ACGME Outcome project
enhancing residency education through outcomes assessment

ABMS

TOOLBOX OF ASSESSMENT METHODS©

A Product of the Joint Initiative

ACGME Outcomes Project
Accreditation Council for Graduate Medical Education

American Board of Medical Specialties (ABMS)

Version 1.1
September 2000

TOOLBOX OF ASSESSMENT METHODS©

A Product of the Joint Initiative

ACGME Outcomes Project
Accreditation Council for Graduate Medical Education

American Board of Medical Specialties (ABMS)

Version 1.1
September 2000

Copyright Disclosure.

General Disclaimer.

The Toolbox includes descriptions of assessment methods that can be used for evaluating residents. It does not include all the tools that can or may be used by a residency program for evaluating residents, or by a program director in verifying that a resident has demonstrated sufficient professional ability to practice competently and independently. Neither ACGME nor ABMS shall be liable in any way for results obtained in applying these assessment methods. The user, and not ACGME or ABMS, shall be soley responsible for the results obtained in applying the assessment methods described herein. Further, the user agrees and acknowledges that, in using the Toolbox, he/she/it is solely responsible for complying with all applicable laws, regulations, and ordinances relating to privacy.

Table of Contents

Preface

Included in this packet are descriptions of assessment methods that can be used for evaluating residents. In addition to a brief description of each method, there is information pertaining to its use, psychometric qualities, and feasibility/practicality.

As a "work in progress", the descriptions reflect the most typical use and research findings related to the method. As this work proceeds, refinements and extensions that reflect the full potential and creative application of the methods can be expected.

The descriptions were developed to assist medical educators with the selection and development of evaluation techniques. They represent a first step in the construction of a more complete toolbox of assessment techniques.

The table on the last pages of this booklet rates assessment tools for robustness and practical use for assessing specific competencies expected of residents. The ratings are based upon a consensus of evaluation experts.

This work is supported in part by a grant from the Robert Wood Johnson Foundation to the Accreditation Council for Graduate Medical Education.

Susan Swing Ph.D.
ACGME Director of Research

Philip G. Bashook Ed.D.
ABMS Director of Evaluation and Education

Glossary

Generalizability – Measurements (scores) derived from an assessment tool are considered generalizable if they can be shown to apply to more than the sample of cases or test questions used in a specific assessment.

Reliability/Reproducibility – A reliable test score means when measurements (scores) are repeated the new test results are consistent with the first scores for the same assessment tool on the same or similar individuals. Reliability is measured as a correlation with 1.0 being perfect reliability and below 0.50 as unreliable. Evaluation measurement reliabilities above 0.65 and preferably near or above 0.85 are recommended.

Validity – Validating assessment measures is a process of accumulating evidence about how well the assessment measures represent or predict a resident's ability or behavior. Validity refers to the specific measurements made with assessment tools in a specific situation with a specific group of individuals. It is the scores not the type of assessment tool that are valid. For example, it is possible to determine if the written exam scores for a group of residents are valid in measuring the residents' knowledge, but it is incorrect to say that "all written exams" are valid to measure knowledge.

Formative Evaluation – In formative evaluation findings are accumulated from a variety of relevant assessments designed for use either in program or resident evaluation. In resident evaluation the formative evaluation is intended to provide constructive feedback to individual residents during their training. In program evaluation the formative evaluation is intended to improve program quality. In neither situation is formative evaluation intended to make a go/no-go decision.

Summative Evaluation – In summative evaluation findings and recommendations are designed to accumulate all relevant assessments for a go/no-go decision. In resident evaluation the summative evaluation is used to decide whether the resident qualifies to continue to the next training year, should be dropped from the program, or at the completion of the residency should be recommended for board certification. In program evaluation the summative evaluation is used to judge whether the program meets the accepted standards for the purpose of continuing, restructuring or discontinuing the program.

360-DEGREE EVALUATION INSTRUMENT

DESCRIPTION

360-degree evaluations consist of measurement tools completed by multiple people in a person's sphere of influence. Evaluators completing rating forms in a 360-degree evaluation usually are superiors, peers, subordinates, and patients and families. Most 360-degree evaluation processes use a survey or questionnaire to gather information about an individual's performance on several topics (e.g., teamwork, communication, management skills, decision-making). Most 360-degree evaluations use rating scales to assess how frequently a behavior is performed (e.g., a scale of 1 to 5, with 5 meaning "all the time" and 1 meaning "never"). The ratings are summarized for all evaluators by topic and overall to provide feedback.

USE

Evaluators provide more accurate and less lenient ratings when the evaluation is intended to give formative feedback rather than summative evaluations. A 360-degree evaluation can be used to assess interpersonal and communication skills, professional behaviors, and some aspects of patient care and systems-based practice.

PSYCHOMETRIC QUALITIES

No published reports of the use of 360-degree evaluation instruments in graduate medical education were found in the literature; however, there are reports of the use of various categories of people evaluating residents at the same time, although with different instruments. Generally the evaluators were nurses, allied health professionals, other residents, faculty/supervisors, and patients. Moderate correlations were found to exist among the scores produced by these evaluators using slightly different assessment tools. Reproducible results were most easily obtainable when five to ten nurses rated residents, while a greater number of faculty and patients were needed for the same degree of reliability. In business, military and education settings, reliability estimates have been reported as great as 0.90 for 360-degree evaluation instruments.

FEASIBILITY/PRACTICALITY

In most clinical settings conducting 360-degree-evaluations will pose a significant challenge. The two practical challenges are: constructing surveys that are appropriate for use by all evaluators in the circle of influence, and orchestrating data collection from a potentially large number of individuals that can be compiled and reported confidentially to the resident. Implementing an electronic system should make the 360-degree-evaluation feasible.

SUGGESTED REFERENCE

Center for Creative Leadership, Greensboro, North Carolina (http://www.ccl.org).

CHART STIMULATED RECALL ORAL EXAMINATION (CSR)

DESCRIPTION
In a chart stimulated recall (CSR) examination patient cases of the examinee (resident) are assessed in a standardized oral examination. A trained and experienced physician examiner questions the examinee about the care provided probing for reasons behind the work-up, diagnoses, interpretation of clinical findings, and treatment plans. The examiners rate the examinee using a well-established protocol and scoring procedure. In efficiently designed CSR oral exams each patient case (test item) takes 5 to 10 minutes. A typical CSR exam is two hours with one or two physicians as examiners per separate 30 or 60-minute session.

USE
These exams assess clinical decision-making and the application or use of medical knowledge with actual patients. Multiple-choice questions are better than CSR at assessing recall or understanding of medical knowledge. Five of the 24 ABMS Member Boards use CSR as part of their standardized oral examinations for initial certification.

PSYCHOMETRIC QUALITIES
Patient cases are selected to be a sample of patients the examinee should be able to manage successfully, for example, as a board certified specialist. One or more scores are derived for each case based upon pre-defined scoring rules. The examinee's performance is determined by combining scores from all cases for a pass/fail decision overall or by each session. If the CSR is used for certification, test scores are analyzed using sophisticated statistical methods (e.g., Item Response Theory (IRT) or generalizability theory) to obtain a better estimate of the examinee's ability. Exam score reliabilities have been reported between 0.65 and 0.88 (1.00 is considered perfect reliability). The physician examiners need to be trained in how to question the examinee and evaluate and score the examinee's responses.

FEASIBILITY/PRACTICALITY
"Mock orals," that use resident's cases but with much less standardization compared to board oral exams, often are used in residency training programs to help familiarize residents with the oral exams conducted for board certification. CSR oral exams can be implemented easily to determine if residents can apply knowledge appropriately in managing patients, but for the exams to be used for high stakes decisions about the resident's abilities such as board certification extensive resources and expertise are required to standardize the exam.

SUGGESTED REFERENCE
Munger, BS. Oral examinations. In Mancall EL, Bashook PG. (editors) *Recertification: new evaluation methods and strategies*. Evanston, Illinois: American Board of Medical Specialties, 1995: 39-42.

CHECKLIST EVALUATION

DESCRIPTION
Checklists consist of essential or desired specific behaviors, activities, or steps that make up a more complex competency or competency component. Typical response options on these forms are a check (□) or "yes" to indicate that the behavior occurred or options to indicate the completeness (complete, partial, or absent) or correctness (total, partial, or incorrect) of the action. The forms provide information about behaviors but for the purpose of making a judgment about the adequacy of the overall performance, standards need to be set that indicate, for example, pass/fail or excellent, good, fair, or poor performance.

USE
Checklists are useful for evaluating any competency and competency component that can be broken down into specific behaviors or actions. Documented evidence for the usefulness of checklists exists for the evaluation of patient care skills (history and physical examination, procedural skills) and for interpersonal and communication skills. Checklists have also been used for self-assessment of practice-based learning skills (evidence-based medicine). Checklists are most useful to provide feedback on performance because checklists can be tailored to assess detailed actions in performing a task.

PSYCHOMETRIC QUALITIES
When observers are trained to use checklists, consistent scores can be obtained and reliability in the range of 0.7 to 0.8 is reported (1.0 is perfect reliability). Performance scores derived from checklists can discriminate between residents in different years of training. Scoring practitioners' behavior using checklists is more difficult when checklists assume a fixed sequence of actions because experienced physicians use various valid sequences and are usually parsimonious in their patient care behaviors.

FEASIBILITY/PRACTICALITY
To ensure validity of content and scoring rules, checklist development requires consensus by several experts with agreement on essential behaviors/actions, sequencing, and criteria for evaluating performance. Checklists require trained evaluators to observe performance and time to complete a checklist will vary depending on the observation period.

SUGGESTED REFERENCES
Noel G, Herbers JE, Caplow M et al. How well do Internal Medicine faculty members evaluate the clinical skills of residents? *Ann Int Med*. 1992; 117: 757-65.

Winckel CP, Reznick RK, Cohen R, Taylor B. Reliability and construct validity of a structured technical skills assessment form. *Am J Surg*. 1994; 167: 423-27.

GLOBAL RATING OF LIVE OR RECORDED PERFORMANCE

DESCRIPTION

Global rating forms are distinguished from other rating forms in that (a) a rater judges general categories of ability (e.g. patient care skills, medical knowledge, interpersonal and communication skills) instead of specific skills, tasks or behaviors; and (b) the ratings are completed retrospectively based on general impressions collected over a period of time (e.g., end of a clinical rotation) derived from multiple sources of information (e.g., direct observations or interactions; input from other faculty, residents, or patients; review of work products or written materials). All rating forms contain scales that the evaluator uses to judge knowledge, skills, and behaviors listed on the form. Typical rating scales consist of qualitative indicators and often include numeric values for each indicator, for example, (a) very good = 1, good =2, fair = 3, poor =4; or (b) superior =1, satisfactory =2, unsatisfactory =3. Written comments are important to allow evaluators to explain the ratings.

USE

Global rating forms are most often used for making end of rotation and summary assessments about performance observed over days or weeks. Scoring rating forms entails combining numeric ratings with comments to obtain a useful judgment about performance based upon more than one rater.

PSYCHOMETRIC QUALITIES

A number of problems with global ratings have been documented: scores can be highly subjective when raters are not well trained; sometimes all competencies are rated the same regardless of performance; and scores may be biased when raters inappropriately make severe or lenient judgments or avoid using the extreme ends of a rating scale. Research reports are mixed about: discriminating between competence levels of different individuals; rating more skilled/experienced physicians better than less experienced physicians; and reproducibility (reliability) of ratings by the same physician/faculty raters, across different physicians/faculty, and variability across physicians/faculty, residents, nurses, and patients ratings of the same resident. Reproducibility appears easier to achieve for ratings of knowledge and more difficult to achieve for patient care and interpersonal and communication skills. A few studies have reported that faculty give more lenient ratings than residents, especially when the residents believe that the ratings will not be used for pass/fail decisions.

FEASIBILITY/PRACTICALITY

Basic global rating forms can be constructed and completed quickly and easily. However, ratings do require time to directly observe performance or interact with the physician being evaluated. Training of raters is important to improve reproducibility of the findings.

SUGGESTED REFERENCE

Gray, J. Global rating scales in residency education. *Acad Med*. 1996; 71: S55-63.

.

OBJECTIVE STRUCTURED CLINICAL EXAMINATION (OSCE)

DESCRIPTION

In an objective structured clinical examination (OSCE) one or more assessment tools are administered at 12 to 20 separate standardized patient encounter stations, each station lasting 10-15 minutes. Between stations candidates may complete patient notes or a brief written examination about the previous patient encounter. All candidates move from station to station in sequence on the same schedule. Standardized patients are the primary assessment tool used in OSCEs, but OSCEs have included other assessment tools such as data interpretation exercises using clinical cases, and clinical scenarios with mannequins, to assess technical skills.

USE

OSCEs have been administered in most US medical schools, many residency programs, and by the licensure boards in Canada for more than five years. The OSCE format provides a standardized means to assess: physical examination and history taking skills; communication skills with patients and family members, breadth and depth of knowledge; ability to summarize and document findings; ability to make a differential diagnosis, or plan treatment; and clinical judgment based upon patient notes.

PSYCHOMETRIC QUALITIES

OSCEs can provide means to obtain direct measures in a standardized manner of a patient-doctor encounter. OSCEs are not useful to measure skills or abilities in continuity of care with repeated patient encounters or invasive procedures. Because OSCEs often use standardized patients the same advantages and limitations apply (See toolbox description of standardized patient examination). A separate performance score is derived for each task performed at a station and scores are combined across stations or tasks to determine a pass/fail score. Statistical weighting of scores on individual tasks is controversial and not recommended. An OSCE with 14 to 18 stations is recommended to obtain reliable measurements of performance.

FEASIBILITY/PRACTICALITY

OSCEs are very useful to measure specific clinical skills and abilities, but are difficult to create and administer. OSCEs are only cost-effective when many candidates are to be examined at one administration. Most OSCEs are administered in medical center outpatient facilities or specially designed patient examining rooms with closed circuit television. A separate room or cubical is needed for each station. For most residency programs developing and administering an OSCE will require the resources and expertise of a consortium of residency programs in an academic institution or metropolitan area.

SUGGESTED REFERENCE

Norman, Geoffrey. *Evaluation Methods: A resource handbook*. Hamilton, Ontario, Canada: Program for Educational Development, McMaster University, 1995: 71-77.

.

PROCEDURE, OPERATIVE, OR CASE LOGS

DESCRIPTION
Procedure, operative, or case logs document each patient encounter by medical conditions seen, surgical operation or procedures performed. The logs may or may not include counts of cases, operations, or procedures. Patient case logs currently in use involve recording of some number of consecutive cases in a designated time frame. Operative logs in current use vary; some entail comprehensive recording of operative data by CPT code while others require recording of operations or procedures for a small number of defined categories.

USE
Logs of types of cases seen or procedures performed are useful for determining the scope of patient care experience. Regular review of logs can be used to help the resident track what cases or procedures must be sought out in order to meet residency requirements or specific learning objectives. Patient logs documenting clinical experience for the entire residency can serve as a summative report of that experience; as noted below, the numbers reported do not necessarily indicate competence.

PSYCHOMETRIC QUALITIES
There are no known studies of case or procedure logs for the purpose of determining accuracy of residents' recording. Unless defined by CPT or other codes, cases or procedures counted for a given category may vary across residents and programs. Minimum numbers of procedures required for accreditation and certification have not been validated against the actual quality of performance of an operation or patient outcomes.

FEASIBILITY/PRACTICALITY
Electronic recording devices and systems facilitate the collection and summarization of patient cases or procedures performed. Although there is considerable cost associated with development, testing, and maintenance of electronic systems, these costs generally are not paid by individual programs and institutions, since systems are available commercially for a relatively small amount (e.g., $2500 annually) or provided free of charge by accrediting or certification bodies. Manual recording is required followed later by data entry unless automated data entry devices are located at or near the point of service. Data entry of manual records typically can be performed by a clerk, but is time consuming depending on the number of residents in the program and log reporting requirements.

SUGGESTED REFERENCE
Watts J, Feldman WB. Assessment of technical skills. In: Neufeld V and Norman G (ed). *Assessing clinical competence.* New York: Springer Publishing Company, 1985: 259-74.

.

PATIENT SURVEYS

DESCRIPTION

Surveys of patients to assess satisfaction with hospital, clinic, or office visits typically include questions about the physician's care. The questions often assess satisfaction with general aspects of the physician's care, (e.g., amount of time spent with the patient, overall quality of care, physician competency (skills and knowledge), courtesy, and interest or empathy). More specific aspects of care can be assessed including: the physician's explanations, listening skills and provision of information about examination findings, treatment steps, and drug side effects. A typical patient survey asks patients to rate their satisfaction with care using rating categories (e.g., poor, fair, good, very good, excellent) or agreement with statements describing the care (e.g., "the doctor kept me waiting," --Yes, always; Yes, sometimes; or No, never or hardly ever). Each rating is given a value and a satisfaction score calculated by averaging across responses to generate a single score overall or separate scores for different clinical care activities or settings.

USE

Patient feedback accumulated from single encounter questionnaires can assess satisfaction with patient care competencies (aspects of data gathering, treatment, and management; counseling, and education; preventive care); interpersonal and communication skills; professional behavior; and aspects of systems-based practice (patient advocacy; coordination of care). If survey items about specific physician behaviors are included, the results can be used for formative evaluation and performance improvement. Patient survey results also can be used for summative evaluation, but this use is contingent on whether the measurement process meets standards of reliability and validity.

PSYCHOMETRIC QUALITIES

Reliability estimates of 0.90 or greater have been achieved for most patient satisfaction survey forms used in hospitals and clinics. Reliability estimates are much lower for ratings of residents in training. The American Board of Internal Medicine reports 20-40 patient responses were needed to obtain a reliability of 0.70 to 0.82 on individual resident ratings using the ABIM Patient Satisfaction Questionnaire. Low per-resident reliability has been associated with surveys that use rating scales; survey questions with response options of "yes, definitely," "yes, somewhat," or "no," may provide more reproducible, and useful results.

FEASIBILITY/PRACTICALITY

A variety of patient satisfaction surveys are available from commercial developers and medical organizations. Creation of new surveys often begins with gathering input from patients using interviews, focus groups, or questionnaires. Physician attitudes and behaviors patients find to be satisfying or dissatisfying are then translated into survey items. Most patient satisfaction surveys are completed at the time of service, and require less than 10 minutes. Alternatively, they may be mailed after the patient goes home or conducted with patients over the phone. Difficulties encountered with patient surveys are: (1) language and

PATIENT SURVEYS

literacy problems; (2) obtaining enough per-resident surveys to provide reproducible results; (3) the resources required to collect, aggregate, and report survey responses; and (4) assessment of the resident's contribution to a patient's care separate from that of the health care team. Because of these concerns, patient satisfaction surveys are often conducted by the institution or by one or more clinical sites and reports specific to the residency program may or may not be prepared. It may be possible to improve feasibility by utilizing effective survey design principles and using computers to collect and summarize survey data.

SUGGESTED REFERENCES

Kaplan SH, Ware JE. The patient's role in health care and quality assessment. In: Goldfield N and Nash D (eds). *Providing quality care (2nd ed): Future Challenge*. Ann Arbor, MI: Health Administration Press, 1995: 25-52.

Matthews DA, Feinstein AR. A new instrument for patients' ratings of physician performance in the hospital setting. *J Gen Intern Med*. 1989:4:14-22.

PORTFOLIOS

DESCRIPTION

A portfolio is a collection of products prepared by the resident that provides evidence of learning and achievement related to a learning plan. A portfolio typically contains written documents but can include video- or audio-recordings, photographs, and other forms of information. Reflecting upon what has been learned is an important part of constructing a portfolio. In addition to products of learning, the portfolio can include statements about what has been learned, its application, remaining learning needs, and how they can be met. In graduate medical education, a portfolio might include a log of clinical procedures performed; a summary of the research literature reviewed when selecting a treatment option; a quality improvement project plan and report of results; ethical dilemmas faced and how they were handled; a computer program that tracks patient care outcomes; or a recording or transcript of counseling provided to patients.

USE

Portfolios can be used for both formative and summative evaluation of residents. Portfolios are most useful for evaluating mastery of competencies that are difficult to evaluate in other ways such as practice-based improvement, use of scientific evidence in patient care, professional behaviors, and patient advocacy. Teaching experiences, morning report, patient rounds, individualized study or research projects are examples of learning experiences that lend themselves to using portfolios to assess residents. The Royal College of Physicians and Surgeons of Canada in the Maintenance of Competence Program (MOCOMPS) has developed a portfolio system for recertification using Internet-based diaries called PCDiary© that could be adapted to residency evaluations.

PSYCHOMETRIC QUALITIES

Reproducible assessments are feasible when there is agreement on criteria and standards for contents of a portfolio. When portfolio assessments have been used to evaluate an educational program (e.g., statewide elementary or high school program) the portfolio products or documentation have been found to be sufficient for program evaluation but are not always appropriate to use in assessing individual students for decisions about promotion to the next grade. However, standard criteria are not necessarily desirable and may be counter-productive when the portfolio purpose is to demonstrate individual learning gains relative to individual goals. The validity of portfolio assessment is determined by the extent to which the products or documentation included in a portfolio demonstrates mastery of expected learning.

FEASIBILITY/PRACTICALITY

Acceptance of portfolios in graduate medical education varies according to preferred learning style. Some residents and practicing physicians have found that by maintaining portfolios credit was allowed for some activities that otherwise would have gone undone or un-noticed. Yet, for others, the time and commitment necessary to create and maintain a portfolio is too great relative to the return.

SUGGESTED REFERENCE

Challis M. AMEE medical education guide no. 11 (revised): Portfolio-based learning and assessment in medical education. *Med Teach.* 1999; 21: 370-86.

RECORD REVIEW

DESCRIPTION

Trained staff in an institution's medical records department or clinical department perform a review of patients' paper or electronic records. The staff uses a protocol and coding form based upon predefined criteria to abstract information from the records, such as medications, tests ordered, procedures performed, and patient outcomes. The patient record findings are summarized and compared to accepted patient care standards. Standards of care are available for more than 1600 diseases on the Website of the Agency for HealthCare Research and Quality (http://www.ahrq.gov/).

USE

Record review can provide evidence about clinical decision-making, follow-through in patient management and preventive health services, and appropriate use of clinical facilities and resources (e.g., appropriate laboratory tests and consultations). Often residents will confer with other clinical team members before documenting patient decisions and therefore, the documented care may not be directly attributed to a single resident but to the clinical team.

PSYCHOMETRIC QUALITIES

A sample of approximately eight to 10 patient records is sufficient for a reliable assessment of care for a diagnosis or procedure. One study in office practice demonstrated that six to eight office records selected randomly are adequate to evaluate care. Missing or incomplete documentation of care is interpreted as not meeting the accepted standard.

FEASIBILITY/PRACTICALITY

Record reviews by trained staff take approximately 20 to 30 minutes per record on average for records of hospitalized patients. The major limitations are: (1) as a retrospective assessment of care the review may not be completed until sufficient patients have been treated which could delay reports about residents' performance for months after a typical one or two month clinical rotation; (2) criteria of care must be agreed-up and translated into coding forms for staff to review records; (3) staff must be trained in how to identify and code clinical data to assure reasonably reliable findings.

SUGGESTED REFERENCE

Tugwell P, Dok, C. Medical record review. In: Neufeld V and Norman G (ed). *Assessing clinical competence.* New York: Springer Publishing Company, 1985: 142-82.

.

SIMULATIONS AND MODELS

DESCRIPTION

Simulations used for assessment of clinical performance closely resemble reality and attempt to imitate but not duplicate real clinical problems. Key attributes of simulations are that: they incorporate a wide array of options resembling reality, allow examinees to reason through a clinical problem with little or no cueing, permit examinees to make life-threatening errors without hurting a real patient, provide instant feedback so examinees can correct a mistaken action, and rate examinees' performance on clinical problems that are difficult or impossible to evaluate effectively in other circumstances. Simulation formats have been developed as paper-and-pencil branching problems (patient management problems or PMPs), computerized versions of PMPs called clinical case simulations (CCX®), role-playing situations (e.g., standardized patients (SPs), clinical team simulations), anatomical models or mannequins, and combinations of all three formats. Mannequins are imitations of body organs or anatomical body regions frequently using pathological findings to simulate patient disease. The models are constructed of vinyl or plastic sculpted to resemble human tissue with imbedded electronic circuitry to allow the mannequin to respond realistically to actions by the examinee. Virtual reality simulations or environments (VR) use computers sometimes combined with anatomical models to mimic as much as feasible realistic organ and surface images and the touch sensations (computer generated haptic responses) a physician would expect in a real patient. The VR environments allow assessment of procedural skills and other complex clinical tasks that are difficult to assess consistently by other assessment methods.

USE

Simulations using VR environments have been developed to train and assess surgeons performing arthroscopy of the knee and other large joints, anesthesiologists managing life-threatening critical incidents during surgery, surgeons performing wound debridement and minor surgery, and medical students and residents responding to cardio-pulmonary incidents on a full-size human mannequin. Written and computerized simulations have been used to assess clinical reasoning, diagnostic plans and treatment for a variety of clinical disciplines as part of licensure and certification examinations. Standardized patients as simulations are described elsewhere.

PSYCHOMETRIC QUALITIES

Studies of high-quality simulations have demonstrated their content validity when the simulation is designed to resemble a real patient. One or more scores are derived for each simulation based upon pre-defined scoring rules set by the experts in the discipline. The examinee's performance is determined by combining scores from all simulations to derive an overall performance score. When included in Objective Structured Clinical Examinations (OSCEs) the case reliabilities are similar to those reported for OSCEs (See OSCEs).

FEASIBILITY/PRACTICALITY

Experts in a specialty carefully craft simulations as clinical scenarios from real patient cases to focus the assessments on specific skills, abilities and "key features" of the case. Technical experts in assessment and simulations then convert the scenarios into simulations as standardized patients, mannequins,

SIMULATIONS AND MODELS

computer-based simulations, and other simulations adding when feasible computer-automated scoring rules to record the examinees' actions. Simulations are expensive to create and often require producing many variations of the pathological conditions or clinical problems to make them economical. Grants and contracts from commercial vendors, foundations, governmental agencies and medical schools continue to be the principle source of funding to develop simulations.

SUGGESTED REFERENCE

Tekian A, McGuire CH, et al (eds.) *Innovative simulations for assessing professional competence*. Chicago, Illinois: University of Illinois at Chicago, Dept. Med. Educ. 1999

STANDARDIZED ORAL EXAMINATION

DESCRIPTION
The standardized oral examination is a type of performance assessment using realistic patient cases with a trained physician examiner questioning the examinee. The examiner begins by presenting to the examinee a clinical problem in the form of a patient case scenario and asks the examinee to manage the case. Questions probe the reasoning for requesting clinical findings, interpretation of findings, and treatment plans. In efficiently designed exams each case scenario takes three to five minutes. Exams last approximately 90 minutes to two and one-half hours with two to four separate 30 or 60-minute sessions. One or two physicians serve as examiners per session. An examinee can be tested on 18 to 60 different clinical cases.

USE
These exams assess clinical decision-making and the application or use of medical knowledge with realistic patients. Multiple-choice questions are better at assessing recall or understanding of medical knowledge. Fifteen of the 24 ABMS Member Boards use standardized oral examinations as the final examination for initial certification.

PSYCHOMETRIC QUALITIES
A committee of experts in the specialty carefully crafts the clinical scenarios from real patient cases to focus the assessment on the "key features" of the case. Cases are selected to be a sample of patients the examinee should be able to manage successfully, for example, as a board certified specialist. One or more scores are derived for each case based upon pre-defined scoring rules. The examinee's performance is determined by combining scores from all cases for a pass/fail decision overall or by each session. Test scores are analyzed using sophisticated statistical methods (e.g., Item Response Theory (IRT) or generalizability theory) to obtain a better estimate of the examinee's ability. Exam score reliabilities have been reported between 0.65 and 0.88 (1.00 is considered perfect reliability). The physician examiners need to be trained in how to provide patient data for each scenario, question the examinee, and evaluate and score the examinee's responses.

FEASIBILITY/PRACTICALITY
A committee of physician specialists develops the examination cases and trains the examiners, often with assistance from psychometric experts. "Mock orals," that use cases but with much less standardization compared to board oral exams, are often used in residency training programs to help familiarize residents with the oral exams conducted for board certification. Extensive resources and expertise are required, however, to develop and administer a standardized oral examination.

SUGGESTED REFERENCE
Mancall EL, Bashook PG. (eds.) *Assessing clinical reasoning: the oral examination and alternative methods*. Evanston, Illinois: American Board of Medical Specialties, 1995.

.

STANDARDIZED PATIENT EXAMINATION (SP)

DESCRIPTION
Standardized patients (SPs) are well persons trained to simulate a medical condition in a standardized way or actual patients who are trained to present their condition in a standardized way. A standardized patient exam consists of multiple SPs each presenting a different condition in a 10-12 minute patient encounter. The resident being evaluated examines the SP as if (s)he were a real patient, (i.e., the resident might perform a history and physical exam, order tests, provide a diagnosis, develop a treatment plan, or counsel the patient). Using a checklist or a rating form, a physician observer or the SPs evaluate the resident's performance on appropriateness, correctness, and completeness of specific patient care tasks and expected behaviors (See description of Checklist Evaluation…). Performance criteria are set in advance. Alternatively or in addition to evaluation using a multiple SP exam, individual SPs can be used to assess specific patient care skills. SPs are also included as stations in Objective Structured Clinical Examinations (See description of OSCE).

USE
SPs have been used to assess history-taking skills, physical examination skills, communication skills, differential diagnosis, laboratory utilization, and treatment. Reproducible scores are more readily obtained for history-taking, physical examination, and communication skills. Standardized patient exams are most frequently used as summative performance exams for clinical skills. A single SP can assess targeted skills and knowledge.

PSYCHOMETRIC QUALITIES
Standardized patient examinations can generate reliable scores for individual stations and total performance useful for pass-fail decisions. Training of raters whether physicians, patients or other types of observers is critical to obtain reliable scores. At least one-half day of testing time (four hours) is needed to obtain reliable scores for assessment of hands-on clinical skills. Research on the validity of some SP exams has found better performance by senior residents than junior residents (construct validity) and modest correlations between SP exam scores and clinical ratings or written exams (concurrent validity).

FEASIBILITY/PRACTICALITY
Development of an examination using standardized patients involves identification of the specific competencies to be tested, training of standardized patients, development of checklists or rating forms and criteria setting. Development time can be considerable, but can be made more time efficient by sharing of SPs in a collaboration of multiple residency programs or in a single academic medical center. A new SP can learn to stimulate a new clinical problem in

STANDARDIZED PATIENT EXAMINATION (SP)

8 to 10 hours; and an experienced SP can learn a new problem in 6 to 8 hours. About twice the training time is needed for SPs to learn to use checklists to evaluate resident performance. Facilities needed for the examination include an examining room for each SP station and space for residents to record medical notes between stations.

SUGGESTED REFERENCE

Van der Vleuten, CPM and Swanson, D. Assessment of clinical skills with standardized patients: State of the art. *Teach Learn Med*. 1990; 2: 58-76.

.

WRITTEN EXAMINATION (MCQ)

DESCRIPTION

A written or computer-based MCQ examination is composed of multiple-choice questions (MCQ) selected to sample medical knowledge and understanding of a defined body of knowledge, not just factual or easily recalled information. Each question or test item contains an introductory statement followed by four or five options in outline format. The examinee selects one of the options as the presumed correct answer by marking the option on a coded answer sheet. Only one option is keyed as the correct response. The introductory statement often presents a patient case, clinical findings, or displays data graphically. A separate booklet can be used to display pictures, and other relevant clinical information. The in-training examinations prepared by specialty societies and boards use MCQ type test items. A typical half-day examination has 175 to 250 test questions.

In computer-based examinations the test items are displayed on a computer monitor one at a time with pictures and graphical images also displayed directly on the monitor. In a computer-adaptive test fewer test questions are needed because test items are selected based upon statistical rules programmed into the computer to quickly measure the examinee's ability.

USE

Medical knowledge and understanding can be measured by MCQ examinations. Comparing the test scores on in-training examinations with national statistics can serve to identify strengths and limitations of individual residents to help them improve. Comparing test results aggregated for residents in each year of a program can be helpful to identify residency training experiences that might be improved.

PSYCHOMETRIC QUALITIES

For test questions to be useful in evaluating a resident's knowledge each test item and the overall exam should be designed to rigorous psychometric standards. Psychometric qualities must be high for pass/fail decisions, but tests used to help residents identify strengths and weaknesses such as in-training examinations need not comply with the same rigorous standards. A committee of experts designing the test defines the knowledge to be assessed and creates a test blueprint that specifies the number of test questions to be selected for each topic. When test questions are used to make pass/fail decisions the test should be pilot tested and statistically analyzed. A higher reliability/reproducibility can be achieved with more test questions per topic. If pass/fail decisions will be made based on test scores a sufficient number of test questions should be included to obtain a test reliability greater than r = 0.85 (1.00 is perfect reliability). Standards for passing scores should be set by a committee of experts prior to administering the examination (criterion referenced exams). If performance of residents is to be compared from year to year at least 25 to 30 percent of the same test questions should be repeated each year.

.

WRITTEN EXAMINATION (MCQ)

FEASIBILITY/PRACTICALITY

A committee of physician specialists develops the examination with assistance from psychometric experts. For in-training examinations each residency program administers an exam purchased from the specialty society or other vendor. Tests are scored by the vendor and scores returned to the residency director for each resident, for each topic, and by year of residency training. Comparable national scores also are provided. All the 24 ABMS Member Boards use MCQ examinations for initial certification.

SUGGESTED REFERENCES

Haladyna TM. *Developing and validating multiple-choice test items*. Hillsdale, New Jersey: L. Erlbaum Associates. 1994.

Case SM, Swanson DB. *Constructing written test questions for the basic and clinical sciences*. Philadelphia, PA: National Board of Medical Examiners, 1996 (www.nbme.org)

List of Suggested References

Case SM, Swanson DB. *Constructing written test questions for the basic and clinical sciences*. Philadelphia, PA: National Board of Medical Examiners, 1996 (www.nbme.org)

Center for Creative Leadership, Greensboro, North Carolina (http://www.ccl.org).

Challis M. AMEE medical education guide no. 11 (revised): Portfolio-based learning and assessment in medical education. *Med Teach.* 1999; 21: 370-86.

Gray, J. Global rating scales in residency education. *Acad Med*. 1996; 71: S55-63.

Haladyna TM. *Developing and validating multiple-choice test items*. Hillsdale, New Jersey: L. Erlbaum Associates. 1994.

Kaplan SH, Ware JE. The patient's role in health care and quality assessment. In: Goldfield N and Nash D (eds). *Providing quality care (2nd ed): Future Challenge.* Ann Arbor, MI: Health Administration Press, 1995: 25-52.

Matthews DA, Feinstein AR. A new instrument for patients' ratings of physician performance in the hospital setting. *J Gen Intern Med*. 1989:4:14-22.

Mancall EL, Bashook PG. (eds.) *Assessing clinical reasoning: the oral examination and alternative methods*. Evanston, Illinois: American Board of Medical Specialties, 1995.

Munger, BS. Oral examinations. In Mancall EL, Bashook PG. (editors) *Recertification: new evaluation methods and strategies*. Evanston, Illinois: American Board of Medical Specialties, 1995: 39-42.

Noel G, Herbers JE, Caplow M et al. How well do Internal Medicine faculty members evaluate the clinical skills of residents? *Ann Int Med*. 1992; 117: 757-65.

Norman, Geoffrey. *Evaluation Methods: A resource handbook*. Hamilton, Ontario, Canada: Program for Educational Development, McMaster University, 1995: 71-77.

Tekian A, McGuire CH, et al (eds.) *Innovative simulations for assessing professional competence*. Chicago, Illinois: University of Illinois at Chicago, Dept. Med. Educ. 1999

.

Tugwell P, Dok, C. Medical record review. In: Neufeld V and Norman G (ed). *Assessing clinical competence.* New York: Springer Publishing Company, 1985: 142-82.

Van der Vleuten, CPM and Swanson, D. Assessment of clinical skills with standardized patients: State of the art. *Teach Learn Med*. 1990; 2: 58-76.

Watts J, Feldman WB. Assessment of technical skills. In: Neufeld V and Norman G (ed). *Assessing clinical competence.* New York: Springer Publishing Company, 1985, 259-74.

Winckel CP, Reznick RK, Cohen R, Taylor B. Reliability and construct validity of a structured technical skills assessment form. *Am J Surg*. 1994; 167: 423-27.