient's consent to proceed | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 3 | Assessment of patient's needs before procedure | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 4 | Preparation for procedure | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 5 | Technical performance of procedure | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 6 | Maintenance of asepsis | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 7 | Awareness of patient's needs during procedure | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 8 | Closure of the procedure including explanation of follow-up care | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 9 | Clinical safety | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 10 | Professionalism | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 11 | Overall ability to perform the procedure (including technical and professional skills) | 1 | 2 | 3 | 4 | 5 | 6 | 7 |

B. How would you rate the candidate's performance (circle one)

Incompetent Borderline Competent

C.

| Demonstrated strengths | Areas for development |
|---|---|
| | |

**FIGURE 7.3.6** Example of technical and nontechnical Likert scale rating, rating of individual, *IPPI*. (Reprinted with permission from Kneebone, R. L., Kidd, J., Nestel, D., Barnet, A., Lo, B., King, R., … Brown, R. [2005]. Blurring the boundaries: Scenario-based simulation in a clinical setting. *Medical Education*, *39*(6), 580–587

engineered to simulate a wide variety of settings, complications, patients, and patient events.

Some procedural and case-specific checklists have been developed for evaluation of crises where specific solutions or "best actions" have been identified by expert analysis and are recognizable in the simulation (Kim et al., 2009). Yudkowsky (2009) notes that checklists are used to convert the examinee's behavior during the observed performance into a number that can be used for scoring. The MBS modality with its range of real-time reactions is an appropriate modality in which to test whether or not a trainee has chosen the "best action" in response to a patient event.

## Simulation Assessment Tools for MBS

Below are examples of assessment tools used in MBS.

*Weighted technical checklist, individual:* Weighted checklist items may be used to identify the items of "key importance

| Doctor actions | P | $D_1$ | $D_2$ | Weight |
|---|---|---|---|---|
| 1. Assess airway | 1.00 | – | – | 1 |
| 2. Assess breathing – respiratory rate and $O_2$ saturation | 0.88 | 0.31 | 0.34 | 1 |
| 3. Assess circulation – blood pressure and heart rate | 0.88 | 0.42 | 0.40 | 1 |
| 4. Establish level of consciousness | 1.00 | – | – | 1 |
| 5. Expose the patient | 0.77 | 0.55 | 0.33 | 1 |
| 6. Above four issues in less than 1 minute | 0.17 | 0.63 | 0.62 | 2 |
| 7. Establish need for IV access | 1.00 | – | – | 1 |
| 8. Initiate fluid replacement | 0.99 | 0.11 | 0.21 | 2 |
| 9. Provide appropriate fluid replacement | 0.86 | 0.30 | 0.35 | 3 |
| 10. Determine need for type and cross | 0.50 | 0.31 | 0.11 | 4 |
| 11. Proper sequencing of survey | 0.42 | 0.70 | 0.69 | 3 |
| 12. All of the above in less than 3 minutes | 0.17 | 0.60 | 0.51 | 3 |
| 13. Premorbid history | | | | |
| : | | | | |
| penicillin anaphylaxis | 0.25 | 0.22 | 0.14 | 1 |
| 14. Examination of lower extremity – circulatory exam | 0.55 | 0.35 | 0.14 | 1 |
| 15. Examination of lower extremity – neurological exam | 0.52 | 0.43 | 0.13 | 1 |
| 16. Determine need for x-ray | 0.81 | 0.26 | 0.08 | 1 |
| 17. Determine and implement immobilization of left leg | 0.23 | 0.47 | 0.30 | 1 |
| 18. Provide analgesia | 0.45 | 0.22 | 0.17 | 1 |

**FIGURE 7.3.7** Example of weighted technical checklist, rating of individual, *Trauma-haemorrhagic hypotension secondary to long bone fracture.* (Reprinted with permission from Murray, D., Boulet, J., Ziv, A., Woodhouse, J., Kras, J., & McAllister, J. [2002]. An acute care skills evaluation for graduating medical students: A pilot study using clinical simulation. *Medical Education*, *36*(9), 833–841.)

**CHECKLIST**

| ACTION | YES (2 points) | With prompting (1 point) | NO (0 points) |
|---|---|---|---|
| **PROBLEM SOLVING** | | | |
| Prompt ABC assessment | | | |
| Implements concurrent management approach (4 points) | | | |
| **SITUATIONAL AWARENESS** | | | |
| Avoids fixation error (4 points) | | | |
| Re-assesses and re-evaluates situation (4 points) | | | |
| **RESOURCE UTILIZATION** | | | |
| Calls for help when indicated | | | |
| Delegates and directs appropriately | | | |
| **LEADERSHIP** | | | |
| Maintains calm demeanor | | | |
| Acts decisively and maintains control of crisis | | | |
| Maintains global perspective | | | |
| **COMMUNICATION** | | | |
| Communicates clearly and concisely | | | |
| Closes the loop and uses names | | | |
| Listens to team input | | | |
| **TOTAL SCORE (30 points)** | | | |

**Resident #:** **Scenario #:**

**Staff #:** **Date :**

**FIGURE 7.3.8** Example of Mannequin-based, nontechnical multipoint scale, rating of the individual within a team, *The Ottawa CRM Checklist*. (Reprinted with permission from Kim, J., Neilipovitz, D., Cardinal, P., & Chiu, M. [2009]. A comparison of global rating scale and checklist scores in the validation of an evaluation tool to assess performance in the resuscitation of critically ill patients during simulated emergencies [abbreviated as "CRM simulator study IB"]. *Simulation in Healthcare*, *4*(1), 6–16.)

**APPENDIX 1 - OTTAWA CRISIS RESOURCE MANAGEMENT (CRM) GLOBAL RATING SCALE**

**EVALUATION CRITERIA:**

This evaluation scale is directed towards assessing competence in crisis management (CM) skills and care of critically ill patients. The standard of competence has been set at the senior resident level, i.e., the third-year resident who has had prior ICU experience, and through experience as a senior housestaff physician, has previous experience in managing crises. As there exists a requisite base of medical knowledge required to effectively manage crises, this will also be evaluated. However, the focus of evaluation will be on crisis management skills. The skills listed below comprise essential aspects of crisis management. In the simulator case scenario sessions. performance in each of these areas will be assessed, in addition to the amount of prompting or guidance required during the case scenario sessions.

The following criteria will be evaluated:

**LEADERSHIP SKILLS**
Stays calm and in control during crisis
Prompt and firm decision-making
Maintains global perspective ("Big picture")

**SITUATIONAL AWARENESS**
Avoids fixation error
Reassesses and re-evaluates situation constantly
Anticipates likely events

**COMMUNICATION SKILLS**
Communicates clearly and concisely
Uses directed verbal/non-verbal communication
Listens to team input

**PROBLEM SOLVING**
Organized and efficient problem solving approach (ABC's)
Quick in implementation (Concurrent management)
Considers alternatives during crisis

**RESOURCE UTILIZATION**
Calls for help appropriately
Utilizes resources at hand appropriately
Prioritizes tasks appropriately

**OVERALL**

**Resident #:** _____ **Date:** _____

**Staff :** _____ **Time:** _____

**FIGURE 7.3.9** Example of nontechnical Global Rating Scale, rating of the individual within a team, *Ottawa Crisis Resource Management (CRM) Global Rating Scale*. (Reprinted with permission from Kim, J., Neilipovitz, D., Cardinal, P., & Chiu, M. [2009]. A comparison of global rating scale and checklist scores in the validation of an evaluation tool to assess performance in the resuscitation of critically ill patients during simulated emergencies [abbreviated as "CRM simulator study IB"]. *Simulation in Healthcare*, *4*(1), 6–16.)

| OVERALL PERFORMANCE | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| Novice; all CM skills require significant improvement | | Advanced novice; many CM skills require moderate improvement | | Competent; most CM skills require minor improvement | | Clearly superior; few, if any CM skills that only require minor improvement |
| **I. LEADERSHIP SKILLS** | | | | | | |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| Loses calm and control for most of crisis; unable to make firm decisions; cannot maintain global perspective | | Loses calm/control frequently during crisis; delays in making firm decisions (or with cueing); rarely maintains global perspective | | Stays calm and in control for most of crisis; makes firm decisions with little delay; usually maintains global perspective | | Remains calm and in control for entire crisis; makes prompt and firm decisions without delay; always maintains global perspective |
| **II. PROBLEM SOLVING SKILLS** | | | | | | |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| Cannot implement ABC's assessment without direct cues; uses sequential management despite cues; fails to consider any alternative in crisis | | Incomplete or slow ABC assessment; mostly uses sequential management approach unless cues; gives little consideration to alternatives | | Satisfactory ABC assessment; without cues; mostly uses concurrent management approach with only minimal cueing; considers some alternatives in crisis | | Thorough yet quick ABC without cues; always uses concurrent management approach; considers most likely alternatives in crisis |

**FIGURE 7.3.9** *(continued)*

**III. SITUATIONAL AWARENESS SKILLS**

| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|
| Becomes fixated easily despite repeated cues; fails to reassess and re-evaluate situation despite repeated cues; fails to anticipate likely events | | Avoids fixation error only with cueing; rarely reassesses and re-evaluates situation without cues; rarely anticipates likely events | | Usuallly avoids fixation error with minimal cueing; reassesses re-evaluates situation frequently with minimal cues; usually anticipates likely events | | Avoids any fixation error without cues; constantly reassesses and re-evaluates situation without cues; constantly anticipates likely events |

**IV. RESOURCE UTILIZATION SKILLS**

| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|
| Unable to use resources and staff effectively; does not prioritize tasks or ask for help when required despite cues | | Able to use resources with minimal effectiveness; only prioritizes tasks or asks for help when required with cues | | Able to use resources with moderate effectiveness; able to prioritize tasks and/or ask for help with minimal cues | | Clearly able to use resources to maximal effectiveness; sets clear task priority and asks for help early with no cues |

**V. COMMUNICATION SKILLS**

| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|
| Does not communicate with staff; does not acknowledge staff communication, never uses directed verbal/non-verbal communication | | Communicates occasionally with staff, but unclear and vague; occasionally listens to but rarely interacts with staff; rarely uses directed verbal/ non- verbal communication | | Communicates with staff clearly and concisely most of time; listens to staff feedback; usually uses directed verbal/ non-verbal communication | | Communicates clearly and concisely at all times. encourages input and listens to staff feedback; consistently uses directed verbal/ non-verbal communication |

**FIGURE 7.3.9** *(continued)*

**CHECKLIST**

| ACTION | YES (2 points) | With prompting (1 point) | NO (0 points) |
|---|---|---|---|
| **PROBLEM SOLVING** | | | |
| Prompt ABC assessment | | | |
| Implements concurrent management approach (4 points) | | | |
| **SITUATIONAL AWARENESS** | | | |
| Avoids fixation error (4 points) | | | |
| Re-assesses and re-evaluates situation (4 points) | | | |
| **RESOURCE UTILIZATION** | | | |
| Calls for help when indicated | | | |
| Delegates and directs appropriately | | | |
| **LEADERSHIP** | | | |
| Maintains calm demeanor | | | |
| Acts decisively and maintains control of crisis | | | |
| Maintains global perspective | | | |
| **COMMUNICATION** | | | |
| Communicates clearly and concisely | | | |
| Closes the loop and uses names | | | |
| Listens to team input | | | |
| **TOTAL SCORE (30 points)** | | | |

**Resident #:** **Scenario #:**

**Staff #:** **Date :**

**FIGURE 7.3.9** *(continued)*



in defining clinical performances." These analytic items were found by Murray et al. (2002) to discriminate between low- and high-ability performers (Figure 7.3.7).

Emergent situations are characterized by dynamic, changing conditions over time depending on mannequin-environmental or team-initiated actions and interactions. Hence, it is difficult to have a single "best action" or specific remedy that can be checked off on a checklist of emergencies that commonly occur in the ICU or ER (e.g., respirator failure, shock, etc.; Kim et al., 2009). Common nontechnical skills assessed via MBS include communication, teamwork, leadership, and decision making (Yule et al., 2006).

*Nontechnical skills multipoint scale for individual rated as part of team*: Crisis resource management (CRM) are nontechnical skills used to keep patients safe in medical emergencies. The Ottawa CRM Checklist assesses CRM using a multipoint scale assessing individuals as part of a team (Figure 7.3.8).

*Nontechnical skills Global Rating Scale for individuals rated as part of team*: The Ottawa CRM Global Rating Scale assesses the individual within the team (Figure 7.3.9).

*Nontechnical skills rating scale rating teams as teams*: The State Obstetric and Pediatric Research Collaboration (STORC) Clinical Teamwork Scale measures teamwork in the clinical setting using a global scale (Figure 7.3.10).

## VR Simulation

Simulation assessment on the VR simulator allows examinees to perform required techniques on virtual patients

**CTS - Clinical Teamwork Scale™ (Global)**

Please note: **Not relevant**- The task was not applicable to the scenario.

| **Overall** | Not Relevant | Unacceptable | Poor | | | Average | | | Good | | | Perfect |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1. How would you rate teamwork during this delivery/emergency?** | ☐ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |

| **Communication** | Not Relevant | Unacceptable | Poor | | | Average | | | Good | | | Perfect |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Overall Communication Rating:** | ☐ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| **1. Orient new members (SBAR)** | ☐ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| **2. Transparent thinking** | ☐ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| **3. Directed communication** | ☐ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| **4. Closed loop communication** | ☐ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |

| **Situational Awareness** | Not Relevant | Unacceptable | Poor | | | Average | | | Good | | | Perfect |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Overall Situational Awareness Rating:** | ☐ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| **1. Resource allocation** | ☐ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| **2. Target fixation** | ☐ Yes | ☐ No | | | | | | | | | | |

| **Decision Making** | Not Relevant | Unacceptable | Poor | | | Average | | | Good | | | Perfect |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Overall Decision Making Rating:** | ☐ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| **1. Prioritize** | ☐ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |

| **Role Responsibility** | Not Relevant | Unacceptable | Poor | | | Average | | | Good | | | Perfect |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Overall Role Responsibility (Leader/Helper) Rating:** | ☐ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| **1. Role clarity** | ☐ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| **2. Perform as a leader/helper** | ☐ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |

| **Other** | Not Relevant | Unacceptable | Poor | | | Average | | | Good | | | Perfect |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1. Patient friendly** | ☐ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |

Additional Notes (Anything regarding individual performance, assertion of position, etc?):

**On-Site**

**Reviewer** *Print Name* *Sign* *Date*

The CTS-Clinical Teamwork Scale™ was developed by the STORC OB Safety Initiative Team (www.storc.org) through support of the Agency for Healthcare Research and Quality (1 U18 HS015800-02). Guise J-M, Deering S, Kanki B, Osterweil P, Li H, Mori T, Lowe N. STORC OB Safety Initiative: Development and Validation of the Clinical Teamwork Scale to Evaluate Teamwork. Simulation in Healthcare, 3 (4): 217-223, 2008

**FIGURE 7.3.10** Example of nontechnical Likert scale rating of the team as a team. (The CTS Clinical Teamwork Scale was developed by STORC OB Safety Initiative Team, [www.storc.org] through support of the Agency for Healthcare Research and Quality [I U18 HS015800-02]). (Guise, J.-M., Deering, S., Kanki, B., Osterwall, P., Li, H., Mori, T., & Lowe, N. [2008]. STORC OB safety initiative: Development and validation of the clinical teamwork scale to evaluate teamwork. *Simulation in Healthcare*, *3*(4), 217–223.)

or simulate a wide variety of procedures from intravenous cannulation to laparoscopic cholecystectomy and endoscopic methods (McGaghie et al., 2010). The most common use of VR simulators is for evaluation of competence in performing procedures including nonoperative invasive techniques and surgeries. McGaghie et al. (2010) note that these techniques require both psychomotor and perceptual skills that are different from traditional open approaches because the practitioner must perform complex invasive procedures on the basis of indirect and limited viewing of 2D images representing the 3D task. The learner may need to overcome reduced depth perception and poor-quality imaging with some simulated displays—manipulate delicate instruments at a distance from the operative site, with consequent limitations on tactile feedback and compensate for a conflict between proprioception and visual feedback. These weaknesses can be addressed in many ways within the VR modality. Haptic touch and pressure feedback technology can convey the feel of the procedure. Simulators with haptic sensors can capture and record trainee "touch" in terms of location and depth of pressure at specific anatomical sites. McGaghie et al. (2010) caution that much more work is needed in "reliability estimation" of haptic data. VR simulators are now in use to educate surgeons, medical subspecialists, clinical and advanced nurses, and other professions in complex procedures that are too dangerous to practice on live patients (McGaghie et al., 2010).

Serious gaming is another type of VR simulation where evaluation of other management and communication skills for individual teams in a virtual environment (e.g., emergency department, trauma bay, delivery room, or community setting) may occur. Remote and simultaneous assessment of multiple participants caring for virtual patients in a computer-generated environment is also possible (Scalese & Issenberg, 2008).

In addition to their current perceptual and psychomotor functionalities, one of the greatest benefits for VR and serious games simulation modality assessment may lie in their "metarealistic" capabilities. Because digital technology can probe subcutaneously, and dynamically show, for example, multiple layers of the skin and organs, requisite knowledge in both quantifiable behaviors and principle-based actions can be tested. An example is an understanding of how to correctly align the patient's head and maneuver the needle during an ultrasound-guided internal jugular central venous catheter placement (US-guided IJ). The procedure of keeping the patient's head at the correct angle and visualizing the distance between the midpoint of the internal jugular and the lateral border of the carotid artery operationalizes the underlying principle that this zone represents the area of nonoverlap between the internal jugular and carotid artery. Through VR functionality, the trainee can dynamically visualize that, relative to this zone, the margin of safety decreases and the percentage overlap increases from 29% to 42% to



**TABLE 7.3.6**

**Healthcare Simulation Technical and Nontechnical Checklist and Rating Scale Assessment Tools**

| | Checklists | | Rating Scales | | |
|---|---|---|---|---|---|
| | Dichotomous | Weighted | Likert | Behaviorally Anchored Rating Scales (BARS) | Multipoint |
| Technical/ Individual | E.g., Operative Structured Assessment of Technical Skills | E.g., the Israeli Board of Anesthesiology Examination Committee | E.g., Global Rating Index for Technical Skills (GRITS) | E.g., OSATS Global Rating of Technical Performance | E.g., Checklist of Expected Actions, Department of Obstetrics and Gynecology, SUMC |
| Technical/Team | E.g., OTAS (Observational Teamwork Assessment of Surgery) checklist | | | | E.g., Pediatric Resuscitation Team Training Checklist Falcone et al. |
| Nontechnical/ Individual | | | E.g., Kalamazoo Doctor-Patient Communication Rating Scale; IPPI | E.g., Calgary-Cambridge Observation Tool | |
| Nontechnical/ Individual within a team | | E.g., BARS Teamwork Scale, Wright et al. | E.g., OTAS (Observational Teamwork Assessment of Surgery) Likert, Crew Resource Management | E.g., Ottawa GRS | E.g., Ottawa CRM checklist |
| | | | E.g., NOTECCHS, Crew Resource Management | | |
| **Non technical/** Team as a whole | | E.g., CATS Communication & Teamwork Skills Assessment | E.g., TEAM (Team Emergency Assessment Measure)E.g., Anesthetists' Nontechnical Skills (ANTS) System; Crew Resource Management | TAS (Teamwork Assessment Scales) | E.g., STORC Clinical Teamwork Scale |

**EXPERT'S CORNER**

## HIGH-STAKES SIMULATION AND PHYSICIAN ASSESSMENT

Adam I. Levine, MD
*Director, Mount Sinai Human Emulation, Education, and Evaluation Lab for Patient Safety (HELPS) Center*

The term "High-Stakes Simulation" is generally reserved to imply simulation-based activities where one's performance has grave or significant consequences on life or livelihood. From someone who has devoted a significant amount of effort to developing, conducting, and reporting on our simulation-based assessment and retraining program, I actually think the term is used too infrequently. Quite frankly, I consider all the simulations I facilitate to be *high-stakes*; simulation is too resource-intensive to be anything less. As educators, we owe it to our learners to make every simulation *high-stakes* for their sake and the sake of their future patients. If, as simulation educators, we weren't convinced of the absolute virtue of simulation, then why would any of us devote so much time and energy to an educational modality that, at times, only impacts a relatively small number of learners? Surely there are more economical ways to educate, but, as they say, "you get what you pay for."

In 1994, I began the simulation program at Mount Sinai. I considered all our simulations to be *high-stakes* but also felt strongly that the simulated environment should not be used for assessment. Even so, the use of standardized patients was on the rise. Their use rapidly morphed from educational "experiences" for our medical students to objective structured clinical examinations (OSCEs). Student enthusiasm also morphed; they were no longer thrilled about going to the Morschand Center (the standardized patient center at Mount Sinai, memorialized in the infamous Seinfeld episode where Jerry asks Kramer "do all schools use these?" and Kramer replies "only the good ones"). Although I was encouraged to assess student performance in simulation and was invited to sit on committees for developing simulation assessment tools, I avoided these like the plague. I had no interest in my *high-stakes* simulation sessions becoming as unpopular as the OSCEs.

To this day, I have fended off suggestions to assess students in my simulation center. I have, however, embraced the concept of assessing residents, fellows, and practicing anesthesiologists in the simulated environment. After all, any attempt to educate a learner must include some degree of judgment of their performance to debrief them constructively and change future behavior. Over time, it became apparent to me that resident performance in simulation correlated well with clinical performance. This realization was later put to use when I was asked to develop simulations for the purpose of assessing a physician whose competency was being questioned; we were to determine whether this physician was "remediateable." Now I know people look at simulation for assessment as if they were facing a firing squad, but it couldn't be any worse than the "traditional" tools used to assess this physician (e.g., multiple-choice and oral examinations), which he failed abysmally. However, once in the simulated operating room, this physician performed admirably and convinced not only me, but also a committee assembled by the New York State Society of Anesthesiologists that he was indeed a candidate for remediation. People are impressed to learn that not only can simulation be a very powerful tool to assess performance, but in this very early example helped salvage a physician's career. Now, 20 years later, we have developed a vibrant simulation-based reentry program for anesthesiologists. The CARE (Clinical Anesthesia RE-Entry) Program, which includes both simulation-based assessment and clinical retraining, is a much needed service for anesthesiologists who have been out of practice for a variety of reasons. Not all anesthesiologists that present for the CARE Program finish successfully and, we believe, society is a little safer for it.

Do I think everyone should dive in and embrace these *high-stakes* activities in their simulation program? Probably not, especially considering the significant impact of a failed remediation. Obviously no one would think that this activity is any thing less than *high-stakes*, but then again nothing in simulation should be anything less.

72% as the head is turned to the contralateral side form 0 (neutral) to 45 and 90 (Troianos et al., 2011). The serious game/VR modality allows the player to pierce the skin at the correct angle, set the head angle, and see through to and navigate around the vein and artery to be sure that the head angle is compatible with these key anatomic areas, which are in turn visible and able to be manipulated. In this way, the player can demonstrate that he/she has a command of the concepts and principles of US-guided IJ needle insertion and can then go on to demonstrate the psychomotor skills within the partial trainer or MBS exercise.

### Simulation Assessment Tools for Virtual Trainers

Standardized assessment metrics are often built into the technology by the vendor and can be generated by computerized reports.

### Sources of Evidence-Based or Best Practices Simulation Assessment Tools

Table 7.3.6 summarizes the types of checklists and rating scales relevant to simulation assessment as described in the previous section. In addition to peer-reviewed literature, other sources of assessment tools can be found through professional organizations (e.g., AORN, ACS, SSH) and agency regulations and mandates (e.g., Joint Commission).

## SUMMARY

Assessment development and implementation is a complex activity. Fairness and accuracy require careful attention to learner (pre- and postlicensure), tool (simulation modality), assessor, and programmatic components. For this reason, institutional simulation programs are encouraged to access institutional statistical and psychometric support,

as well as content experts in the areas of the subject matter case at hand and of assessment itself. Validating assessment tools and testing for their reliability is also a complex process, often requiring over a year of study, analysis, and refinement. For this reason, many educators and researchers choose to reference the peer-reviewed literature for exemplary studies wherein tools were developed. Assessor training is key to best practices simulation assessment. Clear focus on these factors can lead to defensible, reportable outcomes for healthcare performance assessments.

## REFERENCES


Barrows, H., & Abrahamson, S. (1964). The programmed patient: A technique for appraising student performance in clinical neurology. *Academic Medicine*, *39*(8), 802–805.

Boulet, J. R., & Murray, D. J. (2010, April). Simulation-based assessment in anesthesiology: Requirements for practical implementation. *Anesthesiology*, *112*(4), 1041–1052.

Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education*, *37*, 830–837.

Evans, L., Dodge, M., Shah, T., Kaplan, L., Siegel, M., Moore, C., … D'Onofrio, G. (2010). Simulation training in central venous catheter insertion: Improved performance in clinical practice. *Academic Medicine*, *85*(9), 1462–1469.

Feldman, M., Lazzara, E., Vanderbilt, A., & DiazGranados, D. (2012). Rater training to support high-stakes simulation-based assessments. *Journal of Continuing Education in the Health Professions*, *32*(4), 279–286.

Holmboe, E., & Hawkins, R. (2008). Simulation-based assessment. In E. Holmboe & R. Hawkins (Eds.), *Practical guide to the evaluation of clinical competence* (pp. 179–195). Philadelphia, PA: Mosby Elsevier.

Joint Committee on Fair Testing Practices in Education. (2004). *Code of fair testing practices in education*. Washington, DC: American Psychological Association. Retrieved from http://apa.org/science/programs/testing/fair-code.aspx

Kardong-Edgren, S., Adamson, K., & Fitzgerald, C. (2010). A review of currently published evaluation instruments for human patient simulation. *Clinical Simulation in Nursing*, *6*(1), 25–35.

Kim, J., Neilipovitz, D., Cardinal, P., & Chiu, M. (2009). A comparison of global rating scale and checklist scores in the validation of an evaluation tool to assess performance in the resuscitation of critically ill patients during simulated emergencies (abbreviated as "CRM simulator study IB."). *Simulation Healthcare*, *4*(1), 6–16.

Kneebone, R. L., Kidd, J., Nestel, D., Barnet, A., Lo, B., King, R., … Brown, R. (2005). Blurring the boundaries: Scenario-based simulation in a clinical setting. *Medical Education*, *39*(6), 580–587.

Kneebone, R., Nestel, D., Yadolllah, F., Brown, R., Nolan, C., Durack, J., … Darzi, A. (2006). Assessing procedural skills in context: Exploring the feasibility of an integrated procedural performance instrument (IPPI). *Medical Education*, *40*(11), 1105–1114.

Kulik, C., Kulik, J., & Bangert-Drowns, R. (1990). Effectiveness of mastery learning programs: A meta-analysis. *Review of Educational Research*, *60*(2), 265–306.

Lammers, R. L., Davenport, M., Korley, F., Griswold-Theodorson, S., Fitch, M. T., Narang, A., … Robey, W. C. (2008). Teaching and assessing procedural skills using simulation: Metrics and methodology. *Academic Emergency Medicine: Official Journal of the Society for Academic Emergency Medicine*, *15*(11), 1079–1087.

Martin, J., Regehr, G., Reznick, R., Macrae, H. K., Murnaghan, J., Hutchison, C., & Brown, M. (1997). Objective structured assessment of technical skill (OSATS) for surgical residents. *British Journal of Surgery*, *84*(2), 273–278.

McGaghie, W., Butter, J., & Kaye, M. (2009). Observational assessment. In S. M. Downing & R. Yudkowsky (Eds.), *Assessment in health professions education* (pp. 185–215). New York, NY: Routledge.

McGaghie, W. C., Issenberg, S. B., Petrusa, E. R., & Scalese, R. J. (2010). A critical review of simulation-based medical education research: 2003–2009. *Medical Education*, *44*(1), 50–63. doi:10.1111/j.1365-2923.2009.03547.x

McGaghie, W., Miller, G., Sajid, A., & Telder, T. (1978). *Competency-based curriculum development in medical education an introduction*. Geneva, Switzerland: World Health Organization.

McLaughlin, K., Gregor, G., Jones, A., & Coderre, S. (2006). Can standardized patients replace physicians as OSCE examiners? *BMC Medical Education*, *6*, 12.

Murray, D., Boulet, J., Ziv, A., Woodhouse, J., Kras, J., & McAllister, J. (2002). An acute care skills evaluation for graduating medical students: A pilot study using clinical simulation. *Medical Education*, *36*(9), 833–841. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12354246

Rudolph, J. W., Simon, R., Raemer, D. B., & Eppich, W. J. (2008). Debriefing as formative assessment: Closing performance gaps in medical education. *Academic Emergency Medicine*, *15*, 1010–1016.

Scalese, R., & Issenberg, S. B. (2008). Simulation-based assessment. In E. S. Holmboe & R. E. Hawkins (Eds.), *Practical guide to the evaluation of clinical competence*. Philadelphia, PA: Mosby-Elsevier.

Smith, P., & Ragan, J. (1999). *Instructional design*. New York, NY: John Wiley & Sons.

Troianos, C., Hartman, G., Glas, K., Skubas, N., Eberhardt, R., Walaker, J., & Reeves, S. (2011). Guidelines for performing ultrasound guided vascular cannulation: Recommendations of the American Society of Echocardiography and the Society of Cardiovascular Anesthesiologists. *Journal of the American Society of Echocardiography*, *24*, 1291–1318.

Yudkowsky, R. (2009) Assessment in Health Profession Education. In S. M. Downing & R. Yudkowsky (Eds.) *Performance tests* (pp. 217–243). New York, NY: Routledge.

Yudkowsky, R., Downing, S. M., & Tekian, A. (2009). Standard setting. In S. M. Downing & R. Yudkowsky (Eds.), *Assessment in health professions education* (pp. 119–148). New York, NY: Routledge.

Yule, S., Flin, R., Paterson-Brown, S., & Maran, N. (2006). Non-technical skills for surgeons in the operating room: A review of the literature. *Surgery*, *139*(2), 140–149.