Published in final edited form as:
J Contin Educ Health Prof. 2012 ; 32(4): 279–286. doi:10.1002/chp.21156.

Rater Training to Support High-Stakes Simulation-Based Assessments

Dr. Moshe Feldman, PhD [Assistant Professor],
Office of Assessment and Evaluation Studies, Assistant Director of Research and Evaluation, Center for Human Simulation and Patient Safety, School of Medicine, Virginia Commonwealth University

Dr. Elizabeth H. Lazzara, PhD [Graduate Research Associate],
Department of Psychology and Institute for Simulation and Training, University of Central Florida

Dr. Allison A. Vanderbilt, EdD [Director Evaluation Services, Center on Health Disparities, Assistant Professor], and
Office of Assessment and Evaluation Studies, School of Medicine, Virginia Commonwealth University

Dr. Deborah DiazGranados, PhD [Assistant Professor]
Office of Assessment and Evaluation Studies, School of Medicine, Virginia Commonwealth University

Abstract

Competency-based assessment and an emphasis on obtaining higher-level outcomes that reflect physicians' ability to demonstrate their skills has created a need for more advanced assessment practices. Simulation-based assessments provide medical education planners with tools to better evaluate the 6 Accreditation Council for Graduate Medical Education (ACGME) and American Board of Medical Specialties (ABMS) core competencies by affording physicians opportunities to demonstrate their skills within a standardized and replicable testing environment, thus filling a gap in the current state of assessment for regulating the practice of medicine. Observational performance assessments derived from simulated clinical tasks and scenarios enable stronger inferences about the skill level a physician may possess, but also introduce the potential of rater errors into the assessment process. This article reviews the use of simulation-based assessments for certification, credentialing, initial licensure, and relicensing decisions and describes rater training strategies that may be used to reduce rater errors, increase rating accuracy, and enhance the validity of simulation-based observational performance assessments.

Keywords

simulation; assessment; rater training; licensure; certification; credentialing

Correspondence: Moshe Feldman, School of Medicine, Virginia Commonwealth University, PO Box 980466, Richmond, VA, 23298; mfeldman@vcu.edu.

Disclosures: Deborah DiazGranados and Moshe Feldman are partially funded by a grant award (UL1TR000058) from the NIH.

Introduction

The system of licensure, certification, and credentialing of a physician is predicated on an inference made by patients, health care administrators, and all individuals associated with a health care system about the ability of that physician to provide safe and effective patient care based on established guidelines.[1–3] Organizations responsible for regulating the privilege to practice medicine such as the Liaison Committee for Medical Education (LCME), Accreditation Council for Graduate Medical Education (ACGME), National Board of Medical Examiners (NBME), individual states, American Board of Medical Specialties (ABMS), and organizations that deliver patient care are entrusted to maintain a system that applies rigorous assessment methods and criteria when making decisions about physician competence.[3–5] Assessments that poorly predict actual practice behaviors threaten the ability of the regulatory system to assure patient safety and the delivery of high-quality patient care,[6] and may lead to other detrimental consequences such as increased health care costs and higher medical liability insurance premiums.[7,8]

The adoption of the ACGME/ABMS 6 core competencies (professionalism, patient care and procedural skills, medical knowledge, practice-based learning and improvement, interpersonal and communication skills, systems-based practice) and an emerging paradigm emphasizing assessment of behavioral outcomes have spurred a demand for assessments based on direct observations of performance behaviors and skills.[6,9–13] On-the-job assessment is one method for directly observing and rating performance, but it is prone to unreliability because of inconsistency in the type and difficulty of clinical contexts in which skills are demonstrated.[9,14–16] As a result, interest in the use of simulation-based assessments has grown.

Simulation involves a representation of a task or working environment in which physicians can demonstrate their skills without increasing the risks to patients.[17–23] Simulation provides systematic, replicable, and structured environments allowing standardization of behaviorally based skill assessments, thus enabling comparisons among and between trainees and changes in performance over time.[20–21] In instances when minimum competence required to practice medicine is in question such as in remediation, retraining, or physicians reentering the workforce, simulation may be the only way to assess behavioral skills without increasing threats to patient safety.[22] The role of simulation in high-stakes assessments such as those used to make licensure, certification, and credentialing decisions continues to grow with increased demand for assessing behavioral outcomes, adoption of competency based assessment, and more affordable and reliable simulation technology.[4,22]

Simulation affords for opportunities to directly observe and assess physician performance behaviors; however, to use it effectively and make accurate assessments, medical educators need to be aware of and apply best practices of observational performance assessment.[4,10,23] Observational performance assessment is more complex than knowledge tests or attitude surveys in that the rater introduces an additional source of variance due to rater errors or biases that may weaken the quality of inferences one can make about physician competence.[24–27] Simulation-based observational performance assessments have shown mixed levels of reliability and validity. High levels of reliability have been achieved using a checklist to assess competence in cardiac physical examination skills[28] and assessments of board-certified anesthesiologists on technical and nontechnical skills using key behavioral markers.[29] The mini clinical evaluation exercise (mini-CEX)[30] is one assessment tool where some studies have achieved adequate levels of reliability,[31] while others have not.[32] Novice and expert rater assessments of nontechnical skills (eg, situation awareness, decision making, communication and teamwork, leadership) have also been shown to have poor accuracy when minimal rater training was provided.[33] This article describes the role of

simulation-based observational performance assessments for high-stakes assessment and provides rater training strategies to improve accuracy and strengthen the validity of licensure, certification, and credentialing decisions.

Strengths and Limitations of Simulation-Based Observational Assessments

In addition to standardizing the assessment process and reducing risks to patients, simulation-based assessments have several other strengths. They can be used to evaluate multiple aspects of physician performance, including technical, interpersonal, clinical, and teamwork skills.[34,35] For example, observational assessments of central venous catheterization skills using a part task trainer have been used to discern experienced physicians from novices.[36] Standardized patient ratings are now being used for assessing communication and interpersonal skills for high-stakes assessment, and computer-based simulated clinical scenarios are used to assess clinical management skills as part of the USMLE Step 3 exam.[22] Simulation can also improve observational performance assessment by reducing the rater's cognitive workload since learning events and the corresponding responses can be determined a priori, enabling observers to be preemptively prepared.[20] In some instances simulation-based assessments are the only feasible approach for capturing physician performance on rare cases.[35]

Simulation also facilitates assessment at higher levels of evaluation, allowing for stronger inferences about physicians' clinical practice. For example, a physician's skill in providing effective Advanced Cardiac Life Support (ACLS) is often measured using a multiple-choice knowledge test (Moore's Level 3),[12] but assessments at this level may weaken the degree to which that measure is associated with competent practice behaviors and improved patient health.[37] Simulations allow assessments at the level of competence (Moore's Level 4).[12] Competence displayed in the context of a simulation does not necessarily translate into or predict clinical performance.[37] Multiple factors may explain discrepancies between performance in a simulated assessment context and in real practice. For example, learners' motivation to perform may be heightened during formal assessment situations[38–40] or potential barriers such as poor leadership support and existing standards of practice that remain fixed regardless of training may prevent certain skills from manifesting in actual clinical practice.[41–43] However, more direct assessments of clinical performance also have limitations. For example, assessments derived from clinical outcomes may be a product of the team members' collective ability rather than individual competence, as in the case of collaborative care.[9,44] Although reliance solely on assessments derived from a simulation may lead to false inferences about practice behavior,[37] simulation-based assessments can complement other assessment strategies to help decision makers and regulatory bodies make decisions about physician competence and grant privileges to practice medicine.[10,34,37]

Current Uses of Simulation in High-Stakes Assessments

Licensure

Licensure demonstrates that a physician has been found to hold the minimum competency needed to treat and diagnose patients.[45] In the United States, each state sets its own criteria for licensing a physician to practice medicine, but states generally require evidence that the candidate has earned a medical degree, passed a 3-step US Medical Licensure Examination (USMLE), and completed training within an accredited residency program. The USMLE now incorporates simulation in its Step 2 and Step 3 exams. The Step 2 exam uses standardized patients to simulate a patient encounter and to assess clinical skills while the Step 3 exam uses computer-based clinical scenarios in which examinees must make accurate, time-sensitive decisions in managing the case.

Maintenance of Certification

The American Board of Medical Specialties (ABMS) includes 24 specialty boards offering board certification to indicate that a physician holds exceptional expertise in a particular specialty or subspecialty of medical practice.[45] Physicians must complete the requisite predoctoral medical education and 3 to 5 years in an accredited residency training program. They must also pass a board certification examination within 3 to 7 years after completion of residency training depending on the specialty in order to receive initial certification.[46] To maintain their certification, physicians with time-limited cer-tificates must also comply with the ABMS Maintenance of Certification (MOC) process every 6 to 10 years depending on the specialty. The revised ABMS MOC policies emphasize the importance for each certifying physician to demonstrate specialty specific knowledge, skills, and use of best evidence practices across the 6 ACGME/ABMS core competencies, with each specialty board setting its own criteria for meeting these standards.[47–48] Although the concept of using simulation for certifying physicians has been around for over 20 years,[49] it has only recently been formally adopted for certification requirements in anesthesiology, surgery, and internal medicine. Other specialties considering simulation as a required component for MOC include emergency medicine, pediatrics, and radiology.[2,22–23,50–51]

Credentialing

Physicians are now required to demonstrate competence in surgical procedures before being given privileges to perform on real patients.[10,22] For example, simulation can be used to meet some credentialing requirements before performing carotid artery stenting procedures on patients, with assessments at Moore's Level 1 (ie, participation) and Level 4 (ie, demonstration of skills in a simulated setting) based on automated metrics such as procedure time and procedural errors.[52] Technology-enabled assessment methods in simulation is an active area of research,[53] and new methods of assessment (eg, accuracy score) leveraging simulation sensor technologies and physiological model data feeds are being developed.[54] Simulation-based assessments will continue to rely on observational performance assessments to adequately measure ACGME/ABMS core competencies and skills until validity of automated assessments is better established.

Rater Training Strategies

The reliability of observational performance assessments varies across assessment tools, settings, rater types, and competencies being assessed.[55–57] Observing, encoding, retrieving, and evaluating performance can be challenging tasks that are prone to rating errors and biases, especially for complex skills or assessment settings. An example of a rating error is the contrast effect, which occurs when raters spuriously increase their performance ratings after observing very poor performance or decrease their ratings after observing very good performance.[58] TABLE 1 lists several common rater errors and their relevance for the validity of observational performance assessments derived from simulation scenarios.

Rater training aims to improve rater performance by developing the necessary knowledge, skills, and attitudes to accurately evaluate demonstrated skills and competencies. Four core rater training strategies include rater error training (RET), performance dimension training (PDT), frame-of-reference training (FOR), and behavioral observation training (BOT).[31,32,59]

Rater Error Training (RET)

RET generally consists of a brief lecture and discussion of rater errors.[59] The purpose of RET is to reduce the occurrence of rater errors and increase rating accuracy by increasing awareness of potential rater errors such as halo, leniency, and contrast effects, and

encouraging observers to avoid them. Although some evidence has suggested that RET can have an adverse impact on rating accuracy, empirical studies have also demonstrated that RET can be effective when focused on reducing rating errors (eg, halo and leniency) rather than achieving an appropriate distribution of ratings.[59–62]

Performance Dimension Training (PDT)

The purpose of PDT is to train raters to recognize appropriate behaviors associated with each dimension targeted for evaluation. The dimension, competency, or skill is described to raters and behavioral examples of the dimension are provided through a written vignette, video, or role-play. PDT aims to improve accuracy by decreasing variability in ratings due to rater errors or biases unrelated to the targeted performance behaviors. An example of PDT for a simulation-based assessment of teamwork might include a short lecture describing teamwork in health care and specific behaviors that constitute teamwork (eg, situation monitoring, giving information). Video examples or vignettes may be used to provide specific examples of "giving information" so that raters learn to associate similar visual behavioral cues with the dimension being evaluated.[56,59]

Frame-of-Reference Training (FOR)

FOR helps increase accuracy and provides standardization across raters. The objective of FOR is to help raters discriminate variations in the quality of demonstrated skills. An iterative process whereby raters practice rating behaviors, assess reliability, and discuss discrepancies between raters is used to provide practice and feedback to improve rater performance. Again, using teamwork as an example, an FOR approach to preparing raters to assess situation monitoring by team members might ask them to individually assess videotaped examples representing a range of performances from poor (eg, failing to provide critical information to team members) to good (eg, providing critical information urgently) would be developed. Raters would then review the videos as a group and discuss discrepancies between individual raters, often with a facilitator, with the goal of developing a common set of rating rules.

Behavioral Observation Training (BOT)

BOT focuses on developing observation skills such as the detection, perception, recall, and recognition of behavioral events that represent the targeted skill or competency. Behavioral events can be thought of as "triggers" or moments within the scenario associated with the targeted skill, thus providing an opportunity for the rater to assess that skill.[20,63–64] BOT consists of orienting the rater to anticipate assessment events in a simulation scenario in which key behaviors will be demonstrated.

Each rater training strategy is aimed at mitigating a unique threat to the validity of observational performance assessments and is delivered using a variety of methods. TABLE 2 describes each rater training strategy and summarizes its effectiveness for improving validity of performance ratings. Although only one of the studies referenced in TABLE 2 specifically pertains to assessment within the medical context, and the remaining research is derived from the industrial/organizational psychology domain, we posit that all of these studies can inform and improve medical assessments.

Effectiveness of Rater Training

Rater performance is a function of the degree to which raters accurately observe and evaluate the targeted performance dimension or skill and should be consistently monitored. One method for measuring rater accuracy is calculating the percentage of items rated similarly to an established "gold" standard. For example, Evans et al[56] developed a set of

choreographed videos specifically designed to exemplify a certain skill level for central venous catheter insertion. Ratings for each video were scored for accuracy and an overall score calculated. Another method for measuring rater performance is comparing ratings to an expert's ratings of the same performance. In general, 70% agreement is considered acceptable, 80% is considered satisfactory, and 90% and above is considered excellent.[65] Other, more advanced statistical techniques such as Cohen's Kappa or Generalizability Theory[66] can also be used to produce comprehensive estimates of sources of variability contributing to performance ratings.

Rater training has been used to strengthen the reliability and validity of observational performance ratings in the military, industry, and health care.[56,59] Both novice and expert raters have been found to produce accurate and reliable ratings after rater training. Evans et al[56] trained a group of novice raters to observe and evaluate central venous catheter insertion skills with 95% accuracy. Although expert physicians may provide valid and accurate observational performance assessments, they have also been found to be more resistant to rater training when the referent expert rating model differs from their own schemas of good and poor performance.[32,63–64,67] A meta-analysis on the effects of rater training on rating errors, rating accuracy, and sensitivity found that all 4 rater training approaches described above significantly benefited ratings by reducing rater errors or improving rating accuracy.[59] RET is most effective at reducing rater errors when it focuses on awareness of and avoiding rater errors rather than focusing on maintaining a normal distribution of performance ratings across all participants. In other words, RET that instructs raters to maintain a normal distribution across ratees has not been effective in reducing rater errors. FOR training had the biggest positive effect on rating accuracy especially when combined with BOT.[59]

Other studies have failed to find any benefits of rater training. In a randomized control trial using residents, Cook et al[32] assessed the impact of a half-day rater training workshop on rater performance 4 weeks later and found no significant effects of rater training. Iramaneerat et al[68] showed evidence of rater errors during a clinical skills assessment of fourth-year medical students even after receiving rater training, but do not describe the rater training procedures used. Studies have also shown drift effects where raters show high levels of interrater reliability initially, but poor reliability in subsequent performance assessments.[69]

Variability in the effectiveness of rater training may be due to differences across training designs and variation in how rater training is implemented. Rater training programs may occur within a single day in the form of a workshop lasting up to 8 hours. Other models may be continuous, where raters train over the course of several days or weeks in smaller training periods.[56] Technological solutions to rater training may include deploying practice rating videos on digital media, Internet, or learning management systems and having raters discuss discrepancies virtually. Currently, evidence on the effects of rater training supports its utility for enhancing reliability and validity, but more work is needed to delineate the unique effects of the various components and implementation designs.

Guidelines for Improving Rater Accuracy

Simulation-based assessments are increasingly being used to make high-stakes decisions for medical licensure, certification, and credentialing, yet observational performance assessments derived from simulated tasks are often prone to rater errors and biases. This article outlined 4 rater training strategies that can be used to support simulation-based assessments used in high-stakes licensing, certification, and credentialing decisions. All were effective in reducing rater errors and improving rater accuracy in at least some instances. When observers utilize structured observational protocols and undergo sufficient

rater training, simulation-based observational performance assessments can predict the quality of patient care delivered by a physician above and beyond knowledge tests alone.[18] However, as our review highlighted, there are also instances of poor rater performance even after extensive rater training. To optimize raters' training and enhance the accuracy of their assessments, we suggest the following guidelines.

- *Rating training should be planned and conducted with an awareness of the many sources of error in ratings of performance.* Multiple factors unrelated to actual performance can influence variability in observational performance ratings,[27] hence calibrating raters is imperative for ensuring that the optimum level of validity is achieved. These factors include the clinical expertise of raters, instructional design characteristics (eg, training time, method of delivery) of each training strategy, and the number of observations needed to achieve adequate levels of generalizability.
- *Expert clinicians should be involved in the development of rater training and high-stakes assessment process when possible.* However, the reality is that they are often unable to devote adequate time to training and serve as raters for high-stakes simulation assessments that may occur over several days. Experts may also be more resistant to changing their own performance schemas to a different set of behaviors.[33] Training raters without clinical expertise can be an effective approach when experts are unavailable or too cost intensive. In this approach, experts may be used to develop a set of standardized ratings for each practice video so that raters can be trained to use similar decision rules with accuracy measured against the "expert" gold standard.
- *The choice of a rater training approach should take into account the resources available and the nature of the skills being assessed.* RET, PDT, FOR, and BOT have all been shown to improve raters' ability to observe and evaluate performance in some way, but evidence regarding the combination of strategies are required to achieve minimum standards of rater performance remains unclear. FOR training has the largest benefit for rating accuracy[59] and reduces the occurrence of rater errors, but is often the most difficult to develop. Also, rater training strategies may differ depending on the skill being assessed and type of rating tool. For example, a technical skills checklist with specific behaviors for airway management skills using an airway simulator may require less training time than a checklist for rating team performance in a simulated code scenario.
- *Rater training should be considered an ongoing activity.* As with any other skill, training raters requires monitoring and skill maintenance through practice.
- *Reports on the impact of rater training used in interventional studies should include more concise descriptions including training time, rater training strategies used, rater characteristics, and training delivery methods.* Rater training programs are often poorly described in published studies.[15,55] Understanding these details should enable the design of more effective rater training programs.

Simulation-based assessments with adequate rater training are becoming a standard and valuable component of national programs for certification and credentialing.[2,10,36–70] Those involved in high-stakes assessment can use rater training to reduce rating errors in simulation-based assessments and establish more standardized and defensible high-stakes assessments that can strengthen the validity of licensure, certification, and credentialing decisions.

References

1. Blank L, Kimball H, McDonald W, Merino J. ABIM Foundation. ACP Foundation. European Federation of Internal Medicine. Medical professionalism in the new millennium: a physician charter 15 months later. Ann Intern Med. 2003; 138:839–841. [PubMed: 12755556]
2. Boulet JR, Murray DJ. Simulation-based assessment in anesthesiology. Anesthesiology. 2010; 112:1041–1052. [PubMed: 20234313]
3. Nahrwold DL. Continuing medical education reform for competency-based education and assessment. J Contin Educ Health Prof. 2005; 25:168–173. [PubMed: 16173066]
4. Melnick DE. Physician performance and assessment and their effect on continuing medical education and continuing professional development. J Contin Educ Health Prof. 2004; 24:S38–S49. [PubMed: 15712776]
5. Institute of Medicine (IOM). Crossing the Quality Chasm: A New Health System for the 21st Century. Washington, DC: National Academies Press; 2001.
6. Davis DA, Mazmanian PE, Fordis M, Van Harrison R, Thorpe KE, Perrier L. Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. JAMA. 2006; 296:1094–1102. [PubMed: 16954489]
7. Prytowsky JB, Bordage G, Feinglass JM. Patient outcomes for segmental colon resection according to surgeon's training, certification, and experience. Surgery. 2002; 132(4):663–672. [PubMed: 12407351]
8. Goodman JC, Villarreal P, Jones B. The social cost of adverse medical events and what we can do about it. Health Aff. 2011; 30(4):590–595.
9. Evans R, Elwyn G, Edwards A. Review of instruments for peer assessment of physicians. BMJ. 2004; 328:1–5. [PubMed: 14703521]
10. Holmboe ES. Assessment of the practicing physician: challenges and opportunities. J Contin Educ Health Prof. 2008; 28(S1):S4–S10. [PubMed: 19058254]
11. Institute of Medicine (IOM). Redesigning Continuing Education in the Health Professions. Washington, DC: National Academies Press; 2010.
12. Moore DE, Green JS, Gallis HA. Achieving the desired results and improved outcomes: integrating planning and assessment throughout learning activities. J Contin Educ Health Prof. 2009; 29(1):1–15. [PubMed: 19288562]
13. van Hoof TJ, Meehan TP. Integrating essential components of quality improvement into a new paradigm for continuing education. J Contin Educ Health Prof. 2011; 31(3):207–214. [PubMed: 21953662]
14. Lurie SJ, Mooney CJ, Lyness JM. Measurement of the general competencies of the accreditation council for graduate medical education: a systematic review. Acad Med. 2009; 84(3):301–309. [PubMed: 19240434]
15. Kogan JR, Holmboe ES, Hauer KE. Tools for direct observation and assessment of clinical skills of medical trainees. JAMA. 2009; 302(12):1316–1326. [PubMed: 19773567]
16. Gerbert B. Criteria for evaluating methods used to assess physician performance. Mobius. 1984; 4(4):44–47. [PubMed: 10269871]
17. Salas E, Rosen MA. Beyond the bells and whistles: when simulation-based team training works best. Harvard CRICO RMF Forum. 2008; 26(4):6–7.
18. Lievens F, Patterson F. The validity and incremental validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations for predicting job-performance in advanced-level high-stakes selection. J Appl Psychol. 2011; 96(5):927–940. [PubMed: 21480685]
19. Gordon J, Wilkerson W, Shaffer D, Armstrong E. Practicing medicine without risk: students and educators responses to high-fidelity patient simulation. Acad Med. 2001; 76:469–472. [PubMed: 11346525]
20. Rosen MA, Salas E, Silvestri S, Wu TS, Lazzara EH. A measurement tool for simulation-based training in emergency medicine: the simulation module for assessment of resident targeted event responses (SMARTER) approach. Simul Healthc. 2008; 3:170–179. [PubMed: 19088661]

21. McGaghie W, Siddall VJ, Mazmanian PE, Myers J. Lessons for continuing medical education from simulation research in undergraduate and graduate medical education. Chest. 2009; 135(3): 62S–68S. [PubMed: 19265078]
22. Levine AI, Schwartz AD, Bryson EO, Demaria S Jr. Role of simulation in US physician licensure and certification. Mt Sinai J Med. 2012; 79:140–153. [PubMed: 22238047]
23. Boulet JR, Jeffries PR, Hatala RA, Kornfdorffer JR, Feinstein DM, Roche JP. Research regarding methods of assessing learning outcomes. Simul Healthc. 2011; 6:S48–S51. [PubMed: 21705967]
24. Haladyna TM, Kramer GA. The validity of subscores for a credentialing test. Eval Health Prof. 2004; 27(4):349–368. [PubMed: 15492047]
25. Downing SM. Threats to the validity of clinical teaching assessments: what about rater error? Med Educ. 2005; 39:350–355. [PubMed: 15813754]
26. Downing SM, Haladyna SM. Validity threats: overcoming interference with proposed interpretations of assessment data. Med Educ. 2004; 38:327–333. [PubMed: 14996342]
27. Kogan JR, Conforti L, Bernabeo E, Iobst W, Holmboe E. Opening the black box of clinical skills assessment via observation: a conceptual model. Med Educ. 2011; 45:1048–1060. [PubMed: 21916943]
28. Hatala R, Scalese RJ, Cole G, Bacchus M, Kassen B, Issenberg SB. Development and validation of a cardiac findings checklist for use with simulator-based assessments of cardiac physical examination competence. Simul Healthc. 2009; 4(1):17–22. [PubMed: 19212246]
29. Graham J, Mudumbai SC, Gaba DM, Boulet J, Howard SK, Davies MF. External validation of simulation-based assessments with other performance measures of third-year anesthesiology residents. Simul Healthc. 2012; 7(2):73–80. [PubMed: 22374230]
30. Norcini JJ, Blank LL, Arnold GK, Kimball HR. The mini-CEX (clinical evaluation exercise): a preliminary investigation. Ann Intern Med. 1995; 123:795–799. [PubMed: 7574198]
31. Holmboe ES, Hawkins RE, Huot SJ. Effects of training in direct observation of medial residents' clinical competence: a randomized trial. Ann Intern Med. 2004; 140:874–881. [PubMed: 15172901]
32. Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz VS. Effect of rater training on reliability and accuracy of mini-CEX scores: a randomized, controlled trial. J Intern Med. 2008; 24(1):74–79.
33. Yule S, Rowley D, Flin R, Maran N, Youngson G, Duncan J, Paterson-Brown S. Experience matters: comparing novice and expert ratings of non-technical skills using the NOTSS system. Surg Educ. 2009; 79:154–160.
34. van Zanten M, Boulet JR, McKinley D. Using standardized patients to assess the interpersonal skills of physicians: six years' experience with a high-stakes certification examination. Health Commun. 2007; 22(3):195–205. [PubMed: 17967142]
35. DeMaria S Jr, Levin AI, Bryson EO. The use of multimodality simulation in the retraining of the physician for medical licensure. J Clin Anesthesiol. 2010; 22:294–299.
36. Dong Y, Suri HS, Cook DA, et al. Simulation-based objective assessment discerns clinical proficiency in central line placement: a construct validation. Chest. 2010; 137:1050–1056. [PubMed: 20061397]
37. Goldstein, IL.; Ford, JK. Training in Organizations. 4. Belmont, CA: Wadsworth; 2002.
38. DuBois CL, Sackett PR, Zedeck S, Fogli L. Further exploration of typical and maximum performance criteria: definitional issues, prediction, and White-Black differences. J Appl Psychol. 1993; 78(2):205–211.
39. Kleinmann M. Are rating dimensions in assessment centers transparent for participants? Consequences for criterion and construct validity. J Appl Psych. 1993; 78:988–993.
40. Sackett PR, Zedeck S, Fogli l. Relations between measures of typical and maximum job performance. J Appl Psychol. 1988; 73(3):482–486.
41. Baker R, Camosso-Stefinovic J, Gillies C, et al. Tailored interventions to overcome identified barriers to change: effects on professional practice and health care outcomes. Cochrane Database Syst Rev. 2010; 3:CD005470. [PubMed: 20238340]
42. Machin MA, Fogarty GJ. Perceptions of training-related factors and personal variables as predictors of transfer implementation intentions. J Bus Psychol. 2003; 18:51–71.

43. Royer JM. Theories of the transfer of learning. Educ Psychol. 1979; 14:53–70.
44. Landon BE, Normand SL, Blumenthal D, Daley J. Physician clinical performance assessment: prospects and barriers. JAMA. 2003; 290:1183–1189. [PubMed: 12953001]
45. [Accessed March 27, 2012.] American Board of Medical Specialties. http://www.abms.org/AboutBoardCertification/means.aspx
46. [Accessed March 27, 2012.] American Board of Medical Specialties. http://www.abms.org/newsandevents/medianewsroom/releases/releaseBoardEligibility02072012.aspx
47. Havens C, Mallin J. Climate change: It's not about the weather—continuing medical education and maintenance of certification and li-censure. Perm J. 2011; 15(3):88–92. [PubMed: 22058675]
48. Cassel CK, Holmboe ES. Professionalism and accountability: the role of specialty board certification. Trans Am Clin Climatol Assoc. 2008; 119:295–304. [PubMed: 18596848]
49. Maatsch JL. Assessment of clinical competence on the Emergency Medicine Specialty Certification Examination: the validity of examiner ratings of simulated clinical encounters. Ann Emerg Med. 1981; 10(10):504–507. [PubMed: 7283213]
50. Gallagher CJ, Tan JM. The current status of simulation in the maintenance of certification in anesthesia. Int Anesthesiol Clin. 2010; 48:83–99. [PubMed: 20616639]
51. Berkenstadt H, Ziv A, Gafni N, Sidi A. The validation process of incorporating simulation-based accreditation into the anesthesiology Israeli national board exam. Isr Med Assoc J. 2006; 8:728–733. [PubMed: 17125130]
52. Gallagher AG, Cates CU. Approval of virtual reality training for carotid stenting: what this means for procedural-based medicine. JAMA. 2004; 292:3024–3025. [PubMed: 15613672]
53. Amin Z, Boulet JR, Cook DA, et al. Technology-enabled assessment of health professions education: Consensus statement and recommendations from the Ottawa 2010 conference. Med Teach. 2011; 33:364–369. [PubMed: 21517684]
54. Nadler I, Sanderson P, Liley H. The accuracy of clinical assessments as a measure for teamwork effectiveness. Simul Healthc. 2011; 6:260–268. [PubMed: 21705968]
55. Pelgrim EM, Kramer AM, Mokkink HA, van den Elsen L, Grol RM, van der Vleuten CM. In-training assessment using direct observation of single-patient encounters: a literature review. Adv Health Sci Educ. 2011; 16:131–142.
56. Evans LV, Morse JL, Hamann CJ, Osborne M, Lin Z, D'Onofrio G. The development of an independent rater system to assess residents' competence in invasive procedures. Acad Med. 2009; 84(8):1135–1143. [PubMed: 19638785]
57. Borman WC. Format and training effects on rating accuracy and rater errors. J Appl Psychol. 1979; 64(4):410–421.
58. Gaugler BB, Rudolph AS. The influence of assessee performance variation on assessors' judgments. Pers Psychol. 1992; 45(1):77–98.
59. Woehr DJ, Huffcutt AI. Rater training for performance appraisal: a quantitative review. J Occup Organ Psychol. 1994; 67:189–205.
60. Pulakos ED. A comparison of rater training programs: error training and accuracy training. J Appl Psychol. 1984; 69:581–588.
61. Bernardin HJ, Pence EC. Effects of rater training: creating new response sets and decreasing accuracy. J Appl Psychol. 1980; 65:60–66.
62. Hedge JW, Kavanagh MJ. Improving the accuracy of performance evaluations: comparison of three methods of performance appraiser training. J Appl Psychol. 1988; 73:68–73.
63. Fowlkes J, Dwyer DJ, Oser RL, Salas E. Event-based approach to training (EBAT). Int J Aviat Psychol. 1998; 8(3):209–222.
64. Ludbrook J, Marshall VR. Examiner training for clinical examinations. Br J Med Educ. 1971; 5:152–155. [PubMed: 5559490]
65. Streiner, DL.; Norman, GR. Health Measurement Scales: A Practical Guide to Their Development and Use. 3. New York, NY: Oxford University Press; 2003.
66. Iramaneerat C, Yudkowsky R, Myford CM, Downing SM. Quality control of an OSCE using generalizability theory and many-faceted Rasch measurement. Adv Health Sci Educ Theory Pract. 2008; 13(4):479–493. [PubMed: 17310306]

67. Williams RG, Klamen DA, McGaphie WC. Cognitive, social, and environmental sources of bias in clinical performance ratings. Teach Learn Med. 2003; 15:270–292. [PubMed: 14612262]

68. Iramaneerat C, Yudkowsky R. Rater errors in a clinical skills assessment of medical students. Eval Health Prof. 2007; 30:266–283. [PubMed: 17693619]

69. McLaughlin K, Ainslie M, Coderre S, Wright B, Violato C. The effect of differential rater function over time (DRIFT) on objective structured clinical examination ratings. Med Educ. 2009; 43(1): 989–992. [PubMed: 19769648]

70. Ziv A, Rubin O, Sidi A, Berkenstadt H. Credentialing and certifying with simulation. Anesthesiol Clin. 2007; 25:209–223. [PubMed: 17574186]

Lessons for Practice

- Rater training programs should be developed to support simulation-based observational assessments.
- Expert clinicians should be involved in the development of rater training and high-stakes assessment process when possible.
- Rater reliability and accuracy should be consistently monitored for quality assurance purposes.

TABLE 1

Common Rater Errors and Threats to Validity

Rater Error	Description	Threats to Validity
Central tendency	Avoiding extreme positive or negative ratings	Reduces ability to discriminate performance levels between individuals and a standard cutoff
Halo error	All ratings based on one positive or negative observation	Introduces systematic bias into performance ratings and may reduce accuracy
Leniency	Avoiding poor performance scale items	Positively skews ratings and reduces accuracy
Primacy/recency effect	All ratings based on observations made early or late in the scenario	Introduces systematic bias and may reduce accuracy
Contrast effect	Ratings are made relative to performance of previous group	Positively skews ratings when prior group performed very poorly; negatively skews ratings when prior group performed very well
Similar-to-me effect	Ratings based on degree of similarity to rater	Introduces systematic bias and may reduce accuracy
Stereotype effect	Ratings based on group inclusion rather than individual differences	Positively or negatively biases ratings and may reduce accuracy

TABLE 2

Rater Training Strategies and Their Effectiveness for Improving Validity

Strategies	Description	Effectiveness	References
Rater error training (RET)	Familiarizes raters with common rating errors, usually through a didactic lecture	Moderate effects in reducing occurrence of rating errors Mixed effects in improving rating accuracy Degrades rating accuracy when focus is on obtaining normal distribution across rates instead of avoiding rater errors	Woehr, 1994; Hedge & Kavanagh, 1988
Performance dimension training (PDT)	Familiarizes raters with the dimensions being targeted for evaluation using definitions and examples	Mixed effects in reducing occurrence of rater errors Small effects in improving rater accuracy.	Woehr, 1994; Holmboe, 2004
Frame-of-reference training (FOR)	Discriminates between variations in quality of behavioral dimensions being targeted for evaluation Raters practice rating examples, discuss discrepancies between ratings and receive feedback	Small to moderate effects in reducing occurrence of rater errors Moderate to large effects in improving rating accuracy	Woehr 1994; Holmboe, 2004
Behavioral observation training (BOT)	Develops observation skills such as the detection, perception, and recall of specific behavioral events.	Moderate to large effects in improving rating accuracy	Woehr 1994; Holmboe 2004