Advancing Resident Assessment in Graduate Medical Education

Susan R. Swing, PhD
Stephen G. Clyman, MD
Eric S. Holmboe, MD
Reed G. Williams, PhD
For the Accreditation Council for Graduate Medical Education Advisory Committee on Educational Outcome Assessment

Abstract

Background The Outcome Project requires high-quality assessment approaches to provide reliable and valid judgments of the attainment of competencies deemed important for physician practice.

Intervention The Accreditation Council for Graduate Medical Education (ACGME) convened the Advisory Committee on Educational Outcome Assessment in 2007–2008 to identify high-quality assessment methods. The assessments selected by this body would form a core set that could be used by all programs in a specialty to assess resident performance and enable initial steps toward establishing national specialty databases of program performance. The committee identified a small set of methods for provisional use and further evaluation. It also developed frameworks and processes to support the ongoing evaluation of methods and the longer-term enhancement of assessment in graduate medical education.

Outcome The committee constructed a set of standards, a methodology for applying the standards, and grading rules for their review of assessment method quality. It developed a simple report card for displaying grades on each standard and an overall grade for each method reviewed. It also described an assessment system of factors that influence assessment quality. The committee proposed a coordinated, national-level infrastructure to support enhancements to assessment, including method development and assessor training. It recommended the establishment of a new assessment review group to continue its work of evaluating assessment methods. The committee delivered a report summarizing its activities and 5 related recommendations for implementation to the ACGME Board in September 2008.

Introduction

The Accreditation Council for Graduate Medical Education (ACGME) convened the Advisory Committee on Educational Outcome Assessment in 2007–2008 to identify high-quality assessment methods for use in residency programs. These methods would form a core set that could be used by all programs in a specialty. Implementation of assessment methods across programs would enable establishment of national specialty databases of program performance. This, in turn, would set the stage for accomplishing the phase III and IV Outcome Project goals of using educational outcome data in accreditation and identifying benchmark programs.[1]

During the initial phases of the Outcome Project, ACGME invited programs to develop assessment methods as a way to actively engage residency educators and stimulate development of high-quality methods. This "grassroots" approach produced pockets of success but overall was hampered primarily by insufficient resources within residency programs, the extensive testing needed to establish validity, and the unavailability of clear standards for judging the quality of assessment methods.

Advisory Committee Recommendations for Advancing Assessment

The 20-member Advisory Committtee for Educational Outcome Assessment consisted of resident and practicing physicians, resident educators, program directors, designated institutional officials, educational researchers,

Susan R. Swing, PhD, is Vice President, Outcome Assessment with the Accreditation Council for Graduate Medical Education; **Stephen G. Clyman, MD,** is Executive Director, Center for Innovation with the National Board of Medical Examiners; **Eric S. Holmboe, MD,** is Senior Vice President for Quality Research and Academic Affairs with the American Board of Internal Medicine; and **Reed G. Williams, PhD,** is Professor, Department of Surgery with the Southern Illinois University School of Medicine.

This article is derived from the deliberations of the Accreditation Council for Graduate Medical Education (ACGME) Advisory Committee on Educational Outcome Assessment and thereby reflects the contributions of the committee members. Members of the Advisory Committee were as follows: Stephen G. Clyman, MD, Chair; Christopher L. Amling, MD; David Capobianco, MD; Jim Cichon, MSW; Brian Clauser, EdD; Rupa Danier, MD; Pamela L. Derstine, PhD; Paul Doughtery, MD; Lori Goodhartz, MD; Diane Hartmann, MD; Brian Hodges, MD, PhD; Eric Holmboe, MD; Sheldon Horowitz, MD; Michael Kane, PhD; Andrew Go Lee, MD; Paul V. Miles, MD; Richard Neill, MD; Rita M. Patel, MD; William Rodak, PhD; and Reed Williams, PhD. Susan Swing, PhD, was staff to the committee.

Corresponding author: Susan Swing, PhD, Accreditation Council for Graduate Medical Education (ACGME), 515 North State Street, Suite 2000, Chicago, IL 60654, 312.755.7447, srs@acgme.org

DOI: 10.4300/JGME-D-09-00010.1

psychometricians, Residency Review Committee members, staff of certification boards and medical testing organizations, and ACGME staff. The committee performed its work during a 14-month period and delivered a final report to the ACGME Board in September 2008.[2] In its report, the committee presented 5 key recommendations for enhancing assessment of residents, and the processes and frameworks it developed to support their implementation.

Recommendation 1

Standards for evaluating assessment methods should be adopted and implemented. An assessment toolbox containing methods that meet the committee's standards for methods should be established and, when sufficiently equipped, used as the source of assessment methods for residency programs.

Recommendation 2

A goal of the graduate medical education community should be the eventual adoption of a core set of specialty-appropriate assessment methods. Whenever possible, the same methods or method templates should be used across specialties. Specialties should provisionally use and evaluate promising methods if compliant methods are not available.

Recommendation 3

Assessment systems with features defined herein should be implemented within and across residency programs.

Recommendation 4

An Assessment Review Group should be established to refine recommended features for assessment systems, coordinate assessment method development, and manage assessment method review using the standards for methods.

Recommendation 5

Best-evidence guidelines for assessment method implementation and train-the-trainers approaches for assessors and feedback providers should be developed and made available to residency programs.

Furthermore, the committee developed standards, an evidence-based approach for evaluating assessment methods, a report card for displaying results, a provisional set of methods for potential use across programs, assessment system characteristics, and an infrastructure for developing assessment methods and assessor training nationally; these tools are described below.

A Methodology for Evaluating and Selecting Assessment Methods

The committee developed standards and an assessment evaluation methodology to identify high-quality assessment methods. van der Vleuten and Shuwirth's[3] utility equation (ie, utility = validity × reliability × acceptability × educational impact × cost effectiveness) provided a useful framework for guiding extension of standards beyond traditional psychometric considerations. The proposed standards are in 6 areas: reliability, validity, ease of use, resources required, ease of interpretation, and educational impact.

The standards for reliability emphasize the importance of this metric for all scores and subscores used, for interrater and intrarater reliability, and for classifying individuals consistently (eg, as pass or fail) on repeated assessments using the same method. Standards for validity were selected to concur with Messick's view that, "validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment" (emphasis in the original).[4] The standards emphasize establishing validity by providing rationales, evidence, and theory for interpretation of assessment results. The standards also include the extent of a rater's agreement with "gold standard" or consensus ratings and detection of known strengths and weaknesses of a performance.

Standards for ease of use and resources required specify limits on the time it takes to perform an assessment or train assessors, and they favor portable forms that can be completed by an individual assessor without additional resources. Standards for ease of interpretation favor methods that produce simple scores, such as percent correct, and have norms and readily accessible reports for comparing individuals to group performance. The educational impact standard specifies that methods should yield results that stimulate positive change in individual resident performance, knowledge, skills, or atttidues or the educational curriculum, or are actionable and perceived as useful. A more detailed listing of the standards is included in Table 1.

The current ACGME common program requirements state that assessment must be objective.[5] By including the standards for psychometric properties, feasibility, and usefulness, the proposed ACGME standards for assessments are more congruent with those endorsed by other bodies, such as the National Quality Forum,[6] the Joint Commission,[7] and the Postgraduate Medical Education and Training Board,[8] for performance assessments used in the oversight of health care and educational quality.

The standards for assessments represent minimal standards and are intended as guidelines to evaluate and select assessment methods. These standards are evolutionary, and the committee hopes that the proposed Assessment Review Group and others, including medical education assessment researchers, might consider and build on them.

Evidence-Based Grading of Assessment Methods

The committee created a grading scheme for evidence-based evaluation of assessment approaches against the standards for assessment methods; they also devised evidence rules for each of the 6 standards, as well as summary rules that consider the importance of the standards and quality of published evidence about each assessment method being evaluated. An example of the grading rules for educational impact is shown in TABLE 2. Grading rules for other

TABLE 1 **Overview of Standards for Evaluating the Quality of Assessment Methods**[a]

Reliability
1. Reliability indicators must be available for any total score or subscore that will be interpreted.
2. Interrater and intrarater reliability for multiple ratings of the same learner should be provided when scoring or rating entails subjective judgment.
3. For high-stakes decisions, an estimate should be provided of the percentage of learners who would be classified the same on 2 applications of the same method or rating process.
Validity
1. A rationale for each interpretation and use of evaluation results along with evidence and theory should be presented.
2. Processes and procedures used for selection of the content of assessment and for any criteria (eg, importance, frequency, and criticality) used to sample content should be described and justified when validation rests in part on the assessment content.
3. When the rationale for the use and interpretation of an assessment depends on the psychological processes or cognitive operations of the learner or the processes of the evaluator, the theoretical or empirical evidence that supports the interpretation should be provided.
4. When unintended consequences result from use of a specific assessment, an attempt should be made to identify the cause. For example: Is the assessment measuring something other than what it was intended to measure? Did the assessment fail to measure fully the intended construct?
5. The degree of agreement between a single expert rater and "gold standard" or consensus ratings for the same performance should be provided when a single rater using subjective judgments is the basis of the assessment.
6. When a single rater using subjective judgment is the basis of the assessment, the degree to which known strengths and weaknesses of the learner are detected should be provided.
Ease of use
1. The assessment tool is easily carried or accessed in the course of daily clinical or teaching activity.
2. The tool requires little special setup.
3. The tool requires less than 20 minutes for the assessor to complete.
Resources required
1. No additional resources are required beyond the documentation tools.
2. Training requirements for assessors do not exceed an hour.
3. No additional persons other than an individual assessor are required to complete the evaluation.
Ease of interpretation
1. Individual scores are interpretable—for example, on an easily understood scale, such as percent correct or against behavioral or other descriptive criteria—and are accompanied by interpretation guidelines.
2. Normative data are available consisting of: (1) a standard of care; (2) performance of other residents at the same level of training and/or experience; (3) performance of other residents with more or less experience; and (4) the resident's performance level at an earlier stage of education and experience.
3. Preprogrammed, easy-to-read reports and graphs make it simple to compare individual to group performance.
Educational impact
1. The method has been shown to positively affect individual learner performance; that is, there is a change in knowledge, skills, or attitudes.
2. The method has been shown to positively affect or change program curriculum (should be corroborated in at least 2 studies).
3. The method has been shown to provide specific actionable results that are regarded as useful by the learners.

[a] Standards for reliability and validity were derived from Standards for Educational and Psychological Testing.[9]

TABLE 2 **Example of Grading Rules for Educational Impact**

Grading for educational impact
A Meets both standards 1 and 2.
B Meets either standard 1 or 2.
C Standard 3 is met.
NI Not enough information from literature to judge
Standards for educational impact
1. The method has been shown to positively affect individual learner performance (change in knowledge, skills, or attitudes).
2. The method has been shown to positively affect or change program curriculum (should be corroborated in at least 2 studies).
3. The method has been shown to provide specific actionable results that are regarded as useful by the learners.

standards are similar because the most critical aspects of the standard, plus other supportive evidence, are required for the highest grade. TABLE 3 presents the overall summary rules, and TABLE 4 presents example grades for the mini-Clinical Evaluation Excercise in the report card format.

The summary grading approach was adapted from evidence-based medicine practices.[10] Following this approach, a grade is assigned to indicate the strength of evidence for a particular treatment based on prespecified criteria related to the rigor of the research methodology. The results of applying this approach to grading assessment methods will contribute to other ongoing efforts to strengthen the evidence base in medical education, as initiated by Best Evidence Medical Education.[11] The report card is intended as a user-friendly source of evidence to validate use of methods and guide selection of additional assessment approaches.

The committee tested aspects of its emerging evaluation framework and process in a review of selected assessment methods. Nine methods known to have a modicum of published evidence were selected for review. Individual reviewers presented their reviews to the entire committee for discussion and recommendation. Sample methods recommended for inclusion as a starter set in the new ACGME toolbox are listed in TABLE 5. The class 2 methods are recommended for dissemination and use, whereas the class 3 methods are recommended for further development and testing. No class 1 methods were identified.

The quality of assessment methods can only be determined in the context in which they are used. Thus, these methods have only demonstrated the potential for producing quality results as determined through the agreement of available evidence with the proposed standards. Optimal implementation will always include continued validity testing in context.

TABLE 3 **Summary Rules for Evidence-Based Grading of Assessment Methods**

Grading for the overall recommendation	
Class 1	The assessment method is *recommended* as a *core component* of the program's evaluation system.
Class 2	The assessment method *can be considered for use* as *one component* of the program's evaluation system.
Class 3	The assessment method *can be used provisionally* as *a component* of the program's evaluation system.
Criteria for determining level of evidence	
Level A	Published data from methodologically sound evaluation studies of the method in multiple (more than 2) settings provides strong evidence for all components of the modified utility index (reliability, validity, ease of use, resources required, ease of interpretation, and educational impact).
Level B	Published data from methodologically sound evaluation studies of the method in a minimum of 2 settings provide some evidence of acceptable reliability and some evidence of validity, ease of use, and educational impact. Acceptable evidence for ease of interpretation is available for methods used to make high-stakes decisions. Available evidence for ease of use and resources required suggests that the tool is usable by many programs.
Level C	Data from methodologically sound evaluation studies of the method provide evidence of acceptable reliability, validity, and educational impact. Little evidence is available to assist interpretation of performance. Available evidence for ease of use and resources required suggests that the tool is usable by many programs.

TABLE 4 **Assessment Tool Evaluation Framework**[A]

Mini-CEX

Description		Evaluation of Evidence for Each Standard						Summary	
Competency/Skill Domain	Method Type	Reliability	Validity	Ease of Use	Resources Required	Ease of Interpretation	Educational Impact	Evidence	Overall Recommendation
• Patient care	Direct observation and rating	B	B	A	A	C	A	B	Class 2
• Interpersonal and communication skills									
• Professionalism									

Abbreviation: Mini-CEX, Mini–Clinical Evaluation Exercise.

[a] This table illustrates the proposed display format of results for an assessment method evaluation. The As, Bs, and Cs for the 7 categories of standards under "Evaluation of Evidence for Each Standard" would be derived by comparing available evidence about the method being evaluated against the grading rules for the standard. Complete grading rules can be found in the Accreditation Council for Graduate Medical Education Advisory Committee for Educational Outcome Assessment Final Report.[2]

A Residency Program Assessment System

Methods that meet the standards can enhance assessment in residencies. The literature on performance appraisal, however, clearly shows that multiple curricular, social, and cultural aspects of the learning environment influence assessment quality.[27–34] For instance, aligning curricular elements (desired outcomes, learning opportunities, and assessment) so that all target the same competencies is an essential condition for validity.[31–34] Social and cultural factors come into play through assessors, who tend to be more lenient (and less accurate) when assessment is high stakes and may disrupt relationships with the persons assessed; when they feel that assessment is unfair or unimportant;[27–28] or when they lack competence and a sense of efficacy as assessors.[29]

Enhancing assessment, therefore, will require implementation of promising methods within a context of supportive features. The committee selected 9 contextual features from the literature[4,27–37] and organized them as an assessment system to emphasize their collective importance.

1. *Clear purpose and transparency*. This involves clear communication of the purpose, timing, and focus of assessment well before the assessment occurs. The purpose could be communicated as formative (for guiding performance improvement) or summative (for making high-stakes decisions regarding progression, promotion, and graduation).

2. *Blueprint*. Implementation will involve preparation of a blueprint that identifies the knowledge, skills, behaviors, or other outcomes that will be assessed, the learning or patient care context in which assessment will occur, and when the assessment will be done.

3. *Milestones*. Milestones describe, in behavioral terms, learning and performance levels residents are expected to demonstrate for specific competencies by a particular point in residency education.

4. *Tools and processes*. Implementation involves assessment methods that meet the standards for methods and that assess the skills, knowledge, attitudes, behaviors, and outcomes that are specified in the blueprint and milestones.

5. *Qualified assessors*. Qualified assessors are individuals who have observed resident behaviors, have expertise in the areas they are assessing and, where appropriate and feasible, receive training on the assessment methods.

6. *Assessor training*. Assessor training teaches assessors to recognize behaviors characteristic of different levels of performance and associate them with appropriate ratings, scores, and categories (eg, competent or proficient).

7. *Evaluation committee*. Implementation involves review of residents' assessment results by the program's evaluation or competency committee and joint decision making to arrive at a summary assessment.

TABLE 5 **Methods Reviewed and Classified**[a]

Method or Tool Name	Method Type	Competency Domains
Class 2		
Mini-CEX[12]	Direct observation and concurrent rating of real patient encounter	• Patient care
		• Interpersonal and communication skills
		• Professionalism
Medical Record Audit and Feedback[13,14]	National Quality Forum–approved aggregated process and outcome measures derived retrospectively from real patient records	• Patient care
		• Practice-based learning and improvement
Objective Structured Assessment of Technical Skills[15–20]	Direct observation and concurrent rating of simulated operative tasks	• Patient care (operative skills)
		• Medical knowledge
Class 3		
Operative Performance Rating System[21]	Direct observation and concurrent rating in real operative setting	• Patient care
		• Medical knowledge
		• Interpersonal and communication skills
		• Professionalism
Non-Technical Surgical Skills[22,23]	Direct observation and concurrent rating in real operative setting	• Patient care
		• Interpersonal and communication skills
		• Systems-based practice
Anesthesiology Non-Technical Skills[24]	Direct observation and concurrent rating in real operative setting	• Patient care
		• Interpersonal and communication skills
		• Systems-based practice
Communication Assessment Tool[25]	Retrospective patient ratings based on real patient encounter	• Interpersonal and communication skills
SEGUE[26]	Direct observation and concurrent use of checklist in real patient encounter	• Interpersonal and communication skills

Abbreviations: Mini-CEX, Mini-Clinical Evaluation Exercise; SEGUE, Set the stage, Elicit information, Give information, Understand the patient's perspective, End the encounter.

[a] No class 1 methods were identified.

8. *Leadership*. Implementation will involve selection of knowledgeable persons committed to high-quality assessment and capable of engendering faculty commitment to the process.

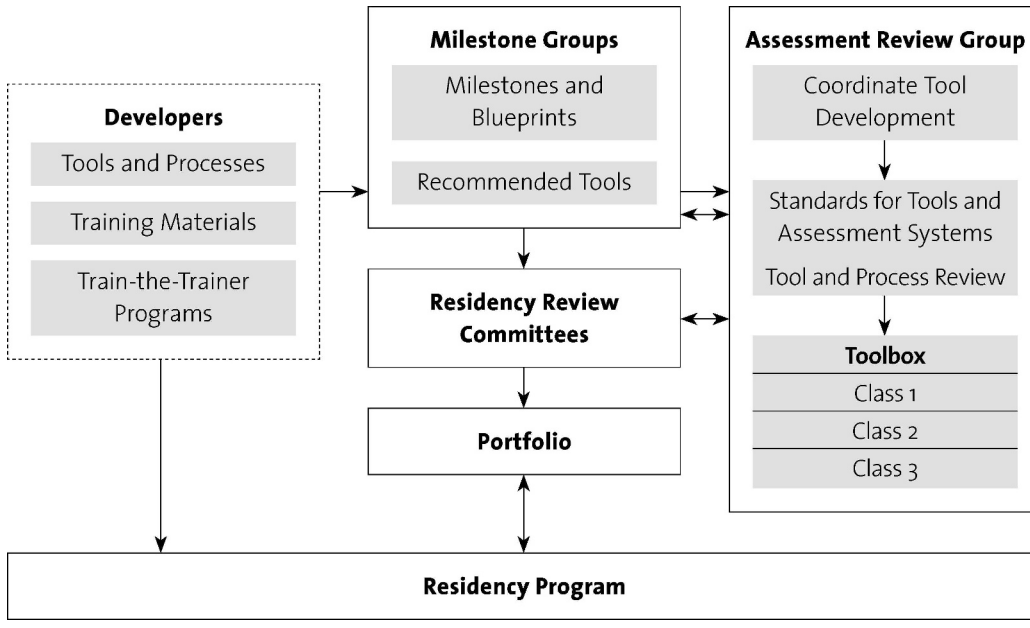
9. *Quality improvement process for assessment*. Implementation will involve periodically reviewing whether assessments are yielding high-quality results that are useful for their intended purposes.

Collectively, these features create conditions for reliable, valid, feasible, and useful assessment. A clear statement of assessment purpose, a blueprint, milestones, and assessment methods, when aligned, establish an essential condition for assessment validity.[4,31–34] A blueprint, milestones, and assessors from different professional roles and work settings facilitate obtaining a broad, representative sample of performance assessments, thereby contributing to assessment validity.[35] Selecting assessors with adequate exposure to resident performance and providing training should enhance their competence and self-efficacy, and thereby improve assessment accuracy and reliability.[28,29,38] Reliability and accuracy of assessment can increase through evaluation committee discussions and the pooling of knowledge about the person being assessed.[39,40] Last, assessment leaders can shape the culture[41] into one in which useful, improvement-oriented, fair assessment is expected and can motivate faculty and deliver needed resources.

An Infrastructure and Process for Assessment System Development

The committee envisioned that improvement to assessment would best be accomplished by a national infrastructure involving groups of experts working in a coordinated and collaborative way. This organized effort would focus on development of milestones, assessment methods, and assessor training. Methods that could be used across specialties would be priorities for development. Through

FIGURE **A Vision for Supporting Outcome Assessment**



this approach, the committee sought to address problems and recurrent issues with the current implementation approach: creation of assessment methods of unknown or variable quality and the resource waste and inefficiency caused by redundant effort.

The committee proposed that a new Assessment Review Group and external developers participate in identifying and developing high-quality assessment methods and assessor training. This work would complement that of the Use Milestone Project groups being convened by certification boards and the ACGME to establish performance-level expectations and identify core assessment methods.[42] An assessment infrastructure including these groups is illustrated in the FIGURE. Functions of the Assessment Review Group and external developers are described below.

1. *Assessment Review Group*. This group will refine the standards for assessment methods, identify assessment method gaps in conjunction with the Use Milestone Project groups and Residency Review Committees, oversee the review of candidate methods for the ACGME toolbox, and facilitate method development through communication with external developers.

2. *External developers*. Professional medical organizations, medical educator collaboratives, or individual medical educators with appropriate expertise could function as developers. Ideally, assessment method developers will create and thoroughly field test high-priority methods and then submit their method and evidence to the Assessment Review Group for evaluation. Assessor training will be developed (as appropriate) for methods adopted for specialty-wide use.

Discussion

The committee's recommendations are designed to increase the quality and value of assessment and to relieve programs of the work of developing new methods. Common tools will make possible the creation of national databases of assessment results and specialty norms. Program directors can use the norms to better interpret resident and program performance scores; Residency Review Committees can use them to better gauge program performance.

However, appropriate caution is urged in using aggregated assessment of individual residents for comparing programs. Factors related to the local context and implementation processes can limit the accuracy and comparability of assessment results. Furthermore, Residency Review Committee collection of performance data for high-stakes use could adversely affect the accuracy of resident assessment and the usefulness of the results for formative purposes within the program. Accreditation review strategies for addressing this potential problem need to be devised and monitored.

Selection and use of assessment methods that meet standards could improve the overall quality of assessment in residencies. Evidence derived from field testing of assessment methods is key to this process. Expansion of field testing and accumulation and publication of evidence will be needed for standards to be applied.

National-level development of assessment methods and assessor training is intended to decrease cost and burden for residency programs. This effort could be hampered if sufficient resources are not allocated and targeted. It will be important to note residency programs' time, effort, and other expenditures on improving assessment, and to ensure that added effort is repaid with information and useful processes for improving resident performance and educational quality.

The move toward a common set of assessment methods should proceed by taking into account the variability of expertise and resources among programs and the substantive differences in competencies required for practice across the specialties. Additional deliberations should discuss how to accommodate and encourage programs in the development and use of more innovative and resource-intensive approaches, such as simulations (in situ or within centers), given the recommendations and standards for use of a core set of methods that require minimum resources.

Assessment is one aspect of improving residents' training and program evaluation. However, it can serve as a key facilitator when integrated judiciously into educational culture and patient care. This integration will require the commitment of faculty, institutions, and communities of practice. It will also require a coordinated national effort that forges efficiencies through use of shared methods among programs and specialties while being attentive to divergent needs and constraints. The use of a small core set of high-quality methods meeting defined standards of excellence, and the appropriate use of the assessment information that flows from them, will do much to ensure that medical training in the United States remains a model for worldwide emulation. The general competencies themselves were an important first step and cornerstone for defining the corpus of professional responsibility for physicians. Assessment is the means by which the attainment of these professional ideals can be assured.

References

1 Accreditation Council for Graduate Medical Education. Accreditation Council for Graduate Medical Education Outcome Project: Timeline–working guidelines. Available at: http://www.acgme.org/outcome/project/timeline/TIMELINE_index_frame.htm. Accessed March 23, 2009.

2 Accreditation Council for Graduate Medical Education Advisory Committee on Educational Outcome Assessment. Final Report: Advancing resident assessment in graduate medical education. September 2008. Unpublished report.

3 van der Vleuten CPM, Shuwirth LWT. Assessing professional competence: from methods to programmes. *Med Educ*. 2005;39:309–317.

4 Kane, MT. Current concerns in validity theory. *J Educ Meas*. 2001;38:319–342.

5 Accreditation Council for Graduate Medical Education. General competency and assessment common program requirements. Available at: http://www.acgme.org/outcome/comp/compCPRL.asp. Accessed August 18, 2008.

6 National Quality Forum. Measure evaluation criteria. August 2008. Available at: http://www.qualityforum.org/uploadedFiles/Quality_Forum/Measuring_Performance/Consensus_Development_Process%E2%80%99s_Principle/EvalCriteria2008-08-28Final.pdf?n=4701. Accessed March 24, 2009.

7 Joint Commission. Attributes of core performance measures and associated evaluation criteria. Available at: http://www.jointcommission.org/NR/rdonlyres/7DF24897-A700-4013-A0BD-154881FB2321/0/AttributesofCorePerformanceMeasuresandAssociatedEvaluationCriteria.pdf. Accessed March 24, 2009.

8 Postgraduate Medical Education and Training Board. Standards for curricula and assessment systems. July 2008. Available at: http://www.pmetb.org.uk/fileadmin/user/Standards_Requirements/PMETB_Scas_July2008_Final.pdf. Accessed March 24, 2009.

9 American Psychological Association, American Educational Research Association, and National Council for Measurement in Education. *The Standards for Educational and Psychological Testing*. Washington, DC: AERA Publishing; 1999.

10 Cochrane Consumer Network. Cochrane and systematic reviews. Available at: http://www.cochrane.org/consumers/sysrev.htm#levels. Accessed June 15, 2009.

11 Association for Medical Education. Best Evidence Medical Education (BEME). Available at: http://www.bemecollaboration.org/beme/pages/index.html. Accessed April 30, 2009.

12 Norcini JJ, Blank LL, Arnold GK, Kimball HR. The mini-CEX (clinical evaluation exercise): a preliminary investigation. *Ann Intern Med*. 1995;123:795–799.

13 Veloski JJ, Boex JR, Grasberger MJ, Evans A, Wolfson DB. Systematic review of the literature on assessment, feedback and physicians' clinical performance: BEME Guide No. 7. *Med Teach*. 2006;28:117–128.

14 Boonyasai RT, Windish DM, Chakraborti C, Feldman LS, Rubin HR, Bass EB. Effectiveness of teaching QI to physicians. *JAMA*. 2007;298:1023–1037.

15 Reznick R, Regehr G, Martin J, McCullogh W. Testing technical skill via an innovative "bench station" examination. *Am J Surg*. 1997;173:226–230.

16 Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg*. 1997;84:273–278.

17 Szaly D, MacRae H, Regehr G, Reznick R. Using operative outcome to assess technical skill. *Am J Surg*. 2000;180:235–237.

18 Ault G, Reznick R, MacRae H, et al. Exporting a technical skills evaluation technology to other sites. *Am J Surg*. 2001;182:254–256.

19 Goff BA, Lentz GM, Lee D, Hournard B, Mandel LS. Development of an objective structured assessment of technical skills for obstetric and gynecology residents. *Obstet Gynecol*. 2000;96:146–150.

20 Goff BA, Lentz, GM, Lee D, Fenner D. Morris J, Mandel LS. Development of a bench station objective structured assessment of technical skills. *Obstet Gynecol*. 2001;98:412–416.

21 Larson JL, Williams RG, Ketchum J, Boehler ML, Dunnington GL. Feasibility, reliability and validity of an operative performance rating system for evaluating surgical residents. *Surgery*. 2005;138:640–647.

22 Yule S, Flin R, Maran N, Rowley D, Youngson G, Paterson-Brown S. Surgeons' non-technical skills in the operating room: reliability testing of the NOTSS Behavior Rating System. *World J Surg*. 2008;32:548–556.

23 Yule S, Flin R, Paterson-Brown S, Maran N, Rowley D. Development of a rating system for surgeons' non-technical skills. *Med Educ*. 2006;40:1098–1104.

24 Fletcher G, Flin R, McGeorge P, Glavin R, Maran N, Patey R. Anaesthetists' non-technical skills (ANTS): evaluation of a behavioral marker system. *Br J Anaesth*. 2003;90:580–588.

25 Makoul G, Krupat E, Chang CH. Measuring patient views of physician communication skills: development and testing of the Communication Assessment Tool. *Patient Educ Couns*. 2007;67:333–342.

26 Makoul G. The SEGUE Framework for teaching and assessing communication skills. *Patient Educ Couns*. 2001;45:23–34.

27 Murphy KR, Cleveland JN. *Understanding Performance Appraisal: Social, Organizational, and Goal-Based Principles*. Thousand Oaks, CA: SAGE Publications; 1995.

28 Levy PE, Williams JR. The social context of performance appraisal: a review and framework for the future. *J Manag*. 2004;30:881–905.

29 Bernadin HJ, Villanova P. Research streams in rater self-efficacy. *Group Org Manag*. 2005;30:61–88.

30 Williams, RG, Klamen DA, McGaghie WC. Cognitive, social, and environmental sources of bias in clinical performance ratings. *Teach Learn Med*. 2003;15:270–292.

31 Wiggins G. *Educative Assessment: Designing Assessments to Inform and Improve Student Performance*. San Francisco, CA: Jossey-Bass; 1998.

32 Harden RM, Crosby JR, Davis MH. AMEE Guide No. 14: outcome-based education: part 1–an introduction to outcome-based education. *Med Teach*. 1999;21:7–13.

33 La Marca PM. Alignment of standards and assessment as an accountability criterion. *Practical Research, Assessment & Evaluation*. 2001;7. Available at: http://pareonline.net/getvn.asp?v=7&n=21. Accessed April 30, 2009.

34 Rothman R, Slattery JB, Vranek JL, Resnick LB. Benchmarking and alignment of standards and testing. Center for Study of Evaluation Technical Report 566. May 2002. Available at: http://www.cse.ucla.edu/products/Reports/TR566.pdf. Accessed April 30, 2009.

35 Accreditation Council for Graduate Medical Education Outcome Project Advisory Group. Model assessment systems for evaluating residents and residency programs. 2000. Unpublished report.

36 Swing SR. Assessing the ACGME general competencies: general considerations and assessment methods. *Acad Emerg Med*. 2002;9:1278–1288.

37 Holmboe ES, Rodak W, Mills G, McFarlane MJ, Schultz HJ. Outcomes-based evaluation in resident education: creating systems and structured portfolios. *Am J Med*. 2006;119:708–714.

38 Miller MD, Linn RL. Validation of performance-based assessments. *Appl Psychol Meas*. 2000;24:367–378.

39 Roch SG. Why convene rater teams: an investigation of the benefits of anticipated discussion, consensus, and rater motivation. *Organ Behav Hum Decis Process*. 2007;104:14–29.

40 Williams RG, Schwind CJ, Dunnington GL, Fortune J, Rogers DA, Boehler ML. The effects of group dynamics on resident progress committee deliberations. *Teach Learn Med*. 2005;17:96–100.

41 Schein E. *Organizational Leadership and Culture*. 3rd ed. San Francisco, CA: Jossey-Bass; 2004.

42 Nasca TJ. The CEO's first column–the next step in the outcomes-based accreditation project. *ACGME Bulletin*. May 2008:2–4.