

Evaluations

Assessing the ACGME General Competencies: General Considerations and Assessment Methods

Susan R. Swing, PhD

Abstract

The Accreditation Council for Graduate Medical Education's (ACGME's) general competency and outcome assessment initiative (i.e., the ACGME Outcome Project) is an effort to enhance residency education and accreditation effectiveness by increasing emphasis on educational outcomes. The Project is also a response to concerns about new graduates' ability to meet the demands of today's practice environment. The competencies emphasize learning in new domains (e.g., Practice-Based Learning and Improvement and Systems-Based Practice) and more traditional ones (e.g., Patient Care and Medical

Knowledge). Outcome assessment will provide evidence of residency program educational effectiveness and information to guide improvement. This paper discusses the development and implementations of assessment methods appropriate to evaluate the performance of residents in each of the core competencies. **Key words:** Accreditation Council for Graduate Medical Education; ACGME; core competencies; assessment; outcomes; residency. *ACADEMIC EMERGENCY MEDICINE* 2002; 9: 1278–1288.

The Accreditation Council for Graduate Medical Education's (ACGME's) general competency and outcome assessment initiative (i.e., the ACGME Outcome Project) is an effort to enhance residency education and accreditation effectiveness by increasing emphasis on educational outcomes. The Project is also a response to concerns about new graduates' ability to meet the demands of today's practice environment. The competencies emphasize learning in new domains (e.g., Practice-Based Learning and Improvement and Systems-Based Practice) and more traditional ones (e.g., Patient Care and Medical Knowledge). Outcome assessment will provide evidence of residency program educational effectiveness and information to guide improvement.

The outcome assessment goal for Phase 2 of the Project (July 2002–June 2006) is to stimulate improvements in residency programs' evaluations of their residents. The desired product is more credi-

ble, accurate, reliable, and useful educational outcome data. Assumptions underlying this increased emphasis on assessment are: 1) exposure to and participation in educational activities do not assure learning; 2) assessment methods used most often in residency programs do not optimally support learning or ascertain how well residents perform; and 3) assessment results can stimulate and direct performance improvement of both individual residents and educational programs.

Crafting a system that provides definitive information about individual and program performance, as well as information useful for improvement, requires thinking beyond what tools to use. Other factors related to the what, who, when, and how of assessment mediate its effectiveness. Some of these are discussed below. A review of selected assessment methods follows the opening discussion.

CONSIDERATIONS WHEN DEVELOPING AN ASSESSMENT SYSTEM

The purpose of assessment is the first consideration for design of an assessment system. From an educational perspective, assessment is done to support the learning process (formative assessment) or to determine the status of learning and performance (summative assessment). The latter is generally used for decision making (e.g., promotion, completion of the residency). Related policy purposes for assessment are improvement and accountability. In residency programs, performance assessment also should be used to protect patients from risk and harm. The what, when, who, and how of assess-

From the Accreditation Council for Graduate Medical Education (ACGME), Chicago, IL (SRS).

Received April 23, 2002; accepted April 23, 2002.

Presented at the Council of Emergency Medicine Residency Directors (CORD) Consensus Conference on the ACGME Core Competencies: "Getting Ahead of the Curve," March 2002, Washington, DC.

Supported in part by a grant from the Robert Wood Johnson Foundation. The views and opinions herein are those of the author.

The ACGME/ABMS Toolbox of Assessment Methods developed by Susan Swing, PhD, and Phil Bashook, EdD, provided the framework for this article.

Address for correspondence and reprints: Susan R. Swing, PhD, Accreditation Council for Graduate Medical Education, 515 North State Street, Chicago, IL 60610. Fax: 312-464-4098; e-mail: srs@acgme.org.

ment will vary at least somewhat depending on purpose. Relevant issues related to purpose are integrated into the discussion of those considerations.

1. What Should Be Assessed? The considerations for what to assess fall into three overlapping conceptual groupings: competencies (e.g., general competencies, specialty-specific knowledge, and skills); knowing and showing how to perform versus actual performance; and processes versus outcomes.

One feature of good assessment is content validity. For assessment of residents, this quality exists to the extent that competencies, care processes, patient conditions, and contexts of care, etc., representative of emergency medicine (EM) practice are assessed. Typically, a blueprint is used to guide selection of a representative sample of content. A blueprint for EM might consist of the framework provided by the ACGME's six general competencies defined more specifically by the core patient care content of EM and essential specific elements of other competencies (e.g., patient-physician communication¹; professional behaviors²; practice analysis and improvement³) as defined by experts in these domains. What to sample for assessment purposes is key to optimizing patient safety and quality of care, as well as making accurate judgments about residents' ability to perform competently. General selection principles include sampling performance on 1) commonly seen, but potentially serious conditions; 2) frequently performed procedures; and 3) infrequently performed, but core procedures and skills, that could put the patient at risk if inappropriately done.

Another consideration involves assessing whether the resident knows, knows how, or can show how to do something (ability to perform) or what he or she actually does when performing real clinical tasks.⁴ Ability to perform is assessed using tests (e.g., written exams, computer-based exams, objective structured clinical exams). Except for high-fidelity simulations, tests remove the effects of daily environmental factors (e.g., multitasking and interruptions, misplaced charts, too many patients, end-of-day fatigue) and gauge what the resident can do in a less-complicated situation. The closer the assessment context resembles a real situation, the more likely it is that an assessment will be able to predict residents' performance as practicing physicians. Assessments that can predict future performance exhibit predictive validity. Tests are most useful for formative purposes, in particular, for determining whether prerequisite knowledge and skills have been obtained. Assessment of performance in real situations is better for ascertaining the level of actual clinical competence.

Determining whether to assess processes of performance or outcomes is the third consideration. Processes of performance include, for example, the steps of a physical exam, steps involved in inserting an endotracheal tube, or communicative components of a resident-patient encounter (e.g., responding to the patient's emotions, checking for understanding). Outcomes are results relative to what the emergency physician wants to accomplish, e.g., patient compliance, restoration of breathing, or x-ray interpretation that agrees with that of the radiologist. (As implied, "outcomes" achieved by the emergency medicine resident may be intermediate relative to the final clinical outcome.) So, what should be assessed—processes or outcomes? The simple answer is both. Concurrent assessment of processes and outcomes informs whether improvement is needed and what steps/processes need to be improved. In general, assessment of processes should be the focus early on in the learning of new skills. Assessment of outcomes best serves the purpose of summative assessment and accountability. Outcome assessment is an appropriate focus as the resident approaches the end of his or her education.

2. When Should Assessment Take Place? When and how often to assess depends in part on the purpose the assessment is intended to serve. The timing of assessment is related to protection of patients, support of learning, and summative assessment for high-stakes decision making.

When residents' first-time performance of procedures puts patients at significant risk of harm even if a supervisor is present, resident practice on models followed by assessment of competence should occur before procedures are attempted on real patients. This much is obvious. Perhaps less obvious is the importance of early assessment of performance of common care processes where an error in the form of a routine omission or incorrectly done diagnostic test, for example, could put numerous patients at risk.

The timing of assessment can significantly and positively influence the efficiency and effectiveness of learning. First, early assessment of prerequisite, foundational, core skills and knowledge potentially enables early identification of deficiencies and the timely provision of educational activities and supervision. Second, during residency many skill sets (e.g., focused interviews and physical exams tailored to specific presenting symptoms) become automatic through repeated application. Assessment early on could preclude the consolidation of inaccurate or incomplete care processes and pre-

vent the need for the effortful process of unlearning and relearning later.⁵ Last, in the interest of residents' uninterrupted progress through the program, assessment conducted before a rotation or other specialized experience ends could inform residents of the need to redouble or refocus their efforts.

Assessment removed in time from initial learning is important for determining retention of knowledge and skills. For this reason, a final summative assessment of residents' performance of key clinical processes toward the end of residency is recommended so that a firm basis for attesting to residents' clinical competence (or ability to practice without supervision) is available.

Performance of clinical tasks is notoriously task-specific. Ongoing, frequent assessment is one way to ensure that a sufficiently large sample of tasks are being assessed to ensure patient safety, support learning, and enable an appropriate inference of overall performance to be made.

3. Who Should Assess Resident Performance?

In addition to faculty or attending supervisors, nurses, medical students, radiologists, consultants from other specialties, translators, social workers, and patients are among the potential evaluators of EM residents. By including members of these groups, the scope of evaluation is broadened to include insights into resident performance in situations not observed by supervisors and from persons with special expertise and different positions in the hierarchy. Evaluators should have direct knowledge through observation or interaction of the to-be-evaluated performance(s) and be knowledgeable of what constitutes effective performance.

Research evidence suggests that "who should assess" might depend on "what is assessed." Typically, faculty supervisors and attendings are considered the obvious choice for assessing patient care skills involving physical exams, diagnostic test ordering and interpretation, patient management, and procedures, etc. However, there is some evidence that they have difficulty providing reliable assessments of humanistic behaviors, interpersonal relations, and communication skills⁶⁻⁹ and that nurses and resident peers are better able to assess these domains reliably.¹⁰⁻¹² Although it seems that patients would be the best evaluators of these competency components, studies have noted that large numbers of patient respondents are needed in order to get stable estimates.^{10,13-15}

4. How Should Assessment Be Done? How refers to the assessment methodology, specifically to as-

essment methods (e.g., ratings, checklists, oral exams) and how they are implemented. Considerable variation is possible in the design and implementation of specific instruments (e.g., a specific clinical performance rating form) for any given type of method (e.g., ratings). Design and implementation features make a difference in the assessment's accuracy, reliability, and usefulness. The discussion below highlights the following features of assessment methodology: psychometric characteristics, scoring criteria, evaluator training, and informational yield.

Indices of psychometric quality related to the consistency of measurement across tasks, observers/raters, situations, or time (i.e., reliability and generalizability) are properties of the results obtained using a particular assessment method or tool in a specific situation. Design features of the method and implementation processes associated with the tools determine the reliability and generalizability of the results. For example, well-defined scoring or rating criteria, the training of observers/raters to use the criteria, and assessment of specific tasks rather than an amalgam of performances are all associated with higher reliability.¹⁶⁻¹⁹ Methods characterized by these particular features, such as objective structured clinical exams (OSCEs), standardized patient (SP) exams, and structured oral exams, tend to be more reliable than global ratings that have none of these features. The implication is that by designing these features into an assessment methodology, reliability can be improved.

Traditionally, validity is the extent to which an assessment measures what it sets out to measure. Two types of validity important for the assessment of resident performance have already been mentioned and defined—content validity and predictive validity. As presented earlier, design features related to content validity and predictive validity, respectively, are assessment of a broadly representative sample of the domain content (i.e., skill and knowledge competencies) and performance on clinical tasks in an authentic setting. Evidence for a third type of validity, construct validity, occurs when observed findings meet expectations determined by a theoretical model of what is being assessed. In the case of resident performance, one model is medical expertise. Investigators have considered better performance by more experienced physicians to be indicative of the construct validity of an assessment. Messick has further articulated key validity issues as ". . . the interpretability, relevance, and utility of scores, the import or value implications of scores as a basis for action, and the functional work of scores in terms of social consequences of their use."²⁰

High reliability and generalizability coefficients (i.e., 0.80) are considered essential when a single assessment method is used for high-stakes decisions. When multiple measures are used, this standard can be relaxed since it is expected that the aggregate of the results will be more reliable than the individual measures. High reliability is neither necessary nor sufficient when selecting methods that will be helpful for learning. Despite the amount of attention given to reliability, it is generally not considered as important as validity.²¹

Assessment methods vary in informational yield, that is, the amount, type, and specificity of information provided back as results. In order to optimally support learning and improvement, the assessment results must indicate, for example, what specific knowledge needs to be learned or what steps of a procedure, patient interview, physical exam, etc., were performed well, left out, or performed incorrectly. Assessment results in the form of scores or ratings of performance on composite categories of knowledge, skills, and behaviors may motivate poor performers' desire to improve but are not very helpful in directing improvement efforts. For example, an assessment consisting of numeric rating or even selection of qualitative labels (such as "good" or "unsatisfactory") for each of the six ACGME general competency areas does not provide information about particular strengths or weaknesses; that is, the scores are difficult to interpret because they are too general, and no "behavioral" meaning has been assigned to the numbers or qualitative labels. In contrast, when assessment on a 36-station OSCE is scored using detailed checklists, the assessment is information-rich, in terms of both specific processes that were performed or performed correctly and how well the resident performed on 36 different tasks. Whereas high informational yield is essential for an assessment method to be helpful for learning, if the results are to be used only for decisions, then scores are the only results required as "feedback" from the assessment.

METHODS FOR ASSESSING RESIDENT PERFORMANCE

Several methods are available to assess resident performance. The best use of the methods vary, as do their features vis-à-vis technical characteristics, implementation characteristics, and informational yield. Their history of use also varies. One of the methods is widely used, another is the criterion ("gold") standard for assessment but is infrequently used, and others are potentially useful but require further development.

Ratings. Global ratings are the most widely used method of assessment in graduate medical education and in EM residencies.²² They involve making a subjective judgment about the quality of behaviors, skills, knowledge, and attitudes exhibited. Typically, in residency programs, residents' clinical skills, medical knowledge, communication skills, and professional behaviors are rated each rotation, quarterly, semiannually, and/or annually. Global ratings involve the mental aggregation of resident performance across numerous patients, situations, and days. However, overall, there is substantial variability in rating forms and how they are used. Important features include categories of performance to be rated (i.e., generic clinical performance or task-specific) and whether the rating is global or focused (based on a single encounter).

Residency directors consider ratings their most important method overall for ascertaining residents' level of competence.²² Research findings related to the psychometric qualities of global ratings make their use in this way highly questionable. Ratings exhibit the systematic rater errors of leniency/severity, range restriction (failure to use the entire rating scale), failure to distinguish among dimensions (halo effect), and cognitive distortion (inappropriate weighting to form judgments).¹⁹ These errors have been reported in studies of ward/in-training/rotation evaluations of residents and medical students by their supervisors.^{6,23,24} Other studies have shown that residents' ratings on ward/in-training evaluations were much higher than their OSCE scores.^{25,26} In studies of the reliability of global ratings, a wide range of findings have been reported. In two different studies of the same global rating form, average reliability ranged from 0.64 to 0.87.^{27,28} In two other studies of other global rating forms, reliability ranged from 0.06 to 0.36.^{7,29}

The content validity of global ratings is also questionable. Since ratings represent an aggregate across time, there is no way of determining what content (knowledge, skills, behaviors) is actually being evaluated or the extent to which it is representative of the larger domain. Also, studies of supervisors' global ratings of residents and medical students report finding halo effects, suggesting that the ratings actually reflect an overall impression rather than performance on the individual categories.^{6,23} Furthermore, global ratings are not very useful in directing improvement and will be difficult to interpret, particularly when the only information available is the numeric ratings on a relatively small number of categories.

Ratings have been used to assess performance in single encounters and to assess competency-specific performance (e.g., humanistic behaviors and com-

munication skills). Research findings related to these uses are somewhat mixed, but overall support the conclusion that reliable ratings can be obtained. When rating forms were tailored to clinical tasks and based on observation of single encounters, interrater reliabilities were in the range of 0.64–0.78 when performance was assessed during medical encounters²³ and surgical procedures.³⁰ In contrast, in other studies of a focused resident–patient encounter during which residents’ history, physical exam, diagnostic, humanistic, and communication skills were assessed using a global rating form, interrater reliabilities were much lower, and in one study, the accuracy of raters in detecting incorrect or omitted maneuvers was also low.^{9,31} In several studies, acceptable levels of interrater reliability were reported when raters used forms consisting of multiple discrete communication skill items to evaluate medical students’ and residents’ video-taped or live encounters with patients.³² Competency or task-specific rating forms used for focused assessment have higher informational yield than global ratings, but are still susceptible to rater errors.

Rating forms could be used to assess resident performance on all the general competencies. The forms are easy to construct and use. Acceptable reliability has sometimes been obtained for global ratings, and has been obtained in many situations when a focused assessment is done using a rating form that is tailored to the competencies being assessed. Even so, the use of ratings and, specifically, global ratings for high-stakes purposes is questionable because of validity and accuracy problems. When compared with assessment results obtained using other methods, for example OSCEs and SPs, ratings often seem inflated. Furthermore, global ratings on a relatively small number of performance dimensions are not helpful for learning and are difficult to interpret if the meaning of the response options are not provided. Ratings done using forms that evaluate how well specific behaviors were performed during a single encounter are potentially more useful, but the results could be misleading because of rater errors and accuracy problems.

Checklists. Behavioral checklists consist of specific actions (e.g., asking about psychosocial issues, checking for understanding) that make up a more complex activity (e.g., the patient interview). The actions are assumed to be important for competent, effective performance. Assessment using a checklist can involve identifying whether the resident performs the actions during an observed performance or whether the actions are performed completely,

partially, and/or correctly. Behavioral checklists are used exclusively to assess performance for single tasks (e.g., a single resident–patient encounter).

Checklists are routinely used as a method for documenting performance on OSCE and SP exams. They also have been used to assess both videotaped and live encounters with patients during which patient interviews, physical exams, test interpretations, and procedures were performed. Studies of checklists report overall good interrater reliability, typically in the range of 0.70–0.80.^{30,33–35} Numerous studies of the use of communication skills checklists have reported reliability/agreement in the range of 0.70–0.90.³² More often than not, evaluators are trained to use the checklists. Some checklists include objective criteria for use in judging correctness or completeness.

The content validity of a given checklist depends on the extent to which the contents represent a consensus of expert opinion about the important behaviors or steps that make up the checklist and whether the score on the checklist is related to a desired outcome. Achieving either of these criteria can be a challenge.³⁶ Nonetheless, some investigators have reported that experts reached consensus on their checklist elements.^{33,37} Other studies have reported obtaining evidence supportive of construct validity based on the predicted findings that more-experienced residents would obtain higher scores than less-experienced residents.^{30,38}

Checklists are information-rich, at least as they pertain to the specific tasks they have been developed to assess (i.e., physical exams, communication skills, and procedures). Checklist results show specifically what was done and done well, and what was not, thus indicating specific improvements needed. One caution in the use of checklists is that they must be constructed to assess performance expected of residents, so that efficiencies learned with experience are not scored as omitted steps.

Checklists are recommended for assessment of observable behaviors and components of work products associated with all the general competencies. However, in order to determine whether the behavior is consistently performed in different situations, multiple assessments in a variety of contexts will be needed. Checklists are more useful for assessing fundamental skills and less useful for assessing nuances and qualitative dimensions of performance that may distinguish the merely competent from the more proficient or expert performer.

360-degree Assessment. In a 360-degree assessment, individuals from the full circle of reporting relationships perform an assessment, usually by

rating designated performance dimensions. The 360-assessment of residents might involve faculty/attending supervisors, consultants, peers, medical students, patients, nurses, allied health professionals, social workers, technicians, and clerks. Self-ratings are an important and recommended part of 360-assessment. Tornow and associates recommend that the 360-assessment instrument consist of the same set of items for all evaluators plus a subset of items designed to capture unique aspects of interaction with particular groups.³⁹ For residents, interpersonal and communication skills, including teamwork and professionalism, are obvious common components for a 360-assessment.

To the best of my knowledge, 360-assessments (as described above) in medical education have not been used. However, inferences about the reliability of 360s can be made from looking at related applications, including global ratings in general and ratings by supervisors, nurses, peers, patients, and self. When designed and implemented like global ratings, 360 ratings by individual evaluators will be subject to the same errors and unreliability as described above. The recommended solution to this is to include enough evaluators to achieve a stable estimate of performance. Research results have indicated that ratings from 20–50 attending physicians,¹⁰ one to five nurses,^{10,11} and 20–50+ patients^{10,13–15} are needed to yield a stable rating of residents' humanistic qualities. No information on the validity of 360-assessments for residents is available. However, it might be assumed that the results of a 360-assessment have relevance and credibility because of the multiple perspectives represented and the number of evaluators involved.

As noted above for ratings, the informational yield of the 360-assessment is in part dependent on the instrument design. To the extent that raters respond to specific behavioral components of a competency, information useful for improvement will result. If the 360-assessment results consist only of numeric ratings without narrative comments or descriptive criteria, the information may be useful for summative assessment (dependent on its reliability). However, it will not be useful for directing improvement, although it likely will stimulate reflection and improvement efforts. The greatest informational yield of the 360-assessment comes from the inclusion of multiple perspectives across groups on the resident's performance and the potential of comparing perspectives and detecting patterns.

At this point in time, 360-assessment of residents is an interesting method for assessing interpersonal and communication skills and professionalism and for the purpose of stimulating reflection and improvement. This method needs to be further devel-

oped and tested. However, feasibility is an issue and is manifest both in obtaining a sufficient number of assessments and perhaps even more so in the administrative challenge of notifying evaluators, and aggregating and reporting the results. Like many assessment methods, computer systems to support collection, aggregation, and reporting of the information would substantially increase feasibility, at least as related to information management. Because of the importance of computer support and the time requirements for crafting an instrument that has suitable common items and items customized for different groups, collaborative, coordinated development of this tool is strongly encouraged. The phasing-in of evaluator groups would also make this process more do-able.

Structured Oral Exams and Structured Case Discussions. Oral exams involve the presentation of a clinical scenario; in some versions, examiners role-play patients and, in still others, SPs are involved.^{40–42} The examinee is asked to manage the case. Depending on the format, the examinee is asked what to look for on exam, how to interpret findings and tests, and how to manage the patient. Individual scenarios last around 5 minutes. In a structured oral exam, the questions and a marking scheme are predetermined. Oral exams have been used to assess Patient Care (information gathering, decision making, patient management), Medical Knowledge, and Interpersonal and Communication Skills.

In traditional oral exams, which are unstructured in type and difficulty of questions as well as in the marking scheme, a wide range of interrater reliability has been reported (–0.04–0.85).⁴⁰ However, numerous investigators have reported that when structure is added in the marking schemes, higher reliability coefficients usually are obtained.^{40,41,43} For example, Anastakis and colleagues reported per case interrater reliabilities of 0.78–0.91 on a four-scenario structured oral exam. Each scenario had five or six questions and a predetermined marking scheme.⁴¹

Residents' case discussions with their supervisors could provide an alternate format for an oral-exam-like assessment. By introducing structured questioning and marking into case discussions for a predetermined sample of cases, various sources of unreliability potentially associated with current use of case discussion performance information (e.g., resident case presentation, variation in cases, subjective differences in rating, forgetting across time) could be reduced. Because the questioning is about residents' own patients, the assessment is highly

relevant. Otherwise, the validity of this method and its overall informational yield will depend on conducting the “structured case discussion oral” for a representative sample of patient cases. This is an untried, untested technique.

Depending on their format, oral exams have elements in common with SP exams and OSCEs, but mostly seem to be an alternative to written exams with clinical scenarios. They are further removed from actual patient care than are OSCEs, SPs, and simulations, and involve “telling how” rather than “showing how.” Their main advantage over these formats is cost. However, further development and investigation of the structured case discussion oral are recommended. This technique would have the advantage of building on a process already performed in most EM residency programs.²² This technique could be used when it is not feasible or desirable to observe the resident–patient encounter. It also could be used to supplement observation through a short set of questions that probe the residents’ decision making. To enhance feasibility related to summarization and display of the results, personal digital assistants (PDAs) could be used to document performance. The greatest challenge will be reliability of the results.

Simulators, Models, and Simulations. In the context of resident performance assessment, simulators, models, and simulations are used to imitate real patients, anatomical regions, clinical tasks, and/or the environment or context in which medical services are provided. For the present purpose, SPs and OSCEs are considered together as one type of simulation, and high-tech simulators and models are discussed together as a second category of assessment. Low-fidelity simulations (e.g., paper-and-pencil and computer-based branching problems) are excluded.

Standardized patients (SPs) and objective structured clinical examinations (OSCEs). Standardized patients are well persons or actual patients trained to simulate a medical condition in a standardized way. An SP exam consists of multiple SPs, each presenting a distinct condition in a 10–12-minute patient encounter. OSCEs are multistation exams that may include several SP stations as well as other stations in which some part of a clinical encounter is replicated, (e.g., x-ray interpretation, full panel of patient laboratory results presented). In both cases, assessment involves the resident performing the required procedure or interpreting test results the same as in a real clinical situation. The SP or faculty observer performs the evaluation typically using a detailed checklist or a rating form.

There is a substantial body of evidence that attests to the psychometric quality of SP exams and OSCEs. Studies of OSCEs across the surgical and medical specialties have provided evidence of content and construct validity and reliability that reaches the benchmark value of 0.80. A range of performance across residents and better performance by more experienced residents are among the published findings, as are findings of a range of performance across residents, including those who perform below minimum pass levels.^{25,26,44–47} The number of stations or SPs in an exam has a greater effect on its reliability/generalizability than do other features; generally, 14–18 SPs or OSCE stations are needed to enable making inferences about a resident’s overall clinical competence. Development of scoring criteria and rater training are typically a part of SP and OSCE development.

When checklists are used and there are a large number of clinical tasks, patient conditions, etc., included in the exam, the exam will be rich in information and have high utility in identifying what the resident does well and what needs improvement. Even performance on a single SP can provide considerable useful feedback when a checklist is used for assessment. The use of global rating forms, while more efficient, will result in an assessment process that can be used for summative assessment, but one that will not be as useful for formative assessment as one utilizing checklists.^{44,48,49}

Objective structured clinical examinations and SP exams can be used for either summative or formative assessment. Some medical educators consider OSCEs the criterion standard of assessment methods.⁴⁷ They have a high degree of authenticity; when unannounced, SPs cannot be distinguished from real patients.⁴⁴ Noninvasive components of Patient Care, Interpersonal and Communication Skill, and components of Professionalism (ethical behavior and cultural sensitivity) can be reliably and validly assessed using these methods. The cost of these methods exceeds that of many other assessment methods.

Simulators and models. Simulators are high-tech, computer-based devices that emulate one or more of the following: the anatomical features, physiological and physical responses, touch, sound, and appearance of the human body. Models are similar in that they imitate at least anatomical features of the human body, but they are not computer-based. Animals, cadavers, and synthetic mannequins, etc., are considered models. For assessment purposes, the simulators and models substitute for the human body as the resident performs medical procedures. Assessment involving high-tech simulators can oc-

cur by way of computer recording of the time to complete procedures, the procedure outcome, and/or the damage done to tissues and organs, etc. Or resident performance can be observed and assessed using a checklist or rating form. For assessments using models, the latter is a required component.

Simulators and models offer a low-risk alternative for assessing performance on high-risk cases or procedures and low-frequency conditions. In addition, for simulators with built-in scoring, technicians can conduct the assessment. For simulators, repetition of case presentation is not a problem. Simulators and models are useful for assessing patient care, including information gathering, diagnosis, patient management, and procedures. Simulators, models, and simulations are available for assessing performance on a number of conditions and situations relevant to EM, including cardiac conditions, accidents (victim stabilization, extrication, limb trauma, head lacerations, compound fractures), airway management, resuscitation, acute stroke, intravenous access, and hazardous material management.⁵⁰

Reports of the reliability of assessments involving simulators, models, and simulations, other than those for OSCEs and SPs, are difficult to find. One study was reported that computed interrater reliability for computer and faculty scoring. It had nonsignificant or weak correlations.⁵¹ A few studies have provided evidence of the validity of simulator and assessment by demonstrating better performance (as expected) by more experienced residents and physicians.^{38,52}

Informational yield of assessments involving simulators and models will vary depending on what is assessed (process or outcome; a single procedure or a representative sample of clinical conditions) and the method used (detailed checklist or global rating form). But in general, the informational yield will indicate whether performance improvement on some specific task is needed. Furthermore, it may specify exactly what improvements are needed.

Simulators and models are promising methods for assessing Patient Care performance in low-frequency but important situations where there is high risk to the patient. To the extent that assessment procedures and documentation are similar to those used for SPs and OSCEs, acceptable reliability and validity can be expected. However, the reliabilities of computer scoring schemes will need to be established.

Portfolios. A portfolio is a collection of work products or other evidence of accomplishment. A port-

folio that will be used for assessment should: 1) describe the amount, type, and quality of evidence required to establish proof of competence; and 2) utilize scoring rubrics, checklists, or rating scales for assessment of individual products. Reflection on one's learning experiences and performance is generally considered an important part of using a portfolio approach and might be assessed along with other performance dimensions.

A portfolio approach could be used to assess all of the general competencies. However, this method is best used to assess competencies or dimensions of competencies difficult to assess in other ways. This could include, for example, analysis and improvement of practice (project report); response to feedback (narrative of reflection and action taken); response to ethical and professional dilemmas experienced or medical errors committed (narrative describing the situation, response, and reflection); changes facilitated to improve system functioning (description of change, before and after data; new protocol developed); and advocacy for patients (audio-tape).

Because of the complexity of portfolios—the diversity of their contents—scoring them reliably is generally considered a significant challenge. Nonetheless, interrater reliability in the range of 0.60–0.82 has been reported.⁵³ Suggested methods for increasing reliability of scoring are use of criteria or a scoring rubric, use of benchmark examples, and rater training. Content validity of the portfolio can be obtained by broad sampling of performance and work samples related to the competency dimensions being assessed.

The portfolio approach provides an opportunity for obtaining unique information about the quality and depth of the residents' learning and performance and how they think about their work. When experiences and reflection for the same topic are collected across time, progression in knowledge and sophistication in thinking about these issues (or lack thereof) can be discerned. Faculty–resident discussions around the portfolio submissions provide an opportunity for further learning.

Although the utility and feasibility of portfolios in resident assessment have not been clearly established, their unique features and a few documented successes in medical education warrant a recommendation for further development of this method.^{53,54} As suggested above, use should be restricted to assessment of competency dimensions difficult to assess in other ways. This includes aspects of Practice-Based Learning and Improvement, Professionalism, and Systems-Based Practice.

CONCLUSIONS AND RECOMMENDATIONS

The following set of simple recommendations for assessment are based on the considerations and findings from research on the assessment methods reviewed.

1. Determine the purpose of the assessment. Assessment is more effective and efficacious when some features of the assessment process—when to assess, psychometric qualities required, and informational yield of assessment—are tailored to fit the specific purpose, either support for learning or summative high-stakes decision making.
2. Assess a broad sample of the most important competencies, care process, and outcomes. Determine what is important by the number of patients potentially affected by omissions and errors and the degree of risk to the patient. Processes are most helpful for learning, but outcomes ultimately matter most. Assessment of both is recommended.
3. Integrate assessment with learning. Assessment conducted during the formative stage and before extensive application facilitates the efficiency and effectiveness of learning by ensuring that only correct processes become automatic. Early assessment also protects patients from harm.
4. Involve informed and invested individuals as evaluators. Evaluators must have significant direct knowledge of the performances they are evaluating and specific knowledge of what constitutes “good, better, best” performance.
5. Assess specific, actual performances. Assessment that depends on the mental aggregation of days, weeks, and months of impressions of performance is fraught with inaccuracy. Assessment by performance sampling may be a better approach.
6. Use well-defined criteria. Availability of criteria and training in how to use them are associated with more reliable assessment.

There can be wide variation in the design and implementation features of specific instruments within a larger class of assessment methodology. The features of the methodology rather than the assessment approach tend to determine the goodness of the assessment. Assessment methods that have included focused assessment of residents performing clinical tasks, instruments designed for the tasks, task-specific performance criteria, and training of evaluators tend to produce results with higher reliability. Assessment approaches that typically exhibit these features include OSCEs, SP ex-

ams, and checklists (most often utilized in OSCEs and SPs). When these features are built into other methods thought to be unreliable, i.e., ratings and oral examinations, higher reliabilities are obtained.

At present, OSCEs and SP exams seem to be the best methods for conducting assessment for high-stakes decisions. However, a one-time administration of an OSCE would do little to protect patients and support learning. Instead, ongoing implementation of snapshot methods of assessment involving observation or interaction focused on specific aspects of resident performance will be more helpful. Numerous checklists already exist and are used in this way, and structured case-discussion orals possibly could be developed to serve this purpose. 360-assessment and portfolios potentially can provide unique insights into resident performance and should be further developed and tested.

Assessment methods that work will take longer to develop and implement than the most feasible, easy-to-implement, and widely used method, i.e., rotation/quarterly ratings. Coordinated, collaborative development of prototype instruments, item banks, and computer technology will ease the difficulty of putting better assessment methods into place. Computer technology can ease the difficulty of collecting, aggregating, and reporting the results.

References

1. Makoul G. Essential elements of communication in medical encounters: the Kalamazoo consensus statement. *Acad Med.* 2001; 76:390–3.
2. Adams J, Schmidt T, Sanders A, Larkin G, Knopp R. Professionalism in emergency medicine. *Acad Emerg Med.* 1998; 5:1193–9.
3. Headrick LA, Richardson A, Priebe GP. Continuous improvement learning for residents. *Pediatrics.* 1998; 101: 768–73.
4. Miller GE. The assessment of clinical skills/competence/performance. *Acad Med.* 1990; 65(9, Sept RIME suppl): S63–S67.
5. Schrifflin RM, Dumais ST. The development of automatism. In: Anderson J (ed). *Cognitive Skills and Their Acquisition.* Hillsdale, NJ: Lawrence Erlbaum Associates, 1981.
6. Gray JD. Global rating scales in residency education. *Acad Med.* 1996; 71(1, Jan RIME suppl):S55–S63.
7. Davis JK, Inamdar S, Stone RK. Inter-rater agreement and predictive validity of faculty ratings of pediatric residents. *J Med Educ.* 1986; 61:901–5.
8. Kaplan CB, Centor RM. The use of nurses to evaluate house officers’ humanistic behavior. *J Gen Intern Med.* 1990; 5:410–4.
9. Noel GL, Hebers JE, Caplow MP, Cooper GS, Pangaro LN, Harvey J. How well do internal medicine faculty members evaluate the clinical skills of residents? *Ann Intern Med.* 1992; 117:757–65.
10. Woollicroft JO, Howell JD, Patel BP, Swanson DB. Resident–patient interactions: the humanistic qualities of internal medicine residents assessed by patients, attending physicians, program supervisors, and nurses. *Acad Med.* 1994; 69:216–24.

11. Butterfield PS, Mazzaferri EL, Sachs LA. Nurses as evaluators of the humanistic behavior of internal medicine residents. *J Med Educ.* 1987; 62:842-9.
12. Gordon MJ. A review of the validity and accuracy of self-assessments in health professions training. *Acad Med.* 1991; 66:762-9.
13. Tamblyn R, Benaroya S, Snell L, McLeod P, Schnarch B, Abrahamowicz M. The feasibility and value of using patient satisfaction ratings to evaluate internal medicine residents. *J Gen Intern Med.* 1994; 9:146-52.
14. Tamblyn R, Schnarch B, Abrahamowicz M, Colliver JA, Snell JA. Can standardized patients predict real patient satisfaction with the doctor-patient relationship? *Teach Learn Med.* 1994; 6:36-44.
15. Webster GD. Final Report on the Patient Satisfaction Questionnaire Project—Executive Summary. Philadelphia, PA: American Board of Internal Medicine, 1989, pp 2-27.
16. Borman WC. Evaluating performance effectiveness on the job: how can we generate more accurate ratings? In: Lloyd JS (ed). *Evaluation of Non-cognitive Skills and Clinical Performance.* Chicago, IL: American Board of Medical Specialties, 1982, pp 179-93.
17. Brennan RL. Performance assessments from the perspective of generalizability theory. *Appl Psych Measure.* 2000; 24:339-53.
18. Dunbar SB, Koretz DM, Hoover HD. Quality control in the development and use of performance assessment. *Appl Measure Educ.* 1991; 4:289-303.
19. Murphy KR, Cleveland JN. *Understanding Performance Appraisal: Social, Organizational, and Goal-Based Perspectives.* Thousand Oaks, CA: Sage, 1995.
20. Messick S. The once and future issues of validity: assessing the meaning and consequences of measurement. In: Weiner H, Braun MI (eds). *Test Validity.* Hillsdale, NJ: Erlbaum, 1998, pp 33-45.
21. Hopkins KD, Stanley JC, Hopkins BR. *Educational and Psychological Measurement and Evaluation (7th ed).* Englewood Cliffs, NJ: Prentice Hall, 1990.
22. Swing SR. ACGME Survey of Resident Evaluation Practices and Resources. Unpublished data, 1998.
23. Turnbull J, McFadyen J, van Barneveld C, Norman G. Clinical work sampling: a new approach to the problem of in-training evaluation. *J Gen Intern Med.* 2000; 15:556-61.
24. Ryan JG, Mandel FS, Sama A, Ward MF. Reliability of faculty clinical evaluations of non-emergency medical residents during emergency department rotations. *Acad Emerg Med.* 1996; 3:1124-30.
25. Schwartz RW, Donnelly MB, Sloan DA, Johnson SB, Strodel WE. Assessing senior residents' knowledge and performance: an integrated evaluation program. *Surgery.* 1994; 116:634-40.
26. Joorabchi B, DeVries JM. Evaluation of clinical competence: the gap between expectations and performance. *Pediatrics.* 1996; 97:179-84.
27. Haber RJ, Avins AL. Do ratings on the American Board of Internal Medicine Resident Evaluation Form detect differences in clinical competence? *J Gen Intern Med.* 1994; 9:140-5.
28. Thompson WG, Lipkin M Jr, Gilbert DA, Guzzo RA, Roberson L. Evaluating evaluation: assessment of the American Board of Internal Medicine Resident Evaluation Form. *J Gen Intern Med.* 1990; 5:214-7.
29. Maxim BR, Dielman TE. Dimensionality, internal consistency, and inter-rater reliability of clinical performance ratings. *Med Educ.* 1996; 21:130-7.
30. Winckel CP, Reznick RK, Cohen R, Taylor B. Reliability and construct validity of a structured technical skills assessment form. *Am J Surg.* 1994; 167:423-7.
31. Kroboth FJ, Hanusa BH, Parker S, et al. The inter-rater reliability and internal consistency of a clinical evaluation exercise. *J Intern Med.* 1992; 7:174-9.
32. Boon H, Stewart M. Patient-physician communication assessment instruments: 1986 to 1996 in review. *Patient Educ Counseling.* 1998; 35:161-76.
33. Miller EV. Performance checklist to evaluate anesthesia skills. In: Lloyd JS (ed). *Evaluation of Noncognitive Skills and Clinical Performance.* Chicago: American Board of Medical Specialties, 1982, pp 139-44.
34. Liu P, Miller E, Herr G, Hardy C, Sivarajan M, Willenkin R. Videotape reliability: a method of evaluation of a clinical performance examination. In: Lloyd JS, Langsley DG (eds). *How to Evaluate Residents.* Chicago: American Board of Medical Specialties, 1986, pp 275-8.
35. Tamblyn RM, Klass DJ, Schnabl GK, Kopelow ML. Sources of unreliability and bias in standardized-patient rating. *Teach Learn Med.* 1991; 3:74-85.
36. Tamblyn R, Barrows H. Data collection and interpersonal skills: the standardized patient encounter. In: Tekian A, McGuire CH, McGaghie WC (eds). *Innovative Simulations for Assessing Professional Competence.* Chicago, IL: University of Illinois at Chicago, Department of Medical Education, 1999, pp 77-112.
37. Woolliscroft JO, Stross JK, Silva J Jr. Clinical competence certification: a critical appraisal. *J Med Educ.* 1984; 59:799-805.
38. Reznick R, Regehr G, McRae H, Martin J, McCulloch W. Testing technical skill via an innovative "bench station" examination. *Am J Surg.* 1997; 173:226-30.
39. Tornow WW, London M, and CCL Associates. *Maximizing the Value of 360-Degree Feedback: A Process for Successful Individual and Organizational Development.* San Francisco: Jossey Bass, 1998.
40. Muzzin LJ, Hart L. Oral examinations. In: Neufeld V, Norman G (eds). *Assessing Clinical Competence.* New York, NY: Springer; 1985, pp 71-93.
41. Anastakis DJ, Cohne R, Reznick RK. The structured oral examination as a method for assessing surgical residents. *Am J Surg.* 1991; 162:67-70.
42. Sawa RJ. Assessing interviewing skills: the simulated office oral examination. *J Fam Pract.* 1986; 23:567-71.
43. Eagle CJ, Martineau R, Hamilton K. The oral examination in anaesthetic resident education. *Can J Anaesth.* 1993; 40:947-53.
44. Van der Vleuten CP, Swanson DB. Assessment of clinical skills with standardized patients: state of the art. *Teach Learn Med.* 1990; 2:58-76.
45. Vu NV, Barrows HS. Use of standardized patients in clinical assessment: recent developments and measurement findings. *Educ Res.* 1994; 23:23-30.
46. Holmboe ES, Hawkins RE. Methods for evaluating the clinical competence of residents in internal medicine: a review. *Ann Intern Med.* 1998; 129:42-8.
47. Sloan DA, Donnelly MB, Schwartz RW, Strodel WE. The objective structured clinical examination: the new gold standard for evaluating postgraduate clinical performance. *Ann Surg.* 1995; 6:735-42.
48. Cohen DS, Colliver JA, Marcy MS, Fried ED, Swartz MH. Psychometric properties of a standardized-patient checklist and rating-scale form used to assess interpersonal and communication skills. *Acad Med.* 1996; 71(1, Jan RIME suppl):S87-S89.
49. Regehr G, McRae H, Reznick RK, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Acad Med.* 1998; 73:993-7.
50. Isenberg SB, McGaghie WC. Assessing knowledge and skills in the health professions: a continuum of simula-

- tion fidelity. In: Tekian A, McGuire CH, McGaghie WC. (eds). *Innovative Simulations for Assessing Professional Competence*. Chicago, IL: University of Illinois at Chicago, Department of Medical Education, 1999, pp 125–46.
51. Paisley AM, Baldwin P, Paterson-Brown S. Validity of surgical simulation for the assessment of operative skill. *Br J Surg*. 2001; 88:1525–32.
 52. Chapman DM, Marx JA, Honigman B, Rosen P, Cavanaugh SH. Emergency thoracotomy: comparison of medical student, resident, and faculty performance on written, computer, and animal model assessments. *Acad Emerg Med*. 1994; 1:373–81.
 53. Friedman Ben David M, Davis MH, Harden RM, Howie PW, Ker J, Pippard MJ. AMEE Medical Education Guide No. 24: Portfolios as a method of student assessment. *Med Teach*. 2001; 23:535–51.
 54. O'Sullivan PS, Cogbill KK, McClain T, Reckase M, Clardy JA. Portfolios as a novel approach for residency evaluation. *Acad Psychiatr*. 2002; in press.