REVIEW ARTICLE

MEDICAL EDUCATION

Malcolm Cox, M.D., and David M. Irby, Ph.D., Editors

Assessment in Medical Education

Ronald M. Epstein, M.D.

As an attending physician working with a student for a week, you receive a form that asks you to evaluate the student's fund of knowledge, procedural skills, professionalism, interest in learning, and "systems-based practice." You wonder which of these attributes you can reliably assess and how the data you provide will be used to further the student's education. You also wonder whether other tests of knowledge and competence that students must undergo before they enter practice are equally problematic.

From the Departments of Family Medicine, Psychiatry, and Oncology and the Rochester Center to Improve Communication in Health Care, University of Rochester School of Medicine and Dentistry, Rochester, NY. Address reprint requests to Dr. Epstein at 1381 South Ave., Rochester, NY 14620, or at ronald_epstein@urmc.rochester.edu.

N Engl J Med 2007;356:387-96.

In one way or another, most practicing physicians are involved in assessing the competence of trainees, peers, and other health professionals. As the example above suggests, however, they may not be as comfortable using educational assessment tools as they are using more clinically focused diagnostic tests. This article provides a conceptual framework for and a brief update on commonly used and emerging methods of assessment, discusses the strengths and limitations of each method, and identifies several challenges in the assessment of physicians' professional competence and performance.

COMPETENCE AND PERFORMANCE

Elsewhere, Hundert and I have defined competence in medicine as "the habitual and judicious use of communication, knowledge, technical skills, clinical reasoning, emotions, values, and reflection in daily practice for the benefit of the individuals and communities being served."[1] In the United States, the assessment of medical residents, and increasingly of medical students, is largely based on a model that was developed by the Accreditation Council for Graduate Medical Education (ACGME). This model uses six interrelated domains of competence: medical knowledge, patient care, professionalism, communication and interpersonal skills, practice-based learning and improvement, and systems-based practice.[2]

Competence is not an achievement but rather a habit of lifelong learning[3]; assessment plays an integral role in helping physicians identify and respond to their own learning needs. Ideally, the assessment of competence (what the student or physician is able to do) should provide insight into actual performance (what he or she does habitually when not observed), as well as the capacity to adapt to change, find and generate new knowledge, and improve overall performance.[4]

Competence is contextual, reflecting the relationship between a person's abilities and the tasks he or she is required to perform in a particular situation in the real world.[5] Common contextual factors include the practice setting, the local prevalence of disease, the nature of the patient's presenting symptoms, the patient's educational level, and other demographic characteristics of the patient and of the physician. Many aspects of competence, such as history taking and clinical reasoning, are also content-specific and not necessarily generalizable to all situations. A student's

clinical reasoning may appear to be competent in areas in which his or her base of knowledge is well organized and accessible[6] but may appear to be much less competent in unfamiliar territory.[7] However, some important skills (e.g., the ability to form therapeutic relationships) may be less dependent on content.[8]

Competence is also developmental. Habits of mind and behavior and practical wisdom are gained through deliberate practice[9] and reflection on experience.[10-14] Students begin their training at a novice level, using abstract, rule-based formulas that are removed from actual practice. At higher levels, students apply these rules differentially to specific situations. During residency, trainees make judgments that reflect a holistic view of a situation and eventually take diagnostic shortcuts based on a deeper understanding of underlying principles. Experts are able to make rapid, context-based judgments in ambiguous real-life situations and have sufficient awareness of their own cognitive processes to articulate and explain how they recognize situations in which deliberation is essential. Development of competence in different contexts and content areas may proceed at different rates. Context and developmental level also interact. Although all clinicians may perform at a lower level of competence when they are tired, distracted, or annoyed, the competence of less experienced clinicians may be particularly susceptible to the influence of stress.[15,16]

GOALS OF ASSESSMENT

Over the past decade, medical schools, postgraduate training programs, and licensing bodies have made new efforts to provide accurate, reliable, and timely assessments of the competence of trainees and practicing physicians.[1,2,17] Such assessments have three main goals: to optimize the capabilities of all learners and practitioners by providing motivation and direction for future learning, to protect the public by identifying incompetent physicians, and to provide a basis for choosing applicants for advanced training.

Assessment can be formative (guiding future learning, providing reassurance, promoting reflection, and shaping values) or summative (making an overall judgment about competence, fitness to practice, or qualification for advancement to higher levels of responsibility). Formative assessments provide benchmarks to orient the learner who is approaching a relatively unstructured body of knowledge. They can reinforce students' intrinsic motivation to learn and inspire them to set higher standards for themselves.[18] Although summative assessments are intended to provide professional self-regulation and accountability, they may also act as a barrier to further practice or training.[19] A distinction should be made between assessments that are suitable only for formative use and those that have sufficient psychometric rigor for summative use. This distinction is especially important in selecting a method of evaluating competence for high-stakes assessments (i.e., licensing and certification examinations). Correspondingly, summative assessments may not provide sufficient feedback to drive learning.[20] However, because students tend to study that which they expect to be tested on, summative assessment may influence learning even in the absence of feedback.

ASSESSMENT METHODS

All methods of assessment have strengths and intrinsic flaws (Table 1). The use of multiple observations and several different assessment methods over time can partially compensate for flaws in any one method.[1,21] Van der Vleuten[22] describes five criteria for determining the usefulness of a particular method of assessment: reliability (the degree to which the measurement is accurate and reproducible), validity (whether the assessment measures what it claims to measure), impact on future learning and practice, acceptability to learners and faculty, and costs (to the individual trainee, the institution, and society at large).

WRITTEN EXAMINATIONS

Written examination questions are typically classified according to whether they are open-ended or multiple choice. In addition, questions can be "context rich" or "context poor."[23] Questions with rich descriptions of the clinical context invite the more complex cognitive processes that are characteristic of clinical practice.[24] Conversely, context-poor questions can test basic factual knowledge but not its transferability to real clinical problems.

Multiple-choice questions are commonly used for assessment because they can provide a large number of examination items that encompass many content areas, can be administered in a relatively short period, and can be graded by com-

puter. These factors make the administration of the examination to large numbers of trainees straightforward and standardized.[25] Formats that ask the student to choose the best answer from a list of possible answers are most commonly used. However, newer formats may better assess processes of diagnostic reasoning. Key-feature items focus on critical decisions in particular clinical cases.[26] Script-concordance items present a situation (e.g., vaginal discharge in a patient), add a piece of information (dysuria), and ask the examinee to assess the degree to which this new information increases or decreases the probability of a particular outcome (acute salpingitis due to *Chlamydia trachomatis*).[27] Because the situations portrayed are ambiguous, script-concordance items may provide insight into clinical judgment in the real world. Answers to such items have been shown to correlate with the examinee's level of training and to predict future performance on oral examinations of clinical reasoning.[28]

Multiple-choice questions that are rich in context are difficult to write, and those who write them tend to avoid topics — such as ethical dilemmas or cultural ambiguities — that cannot be asked about easily.[29] Multiple-choice questions may also create situations in which an examinee can answer a question by recognizing the correct option, but could not have answered it in the absence of options.[23,30] This effect, called cueing, is especially problematic when diagnostic reasoning is being assessed, because premature closure — arriving at a decision before the correct diagnosis has been considered — is a common reason for diagnostic errors in clinical practice.[31,32] Extended matching items (several questions, all with the same long list of possible answers), as well as open-ended short-answer questions, can minimize cueing.[23] Structured essays also preclude cueing. In addition, they involve more complex cognitive processes and allow for more contextualized answers than do multiple-choice questions. When clear grading guidelines are in place, structured essays can be psychometrically robust.

ASSESSMENTS BY SUPERVISING CLINICIANS

Supervising clinicians' observations and impressions of students over a specific period remain the most common tool used to evaluate performance with patients. Students and residents most commonly receive global ratings at the end of a rotation, with comments from a variety of supervising physicians. Although subjectivity can be a problem in the absence of clearly articulated standards, a more important issue is that direct observation of trainees while they are interacting with patients is too infrequent.[33]

DIRECT OBSERVATION OR VIDEO REVIEW

The "long case"[34] and the "mini–clinical-evaluation exercise" (mini-CEX)[35] have been developed so that learners will be directly observed more frequently. In these assessments, a supervising physician observes while a trainee performs a focused history taking and physical examination over a period of 10 to 20 minutes. The trainee then presents a diagnosis and a treatment plan, and the faculty member rates the resident and may provide educational feedback. Structured exercises with actual patients under the observation of the supervising physician can have the same level of reliability as structured examinations using standardized patients[34,36] yet encompass a wider range of problems, physical findings, and clinical settings. Direct observation of trainees in clinical settings can be coupled with exercises that trainees perform after their encounters with patients, such as oral case presentations, written exercises that assess clinical reasoning, and literature searches.[8,37] In addition, review of videos of encounters with patients offers a powerful means of evaluating and providing feedback on trainees' skills in clinical interactions.[8,38]

CLINICAL SIMULATIONS

Standardized patients — actors who are trained to portray patients consistently on repeated occasions — are often incorporated into objective structured clinical examinations (OSCEs), which consist of a series of timed "stations," each one focused on a different task. Since 2004, these examinations have been part of the U.S. Medical Licensing Examination that all senior medical students take.[39] The observing faculty member or the standardized patient uses either a checklist of specific behaviors or a global rating form to evaluate the student's performance.[40] The checklist might include items such as "asked if the patient smoked" and "checked ankle reflexes." The global rating form might ask for a rating of how well the visit was organized and whether the student was appropriately empathetic. A minimum of 10 stations, which the student usually visits over the course of 3 to 4 hours, is necessary to achieve

Table 1. Commonly Used Methods of Assessment.

Method	Domain	Type of Use	Limitations	Strengths
Written exercises				
Multiple-choice questions in either single-best-answer or extended matching format	Knowledge, ability to solve problems	Summative assessments within courses or clerkships; national in-service, licensing, and certification examinations	Difficult to write, especially in certain content areas; can result in cueing; can seem artificial and removed from real situations	Can assess many content areas in relatively little time, have high reliability, can be graded by computer
Key-feature and script-concordance questions	Clinical reasoning, problem-solving ability, ability to apply knowledge	National licensing and certification examinations	Not yet proven to transfer to real-life situations that require clinical reasoning	Assess clinical problem-solving ability, avoid cueing, can be graded by computer
Short-answer questions	Ability to interpret diagnostic tests, problem-solving ability, clinical reasoning skills	Summative and formative assessments in courses and clerkships	Reliability dependent on training of graders	Avoid cueing, assess interpretation and problem-solving ability
Structured essays	Synthesis of information, interpretation of medical literature	Preclinical courses, limited use in clerkships	Time-consuming to grade, must work to establish interrater reliability, long testing time required to encompass a variety of domains	Avoid cueing, use higher-order cognitive processes
Assessments by supervising clinicians				
Global ratings with comments at end of rotation	Clinical skills, communication, teamwork, presentation skills, organization, work habits	Global summative and sometimes formative assessments in clinical rotations	Often based on second-hand reports and case presentations rather than on direct observation, subjective	Use of multiple independent raters can overcome some variability due to subjectivity
Structured direct observation with checklists for ratings (e.g., mini–clinical-evaluation exercise or video review)	Communication skills, clinical skills	Limited use in clerkships and residencies, a few board-certification examinations	Selective rather than habitual behaviors observed, relatively time-consuming	Feedback provided by credible experts
Oral examinations	Knowledge, clinical reasoning	Limited use in clerkships and comprehensive medical school assessments, some board-certification examinations	Subjective, sex and race bias has been reported, time-consuming, require training of examiners, summative assessments need two or more examiners	Feedback provided by credible experts

Clinical simulations				
Standardized patients and objective structured clinical examinations	Some clinical skills, interpersonal behavior, communication skills	Formative and summative assessments in courses, clerkships, medical schools, national licensure examinations, board certification in Canada	Timing and setting may seem artificial, require suspension of disbelief, checklists may penalize examinees who use shortcuts, expensive	Tailored to educational goals; reliable, consistent case presentation and ratings; can be observed by faculty or standardized patients; realistic
Incognito standardized patients	Actual practice habits	Primarily used in research; some courses, clerkships, and residencies use for formative feedback	Requires prior consent, logistically challenging, expensive	Very realistic, most accurate way of assessing clinician's behavior
High-technology simulations	Procedural skills, teamwork, simulated clinical dilemmas	Formative and some summative assessment	Timing and setting may seem artificial, require suspension of disbelief, checklists may penalize examinees who use shortcuts, expensive	Tailored to educational goals, can be observed by faculty, often realistic and credible
Multisource ("360-degree") assessments				
Peer assessments	Professional demeanor, work habits, interpersonal behavior, teamwork	Formative feedback in courses and comprehensive medical school assessments, formative assessment for board recertification	Confidentiality, anonymity, and trainee buy-in essential	Ratings encompass habitual behaviors, credible source, correlates with future academic and clinical performance
Patient assessments	Ability to gain patients' trust; patient satisfaction, communication skills	Formative and summative, board recertification, use by insurers to determine bonuses	Provide global impressions rather than analysis of specific behaviors, ratings generally high with little variability	Credible source of assessment
Self-assessments	Knowledge, skills, attitudes, beliefs, behaviors	Formative	Do not accurately describe actual behavior unless training and feedback provided	Foster reflection and development of learning plans
Portfolios	All aspects of competence, especially appropriate for practice-based learning and improvement and systems-based practice	Formative and summative uses across curriculum and within clerkships and residency programs, used by some U.K. medical schools and specialty boards	Learner selects best case material, time-consuming to prepare and review	Display projects for review, foster reflection and development of learning plans

a reliability of 0.85 to 0.90.[41] Under these conditions, structured assessments with the use of standardized patients are as reliable as ratings of directly observed encounters with real patients and take about the same amount of time.[42]

Interactions with standardized patients can be tailored to meet specific educational goals, and the actors who portray the patients can reliably rate students' performance with respect to history taking and physical examinations. Faculty members who observe encounters with standardized patients can offer additional insights on trainees' clinical judgment and the overall coherence of the history taking or physical examination. Unannounced standardized patients, who with the examinees' prior approval present incognito in actual clinical settings, have been used in health services research to evaluate examinees' diagnostic reasoning, treatment decisions, and communication skills.[43-46] The use of unannounced standardized patients may prove to be particularly valuable in the assessment of higher-level trainees and physicians in practice.

The use of simulation to assess trainees' clinical skills in intensive care and surgical settings is on the rise.[47] Simulations involving sophisticated mannequins with heart sounds, respirations, oximeter readings, and pulses that respond to a variety of interventions can be used to assess how individuals or teams manage unstable vital signs. Surgical simulation centers now routinely use high-fidelity computer graphics and hands-on manipulation of surgical instruments to create a multisensory environment. High-technology simulation is seen increasingly as an important learning aid and may prove to be useful in the assessment of knowledge, clinical reasoning, and teamwork.

Multisource ("360-Degree") Assessments

Assessments by peers, other members of the clinical team, and patients can provide insight into trainees' work habits, capacity for teamwork, and interpersonal sensitivity.[48-50] Although there are few published data on outcomes of multisource feedback in medical settings, several large programs are being developed, including one for all first- and second-year house officers in the United Kingdom and another for all physicians undergoing recertification in internal medicine in the United States. Multisource feedback is most effective when it includes narrative comments as well as statistical data, when the sources are recognized as credible, when the feedback is framed constructively, and when the entire process is accompanied by good mentoring and follow-up.[51]

Recent studies of peer assessments suggest that when trainees receive thoughtful ratings and comments by peers in a timely and confidential manner, along with support from advisers to help them reflect on the reports, they find the process powerful, insightful, and instructive.[51,52] Peer assessments have been shown to be consistent regardless of the way the raters are selected. Such assessments are stable from year to year[53] and predict subsequent class rankings as well as subsequent ratings by supervisors.[54] Peer assessments depend on trust and require scrupulous attention to confidentiality. Otherwise they can be undermining, destructive, and divisive.

Although patients' ratings of clinical performance are valuable in principle, they pose several problems. As many as 50 patient surveys may be necessary to achieve satisfactory reliability.[55] Patients who are seriously ill often do not complete surveys; those who do tend to rate physicians less favorably than do patients who have milder conditions.[56] Furthermore, patients are not always able to discriminate among the elements of clinical practice,[57] and their ratings are typically high. These limitations make it difficult to use patient reports as the only tool for assessing clinical performance. However, ratings by nurses can be valuable. Such ratings have been found to be reliable with as few as 6 to 10 reports,[58] and they correlate with both patients' and faculty members' ratings of the interpersonal aspects of trainees' performance.[59]

Fundamental cognitive limitations in the ability of humans to know themselves as others see them restrict the usefulness of self-assessment. Furthermore, rating oneself on prior clinical performance may not achieve another important goal of self-assessment: the ability to monitor oneself from moment to moment during clinical practice.[10,60] A physician must possess this ability in order to meet patients' changing needs, to recognize the limits of his or her own competence, and to manage unexpected situations.

Portfolios

Portfolios include documentation of and reflection about specific areas of a trainee's competence. This evidence is combined with self-reflection.[61]

In medicine, just as in the visual arts, portfolios demonstrate a trainee's development and technical capacity. They can include chart notes, referral letters, procedure logs, videotaped consultations, peer assessments, patient surveys, literature searches, quality-improvement projects, and any other type of learning material. Portfolios also frequently include self-assessments, learning plans, and reflective essays. For portfolios to be maximally effective, close mentoring is required in the assembly and interpretation of the contents; considerable time can be expended in this effort. Portfolios are most commonly used in formative assessments, but their use for summative evaluations and high-stakes decisions about advancement is increasing.[20]

CHALLENGES IN ASSESSMENT

NEW DOMAINS OF ASSESSMENT

There are several domains in which assessment is in its infancy and remains problematic. Quality of care and patient safety depend on effective teamwork,[62] and teamwork training is emphasized as an essential element of several areas of competence specified by the ACGME, yet there is no validated method of assessing teamwork. Experts do not agree on how to define professionalism — let alone how best to measure it.[63] Dozens of scales that rate communication are used in medical education and research,[64] yet there is little evidence that any one scale is better than another; furthermore, the experiences that patients report often differ considerably from ratings given by experts.[65]

MULTIMETHOD AND LONGITUDINAL ASSESSMENT

The use of multiple methods of assessment can overcome many of the limitations of individual assessment formats.[8,22,36,66] Variation of the clinical context allows for broader insights into competence, the use of multiple formats provides greater variety in the areas of content that are evaluated, and input from multiple observers provides information on distinct aspects of a trainee's performance. Longitudinal assessment avoids excessive testing at any one point in time and serves as the foundation for monitoring ongoing professional development.

In the example at the beginning of this article, a multimethod assessment might include direct observation of the student interacting with several patients at different points during the rotation, a multiple-choice examination with both "key features" and "script-concordance" items to assess clinical reasoning, an encounter with a standardized patient followed by an oral examination to assess clinical skills in a standardized setting, written essays that would require literature searches and synthesis of the medical literature on the basic science or clinical aspects of one or more of the diseases the student encountered, and peer assessments to provide insights into interpersonal skills and work habits.

The combination of all these results into a portfolio resembles the art of diagnosis; it demands that the student synthesize various bits and types of information in order to come up with an overall picture. Although a few medical schools have begun to institute longitudinal assessments that use multiple methods,[8] the best way to deal with the quantity and the qualitatively different types of data that the process generates is not yet clear. New ways of combining qualitative and quantitative data will be required if portfolio assessments are to find widespread application and withstand the test of time.

STANDARDIZATION OF ASSESSMENT

Although accrediting organizations specify broad areas that the curriculum should cover and assess, for the most part individual medical schools make their own decisions about methods and standards of assessment. This model may have the advantage of ensuring consistency between the curriculum and assessment, but it also makes it difficult to compare students across medical schools for the purpose of subsequent training.[67] The ideal balance between nationally standardized and school-specific assessment remains to be determined. Furthermore, within a given medical school, all students may not require the same package of assessments — for example, initial screening examinations may be followed by more extensive testing for those who have difficulties.

ASSESSMENT AND LEARNING

It is generally acknowledged that assessment drives learning; however, assessment can have both intended and unintended consequences.[22] Students study more thoughtfully when they anticipate certain examination formats,[68] and changes in the format can shift their focus to clinical rather than theoretical issues.[69] Assessment by peers seems

to promote professionalism, teamwork, and communication.[52] The unintended effects of assessment include the tendency for students to cram for examinations and to substitute superficial knowledge for reflective learning.

ASSESSMENT OF EXPERTISE

The assessment of trainees and physicians who have higher levels of expertise presents particular challenges. Expertise is characterized by unique, elaborated, and well-organized bodies of knowledge that are often revealed only when they are triggered by characteristic clinical patterns.[70,71] Thus, experts who are unable to access their knowledge in artificial testing situations but who make sound judgments in practice may do poorly on some tests that are designed to assess communication skills, knowledge, or reasoning. Furthermore, clinical expertise implies the practical wisdom to manage ambiguous and unstructured problems, balance competing explanations, avoid premature closure, note exceptions to rules and principles, and — even when under stress — choose one of the several courses of action that are acceptable but imperfect. Testing either inductive thinking (the organization of data to generate possible interpretations) or deductive thinking (the analysis of data to discern among possibilities) in situations in which there is no consensus on a single correct answer presents formidable psychometric challenges.

Table 2. Principles of Assessment.

Goals of assessment
Provide direction and motivation for future learning, including knowledge, skills, and professionalism
Protect the public by upholding high professional standards and screening out trainees and physicians who are incompetent
Meet public expectations of self-regulation
Choose among applicants for advanced training
What to assess
Habits of mind and behavior
Acquisition and application of knowledge and skills
Communication
Professionalism
Clinical reasoning and judgment in uncertain situations
Teamwork
Practice-based learning and improvement
Systems-based practice
How to assess
Use multiple methods and a variety of environments and contexts to capture different aspects of performance
Organize assessments into repeated, ongoing, contextual, and developmental programs
Balance the use of complex, ambiguous real-life situations requiring reasoning and judgment with structured, simplified, and focused assessments of knowledge, skills, and behavior
Include directly observed behavior
Use experts to test expert judgment
Use pass–fail standards that reflect appropriate developmental levels
Provide timely feedback and mentoring
Cautions
Be aware of the unintended effects of testing
Avoid punishing expert physicians who use shortcuts
Do not assume that quantitative data are more reliable, valid, or useful than qualitative data

ASSESSMENT AND FUTURE PERFORMANCE

The evidence that assessment protects the public from poor-quality care is both indirect and scarce; it consists of a few studies that show correlations between assessment programs that use multiple methods and relatively crude estimates of quality such as diagnostic testing, prescribing, and referral patterns.[72] Correlating assessment with future performance is difficult not only because of inadequacies in the assessment process itself but also because relevant, robust measures of outcome that can be directly attributed to the effects of training have not been defined. Current efforts to measure the overall quality of care include patient surveys and analyses of institutional and practice databases. When these new tools are refined, they may provide a more solid foundation for research on educational outcomes.

CONCLUSIONS

Considering all these challenges, current assessment practices would be enhanced if the principles summarized in Table 2 were kept clearly in mind. The content, format, and frequency of assessment, as well as the timing and format of feedback, should follow from the specific goals of the medical education program. The various domains of competence should be assessed in an integrated, coherent, and longitudinal fashion with the use of multiple methods and provision of frequent and constructive feedback. Educators should be mindful of the impact of assessment on learning, the potential unintended effects of assessment, the limitations of each method (including cost), and the prevailing culture of the program or institution in which the assessment is occurring.

Assessment is entering every phase of professional development. It is now used during the medical school application process,[73] at the start of residency training,[74] and as part of the "main-

tenance of certification" requirements that several medical boards have adopted.[75] Multiple methods of assessment implemented longitudinally can provide the data that are needed to assess trainees' learning needs and to identify and remediate suboptimal performance by clinicians. Decisions about whether to use formative or summative assessment formats, how frequently assessments should be made, and what standards should be in place remain challenging. Educators also face the challenge of developing tools for the assessment of qualities such as professionalism, teamwork, and expertise that have been difficult to define and quantify.

No potential conflict of interest relevant to this article was reported.

I thank Tana Grady-Weliky, M.D., Stephen Lurie, M.D., Ph.D., John McCarthy, M.S., Anne Nofziger, M.D., and Denham Ward, M.D., Ph.D., for helpful suggestions.

A video showing an examination involving a standardized patient is available with the full text of this article at www.nejm.org.

REFERENCES

1. Epstein RM, Hundert EM. Defining and assessing professional competence. JAMA 2002;287:226-35.
2. Batalden P, Leach D, Swing S, Dreyfus H, Dreyfus S. General competencies and accreditation in graduate medical education. Health Aff (Millwood) 2002;21(5):103-11.
3. Leach DC. Competence is a habit. JAMA 2002;287:243-4.
4. Fraser SW, Greenhalgh T. Coping with complexity: educating for capability. BMJ 2001;323:799-803.
5. Klass D. Reevaluation of clinical competency. Am J Phys Med Rehabil 2000;79:481-6.
6. Bordage G, Zacks R. The structure of medical knowledge in the memories of medical students and general practitioners: categories and prototypes. Med Educ 1984;18:406-16.
7. Gruppen LD, Frohna AZ. Clinical reasoning. In: Norman GR, Van Der Vleuten CP, Newble DI, eds. International handbook of research in medical education. Part 1. Dordrecht, the Netherlands: Kluwer Academic, 2002:205-30.
8. Epstein RM, Dannefer EF, Nofziger AC, et al. Comprehensive assessment of professional competence: the Rochester experiment. Teach Learn Med 2004;16:186-96.
9. Ericsson KA. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. Acad Med 2004;79:Suppl:S70-S81.
10. Epstein RM. Mindful practice. JAMA 1999;282:833-9.
11. Schon DA. Educating the reflective practitioner. San Francisco: Jossey-Bass, 1987.
12. Epstein RM. Mindful practice in action. II. Cultivating habits of mind. Fam Syst Health 2003;21:11-7.
13. Dreyfus HL. On the Internet (thinking in action). New York: Routledge, 2001.
14. Eraut M. Learning professional processes: public knowledge and personal experience. In: Eraut M, ed. Developing professional knowledge and competence. London: Falmer Press, 1994:100-22.
15. Shanafelt TD, Bradley KA, Wipf JE, Back AL. Burnout and self-reported patient care in an internal medicine residency program. Ann Intern Med 2002;136:358-67.
16. Borrell-Carrio F, Epstein RM. Preventing errors in clinical practice: a call for self-awareness. Ann Fam Med 2004;2:310-6.
17. Leung WC. Competency based medical training: review. BMJ 2002;325:693-6.
18. Friedman Ben-David M. The role of assessment in expanding professional horizons. Med Teach 2000;22:472-7.
19. Sullivan W. Work and integrity: the crisis and promise of professionalism in America. 2nd ed. San Francisco: Jossey-Bass, 2005.
20. Schuwirth L, van der Vleuten C. Merging views on assessment. Med Educ 2004;38:1208-10.
21. Wass V, Van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. Lancet 2004;357:945-9.
22. Van Der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. Adv Health Sci Educ 1996;1:41-67.
23. Schuwirth LW, van der Vleuten CP. Different written assessment methods: what can be said about their strengths and weaknesses? Med Educ 2004;38:974-9.
24. Schuwirth LW, Verheggen MM, van der Vleuten CP, Boshuizen HP, Dinant GJ. Do short cases elicit different thinking processes than factual knowledge questions do? Med Educ 2001;35:348-56.
25. Case S, Swanson D. Constructing written test questions for the basic and clinical sciences. 3rd ed. Philadelphia: National Board of Medical Examiners, 2000.
26. Farmer EA, Page G. A practical guide to assessing clinical decision-making skills using the key features approach. Med Educ 2005;39:1188-94.
27. Charlin B, Roy L, Brailovsky C, Goulet F, van der Vleuten C. The Script Concordance test: a tool to assess the reflective clinician. Teach Learn Med 2000;12:189-95.
28. Brailovsky C, Charlin B, Beausoleil S, Cote S, Van der Vleuten C. Measurement of clinical reflective capacity early in training as a predictor of clinical reasoning performance at the end of residency: an experimental study on the script concordance test. Med Educ 2001;35:430-6.
29. Frederiksen N. The real test bias: influences of testing on teaching and learning. Am Psychol 1984;39:193-202.
30. Schuwirth LW, van der Vleuten CP, Donkers HH. A closer look at cueing effects in multiple-choice questions. Med Educ 1996;30:44-9.
31. Friedman MH, Connell KJ, Olthoff AJ, Sinacore JM, Bordage G. Medical student errors in making a diagnosis. Acad Med 1998;73:Suppl:S19-S21.
32. Graber ML, Franklin N, Gordon R. Diagnostic error in internal medicine. Arch Intern Med 2005;165:1493-9.
33. Pulito AR, Donnelly MB, Plymale M, Mentzer RM Jr. What do faculty observe of medical students' clinical performance? Teach Learn Med 2006;18:99-104.
34. Norman G. The long case versus objective structured clinical examinations. BMJ 2002;324:748-9.
35. Norcini JJ, Blank LL, Duffy FD, Fortna GS. The mini-CEX: a method for assessing clinical skills. Ann Intern Med 2003;138:476-81.
36. Van der Vleuten CP, Norman GR, De Graaff E. Pitfalls in the pursuit of objectivity: issues of reliability. Med Educ 1991;25:110-8.
37. Anastakis DJ, Cohen R, Reznick RK. The structured oral examination as a method for assessing surgical residents. Am J Surg 1991;162:67-70.
38. Ram P, Grol R, Rethans JJ, Schouten B, Van der Vleuten C, Kester A. Assessment of general practitioners by video observation of communicative and medical performance in daily practice: issues of validity, reliability and feasibility. Med Educ 1999;33:447-54.
39. Papadakis MA. The Step 2 clinical-skills examination. N Engl J Med 2004;350:1703-5.
40. Hodges B, McIlroy JH. Analytic global OSCE ratings are sensitive to level of training. Med Educ 2003;37:1012-6.

41. Reznick RK, Blackmore D, Cohen R, et al. An objective structured clinical examination for the licentiate of the Medical Council of Canada: from research to reality. Acad Med 1993;68:Suppl:S4-S6.
42. Wass V, Jones R, Van der Vleuten C. Standardized or real patients to test clinical competence? The long case revisited. Med Educ 2001;35:321-5.
43. Carney PA, Eliassen MS, Wolford GL, Owen M, Badger LW, Dietrich AJ. How physician communication influences recognition of depression in primary care. J Fam Pract 1999;48:958-64.
44. Epstein RM, Franks P, Shields CG, et al. Patient-centered communication and diagnostic testing. Ann Fam Med 2005;3:415-21.
45. Tamblyn RM. Use of standardized patients in the assessment of medical practice. CMAJ 1998;158:205-7.
46. Kravitz RL, Epstein RM, Feldman MD, et al. Influence of patients' requests for direct-to-consumer advertised antidepressants: a randomized controlled trial. JAMA 2005;293:1995-2002. [Erratum, JAMA 2005; 294:2436.]
47. Reznick RK, MacRae H. Teaching surgical skills — changes in the wind. N Engl J Med 2006;355:2664-9.
48. Ramsey PG, Wenrich MD, Carline JD, Inui TS, Larson EB, LoGerfo JP. Use of peer ratings to evaluate physician performance. JAMA 1993;269:1655-60.
49. Dannefer EF, Henson LC, Bierer SB, et al. Peer assessment of professional competence. Med Educ 2005;39:713-22.
50. Violato C, Marini A, Toews J, Lockyer J, Fidler H. Feasibility and psychometric properties of using peers, consulting physicians, co-workers, and patients to assess physicians. Acad Med 1997;72:Suppl 1:S82-S84.
51. Norcini JJ. Peer assessment of competence. Med Educ 2003;37:539-43.
52. Nofziger AC, Davis B, Naumburg EH, Epstein RM. The impact of peer assessment on professional development. Presented at the Ottawa Conference on Medical Education and Assessment, Ottawa, July 15, 2002. abstract.
53. Lurie SJ, Nofziger AC, Meldrum S, Mooney C, Epstein RM. Temporal and group-related trends in peer assessment amongst medical students. Med Educ 2006; 40:840-7.
54. Lurie S, Lambert D, Grady-Weliky TA. Relationship between dean's letter rankings, peer assessment during medical school, and ratings by internship directors. [Submitted for Publication] (in press).
55. Calhoun JG, Woolliscroft JO, Hockman EM, Wolf FM, Davis WK. Evaluating medical student clinical skill performance: relationships among self, peer, and expert ratings. Proc Annu Conf Res Med Educ 1984;23:205-10.
56. Hall JA, Milburn MA, Roter DL, Daltroy LH. Why are sicker patients less satisfied with their medical care? Tests of two explanatory models. Health Psychol 1998; 17:70-5.
57. Chang JT, Hays RD, Shekelle PG, et al. Patients' global ratings of their health care are not associated with the technical quality of their care. Ann Intern Med 2006; 144:665-72.
58. Butterfield PS, Mazzaferri EL. A new rating form for use by nurses in assessing residents' humanistic behavior. J Gen Intern Med 1991;6:155-61.
59. Klessig J, Robbins AS, Wieland D, Rubenstein L. Evaluating humanistic attributes of internal medicine residents. J Gen Intern Med 1989;4:514-21.
60. Eva KW, Regehr G. Self-assessment in the health professions: a reformulation and research agenda. Acad Med 2005;80: Suppl:S46-S54.
61. Carraccio C, Englander R. Evaluating competence using a portfolio: a literature review and Web-based application to the ACGME competencies. Teach Learn Med 2004;16:381-7.
62. Committee on Quality of Health Care in America. Crossing the quality chasm: a new health system for the 21st century. Washington, DC: National Academy Press, 2001.
63. Ginsburg S, Regehr G, Lingard L. Basing the evaluation of professionalism on observable behaviors: a cautionary tale. Acad Med 2004;79:Suppl:S1-S4.
64. Schirmer JM, Mauksch L, Lang F, et al. Assessing communication competence: a review of current tools. Fam Med 2005; 37:184-92.
65. Epstein RM, Franks P, Fiscella K, et al. Measuring patient-centered communication in patient-physician consultations: theoretical and practical issues. Soc Sci Med 2005;61:1516-28.
66. Norman GR, Van der Vleuten CP, De Graaff E. Pitfalls in the pursuit of objectivity: issues of validity, efficiency and acceptability. Med Educ 1991;25:119-26.
67. Colliver JA, Vu NV, Barrows HS. Screening test length for sequential testing with a standardized-patient examination: a receiver operating characteristic (ROC) analysis. Acad Med 1992;67:592-5.
68. Hakstian RA. The effects of type of examination anticipated on test preparation and performance. J Educ Res 1971;64: 319-24.
69. Newble DI, Jaeger K. The effect of assessments and examinations on the learning of medical students. Med Educ 1983; 17:165-71.
70. Schmidt HG, Norman GR, Boshuizen HP. A cognitive perspective on medical expertise: theory and implication. Acad Med 1990;65:611-21. [Erratum, Acad Med 1992;67:287.]
71. Bowen JL. Educational strategies to promote clinical diagnostic reasoning. N Engl J Med 2006;355:2217-25.
72. Tamblyn R, Abrahamowicz M, Dauphinee WD, et al. Association between licensure examination scores and practice in primary care. JAMA 2002;288:3019-26.
73. Eva KW, Reiter HI, Rosenfeld J, Norman GR. The ability of the multiple mini-interview to predict preclerkship performance in medical school. Acad Med 2004; 79:Suppl:S40-S42.
74. Lypson ML, Frohna JG, Gruppen LD, Woolliscroft JO. Assessing residents' competencies at baseline: identifying the gaps. Acad Med 2004;79:564-70.
75. Batmangelich S, Adamowski S. Maintenance of certification in the United States: a progress report. J Contin Educ Health Prof 2004;24:134-8.