Assessment Guidebook

John W. Ragsdale, MD, MS
Carrie A. Elzie, PhD, MA
Lisa D. Howley, PhD, MEd
Mark Speicher, PhD, MHA
Renee L. Williams, MD, MHPE
Pedro Tanaka, MD
Eric S. Holmboe, MD
Kate Hatlak, EdD
Laura Edgar, EdD

Acknowledgment

We wish to acknowledge the following contributors to both the original and current versions of this Guidebook: Daniel Dent, MD; Stephen Durning, MD, PhD; William Iobst, MD; Greg Ogrinc, MD; and Liana Puscas, MD.

General Disclaimer

This revised edition of the Assessment Guidebook provides descriptions of assessment tools and methods along with general guidance and approaches that can be used for assessing medical students, residents, and fellows. The Guidebook does not include all the tools and methods that can or may be used for evaluating learners or for verifying that learners have demonstrated sufficient professional ability to be promoted or to practice without supervision. The Accreditation Council for Graduate Medical Education (ACGME), Association of American Medical Colleges (AAMC), and American Association of Colleges of Osteopathic Medicine (AACOM) shall not be liable in any way for results obtained in applying these assessment methods. The user, and not the ACGME, AAMC, and AACOM, shall be responsible for the results obtained in applying the assessment methods described herein. Further, the user agrees and acknowledges that, in using the Guidebook, the user is solely responsible for complying with all applicable laws, regulations, and ordinances relating to privacy.

ACGME

AAMC
Tomorrow's Doctors, Tomorrow's Cures®

AACOM
American Association of Colleges of Osteopathic Medicine

Preface

This Assessment Guidebook is the result of a collaborative effort from volunteers dedicated to teaching and assessing medical students, residents, and fellows. It is arranged in several sections: ***Section I*** covers basic principles and approaches for developing and choosing valid assessment tools and methods; ***Section II*** provides practical descriptions of assessment methods that can be used in assessment. The description of each method or tool is based on key characteristics for "good assessment" (van der Vleuten, 1996; Norcini 2018); ***Section III*** provides guidance on mapping assessments to the Core Competencies, Foundational Competencies, and the Milestones, a crucial step in creating and managing a program of assessment, also referred to as programmatic assessment.

This Guidebook should be used in conjunction with the other resources listed below that are available on the **Milestones section of the ACGME website:**

1. Clinical Competency Committee Guidebook
2. Milestones Guidebook
3. Milestones Guidebook for Residents and Fellows
4. Milestones Implementation Guidebook

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Table of Contents

Section I

General Principles of Assessment 5
Recommended Assessment Methods by Competency 10
AI in Assessment 17

Section II

Assessment Tools and Methods 22
- Written and Oral Evaluations 22
  - Institutional Written Exams 27
  - Nationally Standardized Written Examinations 29
  - Oral Patient Presentation 34
- Performance-Based Assessments 37
  - Anatomy Lab Practical Exams 39
  - Assessment of Cost-Conscious Care 42
  - Clinical Performance (Record) Review 45
  - Evidence-Based Practice Logs 49
  - Procedure, Operative, or Clinical Case Logs 52
  - Review of Medical Errors and Patient Safety Events 56
  - Research and Scholarly Projects 60
  - Simulation 63
  - Standardized (Simulated) Patients (OSCEs) 66
- Workplace-Based Assessments 69
  - Direct Observation of Clinical Skills 72
  - Direct Observation of Procedural Skills 75
  - Chart-Stimulated Recall and the Assessment of Clinical Reasoning in the Workplace 78
  - Faculty Global Assessments 81
  - Narrative Assessment 85
  - One-Minute Preceptor 89
- Multisource Feedback (360-Degree Feedback) 92
  - Informed Self-Assessment 95
  - Patient Experience Surveys/Patient Feedback 97
  - Peer Assessment 101
  - Portfolio 104

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Section III

Implementation 107

The Role of the Milestones and Entrustable Professional Activities in Programmatic Assessment in GME 110

The Role of the Milestones and Entrustable Professional Activities in Programmatic Assessment in UME 111

Appendix: Van Melle's Core Components Framework for CBME 113

Section I

General Principles of Assessment

Importance of Assessment

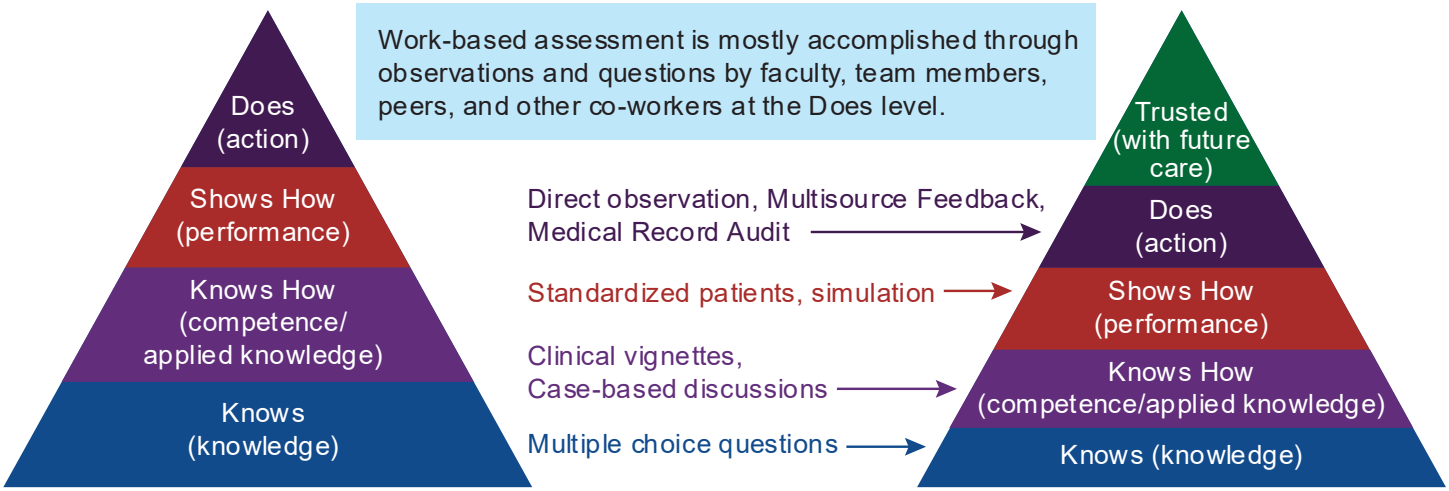
Assessment is vital to the effective professional development of learners, and yet, it is still often neglected, avoided, or performed poorly in medical education. A large body of research has repeatedly confirmed that accurate, robust assessment is essential for effective feedback, coaching, self-regulated learning, and professional growth. This includes assessment for learning and assessment of learning. Assessment culture in medical education has been slow to change despite the introduction of the ACGME Core Competencies and the push toward educational outcomes. The entire medical education system needs to accelerate cultural change to make assessment a valued activity. Valid assessment is a social responsibility between the learner, faculty members, educational programs, the patient, and society. When assessment is done well, learners can more effectively and quickly address gaps to improve and grow.

On the other hand, poor assessment practices can negatively impact learners in several ways. When strengths or weaknesses are incorrectly identified, learners may receive inappropriate feedback or support, leading to misguided development. Additionally, if assessments focus on the wrong learner abilities or needs, critical areas for growth may be overlooked, ultimately hindering progress. Finally, assessments that rely on implicit and explicit assumptions can further compromise accuracy and fairness, resulting in learners being misjudged and potentially disadvantaged in their educational journey. Ultimately, when assessment is done poorly, learners can graduate insufficiently prepared for the next stage of their education or unsupervised practice, and in a worst-case scenario, it can lead to patient harm.

Miller's Assessment Pyramid (Miller 1990) remains a helpful framework for building assessment systems, and while Miller's Pyramid is widely used, the idea of entrustment has been added to the top of the pyramid with the use of entrustment scales in the graphic below.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Assessing for the Desired Outcome



Ten Cate, O, et al, Entrustment Decision Making: Extending Miller's Pyramid. Academic Medicine. 96(2): 199-204, February 2021.

While the first three levels, Knows, Knows How, and Shows How, are important assessment approaches, programs, and institutions should place emphasis on the top of the pyramid, i.e., the Does level. The Does level requires attention to a robust combination of work-based assessments. Likewise, it is critical to recognize that the majority of all assessment is based on two primary activities: asking questions and observing. The Does level is also essential to inform the ultimate entrustment decision to graduate into unsupervised practice. How programs, institutions, and individuals perform these activities varies from assessment method to assessment method.

Utility and Good Assessment

Van der Vleuten (1996) introduced a Utility Index for evaluating the quality of an assessment and helping educators make appropriate tradeoffs when choosing an assessment method or tool. His original index can be represented as:

Utility = Rw x Vw x Aw x EIw x Cw

where

R = reliability; **V** = validity; **A** = acceptability; **EI** = educational impact; **C** = cost

w — the weight a program or institution might place on each element

One thing is quickly obvious: if any one of these five elements are zero, the utility of the tool is also zero. As educators think about what assessment approaches or tools to use, this simple formulation can be helpful.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Framework for Good Assessment: Single Assessments

More recently, an international group produced a list of criteria, often referred to as the Ottawa Criteria for good assessment (Norcini et al. 2018), to help educators judge the quality of assessments:

1. ***Validity or Coherence.*** The results of an assessment are appropriate for a particular purpose as demonstrated by a coherent body of evidence.
2. ***Reproducibility, Reliability, or Consistency.*** The results of the assessment would be the same if repeated under similar circumstances.
3. ***Equivalence.*** The same assessment yields equivalent scores or decisions when administered across different institutions or cycles of testing.
4. ***Feasibility.*** The assessment is practical, realistic, and sensible, given the circumstances and context.
5. ***Educational Effect.*** The assessment motivates those who take it to prepare in a fashion that has educational benefit.
6. ***Catalytic Effect.*** The assessment provides results and feedback in a fashion that motivates all stakeholders to create, enhance, and support education; it drives future learning forward.
7. ***Acceptability.*** Stakeholders find the assessment process and results to be credible.

As is clear to see, many of the elements of the Van der Vleuten Utility Index (1996) are included in Ottawa Criteria. Two differences, however, are important to note. First, cost has been included as a criterion of feasibility, given the largest cost to using an assessment is usually the time required of the assessor. Second, educational impact has been split into two effects, educational and catalytic. The catalytic effect is particularly important and a major focus and goal of the ACGME Milestones.

Primer on Validity

Much work on the validity of assessments has occurred over the past decades. For purposes of this Guidebook, the Messick framework (Messick 1989) is helpful and practical. Cook and colleagues (2006) produced a practical article that describes the Messick framework for medical education. The following example uses a direct observation of clinical care assessment to illustrate each element of the Messick framework.

Element	Example
Content Do items on the assessment completely and properly represent the construct of interest?	Does a tool for direct observation capture the proper domains or components of clinical skills, such as medical interviewing, physical examination, etc.?
Response process The relationship between the intended construct and the thought processes of subjects or observers	Have the faculty members been trained to use the direct observation method? Do the faculty members have a shared mental model of what they are being asked to rate?

Table continues on the next page…

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Element	Example
Internal structure Acceptable reliability and factor structure	Are the assessments reproducible? If faculty members were to re-rate the same patient encounter, would they agree with themselves (intra-rater reliability)? If another faculty member viewed the same encounter, would the two faculty members rate the encounter the same (inter-rater reliability)?
Relations to other variables Correlation with scores from another instrument assessing the same construct	Would direct observation ratings by faculty members correlate with the quality of care delivered to the patient?
Consequences Do scores really make a difference?	Are the direct observation ratings used by the competency committee or promotions committee? Did the learner receive feedback?

Programs of Assessment (Programmatic Assessment)

Assessment programs must be conceptualized through a systems lens. Simply defined, a system consists of two or more interdependent parts that work together to accomplish a shared aim. Assessment programs consist of multiple parts that are interdependent and need to interact effectively and efficiently to produce high-quality care and education within the teaching institution, and to ultimately prepare individuals for twenty-first century practice.

Designing a competency-based medical education (CBME) program using systems thinking requires an understanding of the core components of a CBME-designed program. Van Melle and associates (2019; see Appendix) have identified five core components of CBME:

- defining an outcomes-based competency framework;
- progressive sequencing of competencies within that framework;
- learning experiences tailored to competencies;
- teaching tailored to those competencies; and,
- effective programmatic assessment.

Effective programmatic assessment is essential to ensuring the desired outcome of CBME is achieved.

Effective programmatic assessment can be conceptualized as a subsystem of a program. Programmatic assessment consists of those individuals and groups who work together on a regular basis to perform assessments that enable valid determinations of learner progression towards competence. An effective program of assessment also shares: 1) agreed-upon goals and outcomes; 2) linked individual learner assessment and program evaluation processes; 3) information about learner's performance (both feedback and feed-forward mechanisms); and 4) the desire to produce someone fully prepared to enter a residency or fellowship program or unsupervised practice in the health care system to provide high-quality care. Done accurately and effectively, programmatic assessment optimizes learning, facilitates decision making regarding learner progression towards desired outcomes, and informs quality improvement activities.

No single assessment tool or method is sufficient to judge the overall abilities and readiness for learners to progress to the next stage. The assessment system should include a combination of

assessments to effectively cover the Foundational Competencies for Undergraduate Medical Education, ACGME Core Competencies, and the Milestones in their respective disciplines. In addition to using the above criteria for good assessment and the Van der Vleuten Utility Index, each entity needs to develop assessment map to ensure their assessment programs are robust.

Framework for Good Assessment: Systems of Assessment

Norcini and colleagues (2018) also provided the following guidance for assessment systems:

1. ***Coherent.*** The system of assessment is composed of multiple, coordinated individual assessments and independent performances that are orderly and aligned around the same purposes.
2. ***Continuous.*** The system of assessment is ongoing and individual results contribute cumulatively to system purposes.
3. ***Comprehensive.*** The system of assessment is inclusive and effective, consisting of components that are formative, diagnostic, and/or summative as appropriate to its purposes. Some or all components are authentic and integrative.
4. ***Feasible.*** The system of assessment and its components are practical, realistic, efficient, and sensible, given the purposes, stakeholders, and context.
5. ***Purpose-driven.*** The assessment system supports the purposes for which it was created.
6. ***Acceptable.*** Stakeholders in the system find the assessment process and results to be credible and evidence based.
7. ***Transparent and free from bias.*** Stakeholders understand the workings of the system and its unintended consequences are minimized. Decisions are fair and equitable.

Attention to these recommendations can help programs and institutions build and operate effective programs of assessment to achieve desired outcomes for the programs and institutions, the learner, and ultimately, the public. Creating and improving assessment systems requires a continuous quality improvement mindset. Not every assessment will work as planned; sometimes adjustments will be needed (e.g., faculty development) or a specific approach may need to be removed or changed. Assessment, like all fields, is ever changing as new research informs better practices. Programs and institutions will need to track these changes; this Guidebook is designed to help in that process.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Recommended Assessment Methods by Competency

Patient Care and Procedural Skills

Direct observation is the primary and overarching method to assess patient care and procedural skills. Direct observation can be accomplished through observation of real clinical encounters or through simulation. Simulation can be a powerful approach to teach, observe, assess, and provide feedback to learners. Simulation allows the program to control and standardize both content and context (i.e., the "Shows How" level of Miller's Pyramid [1990]). A mastery-based approach is strongly recommended when using any form of simulation to teach and assess skills, as research has demonstrated this approach to procedural skills training with robust assessment translates into better outcomes for patients (McGaghie et al. 2020).

Direct observation of patient encounters, procedures, family meetings, and other clinical activities is essential, and many tools exist that programs and institutions can use to capture and document these observations (i.e., the "Does" level of Miller's Pyramid [1990]). These tools range from single encounter tools (e.g., mini-clinical evaluation exercise [mini-CEX], objective structured assessment of technical skills) to shift or daily encounter cards, to faculty evaluation forms that serve as a synthesis and compilation of observation over a clinical rotation. To use these assessment tools effectively, a shared mental model of the intent of the tool, meaning of the rating, and impact on the learner must be created by the faculty. Faculty development on assessment and standards for the domains of interest is crucial. When appropriate and with proper consent procedures in place, video recordings can be highly useful for assessment and feedback.

Clinical performance audits of patient outcomes, medical records, discharge summaries, operative notes, and other similar documents are useful for examining performance across specific conditions for groups of patients over time. Most specialties now possess at least some quality and safety performance measures that can be used for this purpose. Quality and safety data are also essential in assessing Systems-Based Practice.

Medical Knowledge

Multiple choice tests have long been used to assess the capability of learners at the "Knows" and "Knows How" levels of Miller's Pyramid (1990). Some standardized knowledge assessments target the "Shows How" level. The most common standardized assessment in GME is the in-training examination, and for UME the licensure exams (National Board of Medical Examiners [NBME] and National Board of Osteopathic Medical Examiners [NBOME]). These in-training examinations have predictive validity for performance on certification examinations.

When assessing medical knowledge, distinguishing between acquisition of knowledge and application of knowledge is critical. When assessing the *acquisition of* medical knowledge, the outcome is to document clinically applicable knowledge of the basic and clinical sciences that underlie the practice of medicine. When assessing the *application of* medical knowledge, the goal is to assess the ability to apply that knowledge to clinical problem solving. These abilities are collectively described as clinical reasoning, the process wherein clinicians observe, collect, and interpret data to diagnose and manage patients.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Components of Clinical Reasoning

- Information gathering
- Hypothesis generation
- Problem representation
- Differential diagnosis
- Leading or working diagnosis
- Diagnostic justification
- Management

Assessment of acquired knowledge is foundational to any assessment program but while essential, it alone is insufficient for determining competence in medical knowledge. Learners must also be able to apply that knowledge in the clinical setting.

Daniel and associates (2019) have categorized the assessment of clinical reasoning into non-workplace-based assessment, assessment in simulated clinical environments, and workplace-based assessment, including descriptions of assessment methods appropriate to each category. In the clinical setting, medical knowledge (both acquired and applied) can also be simultaneously assessed using clinical questioning. Such questioning can be accomplished with multiple formats. Assessment using chart audit and chart-stimulated recall allow for structured clinical questioning and are addressed separately in this guidebook. In-person meetings, such as morning report and morbidity and mortality conferences, also allow for structured questioning.

Clinical questioning is a standard activity during patient rounding, but such questioning is infrequently summarized in writing. A classic questioning strategy in this setting is the one-minute preceptor model. The components of the one-minute preceptor are listed below and can be used to probe a learner's medical knowledge. Faculty members can use any or all of the steps of this time-efficient model in busy clinical settings and use this approach to clinical questioning with all levels of learners.

Clinical reasoning can also be assessed using methods such as the "think aloud" or "Summarize, Narrow, Analyze, Probe, Plan, Select" (SNAPPS) (Wolpaw et al. 2009). These methods prompt the learner to discuss how they arrived at the proposed action while allowing for the assessment of clinical reasoning and the delivery of immediate feedback.

Assessing clinical reasoning and medical knowledge in the clinical workspace ("in vivo" assessment) has grown in importance. A primary reason is the continued and pernicious problem of diagnostic error. As defined by the ACGME in its Common Program Requirements, 4.6, Medical Knowledge extends to include "the application of this knowledge to patient care," in other words, clinical reasoning, is a complex process that is affected by numerous contextual factors. Therefore, assessing clinical reasoning requires faculty members and others to be expert questioners using evidence-based theory. Structured assessment tools, such as chart-stimulated recall and the Assessment of Reasoning Tool, directly target clinical reasoning.

Professionalism

Effective assessment of professionalism requires, at a minimum, a multisource feedback approach. A complete multisource feedback assessment approach must include patient experience surveys, and when appropriate, family experience surveys. While faculty members can assess what they observe with regards to professional behaviors, and several tools exist for

this purpose, faculty members in general are poorer observers and assessors of professionalism unless critical incidents or deficiencies are displayed in their presence. For example, multiple studies have found that professionalism assessments done by physician faculty members often do not correlate with nursing assessments of learners. Ideally, assessment of professionalism goes beyond the multisource feedback and also includes direct observation, simulation, and role play.

Interpersonal and Communication Skills

Effective assessment of interpersonal and communication skills also requires a multisource feedback approach. This is especially critical for assessing interprofessional teamwork. A complete multisource feedback assessment approach must include patient experience surveys, and when appropriate, family experience surveys. Faculty members can assess learners' communication skills with patients and families through assessment tools discussed under Patient Care and Procedural Skills. Direct observation by faculty members, combined with multisource feedback, can provide useful data on this important competency, and standardized patients (SPs) and simulations, using a mastery-based approach, can be used to assess capability in it.

Practice-Based Learning and Improvement

Programs and institutions can assess learners' ability to review, synthesize, and apply literature for patient care by evaluating articles using research quality criteria. A number of tools also exist to judge a learner's ability to perform an evidence-based practice review. Evidence-based practice case logs, using the patient-intervention-comparator-outcome format are an effective tool for capturing application of evidence-based practice skills in patient care (the "Does" of Miller's Pyramid [1990]).

For reflective practice, informed self-assessment and creation of learning goals and individualized learning plans (ILPs) are essential. Informed self-assessment involves learners using their assessment, feedback, and clinical performance data for individual continuous quality improvement. ILPs enable learners to codify their areas for improvement and continued growth and can be tracked over time.

Systems-Based Practice

Systems-based practice focuses on patient safety and quality improvement, and system navigation for patient-centered care, including coordination and transition of care, social and structural drivers of health, local population health, health policy, and emerging technologies. Use of patient safety and quality of care performance data is essential for assessing this competency. It is not possible to effectively assess Systems-Based Practice without clinical performance data. While attribution of clinical performance data to any single learner can be difficult, programs and institutions can use the lens of contribution to review the learner's role in performance measure results. Research has also recently shown that programs and institutions can use a systematic approach to choose quality and safety measures that possess a strong connection to the learner's actions (Schumacher et al. 2019).

A multisource feedback approach is essential for system navigation in patient-centered care. The Institute for Healthcare Improvement Open School is one example of an educational program for assessing knowledge in systems. Chart audit, along with review of sign-out tools and institutional examinations, can round out assessment for these topics. GME programs may also use one of their specialty society's resources, if available.

Building programmatic assessment requires the development of an integrated combination of assessment methods and tools for determining a learner's developmental progression in each of the six Foundational Competencies for Undergraduate Medical Education/ACGME Core Competencies. The tables below provide basic menus of assessment tools/methods appropriate for use in each Foundational Competencies/Core Competency domain.

By Competency

Competency	Competency-Based Assessment Options
Medical Knowledge	• In-training and written examinations • Faculty global assessments • Chart-stimulated recall • Clinical performance (record) review
Patient Care and Procedural skills	• Work-based clinical assessment through direct observation of the individual during care delivery • Faculty and peer assessment • Standardized assessments • Simulation • One-Minute Preceptor • Oral exams • Direct observation
Professionalism	• Informed self-assessment • Multisource feedback, such as a 360-degree evaluation • Patient experience surveys
Interpersonal and Communication Skills	• Patient-reported feedback and experience surveys • Multisource feedback, such as a 360-degree evaluation, especially regarding interprofessional care • Direct observation • Simulation
Practice-based Learning and Improvement	• Evaluation of knowledge, skills, and attitudes from participation in systematic efforts to improve the quality, safety, or value of health care services • Audit and feedback of the medical record • Review of medical errors and patient safety events • Evidence-based practice logs • Research and scholarly projects
Systems-based Practice	• Feedback from multiple faculty evaluations regarding the ability to practice in a complex health care system • Multisource feedback, such as a 360-degree evaluation, especially regarding interprofessional care • Assessment of cost-conscious care • Clinical performance (record) review • Review of medical errors and patient safety events

©2026 Accreditation Council for Graduate Medical Education (ACGME)

By Assessment Tool/Method

Assessment Tool/Method	Targeted Competency
Faculty member assessment (can be interprofessional)	Multiple competencies
Direct observation	• Patient Care and Procedural Skills • Interpersonal and Communication Skills • Medical Knowledge (“in vivo”) • Professionalism
Multisource feedback	• Professionalism • Interpersonal and Communication Skills • Systems-Based Practice • Medical Knowledge
Audit and performance data (clinical and patient safety indicators)	• Patient Care and Procedural Skills • Practice-Based Learning and Improvement • Systems-Based Practice • Medical Knowledge
Simulation (if available)	• Patient Care and Procedural Skills • Interpersonal and Communication Skills • Medical Knowledge
Written and In-training exam (if available)	• Medical Knowledge
Case or procedural logs	• Patient Care and Procedural Skills • Practice-Based Learning and Improvement
Patient experience surveys	• Patient Care and Procedural Skills • Interpersonal and Communication Skills • Professionalism

©2026 Accreditation Council for Graduate Medical Education (ACGME)

References

Cook, David A. and Thomas J. Beckman. 2006. “Current Concepts in Validity and Reliability for Psychometric Instruments: Theory and Application.” *The American Journal of Medicine* 119 (2): 166.e7-166.e-16. https://www.sciencedirect.com/science/article/abs/pii/S0002934305010375.

Daniel, Michelle, Joseph Rencic, Steven J Dunning, et al. 2019. “Clinical Reasoning Assessment Methods: A Scoping Review and Practical Guidance.” *Academic Medicine* 94(6): 902-912. https://pubmed.ncbi.nlm.nih.gov/30720527/.

Edgar, Laura, M. Douglas Jones Jr., Braden Harsy, Morgan Passiment, and Karen E. Hauer. 2021. “Better Decision-Making: Shared Mental Models and the Clinical Competency Committee.” *J Grad Med Educ* 13(2 Suppl):51–8.

Frank, Jason R., Linda S. Snell, Olle ten Cate, et al. 2010. “Competency-Based Medical Education: Theory to Practice.” *Medical Teacher* 32 (8): 638–645. https://pubmed.ncbi.nlm.nih.gov/20662574/.

Messick, Samuel. 1989. “Validity.” In *Educational Measurement,* edited by R.L. Linn. American Council on Education/Macmillan Publishing Co., Inc.

McGaghie, William C., Jeffrey H. Barsuk, and Diane B. Wayne. 2020. *Comprehensive Healthcare Simulation: Mastery Learning in Health Professions Education.* Springer.

Miller, G.E. 1990. “The Assessment of Clinical Skills/Competence/Performance.” *Academic Medicine* 65 (9 Suppl.): S63-67. https://pubmed.ncbi.nlm.nih.gov/2400509/.

Norcini, John, M. Brownell Anderson, Valdes Bollela, et al. 2018. “2018 Consensus Framework for Good Assessment.” *Medical Teacher* 40: (11): 1102-1109. https://pubmed.ncbi.nlm.nih.gov/30299187/.

Schumacher, Daniel J., Danny T.Y. Wu, Karthikeyan Meganathan, et al. 2019. “A Feasibility Study to Attribute Patients to Primary Interns on Inpatient Ward Teams Using Electronic Health Record Data.” *Academic Medicine* 94 (9): 1376-1383. https://pubmed.ncbi.nlm.nih.gov/31460936/.

Ten Cate, O. et al. 2021. “Entrustment Decision Making: Extending Miller’s Pyramid.” *Academic Medicine* 96(2):199-204.

Van Der Vleuten, C.P. 1996. “The Assessment of Professional Competence: Developments, Research and Practical Implications.” *Advances in Health Sciences Education* 1 (1): 41-67. https://pubmed.ncbi.nlm.nih.gov/24178994/.

van der Vleuten C.P.M., L.W. T. Schuwirth, E.W. Driessen, et al. 2012. “A Model for Programmatic Assessment Fit for Purpose.” *Medical Teacher* 34 (3): 205–214. https://pubmed.ncbi.nlm.nih.gov/22364452/.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Van Melle, Elaine, Jason R. Frank, Eric S. Holmboe, et al. 2019. “A Core Components Framework for Evaluating Implementation of Competency-Based Medical Education Programs.” *Academic Medicine* 94 (7): 1002-1009. https://pubmed.ncbi.nlm.nih.gov/30973365/.

Wolpaw, Terry, Klara K. Papp, and Georges Bordage. 2009. “Using SNAPPS to Facilitate Expression of Clinical Reasoning and Uncertainties: A Randomized Comparison Group Trial.” *Academic Medicine* 84 (4): 517-524. https://pubmed.ncbi.nlm.nih.gov/19318792/.

AI in Assessment

Artificial intelligence (AI) is increasingly embedded across assessment modalities in health professions education, primarily as an augmentation layer that enhances scoring, feedback, and pattern detection. Its strongest current contributions lie in text-based analysis, simulation environments, and aggregation of longitudinal performance data, while high-stakes autonomous decision-making remains limited.

Written Tests

Written tests (e.g., institutional written exams, in-training exams) are foundational tools for assessing knowledge and clinical reasoning. AI is most mature in this domain, particularly for constructed-response scoring and item generation. AI applications include automated item generation using large language models (LLMs), short-answer scoring via natural language processing (NLP), and semantic similarity models to compare learner responses with expert exemplars. Machine learning algorithms can also perform distractor analysis and psychometric anomaly detection, identifying poorly functioning items. Predictive analytics may support performance forecasting and adaptive testing. Benefits are substantial: scalability in grading, rapid turnaround for feedback, and improved consistency relative to human raters. AI can also enhance feedback quality by identifying missing concepts and generating tailored explanations. These capabilities reduce faculty burden and enable more frequent formative assessment. However, risks are nontrivial. Construct-irrelevant variance may arise when models reward verbosity or writing style rather than conceptual accuracy. Bias can occur against noncanonical phrasing or non-native English speakers. LLM-generated rationales may be plausible but incorrect (“hallucinations”), and transparency is often limited. For high-stakes use, human oversight, audit trails, and validation studies remain essential.

Oral Exams

Oral exams (e.g., chart-stimulated recall and the assessment of clinical reasoning in the workplace, oral patient presentation) assess clinical reasoning, communication, and real-time decision-making. AI applications here are emerging and largely remain in initial phases. Key methods include speech recognition for transcription, NLP and LLM-based discourse analysis, and argument-mining techniques to map reasoning structure. Prosody analytics can evaluate aspects of delivery such as fluency and pacing. These tools enable automated transcript generation, rubric-assisted scoring, and structured feedback on reasoning and organization.

The primary benefit is enhanced documentation and analytic depth. AI can reveal patterns in reasoning, identify gaps in argumentation, and support more standardized evaluation across examiners. It also reduces administrative burden by automating transcription and initial scoring. Limitations are significant. Speech recognition systems may introduce bias related to accent, dialect, or speech rate. AI may overemphasize linguistic fluency rather than clinical reasoning, introducing construct contamination. Additionally, oral exams are highly interactive; current models struggle to capture examiner probing and contextual nuance. As a result, AI is best positioned as a support tool rather than a primary evaluator in this domain.

Performance Tests

Performance tests (e.g., procedural skills, simulation-based assessments, anatomy practical exam) evaluate technical and applied competencies. AI applications here rely heavily on multimodal data. Computer vision, sensor fusion, and time-series machine learning are central methods. These systems analyze video and motion data to detect task steps, measure

efficiency, and identify errors. In simulation environments, AI can track performance trajectories and generate automated feedback. For anatomy practical exam, image recognition models can assess structure identification. Benefits include objective, fine-grained measurement of performance, particularly for procedural skills. AI can capture metrics (e.g., motion efficiency, error rates) that are difficult for human observers to quantify, enabling more precise feedback and longitudinal tracking. Scalability and standardization are additional advantages. Risks include limited generalizability across contexts and tasks. Models trained on specific procedures or environments may not transfer well. There is also potential for "shortcut learning," in which AI identifies superficial patterns rather than meaningful performance indicators. Bias may arise from differences in equipment, case complexity, or recording conditions. Transparency is often limited, raising concerns about interpretability in high-stakes decisions.

Workplace-Based Assessments

Workplace-based assessments, including direct observation, narrative assessments, and clinical performance reviews, evaluate learners in authentic clinical settings. AI applications in this domain are rapidly evolving, particularly for narrative data. NLP and LLMs are used to classify narrative comments into competency frameworks, extract themes, and assess feedback quality. Machine learning models can identify progression patterns and generate performance dashboards. Predictive analytics may flag learners at risk or detect inconsistencies in assessment data.

Benefits center on synthesis and scalability. AI can aggregate large volumes of qualitative data, transforming fragmented narratives into coherent insights. This enhances feedback quality, supports programmatic assessment, and enables earlier identification of performance concerns. AI can also evaluate the quality of feedback itself, promoting faculty development. However, workplace-based assessments are inherently context-dependent, and AI models may struggle with ecological variability. Bias embedded in assessor narratives can be amplified by AI systems. Narrative sparsity and variability in documentation quality further complicate analysis. There is also a risk of overinterpretation, where probabilistic outputs are treated as definitive judgments. Maintaining traceability to original data and incorporating human judgment are critical safeguards.

Multisource Feedback (360º Feedback)

Multisource feedback integrates perspectives from peers, supervisors, patients, and other health care professionals to assess domains such as professionalism, communication, and teamwork. AI is increasingly used to manage and interpret the large volume of data generated. Core methods include NLP for text analysis, LLM-based summarization, clustering algorithms to identify themes, and anomaly detection to highlight discrepancies across sources. Fairness analytics can evaluate potential bias in feedback patterns. The primary benefit is efficient synthesis. AI can transform hundreds of comments into structured, actionable reports, highlighting strengths, areas for improvement, and discrepancies between self and external assessments. This enhances usability for both learners and educators and supports longitudinal tracking.

Risks include source imbalance and bias. Feedback may reflect social desirability, interpersonal dynamics, or systemic inequities. AI systems may inadvertently amplify these biases if not carefully audited. De-identification challenges also arise, particularly in small teams, raising concerns about confidentiality and psychological safety. Transparency in how summaries are generated and ensuring linkage to source data are essential for trust and validity.

In summary, AI is reshaping assessment across modalities by enhancing scalability, consistency, and analytic depth. Its most robust applications lie in text and data aggregation, while multimodal and high-stakes use remain developmental. Across all domains, the central paradigm remains intelligence augmentation, supporting, rather than replacing, expert human judgment.

Selected Literature

Written Exams

Başaranoğlu, Mert, Erdem Akbay, and Erim Erdem. 2025. “AI-Generated Questions for Urological Competency Assessment: A Prospective Educational Study.” *BMC Medical Education* 25, no. 1: 611.

Cianciolo, Anna T., Noelle LaVoie, and James Parker. 2021. “Machine Scoring of Medical Students’ Written Clinical Reasoning: Initial Validity Evidence.” *Academic Medicine* 96, no. 7: 1026-1035.

Grévisse, Christian. 2024. “LLM-Based Automatic Short Answer Grading in Undergraduate Medical Education.” *BMC Medical Education* 24, no. 1: 1060.

Seneviratne, H.M.T.W., and S.S. Manathunga. 2025. “Artificial Intelligence Assisted Automated Short Answer Question Scoring Tool Shows High Correlation with Human Examiner Markings.” *BMC Medical Education* 25, no. 1: 1146.

Oral Exams

Eachempati, Prashanti, Ramnarayan Komattil, and Arathi Arakala. 2025. “Should Oral Examination Be Reimagined in the Era of AI?” *Advances in Physiology Education* 49, no. 1: 208-209.

Panni, Usman Y., Christa Donald, Jeffrey A. Blatnik, Michael Williams Jr., Jennifer Yu, and Paul E. Wise. 2025. “Evaluating the Quality of AI-Written Scenarios for Virtual Oral Surgical Board Preparatory Examination.” *Journal of Surgical Education* 82, no. 12: 103736.

Performance Tests

Bansal, Saurabh K., Manajyoti Yadav, Jianing Zhou, et al. 2025. “Medical Students' Perception of Automated Note Feedback After Simulated Encounters.” *The Clinical Teacher* 22, no. 6: e70273.

Goldbraikh, Adam, Anne-Lise D’Angelo, Carla M. Pugh, and Shlomi Laufer. 2022. “Video-Based Fully Automatic Assessment of Open Surgery Suturing Skills.” *International Journal of Computer Assisted Radiology and Surgery* 17, no. 3: 437-448.

Holderried, Friederike, Christian Stegemann-Philipps, Anne Herrmann-Werner, et al. 2024. “A Language Model–Powered Simulated Patient with Automated Feedback for History Taking: Prospective Study.” *JMIR Medical Education* 10, no. 1: e59213.

Pedrett, Romina, Pietro Mascagni, Guido Beldi, Nicolas Padoy, and Joël L. Lavanchy. 2023. “Technical Skill Assessment in Minimally Invasive Surgery Using Artificial Intelligence: A Systematic Review.” *Surgical Endoscopy* 37, no. 10: 7412-7424.

Siddiqui, Asad, Zhoujie Zhao, Chuer Pan, Frank Rudzicz, and Tobias Everett. 2023. “Deep Learning Model for Automated Trainee Assessment During High-Fidelity Simulation.” *Academic Medicine* 98, no. 11: 1274-1277.

Yokose, Masashi, Takanobu Hirosawa, Tetsu Sakamoto, et al. 2025. “The Validity of Generative Artificial Intelligence in Evaluating Medical Students in Objective Structured Clinical Examination: Experimental Study.” *JMIR Formative Research* 9: e79465.

Workplace-Based Assessments

Booth, Gregory J., Benjamin Ross, William A. Cronin, et al. 2023. “Competency-Based Assessments: Leveraging Artificial Intelligence to Predict Subcompetency Content.” *Academic Medicine* 98, no. 4: 497-504.

Kwan, Benjamin Y.M., Zier Zhou, Nick Rogoza, et al. 2025. “Leveraging Large Language Models to Evaluate the Quality of Narrative Feedback for Surgery Residents in Competency-Based Medical Education.” *Annals of Surgery Open* 6, no. 4: e608.

Maimone, Christina, Brigid M. Dolan, Marianne M. Green, et al. 2023. “Utilizing Natural Language Processing of Narrative Feedback to Develop a Predictive Model of Pre-Clerkship Performance: Lessons Learned.” *Perspectives on Medical Education* 12, no. 1: 141.

Yilmaz, Yusuf, Alma Jurado Nunez, Ali Ariaeinejad, Mark Lee, Jonathan Sherbino, and Teresa M. Chan. 2022. “Harnessing Natural Language Processing to Support Decisions Around Workplace-Based Assessment: Machine Learning Study of Competency-Based Medical Education.” *JMIR Medical Education* 8, no. 2: e30537.

Multisource Feedback (360º Feedback)

Olex, Amy L., Adam M. Garber, Sally A. Santen, Courtney Blondino, Stephanie Goldberg, and Deborah DiazGranados. 2025. “Utilizing Natural Language Processing to Analyze Student Narrative Reflections for Medical Curriculum Improvement.” *Medical Teacher* 47, no. 6: 970-975.

Turner, Laurah, Danielle E. Weber, Sally A. Santen, et al. 2024. “Making Use of Natural Language Processing to Better Understand Medical Students’ Self-Assessment of Clinical Skills.” *Academic Medicine* 99, no. 3: 285-289.

Reviews

Alizadeh, Majid, and Maryam Jafar Sameri. 2025. “Intelligent Assessment Systems in Medical Education: A Systematic Review.” *Journal of Advances in Medical Education & Professionalism* 13, no. 3: 173.

Feigerlova, Eva, Hind Hani, and Ellie Hothersall-Davies. 2025. “A Systematic Review of the Impact of Artificial Intelligence on Educational Outcomes in Health Professions Education.” *BMC Medical Education* 25, no. 1: 129.

Fuller, Richard, Viktoria C.T. Goddard, Vishna D. Nadarajah, et al. 2022. “Technology Enhanced Assessment: Ottawa Consensus Statement and Recommendations.” *Medical Teacher* 44, no. 8: 836-850.

Gin, Brian C., Patricia S. O'Sullivan, Karen E. Hauer, et al. 2025. "Entrustment and EPAs for Artificial Intelligence (AI): A Framework to Safeguard the Use of AI in Health Professions Education." *Academic Medicine* 100, no. 3: 264-272.

Masters, Ken, Heather MacNeil, Jennifer Benjamin, et al. 2025. "Artificial Intelligence in Health Professions Education Assessment: AMEE Guide No. 178." *Medical Teacher* 47, no. 9: 1410-1424.

Morjaria, Leo, Bhavya Gandhi, Levi Burns, and Matthew Sibbald. 2025. "Generative AI: A Disruptor to Health Professions Learner Assessment." *Cureus* 17, no. 11.

Turner, Laurah, Daniel A. Hashimoto, Shubha Vasisht, and Verity Schaye. 2024. "Demystifying AI: Current State and Future Role in Medical Education Assessment." *Academic Medicine* 99, no. Suppl 1: S42-S47.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Section II

Assessment Tools and Methods

Written and Oral Evaluations

Description

Written and oral evaluations are assessment methods used to evaluate a learner's knowledge, reasoning, and decision-making under structured conditions. Written evaluations are typically paper-or computer-based examinations composed of predetermined questions that do not change during administration. They may include selected-response formats (e.g., multiple choice, true/false, or matching) and constructed-response formats (e.g., short answer or essay questions).

Oral evaluations can be conducted in a controlled testing environment or when an opportunity arises (e.g., scrubbing for a procedure, preparing for the next patient); and allow learners to provide open-ended, spoken responses. They often involve a hypothetical or authentic clinical case and use structured questioning to elicit clinical reasoning, such as forming a differential diagnosis, selecting diagnostic tests, and making management decisions. Unlike written evaluations, oral evaluations allow the examiner to ask follow-up questions and adapt probing based on the learner's responses to more fully assess depth of understanding and reasoning processes. Learner performance in formal oral evaluations is scored using a rubric, and narrative feedback may be provided verbally, in writing, or both.

Together, written and oral evaluations provide complementary approaches to assessing knowledge acquisition, clinical reasoning, and application of principles to defined and evolving problems.

Competencies

Written tests are most commonly used to assess Medical Knowledge, measuring the "Knows" and "Knows How" levels of Miller's Pyramid (1990). Practice-Based Learning and Improvement, Systems-Based Practice, and Patient Care and Procedural Skills can also be assessed using written tests. Oral exams can be used to assess all six competency domains. They are most commonly used to assess Medical Knowledge and Patient Care and Procedural Skills. Assessment of clinical reasoning is a common target for this assessment.

Validity

Validity evidence for written tests can be high, especially for those developed and administered by national organizations. Several studies have shown performance on in-training exams can predict subsequent performance on the associate board certifying exam (Webb et al.1996; Baumgartner and Peterman 1996; Beister 1987).

Validity starts with the examination design. There are several activities to consider, each having a different level of importance depending on the purpose and type of the exam.

- ***Blueprinting:*** Align exam content with curriculum objectives and competency domains.
- ***Item-Writing Best Practices*:** Use case-based and clinically relevant questions, avoid ambiguous language, and ensure a single best answer.

- ***Standard Setting:*** Use evidence-based methods such as Angoff or Hofstee to define passing thresholds.
- ***Continuous Quality Improvement:*** Conduct regular psychometric analyses and faculty development workshops to improve item writing and exam integrity.

By using best practices in exam design, psychometric evaluation, and feedback mechanisms, institutional written exams can serve as robust tools for assessing and enhancing medical education outcomes.

There is limited validity evidence for oral exam. This limited validity is compounded by heterogeneity in how these exams are structured and implemented. Their inter-rater reliability is typically moderate. Since these exams are case-specific, multiple cases are needed in order to make reliable summative assessment decisions. Oral exams often have low to moderate correlation with other assessments, such as faculty global assessments and written exams.

Item Analysis and Psychometric Properties

For written examinations, especially high-stakes exams, rigorous psychometric evaluation is necessary. As faculty and institutions build multiple choice or other written examinations, especially for summative purposes, they should consider these parameters and seek assistance from an assessment expert. This may become more relevant as AI is used to build exams. Key parameters include:

- ***Item Difficulty (P-value):*** Represents the proportion of individuals who answer an item correctly. Values range from 0 (very difficult) to 1 (very easy), with an optimal range of 0.3-0.7 for most high-stakes exams to ensure meaningful differentiation of individual performance.
- ***Item Discrimination (Point-Biserial Correlation):*** Measures how well an item differentiates between high- and low-performing individuals. A positive correlation (typically >0.2) indicates that individuals with higher overall scores are more likely to answer the item correctly, supporting its validity.
- ***Classical Test Theory (CTT) Considerations:*** CTT provides statistical indices such as reliability (Cronbach's alpha) to measure internal consistency, standard error of measurement (SEM) to estimate precision, and total test score distributions to assess overall difficulty and fairness.
- ***Item Response Theory (IRT) Considerations:*** IRT models, such as Rasch, two-parameter logistic (2PL), and the three-parameter logistic (3PL) model, allow a more nuanced understanding of item difficulty, discrimination, and guessing effects. IRT enhances the precision of score interpretation and allows for adaptive testing approaches.

Feasibility

Feasibility for written tests is typically high, especially for selected response formats. The resources needed to administer and score these tests is usually minimal, most notably for in-training exams and other written tests administered by national organizations. Feasibility for written tests with constructed response formats may be a little lower, as more time is needed to score these responses. Oral exams can be used in a variety of courses and settings in the curriculum, from preclinical to clinical. Significant faculty development is required to standardize exam administration and assessment.

Acceptability

Acceptability by faculty members of written tests is high, as they are often administered by national organizations or at the course/department level (so that individual faculty member effort is minimal or nonexistent). Acceptability by learners varies, as some may be better performers on written tests, especially high-stakes exams. Some faculty members may already be familiar with oral exams through their own board certification process, though their familiarity varies by specialty. Generally, acceptability by learners is good. Some learners appreciate the opportunity to talk through their understanding of a clinical case or a topic and may even prefer this method of assessment compared to written exams. Other learners may be intimidated by the oral format and require more preparation than is typically required for a written exam.

Catalytic Effect

Catalytic effect of written tests can be variable. Results of written tests can be used to assist in making decisions regarding an individual's success in attaining curricular objectives or to provide feedback regarding an examinee's knowledge relative to a defined standard. However, while written tests can measure a learner's knowledge level, they do not necessarily predict the application of that knowledge in real clinical settings.

The catalytic effect of oral exams is dependent on the narrative feedback provided by the examiner. With rich narrative feedback, the learner may be prompted to enhance their learning or approach to clinical reasoning. The fact that this assessment is grounded in a patient case may promote transference of the learning to future patient cases. The catalytic effect may also be affected by whether the assessment is used primarily for formative or summative purpose.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

References

Baumgartner, B.R. and S.B. Peterman. 1996. “Relationship Between American College of Radiology In-Training Examination Scores and American Board of Radiology Written Examination Scores.” *Academic Radiology* 3 (10): 873-878. https://pubmed.ncbi.nlm.nih.gov/8923908/.

Biester, T.W. 1987. “The American Board of Surgery In-Training Examination as a Predictor of Success on The Qualifying Examination.” *Current Surgery* 44 (3): 194-198. https://pubmed.ncbi.nlm.nih.gov/3595199/.

Burr, Steven A., John Whittle, Lucy C. Fairclough, Lee Coombes, and Ian Todd. “Modifying Hofstee Standard Setting for Assessments that Vary in Difficulty, and to Determine Boundaries for Different Levels of Achievement.” 2016. *BMC Medical Education* 16: 34. https://pmc.ncbi.nlm.nih.gov/articles/PMC4731915/.

Caldwell, Katharine E., Jorge G. Zarate Rodriguez, Annie Hess, Britta J. Han, Michael M. Awad, and Bethany C. Sacks. 2022. “Standardized Oral Examinations Allow for Assessment of Medical Student Clinical Knowledge and Decrease Racial Grading Differences in a Surgery Clerkship.” *Surgery* 171 (3): 590-597. https://pubmed.ncbi.nlm.nih.gov/34895772/.

Downing, Steven M. 2005. “The Effects of Violating Standard Item Writing Principles on Tests and Students: The Consequences of Using Flawed Test Items on Achievement Examinations in Medical Education.” *Advances in Health Sciences Education: Theory and Practice* 10: 133-143. https://link.springer.com/article/10.1007/s10459-004-4019-5.

Goins, Stacy M., Robert J. French, and Jonathan G. Martin. 2023. “The Use of Structured Oral Exams for the Assessment of Medical Students in their Radiology Clerkship.” *Curriculum Problems in Diagnostic Radiology* 52 (5): 330-333. https://pubmed.ncbi.nlm.nih.gov/37032291/.

Hawkins, Richard E. and David B. Swanson. 2017. “Using Written Examinations to Assess Medical Knowledge and Its Application.” In *Practical Guide to the Evaluation of Clinical Competence* by Eric S. Holmboe, Steven J. Durning, and Richard E. Hawkins. Elsevier.

Jeffery, Michelle, A. Reema Kar, Archana Pradhan, et al. 2024. “Evaluating Clinical Reasoning in Undergraduate Medical Education: The Value of a Virtual Oral Assessment.” *The American Surgeon* 90 (11): 2756-2761. https://pubmed.ncbi.nlm.nih.gov/38676698/.

Miller, G.E. 1990. “The Assessment of Clinical Skills/Competence/Performance.” *Academic Medicine* 65 (9 Suppl.): S63-S67. https://pubmed.ncbi.nlm.nih.gov/2400509/.

O’Donnell, M.J., S.S. Obenshain, and J.B. Erdmann. 1993. “Background Essential to the Proper Use of Results of Step 1 and Step 2 of the USMLE.” *Academic Medicine* 68 (10): 734-739. https://doi.org/10.1097/00001888-199310000-00002.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Pernar, Luise I.M., Reza Askari, and Elizabeth M. Breen. 2020. “Oral Examinations in Undergraduate Medical Education - What Is the ‘Value Added’ to Evaluation?” *American Journal of Surgery* 220 (2): 328-333. https://pubmed.ncbi.nlm.nih.gov/31918844/.

Rajeswaran, Vamana, Luke Devine, Edmund Lorens, Sumitra Robertson, Ella Huszti, and Daniel M. Panisko. 2022. “Types of Clinical Reasoning in a Summative Clerkship Oral Examination.” *Medical Teacher* 44 (6): 657-663. https://pubmed.ncbi.nlm.nih.gov/35000527/.

Rauh, Jessica L., Alison J. Lehane, and Leah M. Sieren. 2023. “Oral ‘Board Style’ Examination for Surgery Clerkship.” *The American Surgeon* 89 (7): 3251-3252. https://pubmed.ncbi.nlm.nih.gov/36815220/.

Rudolph, Michael J., Kimberly K. Daugherty, Mary Elizabeth Ray, Veronica P. Shuford, Lisa Lebovitz, and Margarita M. DiVall. 2019. “Best Practices Related to Examination Item Construction and Post-Hoc Review.” *American Journal of Pharmaceutical Education* 83 (7): 7204. https://pubmed.ncbi.nlm.nih.gov/31619832/.

Saab, Said S., Staci Pollack, Veronica Lerner, Erika Banks, Catherine R. Salva, and Jorie Colbert-Getz. 2023. “Validity Study of an End-of-Clerkship Oral Examination in Obstetrics and Gynecology.” *Journal of Surgical Education* 80 (2): 294-301. https://pubmed.ncbi.nlm.nih.gov/36266228/.

Shulruf, Boaz, Tim Wilkinson, Jennifer Weller, Philip Jones, and Phillipa Poole. 2016. “Insights into the Angoff Method: Results from a Simulation Study.” *BMC Medical Education* 16: 134. https://link.springer.com/article/10.1186/s12909-016-0656-7.

Tolsma, Rachael, Saad Shebrain, Shamsi Daneshvari Berry, and Lisa Miller. 2024. “Medical Student Perceptions of Assessments of Clinical Reasoning in a General Surgery Clerkship.” *BMC Medical Education* 24 (1): 211. https://pubmed.ncbi.nlm.nih.gov/38429706/.

Webb, L.C., D. Juul, C.F. Reynolds III, et al. 1996. “How Well Does the Psychiatry Residency In-Training Examination Predict Performance on the American Board of Psychiatry and Neurology. Part 1. Examination?” *American Journal of Psychiatry* 153 (6): 831-832. https://pubmed.ncbi.nlm.nih.gov/8633702/.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Institutional Written Exams

Description

Institutional written exams are typically used to measure learners' mastery of medical knowledge, problem-solving abilities, and application of foundational and clinical sciences. These exams can take various forms, including multiple-choice questions (MCQs), short-answer questions (SAQs), extended matching items, essay-type questions, and computer-based testing. Well-designed written exams serve both formative and summative purposes, helping to guide learning and ensure competency-based progression. Written quizzes, when well designed, are also useful for formative feedback to the learners.

Competencies

Institutional written exams primarily assess Medical Knowledge at the "Knows" and "Knows How" levels of Miller's Pyramid (1990). They evaluate cognitive skills such as recall, comprehension, and application, and can also assess clinical reasoning when case-based or vignette-style questions are used. Written exams can also contribute to assessing other competencies indirectly, such as Professionalism (via ethics-related questions) and Systems-Based Practice (through scenario-based problem-solving).

Validity

Ensuring validity and reliability in written exams is critical to making fair and defensible inferences about learner competence:

- **Validity**: Content validity is ensured by aligning exam questions with learning objectives and competency frameworks. Construct validity is assessed through correlations with other measures of learner performance (e.g., objective structured clinical examinations (OSCEs), NBME exams). Predictive validity may be evaluated by examining associations between exam scores and future clinical performance**.**
- **Reliability**: Internal consistency is measured using Cronbach's alpha, with values above 0.7 considered acceptable for formative exams and above 0.8 for high-stakes summative assessments. Inter-rater reliability is relevant for exams with subjective scoring (e.g., essays, SAQs).

Feasibility

Institutional written exams are widely used due to their feasibility, cost-effectiveness, and scalability. However, resource considerations include faculty time for item writing and review, the need for psychometric expertise, and technological infrastructure for computer-based testing. Exam security and standardization are essential to maintain integrity.

Acceptability

Written exams are generally well-accepted by faculty and learners, particularly when they are perceived as fair, valid, and aligned with curriculum objectives. Transparent communication regarding grading criteria and remediation policies enhances acceptability.

Catalytic Effect

Well-designed written assessments promote learning by directing learners' focus toward key knowledge areas. The use of post-exam feedback reports, including performance analytics on individual items, enhances their formative value and supports self-directed learning.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

References

Bhat, Subrahmanya K., and Kishan H.L. Prasad. 2021. “Item Analysis and Optimizing Multiple-Choice Questions for a Viable Question Bank in Ophthalmology: A Cross-Sectional Study.” *Indian Journal of Ophthalmology*. 69 (2): 343-346. https://pmc.ncbi.nlm.nih.gov/articles/PMC7933874/.

De Champlain, André F. 2010. “A Primer on Classical Test Theory and Item Response Theory for Assessments in Medical Education.” *Medical Education* 44 (1):109-117. https://pubmed.ncbi.nlm.nih.gov/20078762/.

Hawezy, Dawan J., Saman Taher Barzinjy, Govand Saadadin, Gasjeem Abdilwahid Hawezy, and Waleed Bager. 2024. “Perception of Undergraduate Medical Students and Examiners Towards Grand Objective Structured Clinical Examination.” *BMC Medical Education* 24: 1392. https://link.springer.com/article/10.1186/s12909-024-06388-w.

Hosseinzadeh, Mostafa, and Ki Lynn Matlock Cole. 2024. “Effects of the Quantity and Magnitude of Cross-Loading and Model Specification on MIRT Item Parameter Recovery.” *Educational and Psychological Measurement* 84 (5):1012-1040. https://pubmed.ncbi.nlm.nih.gov/39318481/.

Mitra, Nilesh Kumar, Nagaraja Haleagrahara, Gnanajothy Ponnudurai, and John Judson. 2009. “The Levels of Difficulty and Discrimination Indices in Type A Multiple Choice Questions of Pre-Clinical Semester 1 Multidisciplinary Summative Tests.” *International e-Journal of Science, Medicine, & Education* 3 (1): 2-7. https://www.researchgate.net/publication/227858882.

Rezigalla, Assad Ali, Ali Mohammed Elhassan Seid Ahmed Eleragi, et al. 2024. “Item Analysis: The Impact of Distractor Efficiency on the Difficulty Index and Discrimination Power of Multiple-Choice Items.” *BMC Medical Education* 24: 445. https://pmc.ncbi.nlm.nih.gov/articles/PMC11040895/.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Nationally Standardized Written Examinations

Description

Nationally standardized written examinations are widely used across the continuum of medical education, including undergraduate, graduate, and licensure contexts. These assessments are typically computer-based and use single-best-answer multiple-choice questions. They allow for robust standard-setting across large cohorts of examinees and are often aligned with national licensing standards or specialty board expectations.

These examinations are required for medical licensure and are also used in residency in-training examinations and clinical clerkships (e.g., NBME® Subject or “Shelf” Exams). In addition, many medical schools use them in preclinical curricula as primary, supplementary, or comprehensive assessments.

Designed to reflect real-world clinical scenarios, these exams emphasize the application of knowledge and clinical reasoning rather than rote memorization. They serve as important indicators of readiness for progression and licensure, while providing learners with insight into strengths and areas for improvement and offering educators data to guide curriculum development and individualized support.

Competencies

The Core Competencies most frequently assessed through in-training examinations include Medical Knowledge, Practice-Based Learning and Improvement, Systems-Based Practice, and Patient Care and Procedural Skills.

Validity

Nationally standardized examinations demonstrate strong validity and reliability, supported by large examinee cohorts and rigorous psychometric methods. Evidence supports:

- **Content validity**, due to alignment with national standards and test blueprints
- **Predictive validity**, as performance correlates with outcomes such as licensure and specialty board certification examinations
- **Reliability**, with internal consistency measures (e.g., Cronbach’s alpha) frequently exceeding 0.80 and the use of item response theory to ensure fairness and stability across administrations

There are generally moderate to strong correlations between different standardized examinations (e.g., United States Medical Licensing Examination (USMLE) and board examinations), supporting construct validity. However, correlations with broader outcomes—such as clinical performance ratings, clinical skills assessments, and professionalism indicators—are typically weak to moderate, reflecting the limitation of these exams in assessing real-world clinical performance. Consistent with this, these assessments primarily function at the lower levels of Miller’s pyramid (“knows” and “knows how.”)

Feasibility

These examinations are highly feasible across educational settings and can be used in both formative and summative capacities. They are often developed, administered, and scored by external organizations (e.g., NBME®, licensing boards), reducing the burden on faculty. Key feasibility strengths include scalability to large cohorts, standardized administration and scoring, and the ability to generate structured feedback that can inform both learners and 24 curricula.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

However, feasibility considerations include the cost—particularly for commercially administered examinations—along with administrative requirements such as adherence to standardized testing conditions (e.g., proctoring, room setup, and security protocols). In addition, these examinations are often administered infrequently, which limits opportunities for repeated assessment, and they may introduce unintended consequences, such as an overemphasis on test performance or reduced attention to clinical learning experiences.

Acceptability

These exams are highly acceptable to both faculty members and learners. Learners value them for benchmarking and identifying gaps before high-stakes exams, while faculty members and administrators appreciate their role in curriculum evaluation and learner support. However, some critiques include concerns about test anxiety and overemphasis on standardized test performance. Certainly, licensure exams are by definition higher stakes. Faculty members are typically familiar with this type of examination, having participated as an examinee themselves throughout their career. Both learners and faculty members may see benefit in nationally normed performance data, if such data is provided for the exam.

Catalytic Effect

The catalytic effect of nationally standardized exams is dependent on the data provided to the learner and how the exam is used. If there is robust formative data provided and if the assessment is used in a formative fashion, the learner's performance data may guide further study to strengthen specific content areas. Conversely, if the exam provides only a simple outcome (e.g., a pass/fail decision, a numeric score) and the assessment is used primarily for summative purposes, the exam may not promote further learning.

A formative exam gives learners detailed performance reports, helping them pinpoint knowledge gaps and adjust their study strategies. When data from these assessments is used for documentation of both what has been learned (assessment *of* learning) and what is needed to achieve competence (assessment *for* learning), standardized assessments of medical knowledge can have a strong catalytic effect. Data can inform ILPs as well as the continuous quality improvement of a program/institutions' curriculum and assessment program. When used intentionally as part of a broader assessment system—with actionable feedback and integration into coaching and curriculum improvement—these examinations can have substantial educational impact.

References

American College of Physicians. 2020. “New Study Finds that IM-ITE and USMLE Step Examinations Help to Predict Performance on the Internal Medicine Certifying Examination.” Accessed April 23, 2025. https://www.acponline.org/acp-newsroom/new-study-finds-that-im-ite-and-usmle-step-examinations-help-to-predict-performance-on-the-internal.

Arnhart, K.L., M.M. Cuddy, D. Johnson, M.A. Barone, and A. Young. 2021. “Multiple United States Medical Licensing Examination Attempts and the Estimated Risk of Disciplinary Actions Among Graduates of U.S. and Canadian Medical Schools.” *Acad Med* Sep 1;96(9):1319-1323.

Babbott, Stewart F., B.W. Beasley, K.T. Hinchey, et al. 2007. “The Predictive Validity of the Internal Medicine In-Training Examination.” *The American Journal of Medicine*. 120 (8): 735-740. https://pubmed.ncbi.nlm.nih.gov/17679136/.

Cooke, Brian K., Cynthia Garvan, and Jacqueline A. Hobbs. 2013. “Trends in Performance on the Psychiatry Resident-in-Training Examination (PRITE®): 10 Years of Data from a Single Institution.” *Academic Psychiatry* 37 (4): 261–264. https://pubmed.ncbi.nlm.nih.gov/23652297/.

Cuddy, M.M., C. Liu, W. Ouyang, M.A. Barone, A. Young, and D.A. Johnson. 2022. “An Examination of the Associations Among USMLE Step 3 Scores and the Likelihood of Disciplinary Action in Practice.” *Acad Med* Oct 1;97(10):1504-1510.

Dong, T., C. Zahn C, A. Saguil, et al. 2017. “The Associations Between Clerkship Objective Structured Clinical Examination (OSCE) Grades and Subsequent Performance.” *Teach Learn Med* Jul-Sep;29(3):280-285.

Egan, D.J., A. Husain, M.C. Bond, et al. 2019. “Standardized Video Interviews Do Not Correlate to United States Medical Licensing Examination Step 1 and Step 2 Scores.” *West J Emerg Med* Jan;20(1):87-91.

Filiberto, A.C., L.A. Cooper, T.J. Loftus, S.S. Samant, G.A. Sarosi Jr., and S.A. Tan. 2021. “Objective Predictors of Intern Performance.” *BMC Med Educ* Jan 26;21(1):77.

Goldstein, S.D., B. Lindeman, J. Colbert-Getz, et al. 2014. “Faculty and Resident Evaluations of Medical Students on a Surgery Clerkship Correlate Poorly with Standardized Exam Scores.” *Am J Surg* Feb;207(2):231-5.

Jeyaraju, M., H. Linford, T. Bosco Mendes, C. Caufield-Noll, and S. Tackett. 2023. “Factors Leading to Successful Performance on U.S. National Licensure Exams for Medical Students: A Scoping Review.” *Acad Med* Jan 1;98(1):136-148.

Kane, K.E., D. Yenser, K.R. Weaver, et al. 2017. “Correlation Between United States Medical Licensing Examination and Comprehensive Osteopathic Medical Licensing Examination Scores for Applicants to a Dually Approved Emergency Medicine Residency.” *J Emerg Med* Feb;52(2):216-222.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Kay, C., J.L. Jackson, and M. Frank. 2015. “The Relationship Between Internal Medicine Residency Graduate Performance on the ABIM Certifying Examination, Yearly In-Service Training Examinations, and the USMLE Step 1 Examination.” *Acad Med* Jan;90(1): 100-4.

Lombardi, C.V., N.T. Chidiac, B.C. Record, and J.J. Laukka. 2023. “USMLE Step 1 and Step 2 CK as Indicators of Resident Performance.” *BMC Med Educ* Jul 31;23(1):543.

McCrary, H.C., J.M. Colbert-Getz, W.B. Poss, and B.K. Smith. 2021. “A Systematic Review of the Relationship Between In-Training Examination Scores and Specialty Board Examination Scores.” *J Grad Med Educ* Feb;13(1):43-57.

McDonald, F.S., D. Jurich, L.M. Duhigg, et al. 2020. “Correlations Between the USMLE Step Examinations, American College of Physicians In-Training Examination, and ABIM Internal Medicine Certification Examination.” *Acad Med* Sep;95(9):1388-1395.

NBME. “Taking a Subject Examination.” Accessed April 23, 2025. https://www.nbme.org/examinees/subject-exams

NBME. “Taking an NBME Self-Assessment.” Accessed April 23, 2025. https://www.nbme.org/examinees/self-assessments

NBME. “NBME Subject Exam Program Guide.” Accessed April 23, 2025. https://www.nbme.org/sites/default/files/2022-10/NBME_Subject_Exam_Program_Guide.pdf

Norcini John, Brownell Anderson, Valdes Bollela, et al. 2011. “Criteria for Good Assessment: Consensus Statement and Recommendations from the Ottawa 2010 Conference.” *Medical Teacher* 33 (3): 206–214. https://pubmed.ncbi.nlm.nih.gov/21345060/.

Panda, N., D. Bahdila, A. Abdullah, A.J. Ghosh, S.Y. Lee, and W.B. Feldman. 2021. “Association Between USMLE Step 1 Scores and In-Training Examination Performance: A Meta-Analysis.” *Acad Med* Dec 1;96(12):1742-1754.

Peterson, L.E., J.R. Boulet, and B. Clauser. 2020. “Associations Between Medical Education Assessments and American Board of Family Medicine Certification Examination Score and Failure to Obtain Certification.” *Acad Med* Sep;95(9):1396-1403.

Pinnock, Ralph, Tracy-Lee Fisher, and Jane Astley. 2016. “Think Aloud to Learn and Assess Clinical Reasoning.” *Medical Education* 50 (5): 585-6. https://doi.org/10.1111/medu.13006.

Rayamajhi, Supratik, Prajwal Dhakal, Ling Wang, et al. 2020. “Do USMLE Steps and ITE Score Predict the American Board of Internal Medicine Certifying Exam?” *BMC Medical Education* 20 (1):1–7. https://pubmed.ncbi.nlm.nih.gov/32183789/.

Reid, C.M., D.Y. Kim, J. Mandel, A. Smith, and V. Bansal. 2014. “Correlating Surgical Clerkship Evaluations with Performance on the National Board of Medical Examiners Examination.” *J Surg Res* Jul;190(1):29-35.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Shirkhodaie, C., S. Avila, H. Seidel, R.D. Gibbons, V.M. Arora, J.M. Farnan. 2023. “The Association Between USMLE Step 2 Clinical Knowledge Scores and Residency Performance: A Systematic Review and Meta-Analysis.” *Acad Med* Feb 1;98(2):264-273.

Smeltzer D.J., and B.A. Jones. 1990. “Reliability and Validity of the Psychiatry Resident In-Training Examination.” *Academic Psychiatry* 14 (2): 115–121. https://pubmed.ncbi.nlm.nih.gov/24430340/.

Smith, D.T., A.T. Matelski, M.A.K. Hall, et al. 2025. “USMLE Performance, Subsequent Standardized Testing, and ABIM Certification Exam Preparation for Internal Medicine Residency Programs: A Narrative Review.” *J Gen Intern Med* Mar;40(4):879-891.

Stain, S.C., J.B. Matthews, A. Ata, D.B. Adams, H. Chen, and J.R. Potts. 2021. “US Medical Licensing Exam Performance and American Board of Surgery Qualifying and Certifying Examinations.” *J Am Coll Surg* Dec;233(6):722-729.

Tracy, B.M., B.J. Hazen, C.R. Ward, J.H. Winer, and B.J. Pettitt. 2020. “Sustained Clinical Performance During Surgical Rotations Predicts NBME Shelf Exam Outcomes.” *J Surg Educ* Nov-Dec;77(6):e116-e120.

Webb L.C., D. Juul, C.F. Reynolds, et al. 1996. “How Well Does the Psychiatry Residency In-Training Examination Predict Performance on the American Board of Psychiatry and Neurology Part I Examination?” *American Journal of Psychiatry* 153 (6): 831–832. https://pubmed.ncbi.nlm.nih.gov/8633702/.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Oral Patient Presentation

Description

An oral presentation of a patient case by a learner to a faculty member is common in clinical learning environments. Oral patient presentation typically involves the learner presenting the data they have gathered about the patient (both subjective and objective data), summarizing that information into an assessment of the patient, and providing a plan (which may include diagnosis and/or management decisions). In the oral presentation, the learner communicates patient information to the faculty member to aid in patient care; the learner demonstrates effective use of the correct and appropriate information they have obtained in the context of the presentation (typically from the patient but also from other sources such as the patient's chart). Assessing the oral presentation allows the faculty member to make judgements about the learner's skills in collecting this information, including skills in history-taking, communication, physical exam, and data acquisition. The faculty member is also able to assess the organization of this information within the presentation. Presentations can be organized in a variety of ways; learners are typically expected to structure them according to a particular template. Through the learner's assessment of the patient and ensuing plan, the faculty member can assess the learner's clinical reasoning and decision-making skills. Further elucidation of the learner's decision making may be accomplished with follow-up questions, as demonstrated in the One-Minute Preceptor model. See One-Minute Preceptor

Competencies

An oral patient presentation can be used to assess Patient Care and Procedural Skills, Medical Knowledge, Systems-Based Practice, Practice-Based Learning and Improvement, Professionalism, and Interpersonal and Communication Skills.

Validity

There is limited validity evidence for the oral patient presentation as an assessment in a clinical setting, although there is internal validity evidence for structured assessments of the oral presentation using a tool, and these have shown improved reliability. There is also evidence for using these structured presentation tools as an assessment of clinical reasoning and as a formative tool to improve feedback from faculty. In contrast, there is little evidence to support inferences that are made from the oral presentation about other clinical skills (e.g., history-taking). While the oral patient presentation is often utilized as a surrogate for direct observation, there is no evidence to support this use. Since oral patient presentations are case-specific, multiple assessments are needed for valid summative assessment.

Feasibility

Oral patient presentations can be used in a variety of clinical settings and in a variety of patient situations.

Acceptability

Assessments of oral patient presentations are a common clinical assessment. They are often more acceptable to faculty members when they can be performed as part of the clinical workflow of gathering patient information and making clinical decisions. They can be viewed as burdensome when the presentations are long or unfocused, as may occur more often with junior learners. Learners are likely to have varying acceptance of oral patient presentations. Some learners appreciate the structure and the opportunity to demonstrate their work and abilities. Other learners may find a presentation (particularly in a team setting) brings unwanted attention

©2026 Accreditation Council for Graduate Medical Education (ACGME)

and can be associated with public speaking anxiety. Learners 28 also find oral presentations more acceptable when they are not interrupted with questions during the presentation.

Catalytic Effect

Since oral presentations are linked to other skills (e.g., the data collection that preceded the presentation), a catalytic effect can occur that promotes development of other skills based on feedback from the oral presentation. For example, a learner may be motivated to refine their history-taking skills based on feedback they received regarding their oral presentation of the history they collected. Of course, this effect is dependent on the learner receiving feedback from their oral presentation or being able to self-assess and reflect their performance.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

References

Carter, Cristina, Nicole Akar-Ghibril, Jeff Sestokas, Gabrina Dixon, Wilhelmina Bradford, and Mar Ottolini. 2018. “Problem Representation, Background Evidence, Analysis, Recommendation: An Oral Case Presentation Tool to Promote Diagnostic Reasoning.” *Academic Pediatrics* 18 (2): 228-230. https://pubmed.ncbi.nlm.nih.gov/28807616/.

Cohen, Adam, Moushumi Sur, Carla Falco, Gurpreet Dhaliwal, Geeta Singhal, and Satid Thammasitboon. 2022. “Using the Assessment of Reasoning Tool to Facilitate Feedback About Diagnostic Reasoning.” *Diagnosis (Berlin, Germany)* 9 (4): 476-484. https://pubmed.ncbi.nlm.nih.gov/36073963/.

Haber, R.J., and L.A. Lingard. 2001. “Learning Oral Presentation Skills: A Rhetorical Analysis with Pedagogical and Professional Implications.” *Journal of General Internal Medicine* 16 (5): 308-14. https://pubmed.ncbi.nlm.nih.gov/11359549/.

Landreville, Jeffrey M., Warren J. Cheung, Alexandra Hamelin, and Jason R. Frank. 2019. “Entrustment Checkpoint: Clinical Supervisors' Perceptions of the Emergency Department Oral Case Presentation.” *Teaching and Learning in Medicine* 31 (3): 250-257. https://pubmed.ncbi.nlm.nih.gov/30706726/.

Lewin, Linda Orkin, Lorraine Beraho, Sandra Dolan, Leah Millstein, and David Bowman. 2013. “Interrater Reliability of an Oral Case Presentation Rating Tool in a Pediatric Clerkship.” *Teaching and Learning in Medicine* 25 (1): 31-38. https://www.tandfonline.com/doi/full/10.1080/10401334.2012.741537.

Lingard, L., C. Schryer, K. Garwood, and M. Spafford. 2003. “‘Talking the Talk’: School and Workplace Genre Tension in Clerkship Case Presentations.” *Medical Education* 37 (7): 612-20. https://pubmed.ncbi.nlm.nih.gov/12834419/.

Thammasitboon, Satid, Joseph J. Rencic, Robert L. Trowbridge, Andrew P. J. Olson, Moushumi Sur, and Gurpreet Dhaliwal. 2018. “The Assessment of Reasoning Tool (ART): Structuring the Conversation Between Teachers and Learners.” *Diagnosis (Berlin, Germany)* 27; 5 (4): 197-203. https://pubmed.ncbi.nlm.nih.gov/30407911/.

Thammasitboon, Satid, Moushumi Sur, Joseph J. Rencic, et al. 2021. “Psychometric Validation of the Reconstructed Version of the Assessment of Reasoning Tool.” *Medical Teacher* 43 (2): 168-173. https://pubmed.ncbi.nlm.nih.gov/33073665/.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Performance-Based Assessments

Description

Performance tests are designed to assess a learner's clinical skills "in vitro." Yudkowsky (2009) described performance tests as "...contrived for the purpose of the examination and explicitly invit[ing] the examinee to demonstrate the behavior to be assessed." Such tests are a way for learners to demonstrate their capability in skills in a lower-stakes environment. Examples of performance tests include OSCEs, anatomy lab practical, and evidence-based practice logs.

Competencies

Performance tests measure the "Shows How" level of Miller's Pyramid (1990). Patient Care and Procedural Skills is the most common competency assessed through performance tests, though various assessment tools may assess competence in the other competencies (Systems-Based Practice, Practice-Based Learning and Improvement, Medical Knowledge, Interpersonal Communication Skills, and Professionalism). Performance tests can be an excellent way to assess learners' clinical reasoning skills.

Validity

Validity for performance tests varies by each tool. Especially for tools using human assessors (e.g., SPs), the introduction of human bias can limit the validity of the tool. When explicit criteria are used, a high degree of reliability is attainable for some performance tests.

Feasibility

Feasibility of performance tests also varies by specific tool. Some, such as clinical performance (record) review, can have a high degree of feasibility as medical records and other clinical data are usually available and accessible. Other tools, such as OSCEs, may be less feasible due to the number of resources (time, personnel, money, training, etc.) that are required to properly use the tool.

Acceptability

As with feasibility, acceptability of performance tests can also vary. Clinical performance (record) review may be more acceptable than a standardized patient interaction as less time is needed to conduct a record review. Research and scholarly projects, for example, can also be time intensive. One benefit of OSCEs and standardized patient interactions is that scenarios can be tailored to specific circumstances or to ensure learners demonstrate a particular skill.

Catalytic Effect

Because of the opportunity for learners to demonstrate skills in the "Shows How" level of Miller's Pyramid (1990), catalytic effect can be high. While sometimes time-intensive, feedback given to the learners regarding performance tests can be very specific and actionable, sometimes giving the learner a chance to 'try again' and demonstrate improvement. Being able to demonstrate these skills in a slightly lower-stakes environment can give learners the confidence they need to perform tasks in actual clinical settings.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

References

Boulet, John R., Neena Natt, and Richard E. Hawkins. 2017. “Direct Observation: Standardized Patients.” In *Practical Guide to the Evaluation of Clinical Competence* by Eric S. Holmboe, Steven J. Durning, and Richard E. Hawkins. Elsevier.

Chambers, K.A., J.R. Boulet, and N.E. Gary. 2000. “The Management of Patient Encounter Time in a High-Stakes Assessment Using Standardized Patients.” *Medical Education* 34 (10): 813-817. https://pubmed.ncbi.nlm.nih.gov/11012930/.

Carson, Julie A., Adam Peets, Vincent Grant, and Kevin McLaughlin. 2010. “The Effects of Gender Interactions on Students’ Physical Examination Ratings in Objective Structured Clinical Examination Stations.” *Academic Medicine* 85 (11): 1772-1776. https://doi.org/10.1097/ACM.0b013e3181f52ef8.

Homboe, Eric S. and Daniel Duffy. 2017. “Clinical Practice Review.” In *Practical Guide to the Evaluation of Clinical Competence* by Eric S. Holmboe, Steven J. Durning, and Richard E. Hawkins. Elsevier.

Yudkowsky, Rachel. 2009. “Performance Tests.” In *Assessment in Health Professions Education* edited by Steven Downing and Rachel Yudkowski. Routledge.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Anatomy Lab Practical Exams

Description

The anatomy lab practical exam is a commonly employed tool to assess the practical knowledge of anatomy and is predominantly conducted as a summative assessment; however, similar formative practice opportunities are often given. Anatomy lab practical exams test the ability of learners to relate clinical and functional anatomical knowledge with three-dimensional anatomical structures such as muscles, bones, nerves, blood vessels, organs, and other anatomical landmarks. Typical anatomy lab practical exams involve multiple stations where a student is asked to identify an anatomical structure or answer a question related to that structure. Structures can be labeled on prosections, prosected cadavers, models, radiological images, and other graphics. Students are given a fixed time limit to answer all questions at a given specimen or station. There are typically three ways to gather answers on typical anatomy practical exams: free response, word banks and multiple choice. Several research reports have shown comparability of student performance across all three formats. Because the anatomy lab practical exam can be considered a modified written exam with either short answer or multiple-choice questions, traditional item analysis and psychometrics properties can be applied to assess the validity of the exam (see also “Institutional Written Exam,” pp. 21).

Spotter Exams

A typical spotter anatomy lab practical exam involves learners rotating through various stations, wherein they are presented with numbered pins or pointers identifying specific structures. Learners are required to record the correct anatomical term on an answer sheet or tablet within a short time limit, often 60 seconds per station. They progress through numerous stations to make a comprehensive exam.

Objective Structured Practical Exams (OSPE)

The objective structured practical exam involves multiple stations designed around a specific objective to be tested with the goals of assessing higher cognitive and psychomotor skills, i.e., an ability to relate clinical information with structural material (models, cadaveric dissections, images). After a blueprint of objectives is created, a multidisciplinary approach is used to write questions that test core anatomical knowledge in functional and clinical contexts. Each station is equipped with appropriately labeled cadaveric dissections, models, images, and physiological/pathological data sets to answer relevant clinical questions.

Competencies

Traditional spotter anatomy lab practical exams typically test Medical Knowledge. However, as part of the trend to teach anatomy as more integrated and clinically focused, traditional identify questions are being replaced with clinical vignettes that correspond to the specimens tagged. These objective structured practical exams can assess clinical reasoning in addition to medical knowledge.

Validity

Content validity is generally strong, as anatomy lab practice exams directly assess the identification of anatomical structures central to medical practice. Criterion validity is variable, showing some correlation with written exams but less clear predictive validity for future clinical performance. Furthermore, reliability can be influenced by specimen variability and inter-rater differences. Thus, while traditional anatomy lab practical exams assess anatomical recognition, their ability to accurately reflect clinically relevant anatomical understanding and skills require careful consideration and supplementation with other assessment methods.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Feasibility

Anatomy lab practical exams are typically resource and time demanding. The availability and quality of the specimens can be a limiting variable. Often specimens generated by learners require additional dissection by faculty members for clarity necessitating extra faculty time. The traditional anatomy lab practical exam requires extensive preparation and administration especially for large cohorts, requiring the breakdown of the cohort into manageable groups, several exam cycles and possibly student quarantining to avoid leakage of exam questions. Laboratory space must be cordoned amidst regular class schedules to prepare and administer each exam; the necessary anatomical material also needs to be reserved.

Grading handwritten free response assessments is difficult due to the large number of learners; it is reported to be time consuming, prone to error, and difficult to discern because of handwriting. This type of grading usually requires a round-robin approach with at least two faculty members grading the same sheet to ensure accuracy. However, new applications and artificial intelligence can be used to aid in this process.

Acceptability

While learners recognize the importance of practical anatomical knowledge, the high-pressure, rapid-fire nature of typical anatomy lab practical exam can induce significant anxiety. Many find the focus on rote memorization less engaging than clinically oriented assessments which can assess higher-order thinking skills. However, learners generally accept the practical as a necessary component of their medical education, especially when coupled with clear learning objectives, adequate practice time, and access to high-quality specimens.

Faculty members often view anatomy lab practical exams as a valuable tool for assessing learners’ practical anatomical knowledge and spatial reasoning. They appreciate the direct observation of students' ability to identify structures, which can be challenging to evaluate through written exams alone. Testing learners on their own dissections holds them accountable and provides learner motivation to complete the dissections which maximizes the gift of the anatomical donor.

Catalytic Effect

Anatomy lab practical exams, while often stressful, can have a catalytic effect on learning. The high-stakes nature of these exams, particularly the spotter format, compels learners to engage in intensive, active recall and spatial reasoning. Using multiple specimens reinforces the anatomical variation and the structural uniqueness of humans. The identification of structures on an anatomy practical exam serves as a strong primer for more advanced testing involving anatomical relationships, functional deficits and clinical correlations. Allowing learners to immediately come back and view the correct answers after the exam is an excellent learning tool to deepen their understanding and enhance their retention of anatomical knowledge.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

References

Brenner, Erich, Andy R. M. Chirculescu, Concepcion Reblet, and Claire F. Smith. 2015. "Assessment in Anatomy." *European Journal of Anatomy* 19 (1): 105-124. https://www.researchgate.net/publication/272748570_Assessment_in_anatomy.

Mogali, Sreenivasulu Reddy, Jerome J. Rotgans, Lucy Rosby, Michael Alan Ferenczi, Naomi Low Beer. 2020. "Summative and Formative Style Anatomy Practical Examinations: Do They Have Impact on Students' Performance and Drive for Learning?" *Anatomical Sciences Education* 13 (5): 581-590. https://pubmed.ncbi.nlm.nih.gov/31733172/.

Shaibah, Hassan Sami, Cees P.M. van der Vleuten. 2013. "The Validity of Multiple Choice Practical Examinations as an Alternative to Traditional Free Response Examination Formats in Gross Anatomy." *Anatomical Sciences Education* 6 (3): 149-156. https://pubmed.ncbi.nlm.nih.gov/23109285/.

Yaqinuddin, Ahmed, Muhammad Zafar, Muhammad Faisal Ikram, and Paul Ganguly. 2013. "What Is an Objective Structured Practical Examination in Anatomy?" *Anatomical Sciences Education* 6 (2): 125-133. https://pubmed.ncbi.nlm.nih.gov/22930456/.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Assessment of Cost-Conscious Care

Description

As learners develop their knowledge and skills to manage patients within the healthcare system, they need to also develop an understanding of the financial aspects of healthcare. This includes comparative effectiveness, cost effectiveness, test and medication costs, payment models, and the impacts on individual patients. This type of learning falls under the competency of Systems-based Practice, within the field of health systems science. Sometimes this is referred to as “high value care,” with “value” referring to one of the quintuple aims of healthcare: to reduce costs. To demonstrate learning and, ultimately, competence in providing cost-conscious care, a number of assessment methods and tools described elsewhere in the Guidebook have been developed. These include multiple-choice exams, online case-based modules, chart-stimulated recall, OSCEs, and reflective writing. Existing assessments have also been modified to include an assessment of cost-conscious care, by adding items to faculty global assessment forms, direct observation forms, and oral presentation structures. Despite the development of multiple assessment tools, the assessment of cost-conscious care is not commonly integrated into medical education assessment systems.

Competencies

Assessment of cost-conscious care is primarily used to assess the competency of Systems-based Practice, but it can also involve assessment of Interpersonal and Communication Skills and Practice-based Learning and Improvement.

Validity

Validity evidence is very limited for standalone assessments of cost-conscious care. For situations in which an assessment of cost-conscious care is added to an existing assessment, the validity will be dependent on the validity of the overall assessment.

Feasibility

Feasibility of assessment for cost-conscious care is going to vary based on the nature of the assessment. Multiple-choice exams, online case-based modules, and reflective writing can be added as assessments to an existing curriculum relatively easily. Designing a standardized patient case for an OSCE will require more investment. Adding individual cost-conscious care items to existing global assessment forms and direct observation forms is relatively easy, though there will still be investment required to develop the skills of the faculty who will be using the forms.

Acceptability

While faculty members may be familiar with the principles and practice of cost-conscious care and may value learners developing competence in this area, many faculty members are uncomfortable teaching or assessing these concepts. Providing a structure and tools for faculty members to use can help overcome this barrier. Learners are aware of these issues and interested in developing competence, particularly when teaching and assessment are situated in a developmentally appropriate way (e.g., understanding health system finances when they have sufficient experience with the health system).

Catalytic Effect

Assessment of cost-conscious care can promote increased learner awareness of the concepts (particularly in clinical settings) and provide feedback about their competency which can promote further knowledge acquisition and skill development in this area.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

References

Born, Karen B., Christopher Moriates, Victoria Valencia, Marlou Kerssens, and Brian M. Wong. 2019. "Learners as Leaders: A Global Groundswell of Students Leading Choosing Wisely Initiatives in Medical Education." *Academic Medicine* 94 (11): 1699-1703. https://pubmed.ncbi.nlm.nih.gov/31299673/.

Carey, Magalie E., Brittany L. Gilmore, Michael J. Lawler, Max Hoddwells, Andrew J. Hale, and Allen B. Repp. 2023. "Integrating High-Value Care Concepts into Preclinical Medical Education: A Practical Approach." *Journal of Medical Education and Curricular Development* 10: 23821205231173490. https://pubmed.ncbi.nlm.nih.gov/37163150/.

Cayea, Danelle, Kim Tartaglia, Amit Pahwa, Heather Harrell, Amy Shaheen, and Valerie J. Lang. 2018, "Current and Optimal Training in High-Value Care in the Internal Medicine Clerkship: A National Curricular Needs Assessment." *Academic Medicine* 93 (10): 1511-1516. https://pubmed.ncbi.nlm.nih.gov/29517522/.

Cox, Malcolm, James T. Pacala, Gregory M. Vercellotti, and Judy A. Shea. 2004. "Health Care Economics, Financing, Organization, and Delivery." *Family Medicine* 36 (Suppl.): S20-30. https://fammedarchives.blob.core.windows.net/imagesandpdfs/fmhub/fm2004/January/MalcolmS20.pdf.

El Fadel, Omar, Zachary N. Goldberg, Amiti Jain, et al. 2024. "Integrating Choosing Wisely, Value-Based Care Principles, into Undergraduate Medical Education: A Pilot Study." *Cureus* 25 16 (3): e56912. https://pubmed.ncbi.nlm.nih.gov/38528995/.

Erath, Alexandra, Margaret Mitchell, Sanjana Salwi, Yangzi Liu, and Alexander Sherry. 2019. "The Sooner the Better: High-Value Care Education in Medical School." *Academic Medicine* 94 (11): 1643-1645. https://pubmed.ncbi.nlm.nih.gov/31335820/.

Huang, Grace C., Carrie D. Tibbles, Lori R. Newman, and Richard M. Schwartzstein. 2016. "Consensus of the Millennium Conference on Teaching High Value Care." *Teaching and Learning in Medicine* 28 (1): 97-104. https://pubmed.ncbi.nlm.nih.gov/26787090/.

Moser, Eileen M., Grace C. Huang, Clifford D. Packer, et al. 2016. "SOAP-V: Introducing a Method to Empower Medical Students to Be Change Agents in Bending the Cost Curve." *Journal of Hospital Medicine* 11 (3): 217-220. https://pubmed.ncbi.nlm.nih.gov/26416013/.

Natt, Neena, Michael Dekhtyar, Yoon Soo Park, et al. 2020. "Promoting Value Through Patient-Centered Communication: A Multisite Validity Study of Third-Year Medical Students." *Academic Medicine* 95 (12): 1900-1907. https://pubmed.ncbi.nlm.nih.gov/32459676/.

Nundy, Shantanu, Lisa A. Cooper, and Kedar S. Mate. 2022. "The Quintuple Aim for Health Care Improvement: A New Imperative to Advance Health Equity." *JAMA* 327 (6): 521-522. https://pubmed.ncbi.nlm.nih.gov/35061006/.

Oza, Sandra K., Pablo Joo, Joseph H. Grochowalski, Steven Rougas, Paul George, and Felise Milan. 2022. “Novel Use of an OSCE to Assess Medical Students' Responses to a Request for a Low Value Diagnostic Imaging Test: A Mixed Methods Analysis.” *Patient Education and Counseling* 105 (7): 2264-2269. https://pubmed.ncbi.nlm.nih.gov/34716052/.

Steele, Christopher, Danelle Cayea, Justin Berk, et al. 2019. “Novel First-Year Curriculum in High-Value Care.” *The Clinical Teacher* 16 (5): 513-518. https://pubmed.ncbi.nlm.nih.gov/30644162/.

Tartaglia, Kimberly M., Nicholas Kman, and Cynthia Ledford. 2015. “Medical Student Perceptions of Cost-Conscious Care in an Internal Medicine Clerkship: A Thematic Analysis.” *Journal of General Internal Medicine* 30 (10):1491-1496. https://pubmed.ncbi.nlm.nih.gov/25931005/.

Clinical Performance (Record) Review

Description

Critique of the clinical (i.e., medical) record is a time-honored approach in the assessment of learners and can also be used by learners themselves. Medical records review serves a number of important functions as: 1) an archive of important patient medical information for use by other health care providers and patients; 2) a source of data to assess performance in practice for specific acute and chronic medical conditions (e.g., pneumonia, diabetes), pre- or post-operative care, or prevention; 3) monitoring of patient safety and complications; and 4) documentation of diagnostic and therapeutic decisions. One can readily see how these patient care functions of the medical record can be used for educational and evaluative purposes. Learners must be involved in monitoring their own clinical practice and improving the quality of care based on a systematic review of the care they provide. Practice review, using the medical record, can promote self-reflection and support self-regulated learning, which are important skills needed for lifelong learning and for improving the delivery of care to patients, families, and communities.

Competencies

Practice audits are an essential method in the assessment of Practice-Based Learning and Improvement, Interpersonal and Communication Skills, and Systems-Based Practice. Clinical records can also serve as the content for chart-stimulated recall to explore clinical reasoning, particularly for Patient Care and Procedural Skills and Medical Knowledge.

Validity

An extensive body of literature exists on the validity and reliability of a host of clinical performance measures. Both the Agency for Healthcare Research and Quality (AHRQ) and the National Committee for Quality Assurance (NCQA) house a large warehouse of clinical performance measures with supporting evidence available for those interested in the science behind performance and safety measures. Many specialty societies also publish and support clinical performance and safety measures within their discipline. GME programs, in particular, are encouraged to check with their respective specialty societies to see what is available.

The major challenge for using clinical performance review as a tool is attribution of performance measure results to a single learner. It is unnecessary to attribute patient outcomes to any individual when the patient receives care in a complex system of interprofessional care. Using the lens of contribution is more useful (i.e., what was the learner's contribution to the performance on this set of measures?). For formative purposes, a review of eight to 10 medical records of the learner provides sufficient reliability.

Higher levels of reliability are attainable by aggregating performance measure results across learners within a specific clinical setting. Examining performance at a program and institutional level is also critically important as a growing body of research has found associations between performance by the clinical site and the performance of program graduates in autonomous practice (Asch et al. 2009; Chen et al. 2014; Bansal et al. 2016).

Feasibility

Feasibility depends heavily on the purpose for and method used to conduct the review. Almost all teaching hospitals and clinics now use an electronic health record (EHR). Quality and safety measures are commonly embedded into EHRs. Programs and institutions should check with their quality and information technology departments both about which quality and safety measures can be pulled from the EHR and how they can be obtained to benefit the learners'

performance. One major challenge for most EHR systems is how to properly identify or attribute a patient to a learner. Because mechanisms exist to accomplish this process, programs and institutions are encouraged to work with their information technology departments to choose appropriate clinical settings and processes through which to pull data for individual learners.

While the ease of using extracted EHR data is appealing, manual medical record audits can be an extremely meaningful educational and assessment experience. These manual audits can be performed by the individual learner. This approach supports learning about how a quality or safety measure is constructed, enables immediate learner reflection on their own practice, and generates data on their practice for formative assessment, feedback and coaching. If a manual process is used, the program or institution should use a standardized collection form to guide the data extraction. A wealth of information can be garnered from the EHR review outside of a simple performance outcome measure that adds to learning and reflection. Depending on the focus and number of performance measures and settings targeted, record reviews can take anywhere from 10 to 30 minutes per record on average.

Acceptability

Use of clinical performance measures is now standard practice in health care and should also be a standard aspect of medical education. Factors that diminish acceptability include:

- receipt of performance data for patients not primarily cared for by the learner (i.e., learner's contribution to the outcome measure is missing);
- lack of understanding the measure;
- lack of efficacy in how to improve performance on the measure;
- training in dysfunctional clinical settings or systems where improving performance is challenging;
- lack of meaningful feedback or coaching; and,
- lack of a pathway for making changes to practice one a quality gap is identified.

Catalytic Effect

When results from performance measures are combined with effective feedback, coaching, and systematic quality improvement efforts, the catalytic effects can be powerful. A number of studies show that learners can be the primary leaders and drivers for quality improvement and patient safety efforts, in addition to using the data for their own professional growth (Lynn et al. 2012; Myers and Wong 2019; Schumacher et al. 2018).

References

Asch, David A., Sean Nicholson, Sindu Srinivas, Jeph Herrin, and Andrew J. Epstein. 2009. "Evaluating Obstetrical Residency Programs Using Patient Outcomes." *Journal of the American Medical Association* 302 (12): 1277-1283. https://pubmed.ncbi.nlm.nih.gov/19773562/.

Bansal, Neha, Kristine D. Simmons, Andrew J. Epstein, Jon B. Morris, and Rachel R. Kelz. 2016. "Using Patient Outcomes to Evaluate General Surgery Residency Program Performance." *JAMA Surgery* 151 (2): 111-119. https://jamanetwork.com/journals/jamasurgery/fullarticle/2466606.

Chen, Candice, Stephen Petterson, and Robert Phillips. 2014. "Spending Patterns in Region of Residency Training and Subsequent Expenditures for Care Provided by Practicing Physicians for Medicare Beneficiaries." *Journal of the American Medical Association* 312 (22): 2385-2393. https://jamanetwork.com/journals/jama/fullarticle/2020373.

Govaerts, Marjan J.B., Cees P.M. van der Vleuten, Eric S. Holmboe. 2019. "Managing Tensions in Assessment: Moving Beyond Either-Or Thinking." *Medical Education,* 53 (1): 64-75. https://pubmed.ncbi.nlm.nih.gov/30289171/.

Norcini, John, M. Brownell Anderson, Valdes Bollela, et al. 2018. "2018 Consensus Framework for Good Assessment." *Medical Teacher* 40 (11): 1102-1109. https://pubmed.ncbi.nlm.nih.gov/30299187/.

Lynn, Lorna, Brian J. Hess, Weifeng Weng, et al. 2012. "Gaps in Quality of Diabetes Care in Internal Medicine Residency Clinics Suggest Better Training is Needed in Ambulatory Settings." *Health Affairs* 31 (1): 150-158. https://doi.org/10.1377/hlthaff.2011.0907.

Myers, Jennifer S., and Brian M. Wong. 2019. "Measuring Outcomes in Quality Improvement Education: Success Is in the Eye of the Beholder." *BMJ Quality and Safety* 28 (5): 345-348. https://pubmed.ncbi.nlm.nih.gov/30886120/.

Sargeant, Joan, Karen Mann, Cees van der Vleuten, et al. 2008. "'Directed' Self-Assessment: Practice and Feedback Within a Social Context." *Journal of Continuing Education in the Health Professions* 28 (1): 47-54. https://pubmed.ncbi.nlm.nih.gov/18366127/.

Schumacher, Daniel J., Eric S. Holmboe, Cees van der Vleuten, et al. 2018. "Developing Resident-Sensitive Quality Measures: A Model from Pediatric Emergency Medicine." *Academic Medicine* 93 (7): 1071-1078. https://pubmed.ncbi.nlm.nih.gov/29215378/.

Schumacher, Daniel J., Danny T.Y. Wu, Karthikeyan Meganathan, et al. 2019. "A Feasibility Study to Attribute Patients to Primary Interns on Inpatient Ward Teams Using Electronic Health Record Data." *Academic Medicine* 94 (9): 1376-1383. https://pubmed.ncbi.nlm.nih.gov/31460936/.

Schumacher, Daniel J., Abigail Martini, Eric S. Holmboe, et al. 2020. "Initial Implementation of Resident-Sensitive Quality Measures in the Pediatric Emergency Department: A Wide

Range of Performance." *Academic Medicine* 95 (8): 1248-55. https://pubmed.ncbi.nlm.nih.gov/31913878/.

Schuwirth, Lambert W.T., and Cees P.M. van der Vleuten. 2011. "Programmatic Assessment: From Assessment of Learning to Assessment for Learning." *Medical Teacher* 33 (6): 478-485. https://pubmed.ncbi.nlm.nih.gov/21609177/.

van Der Vleuten, C.P.M., L.W.T. Schuwirth, E.W. Driessen, et al. 2015. "Twelve Tips for Programmatic Assessment." *Medical Teacher* 37 (7): 641-646. https://pubmed.ncbi.nlm.nih.gov/25410481/.

Watling, Christopher J., and Shiphra Ginsburg. 2019. "Assessment, Feedback and the Alchemy of Learning." *Medical Education* 53 (1): 76-85. https://pubmed.ncbi.nlm.nih.gov/30073692/.

Wong, Brian M., Karyn D. Baum, Linda A. Headrick, et al. 2020. "Building the Bridge to Quality: An Urgent Call to Integrate Quality Improvement and Patient Safety Education with Clinical Care." *Academic Medicine* 95 (1): 59-68. https://pubmed.ncbi.nlm.nih.gov/31397709/.

Xu, Jacqueline, Matthew A. Silver, Jung Kim, J., et al. 2024. "Using the Electronic Health Record to Provide Audit and Feedback in Medical Student Clerkships." *JAMIA Open* 7 (3): ooae090. https://pubmed.ncbi.nlm.nih.gov/39314672/.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Evidence-Based Practice Logs

Description

Evidence-based practice logs are reflective documents used to assess the learner's ability to integrate patient preferences, clinical expertise and evidence-based process into patient care. Learners document their engagement with the process including literature searches, clinical questions and application of evidence-based medicine to their clinical cases. The Sicily statement (Dawes, et al. 2005) on evidence-based practice provides a five-step model for evidence-based practice:

1. Ask a clinical question.
2. Collect the most relevant evidence.
3. Critically appraise the evidence.
4. Integrate the evidence with one's clinical expertise, patient preferences and values to make a practice decision.
5. Evaluate the change or outcome.

These logs enable the faculty member to assess the learner's ability to practice medicine in a scientifically rigorous, patient-centered, and reflective manner. They serve as a longitudinal assessment tool that promotes ongoing reflection and feedback on learner's professional development and clinical reasoning.

Competencies

Evidence-based practice logs can be used to assess Medical Knowledge, Patient Care and Procedural Skills, Practice-Based Learning and Improvement, and Professionalism.

Validity

Evidence-based practice logs can provide valid assessments on clinical reasoning and medical knowledge. These logs capture the EBP process that include formulating the clinical question through interpreting research findings. They require the learners to apply skills such as critical thinking, reviewing the scientific literature, and applying these processes to their clinical decisions, which, in turn, are valid measures of their clinical reasoning. These logs can serve as a proxies for how well learners apply the evidence-based process to real world practice. When implemented properly, they encourage reflective practice and promote continuous improvement.

Feasibility

The feasibility of this method of assessment is due to its low resource requirements and easy integration into clinical practice. It offers a framework for learners to apply to their clinical decision making during patient encounters. The use of a digital platform makes the logs more accessible and scalable and offers the ability to collect and track data and give feedback in a simplified manner. This assessment method is also low cost and does not require a financial investment. Faculty development is integral for making this a smooth process. Faculty members should be trained in how to review these logs and provide consistent actionable feedback. The main challenge to feasibility is the time commitment required to conduct a quality review and to give feedback.

Acceptability

Evidence-based practice logs are relevant to clinical practice, encourage self-directed learning and provide actionable feedback. The feedback received is direct and actionable by providing examples of where learners can improve their clinical reasoning skills. This method promotes lifelong learning and evidence-based thinking ─ critical skills for any medical professional. Challenges may include increased workload for the faculty and learners, and inconsistent feedback. Faculty development on providing feedback is necessary to enhance acceptability.

Catalytic Effect

Evidence-based practice logs promote critical thinking, enhance professionalism and accountability, and encourage integrating evidence-based decision making into clinical practice. These effects allow learners to examine the literature to analyze research studies and provide proof of clinical practices, which in turn fosters development of clinical reasoning and critical thinking skills. Use of evidence-based practice logs also promotes the culture of quality care within medicine by encouraging learners to provide the best clinical care by integrating the evidence into practice.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

References

Dawes, Martin, William Summerskill, Paul Glasziou, et al. 2005. “Sicily Statement on Evidence-Based Practice.” *BMC Medical Education*, 5 (1): 1. https://pubmed.ncbi.nlm.nih.gov/15634359/.

Greenhalgh, Trisha, Jeremy Howick, Neal Maskrey, and Evidence-Based Medicine Renaissance Group. 2014. “Evidence Based Medicine: A Movement in Crisis?” *BMJ (Clinical Research Edition)* 348: g3725. https://pubmed.ncbi.nlm.nih.gov/24927763/.

Larsen, Camilla M., Anne Seneca Terkelsen, Anne-Marie Fiala Carlsen, and Hanne Kaae Kristensen. 2019. “Methods for Teaching Evidence-Based Practice: A Scoping Review.” *BMC Medical Education* 19 (1): 259. https://pubmed.ncbi.nlm.nih.gov/31296212/.

Sackett, D.L., W.M. Rosenberg, J.A. Gray, R.B. Haynes, and W.S. Richardson. 1996. “Evidence Based Medicine: What It Is and What It Isn't.” *BMJ (Clinical Research Edition)* 312 (7023): 71-72. https://pubmed.ncbi.nlm.nih.gov/8555924/.

van Der Vleuten, C.P.M., L.W.T. Schuwirth, E.W. Driessen, et al. 2015. “Twelve Tips for Programmatic Assessment.” *Medical Teacher* 37 (7): 641-646. https://pubmed.ncbi.nlm.nih.gov/25410481/.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Procedure, Operative, or Clinical Case Logs

Description

Procedure, operative, or case logs (logs) document each patient encounter/patient care experience, and/or procedures and surgical operations performed. The logs may involve recording of patient presentations or diagnoses and the learner's role in them (e.g., observed, participated). 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. They may document each procedure performed, assisted with, and/or observed, depending on the objectives of logging and directions given to the learner. 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. The ACGME Case Log System is one example of a case log system. Case logs can also be completed via mobile apps or in learning management systems. Proprietary mobile apps that facilitate simultaneous case logging and request for operative performance evaluation are also in use.

Logs of types of cases seen or procedures performed are useful for determining the scope of patient care experiences for a learner. Regular review of logs can be used to help learners track what cases or procedures must be sought out in order to meet program requirements or specific learning objectives. Patient logs documenting clinical experience for the entire residency or fellowship program can serve as a summative report of that experience; as noted below, the numbers reported do not necessarily indicate competence. Some institutions may be able to develop tools to pull data from electronic medical records.

Competencies

Procedural or operative case logs can be useful for assessing Patient Care and Procedural Skills Competencies. Depending on how the case log tool is constructed, it could be useful for Practice-Based Learning and Improvement and Systems-Based Practice (if patient safety issues and complications are included in the assessment). However, volume of procedures is a crude proxy for competence. Procedure and case logs should be accompanied by assessment of the learner's competence in a procedure or case. Any complications that may have occurred during the procedure can also be documented.

Informed consent is an essential communication skill that could also be combined as part of a case log type of assessment.

Validity

There are few studies of case or procedure logs for the purpose of determining accuracy of learners' recording. Unless defined by CPT or other codes, the cases or procedures counted in a given category may vary across learners and programs. Minimum numbers of procedures required for accreditation and certification have not been rigorously validated against the actual quality of performance of a procedure or patient outcomes. Research for some procedures (e.g., central lines, colonoscopy) has demonstrated validity for some patient outcomes (Rosvall et al. 2024).

Assessing actual performance along a growth curve with a mastery-based approach to teaching and assessing procedural skills should be the standard for assessing procedural and operative skills.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Feasibility

Electronic systems can facilitate the collection and summarization of patient cases or procedures performed. Several specialties have developed smart phone and web-based apps for collecting data and assessing procedural skills. Most learning management systems include a mechanism for collecting and entering case log data. If manual recording is used (i.e., using a paper-based form), the users are required to enter the data into their electronic system. Data entry of manual records typically can be performed by a program coordinator. but such activity is time consuming depending on the number of learners in the program and case log reporting requirements.

Smart phone apps and/or web-based tools substantially enhance the feasibility of gathering case log data. These tools also can provide instant access for learners to their data and progress on specific procedures.

Acceptability

The acceptability of procedure and operative case logs is usually not a major barrier to their use. The main challenge usually involves the feasibility of collecting, entering, and tracking the data. Another issue in evaluating individual learners and programs is that some learners may not understand the value of logging and only record cases up to the minimum requirements. For this reason, it is important to communicate clearly with learners about the goals and expectations for logging cases/procedures. Faculty development is also integral to the acceptability of procedure/case logs to ensure consistency in the feedback process.

Catalytic Effect

If the specific tool is well-designed and includes data (through ratings and narrative text) on the learner's performance, these tools can facilitate feedback conversations between faculty members and learners to support longitudinal learner development.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

References

Abdelstattar Jad M., Yazan N. Aljamal, Raaj K. Ruparel, et al. 2018. "Correlation of Objective Data with General Surgery Resident In-Training Evaluations and Operative Volumes." *Journal of Surgical Education* 75 (6): 1430-1436. https://pubmed.ncbi.nlm.nih.gov/29773409/.

Albert, Dara. V., James R. Brorson, Christina Amidei, et al. 2014. "Education Research: Case Logs in the Assessment of Medical Students in the Neurology Outpatient Clinic." *Neurology* 82 (16): e138-e141. https://www.neurology.org/doi/10.1212/wnl.0000000000000333.

Dehmer, Jeffrey J., Keith D. Amos, Timothy M. Farrell, et al. 2013. "Competence and Confidence with Basic Procedural Skills: The Experience and Opinions of Fourth-Year Medical Students at a Single Institution." *Academic Medicine* 88 (5): 682-687. https://pubmed.ncbi.nlm.nih.gov/23524922/.

Ghaderi, Iman, Farouq Manji, Yoon Soo Park, et al. 2015. "Technical Skills Assessment Toolbox: a Review Using the Unitary Framework of Validity." *Annals of Surgery* 261: (2) 251-262. https://pubmed.ncbi.nlm.nih.gov/24424150/.

Gisondi, Michael A., Linda Regan, Jeremy Branzetti, et al. 2018. "More Learners, Finite Resources, and the Changing Landscape of Procedural Training at the Bedside." *Academic Medicine* 93 (5): 699-704. https://pubmed.ncbi.nlm.nih.gov/29166352/.

Govaerts, Marjan J.B., Cees P.M. van der Vleuten, Eric S. Holmboe. 2019. "Managing Tensions in Assessment: Moving Beyond Either-Or Thinking." *Medical Education* 53 (1): 64-75. https://pmc.ncbi.nlm.nih.gov/articles/PMC6586064/.

Kogan, Jennifer R., Eric S. Holmboe, and Karen E. Hauer. 2009. "Tools for Direct Observation and Assessment of Clinical Skills of Medical Trainees: A Systematic Review." *JAMA* 23 302 (12):1316-1326. https://pubmed.ncbi.nlm.nih.gov/19773567/.

Lopez, B., P.F. Kolecki, D.F. Louis, et al. 2004. "The Use of a Personal Digital Assistant Patient Encounter Log System to Track Procedures Performed by Students During a Mandatory Emergency Medicine Clerkship." *Annals of Emergency Medicine* 2004 44 (4, Suppl.): S48-S49. https://www.sciencedirect.com/science/article/abs/pii/S0196064404008820.

Naik, Nimesh D., Eduardo F. Abbott, Johnathan M. Aho, et al. 2017. "The ACGME Case Log System May Not Accurately Represent Operative Experience Among General Surgery Interns." *Journal of Surgical Education* 74 (6): e106-110. https://www.sciencedirect.com/science/article/abs/pii/S1931720417302428.

Norcini, John, M. Brownell Anderson, Valdes Bollela, et al. 2018. "2018 Consensus Framework for Good Assessment." *Medical Teacher* 40 (11): 1102-1109. https://pubmed.ncbi.nlm.nih.gov/30299187/.

Penciner, Rick, Sanam Siddiqui, and Shirley Lee. 2007. "Emergency Medicine Clerkship Encounter and Procedure Logging Using Handheld Computers." *Academic Emergency Medicine* 14 (8): 727-31. https://pubmed.ncbi.nlm.nih.gov/17601997/.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Ragsdale, John W., Elizabeth B. Seelbach, Sarah Vick, et al. 2022. “Practice Doesn’t Make Perfect: Clinical Experience with Procedures Does Not Correlate Well with Competence in Third-Year Medical Students.” *Journal of Surgical Education* 79 (6): 1441-1446. https://pubmed.ncbi.nlm.nih.gov/35933309/.

Rosvall, Annica, Malin Axelsson, Ervin Toth, Christine Kumlien, Magdalena Annersten Gershater. 2024. “Development and Content Validity Testing of a Colonoscopy-Specific Patient-Reported Experience Measure: The Patient Experience Colonoscopy Scale (PECS).” *Journal of Patient-Reported Outcomes* 8 (32). https://pmc.ncbi.nlm.nih.gov/articles/PMC10948700/.

Sargeant, Joan, Karen Mann, Cees van der Vleuten, et al. 2008. “‘Directed’ Self-Assessment: Practice and Feedback Within a Social Context.” *Journal of Continuing Education in the Health Professions* 28 (1): 47-54. https://pubmed.ncbi.nlm.nih.gov/18366127/.

Schuwirth, Lambert W.T., and Cees P.M. van der Vleuten. 2011. “Programmatic Assessment: from Assessment of Learning to Assessment for Learning.” *Medical Teacher* 33 (6): 478-485. https://pubmed.ncbi.nlm.nih.gov/21609177/.

Stulberg, Jonah J., Reiping Huang, Lindsey Kreutzer, et al. 2020. “Association Between Surgeon Technical Skills and Patient Outcomes.” *JAMA Surgery* 155 (10): 960-968. https://pubmed.ncbi.nlm.nih.gov/32838425/.

van Der Vleuten, C. P. M., L.W.T. Schuwirth, E.W. Driessen, et al. 2015. “Twelve Tips for Programmatic Assessment.” *Medical Teacher* 37 (7): 641-646. https://pubmed.ncbi.nlm.nih.gov/25410481/.

Williams, Reed G., Michael J. Kim, and Gary L. Dunnington. 2016. “Practice Guidelines for Operative Performance Assessments.” *Annals of Surgery* 264 (6): 934-948. https://journals.lww.com/annalsofsurgery/Abstract/2016/12000/Practice Guidelines for Operative Performance.13.aspx.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Review of Medical Errors and Patient Safety Events

Description

The review of medical errors and patient safety events as an assessment method in medical education involves engaging learners in the analysis, discussion, or reflection on real or simulated clinical errors**.** These reviews may occur through:

- morbidity and mortality (M and M) conferences;
- root cause analyses (RCAs);
- patient safety incident reports; and
- simulated adverse event debriefings

Learners are asked to analyze contributing factors, propose prevention strategies, and reflect on their roles and responses. Assessments can include written reflections, presentations, or structured debriefings evaluated using rubrics aligned with safety and professionalism domains.

Competencies

Review of medical errors and patient safety events can be used to assess Medical Knowledge, Practice-Based Learning and Improvement, and Systems-Based Practice.

Validity

The review of medical errors can be a valid assessment of how learners think about safety, accountability, and improvement — especially when linked to observed behaviors or structured reflections. Validity improves when using standardized cases, clear scoring criteria, and trained evaluators. There is strong construct validity when tied to authentic tasks and decisions faced in clinical settings.

Feasibility

Feasibility varies depending on the format:

- Simulated cases are highly feasible and scalable across institutions.
- Participation in real patient safety reviews may be limited by legal, institutional, or privacy barriers.
- Written reflections or analyses of anonymized events are commonly used and require moderate faculty oversight.

Integration into interprofessional training or quality improvement initiatives can improve feasibility and relevance.

Acceptability

Acceptability among educators is generally high due to alignment with national safety goals. Among learners, acceptability increases when the process:

- Emphasizes learning over punishment.
- Occurs in psychologically safe environments.
- Is supported by coaching or mentorship.

Concerns may arise if learners fear being evaluated for admitting mistakes or if institutional culture is punitive.

Catalytic Effect

Reviewing medical errors and safety events has strong catalytic potential, especially when learners engage in structured reflection and feedback:

- Promotes a culture of safety and openness.
- Encourages systems thinking and responsibility.
- Builds resilience and strategies for error disclosure and emotional response.
- Reinforces the importance of patient advocacy and accountability.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

References

Association of American Medical Colleges. *Quality Improvement and Patient Safety Competencies Across the Learning Continuum (Version 2)*. 2024. Accessed April 23, 2025. https://www.aamc.org/data-reports/report/qipscompetencies.

Fischer, Melissa A., Kathleen M. Mazor, Joann Baril, Eric Alper, Deborah DeMarco, and Michele Pugnaire. 2006. "Learning from Mistakes: Factors That Influence How Students and Residents Learn from Medical Errors." *Journal of General Internal Medicine* 21 (5): 419–423. https://pmc.ncbi.nlm.nih.gov/articles/PMC1484785/.

Kiesewetter, Isabel, Karen D. Könings, Moritz Kager, and Jane Kiesewetter. 2018. "Undergraduate Medical Students' Behavioural Intentions Towards Medical Errors and How to Handle Them: A Qualitative Vignette Study." *BMJ Open* 8 (3):e019500. https://pmc.ncbi.nlm.nih.gov/articles/PMC5857650/.

Leape, Lucian L. 1994. "Error in Medicine." *JAMA* 272 (23):1851–1857. https://jamanetwork.com/journals/jama/article-abstract/384554.

Lee, Sungjoon, HyeRin Roh, Myounghun Kim, and Ji Kyoung Park. 2022. "Evaluating Medical Students' Ability to Identify and Report Errors: Finding Gaps in Patient Safety Education." *Medical Education Online* 27 (1): 2011604. https://pubmed.ncbi.nlm.nih.gov/35129092/.

Martinez, William, and Bernard Lo. 2008. "Medical Students' Experiences with Medical Errors: An Analysis of Medical Student Essays." *Medical Education* 42 (7): 733–741. https://pubmed.ncbi.nlm.nih.gov/18507766/.

Mohsin, Syed Umer, Yahya Ibrahim, and Diane Levine. 2019. "Teaching Medical Students to Recognise and Report Errors." *BMJ Open Quality* 8 (2): e000558. https://pubmed.ncbi.nlm.nih.gov/31276054/.

Murphy, Martina, Jennifer Duff, Julie Whitney, Benjamin Canales, Merry-Jennifer Markham, and Julia Close. 2017. "Implementation of a Mock Root Cause Analysis to Provide Simulated Patient Safety Training." *BMJ Open Quality* 6 (2): e000096. https://pmc.ncbi.nlm.nih.gov/articles/PMC5699135/.

Patey, Rona, Rhona Flin, Brian H. Cuthbertson, et al. 2007. "Patient Safety: Helping Medical Students Understand Error in Healthcare." *Quality & Safety in Health Care* 16 (4): 256–259. https://pmc.ncbi.nlm.nih.gov/articles/PMC2464940/.

Ryder, Hilary F., Jonathan T. Huntington, Alan West, and Greg Ogrinc. 2019. "What Do I Do When Something Goes Wrong? Teaching Medical Students to Identify, Understand, and Engage in Reporting Medical Errors." *Academic Medicine* 94 (12): 1910–1915. https://pubmed.ncbi.nlm.nih.gov/31335816/.

Safarpour, Hamid, Maryam Tofighi, Leila Malekyan, Jafar Bazyar, Saeideh Varasteh, and Rahman Anvary. 2017. *International Journal of Clinical Medicine* 8 (1): 1–11. https://file.scirp.org/pdf/IJCM_2017012014404352.pdf.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Weiss, Kevin B., James P. Bagian, and Thomas J. Nasca. 2013. “The Clinical Learning Environment: The Foundation of Graduate Medical Education.” *JAMA* 309 (16): 1687–1688. https://pubmed.ncbi.nlm.nih.gov/23613072/.

Wu, A.W. 2000. “Medical Error: The Second Victim. The Doctor Who Makes the Mistake Needs Help Too.” *BMJ (Clinical Research Edition)*. 320 (7237):726–727. https://pubmed.ncbi.nlm.nih.gov/10720336/.

Research and Scholarly Projects

Description

Research and scholarly projects often involve original research, quality improvement initiatives, educational innovations, or community health interventions. Assessment methods typically include:

- written proposals and final papers;
- poster or oral presentations;
- research portfolios;
- reflections or peer evaluations; and
- structured rubrics aligned with research competencies.

Faculty members assess these projects based on criteria such as question formulation, methodological rigor, ethical conduct, data interpretation, and communication of findings.

Competencies

Research and scholarly projects often include one or more Core Competencies including Medical Knowledge, Practice-Based Learning and Improvement, and Interpersonal and Communication Skills.

Validity

The validity of research project assessments is supported when clear criteria and structured rubrics are used, and when multiple dimensions (e.g., process, product, and presentation) are evaluated. Construct validity is strong when assessments reflect real-world scholarly tasks. Including mentor and peer feedback strengthens the authenticity and depth of the assessment.

Feasibility

Feasibility depends on institutional resources, faculty availability, and project scope. Programs with longitudinal research tracks, scholarly concentrations, or summer programs make this more feasible. Use of standardized templates, virtual presentations, or poster days improves efficiency. Faculty mentorship demands are a known challenge, but are often offset by shared interest in scholarship.

Acceptability

Research projects are generally well-accepted by:

- learners**,** who appreciate applying classroom learning in authentic contexts;
- faculty members**,** who value mentoring and scholarly engagement; and
- institutions, which often integrate these into graduation requirements or capstones.

Concerns may include perceived disparities in mentorship quality or unequal access to high-quality project opportunities, which can be mitigated by structured support.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Catalytic Effect

Scholarly projects have a strong catalytic effect, including:

- fostering curiosity and inquiry;
- improving research literacy and critical appraisal skills;
- encouraging academic careers and lifelong learning; and
- developing confidence in addressing complex clinical or systems problems.

When projects are well-mentored, they enhance identity formation as physician-scholars and can influence specialty choice or engagement in advocacy.

References

Bierer, S. Beth, and Huiju Carrie Chen. 2010. “How to Measure Success: The Impact of Scholarly Concentrations on Students—A Literature Review.” *Academic Medicine* 85 (3): 438–452. https://pubmed.ncbi.nlm.nih.gov/20182116/.

Chang, Youjin, and Christopher J. Ramnanan. 2015. “A Review of Literature on Medical Students and Scholarly Research: Experiences, Attitudes, and Outcomes.” *Academic Medicine* 90 (8): 1162–1173. https://pubmed.ncbi.nlm.nih.gov/25853690/.

Glassick, Charles E, Mary Taylor Huber, and Gene I. Maeroff. 1997. *Scholarship Assessed: Evaluation of the Professoriate*. Jossey-Bass.

Green, Emily P., Jeffrey M. Borkan, Susan H. Pross, et al. 2010. “Encouraging Scholarship: Medical School Programs to Promote Student Inquiry Beyond the Traditional Medical Curriculum.” *Academic Medicine* 85 (3): 409–418. https://pubmed.ncbi.nlm.nih.gov/20182113/.

Houlden, Robyn L., Jamila B. Raja, Christine P. Collier, Albert F. Clark, and Jennifer M. Waugh. 2004. “Medical Students’ Perceptions of an Undergraduate Research Elective.” *Medical Teacher* 26 (7): 659–661. https://pubmed.ncbi.nlm.nih.gov/15763861/.

Simpson, Deborah, Ruth-Marie Fincher, Janet P. Hafler, et al. 2007. “Advancing Educators and Education by Defining the Components and Evidence Associated with Educational Scholarship.” *Medical Education* 41 (10): 1002–1009. https://pubmed.ncbi.nlm.nih.gov/17822412/.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Simulation

Description

In general terms, medical simulations aim to imitate real patients, anatomic regions, or clinical tasks, and to mirror the real-life situations in which medical services are rendered. Such simulations range from static anatomic models and single-task trainers (e.g., venipuncture arms and intubation mannequin heads) to dynamic computer-enhanced systems that can respond to user actions (e.g., full-body patient simulators); from relatively low-technology standardized patient (SP) encounters to very high-tech virtual reality surgical simulators; and from individual trainers for evaluating the performance of a single user to interactive role-playing scenarios involving teams of health professionals. "Simula*tion*" refers broadly to any device or set of conditions – including, for example, standardized patient-based exams – that attempts to present evaluation problems authentically, whereas a "simula*tor*," more narrowly defined, is a simulation *device*.

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.

Mannequins are imitations of body organs or anatomical body regions frequently using pathological findings to simulate patient disease. The models are constructed of various materials sculpted to resemble human tissue with embedded electronic circuitry to allow the mannequin to respond realistically to actions by the examinee. Virtual reality simulations or environments use computers sometimes combined with anatomical models to mimic as much as is feasible realistic organ and surface images and the touch sensations (computer-generated haptic responses) a physician would expect in a real patient. Virtual reality environments allow assessment of procedural skills and other complex clinical tasks that are difficult to assess consistently by other assessment methods.

Competencies

Simulation can potentially assess a range of competencies at the "Shows How" level of Miller's Pyramid (1990). Simulation is becoming increasingly important as part of mastery-based learning for Patient Care and Procedural Skills (see below). Assessment of clinical reasoning (Medical Knowledge) is possible and can also be incorporated into simulation scenarios. Simulations of interprofessional teamwork are growing in interest and can assess capability in Interpersonal and Communication Skills. A collection of simulation exercises has potential to better discriminate between learner performance than clinical evaluations.

Validity

Substantial progress has been made in validity research of simulations over the past decade. Studies of high-quality simulations have demonstrated their content validity when the simulation is designed to resemble a real patient, including objective structured clinical exams (OSCEs) discussed in another section. Studies on mastery-based learning using various forms of

simulation have found some positive correlations between performance in the simulations and quality of care for patients. Virtual reality and partial-task trainers have also been shown to enhance training and performance in procedures. A full treatise of the validity of simulation is beyond the scope of this Guidebook; the reader is encouraged to access the references provided below.

Feasibility

It is strongly recommended that programs and institutions interested in using simulation collaborate with experts at a simulation center. If an institution does not possess a simulation center or specific simulation tools and materials, it should reach out to a simulation center in the local or regional area. Simulation centers are increasingly creating consortia to share expertise and provide services to outside programs or institutions. The most critical feasibility challenges in simulation, depending on the specific method or approach used, are cost and access to the requisite expertise, materials, and setting.

Acceptability

Acceptability of simulation, especially when used for formative purposes, is quite high. When simulation is used for high-stakes, summative purposes, acceptability can vary depending on the nature of the simulation and purpose.

Catalytic Effect

Well-designed and executed simulations can produce a powerful learning effect. As a result, simulation can be an especially meaningful supplement to clinical experience when it involves a clinical scenario that is rare, but serious. This finding appears to be especially true when mastery-based learning principles are incorporated into the simulation design and experience.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

References

Abdelstattar Jad M., Yazan N. Aljamal, Raaj K. Ruparel, et al. 2018. “Correlation of Objective Data with General Surgery Resident In-Training Evaluations and Operative Volumes.” *Journal of Surgical Education* 75 (6): 1430-1436. https://pubmed.ncbi.nlm.nih.gov/29773409/.

McGaghie, William C., Jeffrey H. Barsuk, and Diane B. Wayne. 2020. *Comprehensive Healthcare Simulation: Mastery Learning in Health Professions Education.* Springer.

Motola, Ivette, Luke A. Devine, Hyan Soo Chung, et al. 2013. “Simulation in Healthcare Education: A Best Evidence Practical Guide. AMEE Guide No. 82.” *Medical Teacher* 35 (10): e1511-1530. https://pubmed.ncbi.nlm.nih.gov/23941678/.

Scalese R. 2018. “Simulation-Based Assessment.” In *Practical Guide to the Evaluation of Clinical Competence* by Eric S. Holmboe, Steven J. Durning, and Richard E. Hawkins. Elsevier.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Standardized (Simulated) Patients (OSCEs)

Description

Direct observation that occurs via simulation of a patient by lay individuals (i.e., “actors”) uses what are known as standardized or simulated patients (SPs). An SP is an individual trained to portray a patient (or sometimes a family member, other health care professional, bystander, etc.) who can provide education and training and may also perform assessment of learners for specific competencies. SPs can also be trained to document and report back learner actions and behaviors, teach learners via role-play and repeated practice, rate learner Interpersonal and Communication Skills, and also provide detailed and timely feedback. SPs are an excellent means to provide first exposure—combined with assessment, feedback, and coaching—to difficult encounters, such as breaking bad news, working with agitated or upset patients and families, disclosing medical errors, making informed decisions in complex clinical situations, conducing hand-offs, etc.

OSCEs are a formal, standardized use of SPs in series of stations to assess and rate the clinical skills of trainees. OSCEs can be high- or low-stakes, depending on their purpose. Some GME programs use OSCEs during residency orientation to perform a baseline needs assessment for the incoming resident.

There is also an increasing body of evidence regarding use of unannounced SPs (“secret shoppers”) in clinical settings. This enables an SP to be embedded into the context of actual clinical practice. Multiple studies have shown this is a useful assessment approach (Rethans et al. 2007; Weiner and Schwartz 2014; Zabar et al. 2009).

Competencies

SPs assess the “Shows How” level of Miller’s Pyramid (1990) and can be a very useful supplement to direct observation of actual clinical care. SPs can be especially useful for assessing the competencies of Patient Care and Interpersonal and Communication Skills for the one-on-one patient encounter. SPs and OSCEs can be combined with assessment of Medical Knowledge (either through written notes or in-person questioning), Systems-Based Practice when scenarios include care coordination, need to include community resources, etc., and professionalism when the scenario includes ethical issues, dealing with error disclosure, conflict, etc.

Validity

There is extensive literature on SPs regarding the validity elements of content, response process, and reliability (Harden et al. 2015; Motola et al. 2013; Pell et al. 2010). At one time, OSCEs were part of the USMLE® Step 2 clinical skills exam. Less evidence is available around the relationship of OSCEs to other assessment data, especially high-stakes OSCEs and future performance in education and training or practice.

Feasibility

Development of an exam using SPs involves identifying the specific competencies to be assessed, training SPs, developing checklists or rating forms, and setting criteria. Development time can be considerable but can be made more efficient by sharing SPs in collaboration with other programs within the same institution or at other institutions. A new SP can learn to stimulate a new clinical problem in eight to 10 hours; an experienced SP can learn a new problem in six to eight hours. About twice the amount of training time is needed for SPs to learn to use checklists to evaluate learner performance. Facilities needed for the exam include

an examining room for each SP station/clinical setting and space for learners to record medical notes between stations. Additional logistical considerations will be needed for unannounced SPs used in clinical settings (scheduling, ensuring the SP is not identifiable, etc.).

Acceptability

SPs have become a staple of education, training, and assessment in undergraduate medical education, and increasingly so in GME. Both educators and learners see value in learning from and being assessed by SPs.

Catalytic Effect

SPs can provide excellent real-time feedback when used mostly for teaching and feedback. Score reports from more formal SP encounters (i.e., OSCEs, unannounced SPs) can also be valuable provided sufficiently detailed information about performance is provided. Providing learners with only numeric ratings or scores will not be very useful for guiding ILPs.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

References

Harden, Ronald M., Pat Lilley, and Madalena Patricio. 2015. *The Definitive Guide to the OSCE.* Elsevier.

Motola, Ivette, Luke A. Devine, Hyun Soo Chung, et al. 2013. “Simulation in Healthcare Education: A Best Evidence Practical Guide. AMEE Guide No. 82.” *Medical Teacher* 35 (10): e1511-1530. https://pubmed.ncbi.nlm.nih.gov/23941678/.

Pell, Godfrey, Richard Fuller, Matthew Homer, et al. 2010. “How to Measure the Quality of the OSCE: A Review of Metrics - AMEE Guide No. 49.” *Medical Teacher* 32 (10): 802-811. https://pubmed.ncbi.nlm.nih.gov/20854155/.

Rethans, Jan-Joost, Simone Gorter, Lonneke Bokken, and Linda Morrison. 2007. “Unannounced Standardised Patients in Real Practice: A Systematic Literature Review.” *Medical Education* 41 (6): 537-549. https://pubmed.ncbi.nlm.nih.gov/17518833/.

Weiner, Saul J., and Alan Schwartz. 2014. “Directly Observed Care: Can Unannounced Standardized Patients Address a Performance Gap in Performance Measurement?” *Journal of General Internal Medicine* 29 (8): 1183-1187. https://pmc.ncbi.nlm.nih.gov/articles/PMC4099461/.

Zabar, Sondra, Tavinder Ark, Colleen Gillespie, et al. 2009. “Can Unannounced Standardized Patients Assess Professionalism and Communication Skills in the Emergency Department?” *Academic Emergency Medicine* 16 (9): 915-918. https://pubmed.ncbi.nlm.nih.gov/19673703/.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Workplace-Based Assessments

Description

Assessment of the “Does” level of Miller’s pyramid (1990) typically requires direct workplace observation, i.e., evaluating performance in the context of actual clinical practice. Workplace-based assessment refers to an “assessment of what doctors actually do in practice.” (Swanwick and Chana 2009). Workplace-based assessments are a vital part of medical education, designed to evaluate clinical or procedural skills and professional behaviors in real-world settings either in a summative or formative manner. The three essential components of workplace-based assessments are: direct observation, conduction at workplace, and contextual, constructive, immediate feedback.

Various types of workplace-based assessments exist and fall into the following general categories:

1. *Observation of clinical activities*, including the mini-clinical evaluation exercise (mini-CEX, clinical encounter cards, clinical work sampling) and direct observation of procedural skills
2. *Clinical Case Discussion,* including chart-stimulated recall and case-based discussions
3. *Feedback from peers, coworkers, and patients, or multiple sources (360-degree assessment*), including the mini-assessment tool team assessment of behaviors and the patient satisfaction questionnaire.
4. *Documentation of Work*, including logbooks, clinical encounter cards, and portfolios.

Competencies

Depending on the goal of workplace-based assessments, they may be effective at addressing several Core Competencies. Given that they are done in a clinical setting, workplace-based assessments usually focus on Patient Care and Procedural Skills. However, if gathering multisource feedback, they may also be effective at assessing Medical Knowledge, Professionalism, Interpersonal and Communication Skills or Systems-Based Practice.

Validity

Workplace-based assessments have the potential to be highly valid assessments, but their validity is contingent on careful attention to design, implementation, and interpretation. Standardization of assessment tools and rater training are crucial for improving reliability. Assessing learners in a wide range of clinical experiences is also essential for ensuring generalizability across varying sites. The content validity of work-based assessments can be increased by incorporating all domains (attitudes, skills and knowledge) and all competencies. They need to be implicit, standardized, and based on objective judgments.

Feasibility

Implementing workplace-based assessments can be resource-intensive, requiring significant time and effort from assessors. Unlike standardized settings, workplaces are not universally welcoming environments for learners and do not necessarily offer equal opportunities to all learners. Changes in clinical service organization and delivery have resulted in shorter care time, leaving learners to compete for training time, patient contact, and appropriate supervision. As a result, ensuring consistency and adequate supervision and assessment opportunities for all learners can be challenging. Faculty development training, student sensitization, and blueprinting are also needed.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Acceptability

Learners value work-based learning experiences and seek opportunities to engage in work-based learning. They want a structured environment where work-based learning is valued, protected and appropriately supervised. However, work-based assessments are often considered less informal learning, and may be less valued by teachers and learners. Creating an atmosphere of mutual trust to ensure both delivery and reception of constructive, nonjudgmental feedback is critical to the success of any workplace-based assessment.

Catalytic Effect

Workplace-based assessments promote active learning through direct engagement with patients, clinical tasks, and problem-solving. Feedback provided during workplace-based assessments helps learners identify their strengths and weaknesses, enabling them to target specific areas for improvement and refine their clinical skills. Overall, these types of assessments empower learners to function with appropriate responsiveness in real-life patient scenarios.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

References

Norcini, John, and Vanessa Burch. 2007. “Workplace-Based Assessment As an Educational Tool: AMEE Guide No. 31.” *Medical Teacher* 29 (9): 855-871. https://pubmed.ncbi.nlm.nih.gov/18158655/.

Singh, Kirti, and Arshi Singh. 2021. “Workplace-Based Assessment: A Real-Time Assessment Tool.” *MAMC Journal of Medical Sciences* 7 (3): 204-210. https://journals.lww.com/mamc/fulltext/2021/07030/workplace_based_assessment__a_real_time_assessment.3.aspx.

Swanwick, Tim, and Nav Chana. 2009. “Workplace-Based Assessment.” *British Journal of Hospital Medicine* 70 (5): 290-293. https://pubmed.ncbi.nlm.nih.gov/19451876/.

Direct Observation of Clinical Skills

Description

Direct observation of clinical skills refers to observing a learner performing a task (e.g., taking a history from a patient, performing a physical exam, discussing informed consent, or leading a shared decision-making conversation.

Direct observation of clinical skills is a workplace-based assessment daily strategy that evaluates the learner in authentic clinical settings. Thus, direct observation of clinical skills is a critically important tool to assess the actual behavior of the learner in these settings, and it addresses the top of Miller's Pyramid (1990). Since the assessor is watching the learner directly, this type of assessment can provide rich formative feedback. A possible limitation of is that the act of being observed may change the learner's performance (i.e., the Hawthorne effect).

Competencies

Direct observation is essential for assessing Patient Care and Interpersonal and Communication Skills. It can also be used to assess Professionalism in interactions with patients, family members, and members of the interprofessional team. When combined with questioning, it can also be used to assess the underlying Medical Knowledge and clinical reasoning that informed the observed behaviors.

Validity

Substantial literature exists regarding the reliability of certain direct observation tools, particularly the mini-clinical evaluation exercise (mini-CEX). Many tools such as this one have evidence of internal construct validity. High levels of reliability can typically be achieved when a sufficient number of assessments are performed by multiple observers who have received sufficient training. Evidence for the other validity components, such as correlations with other measures of clinical performance, is dependent on the specific tool. Faculty development can be one of the biggest challenges to validity. Faculty members may lack sufficient competence in the clinical skills they are assessing, or they may lack familiarity with the assessment tool or best practices for conducting a direct observation.

Feasibility

While time is always a factor, direct observation can be done in smaller aliquots, sometimes called "snapshots," as part of the routine work of faculty members. Technological solutions (e.g., apps) can help faculty members complete assessments quickly and in the moment. Some examples of snapshots are provided below.

Interview	Physical Exam	Counseling	Procedures
1. Agenda setting for outpatient visit 2. Portion of history 3. Pre-rounds	1. Focused physical exam maneuver 2. Part of physical exam 3. Pre-rounds	1. Post-rounds 2. Discharge 3. Starting a medication/therapy 4. Counseling for behavior change	1. Informed consent 2. Procedure 3. Procedural post-check

Acceptability

Acceptability among faculty members and learners depends heavily on culture. Program and institutional cultures that: 1) discourage direct observation/have direct observation occur infrequently; 2) make each observation high stakes; 3) lack faculty member buy-in and/or

4) provide infrequent or poor feedback, can undermine the acceptability of direct observation. In particular, infrequent direct observation can lead to each encounter feeling high stakes to the learner. Programs and institutions that build frequent direct observation into the program of assessment and support faculty members in developing this assessment skill achieve greater acceptability.

Catalytic Effect

Direct observation of clinical skills can be a catalyst to a discussion with the learner that involves detailed feedback and coaching. Too often faculty members assess clinical skills through proxy measures (e.g., assessing a learner's history taking through an oral presentation of that history). When done well, direct observation allows the faculty member to see the learner's performance firsthand, discuss their observations, and provide concrete guidance. This process can have a catalytic effect to promote better implementation of the feedback by the learner in future encounters.

Suggested Direct Observation Tools

Direct Observation of Clinical Care (DOCC)

- A tool designed to assess: performing a history and physical exam, effective clinical reasoning, informed decision making, breaking bad news, and safe hand-offs
- Available at: https://dl.acgme.org/pages/assessment

Mini-Clinical Evaluation Exercise (Mini-CEX)

- A tool designed to assess: interviewing, physical exam, professionalism, clinical judgment, counseling, and organization
- Available at: https://www.abim.org/media/rybg2qlg/mini-cex.pdf.

Observation Rating Tool (ORT)

- A tool designed to assess: communication, history-taking, clinical reasoning
- Available here: https://bmcmededuc.biomedcentral.com/articles/10.1186/s12909-020-02110-8

TeamSTEPPS Team Performance Observation Tool

- A tool designed to assess: team structure, communication, leadership, situation monitoring, and mutual support
- Available here: https://www.ahrq.gov/teamstepps-program/resources/tools/index.html

References

Alex, Capri P., H. Barrett Fromme, Larrie Greenberg, et al. 2024. “Exploring Medical Student Experiences with Direct Observation During the Pediatric Clerkship.” *Academic Medicine* 99 (9): 997-1006. https://pubmed.ncbi.nlm.nih.gov/38696720/.

Fromme, H. Barrett, Reena Karani, and Steven M. Downing. 2009. “Direct Observation in Medical Education: A Review of the Literature and Evidence for Validity.” *The Mount Sinai Journal of Medicine* 76 (4): 365-371. https://pubmed.ncbi.nlm.nih.gov/19642150/.

Haring, Catharina M., Claudia C.R. Klaarwater, Geert A. Bouwmans, et al. 2020. “Validity, Reliability and Feasibility and a New Observation Rating Tool and a Post Encounter Rating Tool for the Assessment of Clinical Reasoning Skills of Medical Students During Their Internal Medicine Clerkship: A Pilot Study.” *BMC Medical Education* 20 (1):198. https://pubmed.ncbi.nlm.nih.gov/32560648/.

Hauer, Karen E., Eric S. Holmboe, Jennifer R. Kogan. 2011. “Twelve Tips for Implementing Tools for Direct Observation of Medical Trainees’ Clinical Skills During Patient Encounters.” *Medical Teacher* 33 (1): 27-33. https://pubmed.ncbi.nlm.nih.gov/20874011/.

Kogan, Jennifer R., Eric S. Holmboe, and Karen E. Hauer. 2009. “Tools for Direct Observation and Assessment of Clinical Skills of Medical Trainees: A Systematic Review.” *JAMA* 302 (12): 1316-1326**.** https://pubmed.ncbi.nlm.nih.gov/19773567/.

Kogan, Jennifer R., Rose Hatala, Karen E. Hauer, et al. 2017. “Guidelines: The Do’s, Don’ts and Don’t Knows of Direct Observation of Clinical Skills in Medical Education.” *Perspectives in Medical Education* 6 (5): 286–305. https://pmejournal.org/articles/10.1007/S40037-017-0376-7.

Mazor, Kathleen M., Matthew C. Holtman, Yakov Shchukin, et al. 2011. “The Relationship Between Direct Observation, Knowledge, and Feedback: Results of a National Survey.” *Academic Medicine* 86 (10, Suppl.): S63-S68. https://pubmed.ncbi.nlm.nih.gov/21955772/.

Mortaz Hejri, Sara, Mohammad Jalili, Rasoul Masoomi, et al. 2020. “The Utility of Mini-Clinical Evaluation Exercise in Undergraduate and Postgraduate Medical Education: A BEME Review: BEME Guide No. 59.” *Medical Teacher* 42 (2): 125-142. https://pubmed.ncbi.nlm.nih.gov/31524016/.

Norcini, John J., Linda L. Blank, F. Daniel Duffy, et. al. 2003. “The Mini-CEX: A Method for Assessing Clinical Skills.” *Annals of Internal Medicine* 138 (6): 476-481. https://pubmed.ncbi.nlm.nih.gov/12639081/.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Direct Observation of Procedural Skills

Description

Direct observation of procedural skills refers to observing a learner performing a range of procedures, (e.g., lumbar puncture, central venous catheters, endoscopy, complex surgical procedures). Each specialty has procedures considered essential to that discipline. In GME, for many cases, core procedures are codified by the ACGME Review Committees and the various certifying boards.

As is the case with clinical skills, direct observation is essential to the assessment of procedural skills. Faculty member proficiency in both the procedure being observed and assessment is necessary for effective direct observation of procedural skills.

Competencies

Direct observation of procedural skills focuses on assessing Patient Care and Procedural Skills and Interpersonal and Communication Skills. Performing procedures almost always involves a team. Direct observation by faculty members and other members of the healthcare team can also inform learner competence in clinical reasoning and interprofessional teamwork.

Validity

Substantial literature exists regarding the reliability of procedurally based direct observation tools, such as the objective structured assessment of technical skills (OSATS), the Zwisch Scale (George et al. 2014), and other tools. Similar to assessment tools for clinical skills, it must be emphasized that high levels of reliability can be achieved when a sufficient number of assessments are performed by multiple observers. Faculty members should be primarily encouraged to do large numbers of assessments in order to increase reliability and to overcome angst about the need for perfect accuracy on each assessment. However, the same variability problem that hampers tools for clinical skills also affects assessment tools for procedural skills. Faculty members often lack shared mental models about optimal approaches to procedures even when effective practice has been codified. Faculty members may also lack sufficient levels of skill in the procedural skills they are assessing. While faculty development cannot fix all challenges, training in both the procedural skills of interest and assessment can help.

Feasibility

Procedures enable opportunities for direct observation, since faculty members are often participating in the procedure in some manner (i.e., supervising, assisting). Smartphone assessment apps can help faculty members complete assessments more efficiently.

Acceptability

Acceptability among faculty members and learners depends heavily on program and institutional culture and frequency of observation. Program and institutional cultures that discourage direct observation, make each observation high stakes, lack faculty member buy-in, and/or provide infrequent or poor feedback all undermine acceptability of direct observation. Infrequent observation of procedural skills can lead to each encounter feeling high stakes to the learner. Programs and institutions that make direct observation of procedural skills a habit, train and support faculty members in this assessment skill, and provide a safe learning environment for learners to seek direct observation of procedural skills achieve greater acceptability.

The learner must be made aware of the criteria that will be used when reviewing the chart(s). A checklist of questions should be developed. Faculty assessors should be trained regarding the rational for specific questions and the desired responses to those questions. Comments should be provided by the faculty assessor documenting the learner's response to questions. Face-to-face discussion and feedback to the learner should occur, including discussion about what the learner will do differently moving forward.

Catalytic Effect

Direct observation of procedural skills is essential in enabling effective feedback and coaching. Too often faculty members judge clinical skills through proxies, such as a learner presenting a patient at morning report or morbidity and mortality conferences. When done well, direct observation can be very impactful in helping learners improve procedural skills and outcomes for patients.

References

Abdelsattar, Jad M., Yazan.N. Aljamal, Raaj K. et al. 2018. “Correlation of Objective Data with General Surgery Resident In-Training Evaluations and Operative Volumes.” *Journal of Surgical Education* 75: 1430-6. https://pubmed.ncbi.nlm.nih.gov/29773409/.

George, Brian C., Ezra N. Teitelbaum, Shari L. Meyerson, et al. 2014. “Reliability, Validity, and Feasibility of the Zwisch Scale for the Assessment of Intraoperative Performance.” *Journal of Surgical Education* 71(6): e90-6. https://pubmed.ncbi.nlm.nih.gov/25192794/.

George, Brian C., Jordan D. Bohnen, Reed C. Williams, et al. 2017. “Readiness of US General Surgery Residents for Independent Practice.” *Annals of Surgery* 266(4):582-594. https://pubmed.ncbi.nlm.nih.gov/28742711/.

George, Brian C., Jordan D. Bohnen, Mary C. Schuller, Jonathan P. Fryer. 2020. “Using Smartphones for Trainee Performance Assessment: A SIMPL Case Study.” *Surgery* 167(6): 903-906. https://pubmed.ncbi.nlm.nih.gov/31668358/.

Hatala, Rose, David A. Cook, Ryan Brydges, and Richard Hawkins. 2015. “Constructing a Validity Argument for the Objective Structured Assessment of Technical Skills (OSATS): A Systematic Review of Validity Evidence.” *Advances in Health Sciences Education and Theory Practice* 20(5): 1149-75. https://pubmed.ncbi.nlm.nih.gov/25702196/.

Husk, Katherine E., Lee A. Learman, Christine Field, AnnaMarie Connolly. 2020. “Implementation and Initial Construct Validity Evidence of a Tool, myTIPreport, for Interactive Workplace Feedback on ACGME Milestones.” *Journal of Surgical Education* 7(6): 1334-1340. https://pubmed.ncbi.nlm.nih.gov/32546386/.

Thanawala, Ruchi M., Jonathan L. Jesneck, Neal E. Seymour. 2019. “Education Management Platform Enables Delivery and Comparison of Multiple Evaluation Types.” *Journal of Surgical Education* 76(6): e209-216. https://pubmed.ncbi.nlm.nih.gov/31515199/.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Chart-Stimulated Recall and the Assessment of Clinical Reasoning in the Workplace

Description

Chart-stimulated recall is a work-based assessment tool that complements chart audit by combining the chart review of a patient encounter with a structured oral exam. Chart-stimulated recall provides a structured approach to the use of clinical questioning. With appropriate faculty development, a faculty member uses a set of predetermined questions to probe a learner's thought process in areas including clinical assessment (Medical Knowledge; Patient Care and Procedural Skills), selection and interpretation of clinical finding/labs (Medical Knowledge), and treatment and management plans, including the use of referrals (Medical Knowledge; Patient Care and Procedural Skills; Systems-Based Practice; Interpersonal and Communications Skills). Using reflective questioning and prompted self-assessment, chart-stimulated recall can also assess some aspects of Professionalism and Practice-Based Learning and Improvement.

The assessment of reasoning tool by the Society to Improve Diagnosis in Medicine can be used to directly assess clinical reasoning skills during patient care activities (e.g., precepting, rounding) or guide questioning as part of a chart-stimulated recall exercise. This tool provides a structured approach to assessing hypothesis-directed data collection, problem representation, prioritized differential diagnosis, high-value testing, and metacognition. The Society to Improve Clinical Reasoning in Medicine has also provided a just-in-time faculty tutorial for use of this tool.

When conducting a chart-stimulated recall exercise, the following steps should be taken:

1. The learner must be made aware of the criteria that will be used when reviewing the chart(s).
2. A checklist of questions should be developed.
3. Faculty assessors should be trained regarding the rational for specific questions and the desired responses to those questions.
4. Comments should be provided by the faculty assessor documenting the learner's response to questions.
5. Face-to-face discussion and feedback to the learner should occur, including discussion about what the learner will do differently moving forward.

Competencies

Chart-stimulated recall and the assessment of reasoning tool predominantly assess Medical Knowledge and Patient Care and Procedural Skills. Chart-stimulated recall can also address other Core Competencies based on the structure and desired outcome of the tool/exercise, including Interpersonal and Communication Skills and Professionalism.

Validity

When structured appropriately, chart-stimulated recall can provide valid assessments of clinical reasoning. However, there can be validity issues with memory decay, recall bias, and the potential for post-hoc reflection to cloud in-action thought processes. As early as 1982, the American Board of Emergency Medicine demonstrated that well-designed chart-stimulated recall exercises were valid and reliable enough to inform board certification decisions (Munger et al. 1982). Educators must consider the timing of stimulus and interviews, the type of cues used, and the level of interaction to ensure the validity of the data. The validity and reliability can

be maximized by adhering to strict protocols such as the use of non-directive questioning or prompts, immediacy of interview after the activity, and the initiation of pauses within the activity for recall. This degree of structure and preparation is not easily reproduced in programs and institutions. Programs and institutions may, therefore, use chart-stimulated recall as a preferred formative assessment tool to guide ILPs and feedback to learners regarding clinical reasoning skills.

Feasibility

Successful use of chart-stimulated recall requires that the specifics of "how," "when," and "why" are clearly understood by everyone participating in the exercise. Programmatic use of chart-stimulated recall requires advanced planning and is not an assessment that can be used "on the fly." The assessment of clinical reasoning tool can be used during precepting and rounds, and any time a diagnosis is being discussed, making it a good tool to combine with other direct observation assessment methods. When using chart-stimulated recall, programs should determine the rotation(s)/learning venues that will use chart-stimulated recall, the location, the learners who will be assessed, the faculty members who will conduct chart-stimulated recall, and the desired outcome of the exercise. The specific questions used in the exercise and the frequency with which chart-stimulated recall is completed should also be standardized. Finally, adequate time (i.e., 30 to 60 minutes) must also be allocated to complete the chart-stimulated recall exercise. Given the complexities of chart-stimulated recall, care must be taken to ensure its value is fully defined and explained.

Acceptability

With appropriate preparation and communication, the acceptability of chart-stimulated recall can be high. Learners often appreciate the opportunity to reflect on their performance and receive personalized feedback, while faculty members value the in-depth understanding it offers. Ensuring learner and patient comfort with the recording process, providing adequate faculty member training, and integrating chart-stimulated recall with other assessment methods can further advance its acceptability and effectiveness in promoting meaningful learning and professional development. Given the time commitment and preparation required to conduct chart-stimulated recall, its role in the program/institution's assessment system must be clearly understood by both learners and faculty members. If this is clearly communicated and understood, chart-stimulated recall can be developed as an acceptable and essential part of a program/institution's assessment portfolio. The assessment of reasoning tool is more portable and can be completed as part of daily educational activities.

Catalytic Effect

Both chart-stimulated recall and the art of reasoning tool offer many benefits for assessing clinical reasoning, including timely feedback in authentic practice, structured exploration of diagnostic and treatment decisions, adaptability to multiple levels of learners, and the opportunity for formative and summative assessment. The catalytic effect of chart-stimulated recall can be profound if learners and faculty members are actively involved in using the generated assessment data (through co-production) in the development of ILPs and the ongoing improvement of the curriculum.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

References

Ilgen, Jonathan S., Aloysius J. Humbert, Gloria Kuhn, et al. 2012. “Assessing Diagnostic Reasoning: A Consensus Statement Summarizing Theory, Practice, and Future Needs.” *Academic Emergency Medicine* 19 (12): 1454-61. https://doi.org/10.1111/acem.12034.

Lyle John. 2003. “Stimulated Recall: A Report on Its Use in Naturalistic Research.” *British Educational Research Journal* 29 (6):861-78. https://psycnet.apa.org/doi/10.1080/0141192032000137349.

Munger, Brendan S., Ronald L. Krome, Jack C. Maatsch, and George Podgorny. 1982. “The Certification Examination in Emergency Medicine: An Update.” *Annals of Emergency Medicine* 11 (2): 91-96. https://www.sciencedirect.com/science/article/abs/pii/S0196064482803041.

Philibert, Ingrid. 2018. “Using Chart Review and Chart-Stimulated Recall for Resident Assessment.” *Journal of Graduate Medical Education* 10 (1): 95-96. https://doi.org/10.4300/JGME-D-17-01010.1.

Sanchez, Hugo Santiago, and Trevor Grimshaw. 2020. “Stimulated Recall.” In *The Routledge Handbook of Research Methods in Applied Linguistics*, edited by Jim McKinley and Heath Rose. Routledge.

Sinnott, Carol, Martina A. Kelly, and Colin P. Bradley. 2017. “A Scoping Review of the Potential for Chart Stimulated Recall as a Clinical Research Method.” *BMC Health Services Research* 17 (583). https://doi.org/10.1186/s12913-017-2539-y.

Thammasitboon, Satid, Joseph J. Rencic, Robert L. Trowbridge, et al. 2018. “The Assessment of Reasoning Tool (ART): Structuring the Conversation Between Teachers and Learners.” *Diagnosis* 5 (4): 197-203. https://doi.org/10.1515/dx-2018-0052.

Zhai, Xuesong, Xiaoyan Chu, Minjuan Wang, et al. 2024. “A Systematic Review of Stimulated Recall (SR) in Educational Research from 2012 to 2022.” *Humanities and Social Sciences Communications* 11: 489. https://doi.org/10.1057/s41599-024-02987-6.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Faculty Global Assessments

Description

Faculty assessment of learners remains a pillar of any programmatic assessment system. While faculty members may have responsibility for additional assessments, all faculty members supervising learner educational experiences typically complete an assessment of each learner. These are usually global assessments that include a rating scale to assess a learner's knowledge, skills, and attitudes using specific frameworks for defining the anchors for what is being assessed. Common frameworks include assessments of specific, individual components of a task (e.g., physical exam developing a differential diagnosis); developmental frameworks such as the Dreyfus and Dreyfus model (Dreyfus 2001) (i.e., progression from novice to master); or synthetic frameworks that assess the learner's ability to integrate necessary knowledge, skills, and attitudes. With the advent of the Milestones and entrustable professional activities (EPAs), "entrustability scales" have also become common. These are behaviorally anchored ordinal scales based on the amount of contribution a faculty supervisor provides in the care of a patient or the amount of supervision required. Commonly used or adapted scales include the Chen scale (ranging from "not allowed to practice" to "allowed to supervise others") and the Ottawa co-activity scale (ranging from faculty "had to do" to faculty "did not need to be there.") (Chen et al. 2015; Rekman et al. 2016).

Competencies

Faculty global assessment forms can be used to assess Patient Care and Procedural Skills, Professionalism, Medical Knowledge, Practice-Based Learning and Improvement, Systems-Based Practice, and Interpersonal and Communication Skills.

Validity

Multiple studies have demonstrated intra- and inter-rater reliability issues with faculty global assessments (Murphy and Cleveland 1995; Beeson et al. 2015). When using faculty global assessment forms, it is important to remember the major source of variability is the faculty member, not the form. The "horn" and "halo" effects, leniency (i.e., dove), or severity (i.e., hawk) errors, and straight-line and central tendency errors can all reduce the utility of faculty global assessment. Such variability occurs when faculty members do not know or apply assessment criteria inaccurately; it also highlights the importance of faculty development to grow consensus on what is being assessed and a shared mental model of educational goals, objectives, and outcomes. Research additionally suggests that a minimum of seven to 11 evaluations from multiple faculty members should be collected to reduce the impact of rater variability and allow for a reproducible, holistic assessment of general clinical competence (Thomas et al. 2011). It is better to focus time and energy on faculty development efforts rather than on the creation of "new and better" assessment forms.

Feasibility

Faculty global assessment forms are relatively time efficient when compared to other assessment methods and tools and are generally well accepted by both faculty members and learners. However, many programs and institutions still struggle to get faculty assessment forms completed and returned in a timely fashion. Mobile tools and apps may help in these areas. Faculty members should also be encouraged to capture assessment notes and observations during their time with the learner. This approach can help faculty members remember key details when the time comes to complete the assessment form.

Acceptability

Faculty global assessments are accepted as routine and necessary in virtually all clinical settings. Certain design characteristics of these forms can impact acceptability. Form content should reflect that there is a limit to how much a faculty member can be asked to observe and report. In general, long forms with too many items, coupled with short rotation or observation durations, result in less useful ratings and assessment information. Content on the form should also be “fit for purpose.” If a specific item or line of questioning does not apply to a rotation or learning experience, faculty members should be able to select a “not observed” or “not applicable” option. Likewise, forms will be more acceptable if they are well understood by faculty members, are simple to use, and are acted upon appropriately by the program or institution when critical feedback is provided.

Catalytic Effect

The catalytic effect of faculty global assessments is dependent upon how the results are shared with learners. If time is taken to review the assessment in a face-to-face meeting and the learner is asked to make a commitment to improving performance in a specific area, the process can promote future learning. Given the duration of clinical rotations and the frequent changing of faculty preceptors, the catalytic effect of this type of assessment may be diminished through lack of continuity with the faculty member who initiated this commitment to change.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

References

Beeson, Michael S., Eric S. Holmboe, Robert C. Korte et al. 2015. “Initial Validity Analysis of the Emergency Medicine Milestones.” *Academic Emergency Medicine* 22 (7): 838-844. https://pubmed.ncbi.nlm.nih.gov/26112031/.

Chen, H. Carrie, W.E. Sjoukje van den Broek, and Olle ten Cate. 2015. “The Case for Use of Entrustable Professional Activities in Undergraduate Medical Education.” *Academic Medicine* 90 (4): 431-6. https://pubmed.ncbi.nlm.nih.gov/25470310/.

Crossley, Jim, Gavin Johnson, Joe Booth, et al. 2011. “Good Questions, Good Answers: Construct Alignment Improves the Performance of Workplace-Based Assessment Scales.” *Medical Education* 45: 560–569. https://pubmed.ncbi.nlm.nih.gov/21501218/.

Dreyfus, Hubert L. 2001. *On the Internet*. Routledge.

Murphy, Kevin R. and Jeanette N. Cleveland. 1995. *Understanding Performance Appraisal.* Sage Publications.

Pulito, Andrew R., Michael M. Donnelly, Margaret Plymale, et al. 2006. “What Do Faculty Observe of Medical Students’ Clinical Performance?” *Teaching and Learning in Medicine* 18 (2): 99-104. https://pubmed.ncbi.nlm.nih.gov/16626266/.

Rekman, Janelle, Stanley J. Hamstra, Nancy Dudek, et al. 2016. “A New Instrument for Assessing Resident Competence in Surgical Clinic: The Ottawa Clinic Assessment Tool.” *Journal of Surgical Education* 73 (4): 575-82. https://pubmed.ncbi.nlm.nih.gov/27052202/.

Rekman, Janelle, Wade Gofton, Nancy Dudek, Tyson Gofton, et al. 2016. “Entrustability Scales: Outlining Their Usefulness for Competency-Based Clinical Assessment.” *Academic Medicine* 91 (2): 186-90. https://pubmed.ncbi.nlm.nih.gov/26630609/.

Rodrigues, Rechell G., and Louis N. Pangaro. 2012. “AM Last Page: Mapping the ACGME Competencies to the RIME Framework.” *Academic Medicine* 87 (12): 1781. https://pubmed.ncbi.nlm.nih.gov/23187923/.

Ryan, Michael S., Asra R. Khan, Yoon Soo Park, et al. 2022. “Workplace-Based Entrustment Scales for the Core EPAs: A Multisite Comparison of Validity Evidence for Two Proposed Instruments Using Structured Vignettes and Trained Raters.” *Academic Medicine* 97 (4): 544-551. https://pubmed.ncbi.nlm.nih.gov/34192721/.

Silber, Cynthia G., Thomas J. Nasca, David L. Paskin, et al. 2004. “Do Global Rating Forms Enable Program Directors to Assess the ACGME Competencies?” *Academic Medicine* 79 (6): 549-556. https://pubmed.ncbi.nlm.nih.gov/15165974/.

Ten Cate, Olle, Alan Schwartz, and H. Carrie Chen. 2020. “Assessing Trainees and Making Entrustment Decisions: On the Nature and Use of Entrustment-Supervision Scales.” *Academic Medicine* 95 (11): 1662-1669. https://pubmed.ncbi.nlm.nih.gov/32324633/.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Thomas, Matthew R., Thomas J. Beckman, Karen F. Mauck, Stephen S. Cha, and Kris G. Thomas. 2011 “Group Assessments of Resident Physicians Improve Reliability and Decrease Halo Error.” *Journal of General Internal Medicine* 26 (7): 759-764. https://pmc.ncbi.nlm.nih.gov/articles/PMC3138588/.

Tolsgaard, Martin G., Henrick Arendrup, Bjarne O. Lindhard, et al. 2012. “Construct Validity of the Reporter-Interpreter-Manager-Educator Structure for Assessing Students’ Patient Encounter Skills.” *Academic Medicine* 87 (6): 799-806. https://pubmed.ncbi.nlm.nih.gov/22534594/.

Williams, Reed G., Debra A. Klamen, William C. McGaghie. 2003. “Cognitive, Social and Environmental Sources of Bias in Clinical Performance Ratings.” *Teaching and Learning in Medicine* 15 (4): 270-292. https://pubmed.ncbi.nlm.nih.gov/14612262/.

Narrative Assessment

Description

Narrative assessment provides qualitative assessment of a learner's performance. It can be used either as a standalone assessment tool or paired with a rating tool. When used as a standalone assessment tool, the narrative description is intended to capture the assessment of the learner purely in a qualitative format. When paired with a rating tool, the narrative description is intended to provide supporting details for the quantitative data or add information not captured by the data. Narrative assessments may have specific prompts (e.g., "describe the learner's ability to…") or they may be unstructured to allow for open-ended comments. Narrative assessments can be a component of a variety of assessments, including (but not limited to) faculty global assessments, direct observations, and OSCEs. Narrative assessment can be utilized for formative or summative purposes.

Learners often find particular value in the formative nature of the written comments, seeing details of their performance highlighted in a reinforcing or constructive way. This can serve as a form of feedback, particularly when the comments follow best practices for feedback (e.g., comments that are balanced, specific, behavioral). This written feedback should ideally mimic verbal feedback from the faculty member to the learner (when applicable) to reinforce the message and to avoid any perceived discrepancy between the two modes of feedback.

Competencies

Narrative assessment can be used to assess Patient Care and Procedural Skills, Professionalism, Medical Knowledge, Practice-Based Learning and Improvement, Systems-Based Practice, and Interpersonal and Communication Skills.

Validity

Multiple studies have demonstrated reliability issues with faculty assessment. These studies have primarily focused on quantitative ratings. However, it would be expected that reliability would be similarly, if not more, problematic with qualitative comments. Multiple studies utilizing frameworks based on feedback literature have developed tools to assess the quality of narrative comments. Some of these studies have shown an increase in quality (defined as alignment with feedback best practices) when utilizing these tools in faculty development. Very little literature exists to support the validity of narrative assessment in capturing performance accurately. Whether the goal of narrative assessment is formative or summative, development of the faculty members writing the narrative is paramount.

Feasibility

Narrative assessment can be conducted both effectively and efficiently with faculty member training. It is important for faculty members to understand the features of narrative comments which are high-quality as well as these features are distinct from the quantity of comments. Similar to global assessment ratings, challenges exist in expecting forms to be returned in a timely fashion. These challenges can be particularly problematic when the narrative assessment is one of the primary sources of feedback to the learner to shape their performance. Similar to global assessments, faculty members should be encouraged to capture assessment notes and observations during their time with the learner. This can be even more critical for narrative assessment to help the faculty member recall details and examples that can illustrate and explain assessment statements.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Acceptability

Narrative assessments are common in a variety of assessment situations, particularly in global assessments used in clinical education and in small group settings. Faculty members and learners usually accept the utility of narrative assessment; learners, in particular, often view it as an important source of feedback. However, when narrative assessment is utilized in a summative manner, particularly when the narrative comments are used in high-stakes decision-making (e.g., including narrative comments on a medical student performance evaluation), learner acceptability may be more variable.

Catalytic Effect

One important determination of the catalytic effect of narrative assessments is the relationship of the narrative assessment to other assessments. If learners perceive congruence between the written comments and the quantitative ratings, or between the written comments and any verbal feedback, then the narrative assessment can further support learner insight and development. Another important determination of the catalytic effect is how the narrative assessment is utilized to support learning. If the focus is formative and the learner is encouraged to reflect on the feedback for growth or is aided in doing so through coaching, the narrative assessment can lead to meaningful learner development. However, if there is incongruence among assessments or if the summative nature of narrative assessment is overemphasized, the learner may focus on the legitimacy of the comments, resulting in situations wherein the catalytic effect may be minimized.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

References

Bartlett, Maggie, James Crossley, and Robert McKinley. 2017. “Improving the Quality of Written Feedback Using Written Feedback.” *Education for Primary Care* 28 (1): 16-22. https://pubmed.ncbi.nlm.nih.gov/27499463/.

Chakroun, Molk, Vincent R. Dion, Kathleen Ouellet, et al. 2022. “Narrative Assessments in Higher Education: A Scoping Review to Identify Evidence-Based Quality Indicators.” *Academic Medicine* 97 (11): 1699-1706. https://pubmed.ncbi.nlm.nih.gov/35612917/.

Chan, Teresa M., Stefanie S. Sebok-Syer, Christoper Sampson, and Sandra Monteiro. 2020. “The Quality of Assessment of Learning (Qual) Score: Validity Evidence for a Scoring System Aimed at Rating Short, Workplace-Based Comments on Trainee Performance.” *Teaching and Learning in Medicine* 32 (3): 319-329. https://pubmed.ncbi.nlm.nih.gov/32013584/.

Ginsburg, Shiphra, Cees P.M. van der Vleuten, and Kevin W. Eva. 2017. “The Hidden Value of Narrative Comments for Assessment: A Quantitative Reliability Analysis of Qualitative Data.” *Academic Medicine* 92 (11): 1617-1621. https://pubmed.ncbi.nlm.nih.gov/28403004/.

Ginsburg, Shiphra, Glenn Regehr, Lorelei Lingard, and Kevin W. Eva. 2015. “Reading Between the Lines: Faculty Interpretations of Narrative Evaluation Comments.” *Medical Education* 49 (3): 296-306. https://pubmed.ncbi.nlm.nih.gov/25693989/.

Hall, Alan M., Adam Gray, and John W. Ragsdale. 2024. “Making Narrative Feedback Meaningful.” *The Clinical Teacher* 21 (5): e13766. https://pubmed.ncbi.nlm.nih.gov/38651603/.

Kogan, Jennifer R., C. Jessica Dine, Lisa N. Conforti, and Eric S. Holmboe. 2023. “Can Rater Training Improve the Quality and Accuracy of Workplace-Based Assessment Narrative Comments and Entrustment Ratings? A Randomized Controlled Trial.” *Academic Medicine* 98 (2): 237-247. https://pubmed.ncbi.nlm.nih.gov/35857396/.

McGuire, Natalie, Anita Acai, and Ranil R. Sonnadara. 2025. “The McMaster Narrative Comment Rating Tool: Development and Initial Validity Evidence.” *Teaching and Learning in Medicine* 37 (1): 86-98. https://pubmed.ncbi.nlm.nih.gov/37964518/.

Parsons, Andrew S., Kelley Mark, James R. Martindale, et al. 2022. “Concordance of Narrative Comments with Supervision Ratings Provided During Entrustable Professional Activity Assessments.” *Journal of General Internal Medicine* 37 (9): 2200-2207. https://pubmed.ncbi.nlm.nih.gov/35710663/.

Sebok-Syer, Stefanie S., Don A. Klinger, Jonathan Sherbino, and Teresa M. Chan. 2017. “Mixed Messages or Miscommunication? Investigating the Relationship Between Assessors’ Workplace-Based Assessment Scores and Written Comments.” *Academic Medicine* 92 (12): 1774-1779. https://pubmed.ncbi.nlm.nih.gov/28562452/.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Wilby, Kyle John, Diana H.J.M. Dolmans, Zubin Austin, and Marjan J. B. Govaerts. 2019. "Assessors' Interpretations of Narrative Data on Communication Skills in a Summative OSCE. *Medical Education* 53 (10): 1003-1012. https://pubmed.ncbi.nlm.nih.gov/31304615/.

Wilby, Kyle John, Marjan J.B. Govaerts, Zubin Austin, and Diana Dolmans. 2019. "Discriminating Features of Narrative Evaluations of Communication Skills During an OSCE." *Teaching and Learning in Medicine* 31 (3): 298-306. https://pubmed.ncbi.nlm.nih.gov/30755046/.

One-Minute Preceptor

Description

The One-Minute Preceptor (originally published as the five “microskills”) is a model that can be used to structure a teaching encounter with a learner. The five steps of the model are:

1. Get a commitment: the learner is asked to commit to a specific step with a patient.
2. Probe for supporting evidence: the learner is asked to justify the commitment they have just made.
3. Teach general rules: the faculty member provides the learner with a brief teaching point that is relevant to the current patient encounter.
4. Reinforce what was done right: the faculty member provides reinforcing feedback to the learner about their performance.
5. Correct mistakes: the faculty member provides constructive feedback to the learner about their performance.

Using this model in a clinical interaction with a learner, faculty can support the learner in making decisions, understand the learner’s decision-making process, provide teaching, and provide feedback. This structured teaching encounter is anchored in a clinical case, often following an oral presentation by the learner. This model is one of the most widely taught methods to help faculty and residents improve their teaching skills.

Competencies

The One-Minute Preceptor model can be used to assess Patient Care and Procedural Skills and Interpersonal and Communication Skills.

Validity

There is limited evidence of the validity of this model to improve the effectiveness of learner assessment. However, there is evidence that the use of this model improves teaching behaviors (including the provision of feedback) and learner perceptions of teaching. Effective implementation of this model is dependent on effective faculty development.

Feasibility

The One-Minute Preceptor can be used in a variety of clinical settings. Effective use by faculty members requires training, although this training can accomplished in a relatively short time.

Acceptability

The One-Minute Preceptor is appealing to faculty members because it provides a structure to guide their interactions with learners. Faculty members also appreciate that it can be completed quickly, especially in the midst of other clinical work. Some faculty members may find the structure overly rigid, especially if they are newly utilizing it. Learners are typically accepting of the model as it ensures they receive feedback in the moment and receive focused teaching. Studies have shown that learners prefer use of the One-Minute Preceptor to traditional precepting.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Catalytic Effect

The One-Minute Preceptor can have a powerful catalytic effect as it can provide actionable feedback to shape learner performance. It also grounds teaching concepts in the context of a patient case, and a case in which the learner was actively involved, which can aid retention and spur additional learning.

References

Aagaard, Eva, Arianne Teherani, and David M. Irby. 2004. “Effectiveness of the One-Minute Preceptor Model for Diagnosing the Patient and the Learner: Proof of Concept.” Academic Medicine 79 (1): 42-49. https://pubmed.ncbi.nlm.nih.gov/14690996/.

Farrell, Susan E., Laura R. Hopson, Margaret Wolff, Robin R. Hemphill, and Sally A. Santen. 2016. “What’s the Evidence: A Review of the One-Minute Preceptor Model of Clinical Teaching and Implications for Teaching in the Emergency Department.” *The Journal of Emergency Medicine* 51 (3): 278-283. https://pubmed.ncbi.nlm.nih.gov/27377967/.

Furney, Scott L., Alex N. Orsini, Kym E. Orsetti, David T. Stern, Larry D. Gruppen, and David M. Irby. 2001. “Teaching the One-Minute Preceptor. A Randomized Controlled Trial.” *Journal of General Internal Medicine* 16 (9): 620-624. https://pmc.ncbi.nlm.nih.gov/articles/PMC1495264/.

Irby, David M., Eva Aagaard, and Arianne Teherani. 2004. “Teaching Points Identified by Preceptors Observing One-Minute Preceptor and Traditional Preceptor Encounters.” *Academic Medicine* 79 (1): 50-55. https://pubmed.ncbi.nlm.nih.gov/14690997/.

Neher, J.O., K.C. Gordon, B. Meyer, and N. Stevens. 1992. “A Five-Step “‘Microskills’ Model of Clinical Teaching.” *The Journal of the American Board Family Practice* 5 (4): 419-424. https://pubmed.ncbi.nlm.nih.gov/1496899/.

Post, Robert E., R. Glen Quattlebaum, and Joseph J. Benich 3rd. 2009. “Residents-As-Teachers Curricula: A Critical Review.” *Academic Medicine* 84 (3): 374-380. https://pubmed.ncbi.nlm.nih.gov/19240450/.

Servey, Jessica, and Kristen Wyrick. 2018. “Teaching Clinical Precepting: A Faculty Development Workshop Using Role-Play.” *MedEdPORTAL* 14: 10718. https://pmc.ncbi.nlm.nih.gov/articles/PMC6342365/.

Teherani, Arianne, Patricia O’Sullivan, Eva M. Aagaard, Elizabeth M. Morrison, and David M. Irby. 2007. “Student Perceptions of the One Minute Preceptor and Traditional Preceptor Models.” *Medical Teacher* 29 (4): 323-327. https://pubmed.ncbi.nlm.nih.gov/17786745/.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Multisource Feedback (360-Degree Feedback)

Description

Multisource feedback, also referred to as 360-degree feedback, consists of measurement tools, usually surveys or other rating forms, completed by multiple individuals who interact and work with a learner. Assessors completing multisource feedback forms could include health care professionals, such as nurses, therapists, pharmacists, social workers, and others, who work with the learner on a regular basis (e.g., in the clinic, operating room, or inpatient unit) or during a specific rotation. Peers and other learners should also be part of multisource feedback, as should patients and families; however, peer and patient evaluations are covered in more detail in separate sections. Having the learner complete the same multisource feedback form provides valuable insight into the learner's perception of themselves compared to others using the same tool. Most multisource approaches use a survey, rating scale, or questionnaire to gather information about a learner's performance. The power of multisource feedback is the opportunity to gather assessments on key competencies (e.g., teamwork, communication, management skills, decision making) from multiple perspectives.

Competencies

Multisource feedback assessments must be a core component of any program of assessment. It is essential for assessing Professionalism, Interpersonal and Communication Skills (especially interprofessional teamwork knowledge, skills, and attitudes), and Systems-Based Practice.

Validity

Substantial literature now exists on the use of multisource feedback in medicine and medical education. Studies of practicing physicians have found associations between multisource feedback ratings and patient complaints and malpractice claims. Research supports the use of multisource feedback for mostly formative purposes; when multisource feedback is used as a stand-alone assessment for high-stakes decisions, assessors tend to inflate their ratings and reduce the educational and feedback value of multisource feedback. Data from multisource feedback assessments can and should be used by the Clinical Competency Committee or progress and promotion committee to make judgments on the Professionalism, Interpersonal and Communication Skills, and Systems-Based Practice subcompetencies, depending on who completes the multisource feedback and what specific multisource feedback tool is used. Three systematic reviews on multisource feedback are provided in the references below.

Feasibility

Multisource feedback requires some effort and coordination to be used effectively. Programs and institutions should pick either a longitudinal clinical site (e.g., ambulatory clinic, intensive care unit for critical care fellows, emergency department, etc.), or a specific rotation where a team of health care professionals can assess professionalism and interpersonal teamwork skills. Ideally a program should try to use an existing tool with validity evidence, but if programs or institutions develop their own multisource feedback instruments, they should evaluate the quality of the instrument using the Utility Index (Van der Vleuten 1996) or Ottawa Criteria (Norcini 2018) for good assessment.

Acceptability

When used properly, educational programs and health systems have found multisource feedback to be highly useful and impactful. Using a web-based multisource feedback tool focused on interprofessional teamwork, Chesluk and colleagues (2015) found that participating physicians valued the feedback, especially the narrative comments.

Catalytic Effect

Multisource feedback can provide valuable insights into professionalism, interpersonal communication, and interprofessional teamwork competencies. Research has shown the effectiveness of multisource feedback for professional development requires that the receiver of the feedback review the results with a mentor, advisor, or trusted peer (Donnon et al. 2014). This conversation about the results of multisource feedback enables sense-making for the learner and creation of a more effective ILP. Comparison of learners' perceptions of themselves against the opinions of others can spark meaningful introspection and change if the results differ.

Suggested Assessment Tool Available from the ACGME

The Teamwork Effectiveness Assessment Module (TEAM) can be used by programs and institutions to gather and interpret feedback from the interprofessional team with whom they work to care for patients in the hospital or clinic. TEAM is designed for use even in work settings and clinical rotations that do not provide formal support or training for interprofessional teamwork. This multisource feedback tool can assist residents and fellows in assessing key competencies, such as Interpersonal and Communication Skills and Professionalism, as well as the Milestones overall.

For example, TEAM is particularly well suited for assessing the Milestones in the *interprofessional and team communication* subcompetency of Interpersonal and Communication Skills. TEAM will also be helpful in assessing and providing feedback for *professional behavior and ethical principles,* (a subcompetency in Professionalism); *accountability/conscientiousness* (a subcompetency in Professionalism); *and physician role in health care systems* (a subcompetency in Systems-Based Practice).

Access information about **TEAM** on Learn at ACGME at **https://dl.acgme.org/pages/assessment**.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

References

Chesluk Benjamin J., Siddharta Reddy, Brian Hess, et al. 2015. “Assessing Interprofessional Teamwork: Pilot Test of a New Assessment Module for Practicing Physicians.” *Journal of Continuing Education in the Health Professions* 35 (1): 3-10. https://pubmed.ncbi.nlm.nih.gov/25799967/.

Donnon, Tyrone, Ahmed Al Ansari, Samah Al Alawi, et al. 2014. “The Reliability, Validity, and Feasibility of Multisource Feedback Physician Assessment: A Systematic Review.” *Academic Medicine* 89 (3): 511-516. https://pubmed.ncbi.nlm.nih.gov/24448051/.

Frost, Jody S., Dana P. Hammer, Loretta M. Nunez, et al. 2019. “The Intersection of Professionalism and Interprofessional Care: Development and Initial Testing of the Interprofessional Professionalism Assessment (IPA).” *Journal of Interprofessional Care* 33 (1): 102-115. https://pubmed.ncbi.nlm.nih.gov/30247940/.

Laggo, Janaka, William R. Berry, Kate Miller, et al. 2019. “Multisource Evaluation of Surgeon Behavior is Associated with Malpractice Claims.” *Annals of Surgery* 270 (1): 84-90. https://pubmed.ncbi.nlm.nih.gov/29578910/.

Lockyer, Jocelyn. 2013. “Multisource Feedback: Can it Meet Criteria for Good Assessment?” *Journal of Continuing Education in the Health Professions* 33 (2): 89-98. https://doi.org/10.1002/chp.21171.

Norcini, John. M., M. Brownell Anderson, Valdes Bollela, et al. 2018. “Consensus Framework for Good Assessment.” *Medical Teacher* 40: (11): 1102-1109. https://pubmed.ncbi.nlm.nih.gov/30299187/.

Stevens, Sebastian, James Read, Rebecca Baines, et al. 2018. “Validation of Multisource Feedback in Assessing Medical Performance: A Systematic Review.” *Journal of Continuing Education in the Health Professions* 38 (4): 262-268. https://pubmed.ncbi.nlm.nih.gov/30157152/.

Van Der Vleuten, C.P. 1996. “The Assessment of Professional Competence: Developments, Research and Practical Implications.” *Advances in Health Sciences Education* 1 (1): 41-67. https://pubmed.ncbi.nlm.nih.gov/24178994/.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Informed Self-Assessment

Description

Informed self-assessment refers to a deliberate, evidence-informed process in which learners evaluate their own performance, knowledge, skills, and behaviors to guide their learning and professional development. Unlike casual or intuitive self-reflection, informed self-assessment incorporates external data and feedback to help learners make accurate judgments about their abilities and identify areas for improvement. Informed self-assessments catalyze learning by turning reflection into action. When grounded in feedback and meaningful data, they empower learners to take ownership of their growth, thereby enhancing both short-term academic performance and long-term professional development.

Validity

Research shows that self-assessment alone is often inaccurate, particularly among novice learners. However, when learners are guided to use external data—such as exam scores, structured feedback, or performance metrics—the accuracy improves. Informed self-assessment is most valid when used to identify relative strengths and weaknesses rather than absolute competence levels.

Reliability

Informed self-assessment, especially when done informally, tends to show low reliability due to biases like overconfidence or underestimation. However, the use of structured tools such as rubrics or guided reflection templates can enhance reliability, particularly when supported by coaching and external benchmarks.

Feasibility

It is generally feasible when embedded in learning portfolios, coaching sessions, milestone reviews, or longitudinal assessments. The approach requires minimal financial resources but does require intentional faculty development and support to ensure meaningful feedback is integrated.

Acceptability

Informed self-assessment is highly acceptable when framed as a formative, growth-oriented process rather than a summative judgment. Learners value opportunities to direct their own learning, especially when they trust the feedback provided and perceive it as useful. Faculty members also appreciate its potential to foster learner reflection and accountability.

The Catalytic Effect

Promoting metacognition encourages learners to reflect on how they learn, not just what they know, fostering strategic, self-regulated learning. When this reflection is supported by structured feedback and data—such as performance dashboards or narrative feedback—learners can identify specific gaps and develop targeted, personalized learning plans. Actively engaging in self-assessment also enhances motivation and ownership, increasing learners' investment in their education and sense of responsibility for their competence. As learners recognize the value of feedback for improving future performance, they become more open to and proactive in seeking it. Collectively, these practices support professional identity formation by fostering habits essential to lifelong learning and professionalism which are critical for success in clinical practice.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

References

Davis, David A., Paul E. Mazmanian, Michael Fordis, R. Van Harrison, Kevin E. Thorpe, and Laura Perrier. 2006. “Accuracy of Physician Self-Assessment Compared with Observed Measures of Competence.” *JAMA* 296 (9): 1094–1102. https://pubmed.ncbi.nlm.nih.gov/16954489/.

Eva, Kevin, and Glenn Regehr. 2005. “Self-Assessment in the Health Professions: A Reformulation and Research Agenda.” *Academic Medicine* 80 (10 Suppl):S46–S54. https://pubmed.ncbi.nlm.nih.gov/16199457/.

Mann, Karen, Cees van der Vleuten, Kevin Eva, et al. 2011. “Tensions in Informed Self-Assessment: How the Desire for Feedback and the Desire to Be Seen as Competent Can Conflict.” *Academic Medicine* 86 (9): 1120–1127. https://pubmed.ncbi.nlm.nih.gov/21785309/.

Norcini, John, Brownell Anderson, Valdes Bollela, et al. 2011. “Criteria for Good Assessment: Consensus Statement and Recommendations from the Ottawa 2010 Conference.” *Medical Teacher* 33 (3): 206–214. https://pubmed.ncbi.nlm.nih.gov/21345060/.

Sargeant, Joan, Heather Armson, Ben Chesluk, et al. 2010. “The Processes and Dimensions of Informed Self-Assessment: A Conceptual Model.” *Academic Medicine* 85 (7): 1212–1220. https://pubmed.ncbi.nlm.nih.gov/20375832/.

Sargeant, Joan, Kevin W. Eva, Heather Armson, et al. 2011. “Features of Assessment Learners Use to Make Informed Self-Assessments of Clinical Performance.” *Medical Education* 45 (6): 636–647. https://pubmed.ncbi.nlm.nih.gov/21564201/.

Schön, Donald A. 1984. *The Reflective Practitioner: How Professionals Think in Action*. Basic Books.

Schumacher, David J., Robert Englander, and Carol Carraccio. 2013. “Developing the Master Learner: Applying Learning Theory to the Learner, the Teacher, and the Learning Environment.” *Academic Medicine* 88 (11): 1635–1645. https://pubmed.ncbi.nlm.nih.gov/24072107/.

van der Vleuten C.P.M., L.W.T. Schuwirth, E.W. Driessen, Jet al. 2012. “A Model for Programmatic Assessment Fit for Purpose.” *Medical Teacher* 34 (3):205–214. https://pubmed.ncbi.nlm.nih.gov/22364452/.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Patient Experience Surveys/Patient Feedback

Description

Patient feedback is a formative and summative assessment method in which patients (real or standardized) provide structured evaluations of learners' clinical interactions. Patient experience surveys are essential if a program or institution truly wants to assess the breadth of interpersonal and communication skills. Without the input of patients and families, it is simply not possible to assess patient- and person-centeredness. Substantial research over the last 20 years has led to a number of useful surveys. Instruments typically include validated questionnaires (e.g., the CARE Measure [Mercer et al. 2004], multisource format tools) administered after encounters in clinical or simulated settings. Feedback can be shared immediately or collected over time and used within portfolios, coaching sessions, or clerkship evaluations. Survey questions often assess specific aspects of patients' interactions with the health care system, such as whether their questions were satisfactorily answered and whether the patients felt they were treated respectfully. While satisfaction with care is still important to measure, surveys now more effectively measure specific aspects of care that affect satisfaction ratings and are more actionable. These include the learner's explanations, listening skills, and provision of information about exam findings, treatment steps, and drug side effects. A typical patient survey asks patients to rate their agreement with statements describing the care or to rate aspects of their experience on a scale (e.g., how often the doctor showed you respect: always, sometimes, rarely, never). While some scales still may use "quality adjectives," such as scales ranging from "poor" to "outstanding," research conducted by the Consumer Assessment of Healthcare Providers and Systems suggests these scales are less useful and helpful. For any instrument, each rating is given a value and a score may be calculated by averaging across responses to generate a single score overall or separate scores for different clinical care activities or settings.

Competencies

Patient feedback accumulated from single-encounter questionnaires can assess the quality of the patient experience and provide insights into the Patient Care and Procedural Skills (e.g., aspects of data gathering, treatment, and management; counseling, and education; preventive care); Interpersonal and Communication Skills (e.g., explanation of care plans); Professionalism (e.g., respect for the patient/family); and Systems-Based Practice (e.g., patient advocacy; coordination of care). If survey items about specific physician behaviors are included, the results can be used for formative assessment and performance improvement.

Validity

Patient feedback has demonstrated content and construct validity when aligned with observable behaviors and using validated tools. Research shows that patient evaluations often correlate with faculty ratings of similar domains. However, validity may be limited if patients lack medical knowledge or are influenced by personal bias or single encounters. Aggregating feedback across multiple patients increases reliability and interpretability.

The recommended reliability coefficient for making higher-stakes decisions about physicians based on patient experience survey ratings is 0.70. Research has found that for commonly used surveys, such as the Consumer Assessment of Healthcare Providers and Systems, 45 surveys are needed for a high degree of reliability.

These numbers are very difficult to achieve for individual learners; thus, studies of reliability estimates for learners are lower. For any type of learner, attribution of patient responses to that learner is a significant challenge, particularly in a team-based environment. Despite the

challenges posed by the number of responses needed and attribution, there is evidence that patient experience surveys correlate with patient outcomes, which is a strong measure of validity. There is limited data that patient surveys correlate with other measures for learners, such as communication skill ratings.

Feasibility

Patient feedback is moderately feasible, particularly when integrated into existing feedback structures (e.g., outpatient clinics, OSCEs). Standardized tools like the CARE Measure (Mercer et al. 2004) are short and easy to administer. However, logistical challenges include training patients, ensuring anonymity, and collecting sufficient feedback for meaningful evaluation. Electronic platforms have improved data collection and reporting.

A variety of patient experience surveys are available from multiple sources. Programs and educational leaders are encouraged to check with their hospital's Quality Improvement Office as a first step to see what surveys are already being used locally. The Consumer Assessment of Healthcare Providers and Systems can be obtained from the Agency for Healthcare Research and Quality (AHRQ); programs, and institutions can access and download an extensive portfolio of surveys for different settings. Creation of new surveys requires substantial effort and should only be undertaken if an existing survey with research evidence cannot be identified. Ideally, the patient experience surveys would be completed at the time of service and should require less than 10 minutes to complete.

Electronic means of collecting patient survey data (e.g., smartphone app, tablet, or easily accessible computer terminal) are recommended. Surveys can be mailed after the patient goes home, or conducted with patients over the phone for those patients or resource settings where electronic means are not feasible. Difficulties encountered with patient surveys include: 1) language and literacy problems; 2) obtaining sufficient surveys to provide reliable results; 3) the resources required to collect, aggregate, and report survey responses; and 4) assessment of the learner's contribution to a patient's care separate from that of the health care team. Because of these concerns, patient experience surveys are often conducted by the institution or by one or more clinical sites and reports specific to one GME program may or may not be prepared.

Acceptability

Acceptability is generally high among patients and growing among educators and learners. Patients appreciate being part of the educational process. Learners often value authentic feedback from the patient perspective, especially when it highlights interpersonal strengths or blind spots. Some concern exists about fairness and patients' understanding of performance expectations, especially in early clinical learners.

Patient experience surveys have become standard practice and there is good evidence that physicians and health care professionals find the feedback useful, especially when written comments are provided. Less is known about acceptability in GME as patient experiences are usually limited to multisource feedback where only a few patients may be sampled. Some studies have found that patient feedback surveys do provide valuable information to programs and institutions which should work to develop a "culture of acceptability" around patient experience surveys. As noted above, it is simply not possible to assess patient and person-centeredness without this assessment method.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Catalytic Effect

When the challenges such as sufficient number and attribution can be addressed, patient experience surveys can have a strong catalytic effect, supporting improvement plans and professional growth for learners. These survey results promote:

- Reflection and emotional intelligence
- Improved communication and rapport-building
- Patient-centered care orientation
- Motivation to improve interpersonal skills

When coupled with coaching or narrative feedback, it helps learners translate input into specific behavior changes and professional growth. Patient experience surveys can definitely produce and support improvement and professional growth While not a typical psychometric approach, use of a narrative, clinometric approach where the patient is simply asked a few open-ended questions at the end of a clinical encounter (e.g., “What did you like about your visit today?” “What did you dislike?” “What would you change?”) can facilitate timely changes in practice.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

References

Agency for Healthcare Quality and Research. 2011. “About CAHPS®.” Page last reviewed July 2024. https://www.ahrq.gov/cahps/about-cahps/index.html.

Agency for Healthcare Quality and Research. 2021. “Guidelines for Using the CAHPS® Clinician & Group Survey.” https://www.ahrq.gov/sites/default/files/wysiwyg/cahps/surveys-guidance/cg/guidance-cg-cahps.pdf.

Campbell, J.L., S.H. Richards, A. Dickens, M. Greco, A Narayanan, and S. Brearley. “Assessing the Professional Performance of UK Doctors: An Evaluation of the Utility of the General Medical Council Patient Questionnaire.” 2008. *Quality & Safety in Health Care* 17 (3):187–193. https://pubmed.ncbi.nlm.nih.gov/18519625/.

Concato, J., and A.R. Feinstein. 1997. “Asking Patients What They Like: Overlooked Attributes of Patient Satisfaction with Primary Care.” *American Journal of Medicine* 102: 399-406. https://pubmed.ncbi.nlm.nih.gov/9217623/.

Godillot, Clothilde, Fatma Jendoubi, Maria Paulina Konstantinou, et al. 2021. “How to Assess Patient Satisfaction Regarding Physician Interaction: A Systematic Review.” *Dermatologic Therapy* 34 (2): e14702. https://pubmed.ncbi.nlm.nih.gov/33368997/.

Greco, M., A. Brownlea, and J. McGovern. 2001. “Impact of Patient Feedback on the Interpersonal Skills of General Practice Registrars: Results of a Longitudinal Study.” *Medical Education* 35 (8): 748–756. https://pubmed.ncbi.nlm.nih.gov/11489102/.

Hess, Brian J., Lorna A. Lynn, Lisa N. Conforti, and Eric S. Holmboe. 2011. “Listening to Older Adults: Elderly Patients’ Experience of Care in Residency and Practicing Physician Outpatient Clinics.” *Journal of the American Geriatrics Society* 59 (5): 909-915. https://pubmed.ncbi.nlm.nih.gov/21517787/.

Kojima, Lisa, Amy Y. Han, Judith C. French, and Jeremy M. Lipman. 2022. “The Patient’s Voice in Assessing Resident Communication Skills.” *Journal of Surgical Education* 79 (6): e220-e224. https://pubmed.ncbi.nlm.nih.gov/36075805/.

Mercer, Stewart W., Margaret Maxwell, David Heaney, and Graham Cm Watt. 2004. “The Consultation and Relational Empathy (CARE) Measure: Development and Preliminary Validation and Reliability of an Empathy-Based Consultation Process Measure.” *Family Practice* 21(6): 699-705. https://pubmed.ncbi.nlm.nih.gov/15528286/.

Price, Rebecca Anhang, Marc N. Elliott, Alan M. Zaslavsky, et al. 2014. “Examining the Role of Patient Experience Surveys in Measuring Health Care Quality.” *Medical Care Research and Review* 71 (5): 522-554. https://pmc.ncbi.nlm.nih.gov/articles/PMC4349195/.

Peer Assessment

Description

Peer assessment in medical education is a powerful tool that enhances learning and professional development. Adoption of this type of assessment helps move from an individualistic and teacher-led assessment approach to a more collaborative, learner-centered approach that aligns with social constructivism. By evaluating their peers, learners can actively engage with the material, improve their self-assessment skills, and identify areas for improvement. Peer assessment fosters crucial professional skills such as giving and receiving constructive feedback, communication, and teamwork. Peer assessment can also be used to supplement faculty members' feedback and contribute to a more comprehensive evaluation. Within patient care, peer assessments can be utilized to evaluate direct patient encounters, collaboration within interprofessional teams, and clinical reasoning skills. Furthermore, it can promote a culture of accountability and shared standards, encouraging learners to take responsibility for their learning and strive for excellence. When implemented effectively, peer assessment can significantly enhance the medical education experience and prepare learners for collaborative practice in healthcare.

Competencies

Peer assessments can be used to assess virtually all competencies, particularly around evaluation of any skill that requires multisource feedback, including Patient Care and Procedural Skills, Professionalism, Interpersonal and Communication Skills, and Systems-Based Practice. Peer assessment is also an important tool for Practice-Based Learning and Improvement. whereby learners seek and apply feedback as well as engage in self-reflection. Additionally, it can also be utilized to assess Professionalism attributes such as empathy, accountability, and respect. Systems-Based Practice employs peer feedback related to how learners work and contribute to system improvements.

Validity

The quality of peer assessment can be influenced by a variety of factors such as the reliability of the assessment, interaction between peers, the stakes of the assessment, and the assumption of equivalence between the evaluations of each peer. Analyzing the results for patterns and correlations, while considering potential biases, helps in accurate interpretation.

Feasibility

Peer assessments are widely used due to their feasibility, cost-effectiveness, and scalability. The assessment instrument and process should be easy to understand and use for both learners and faculty members. It is important to provide clear guidelines and rubrics to ensure consistency and fairness, especially for summative evaluations. Peer assessment can be time-consuming for both learners and faculty members, requiring efficient processes, dedicated time allocation and resources such as technology. Extensive training for assessors to provide constructive feedback and ensure inter-rater reliability should be required. Consideration should be given to using anonymous assessments that can reduce bias and encourage honest feedback. As with any new feedback instrument, pilot testing is critical prior to implementation.

Acceptability

The acceptability of using peer assessments in medical education is generally high, with growing recognition of their value in promoting learning and professional development. Regardless, it is always important to ensure that both learners and faculty members understand the value and purpose of peer assessment and are willing to participate in the process. When

using peer assessment as a summative evaluation, concerns about potential bias and halo/horn effects due to personal relationships or group dynamics need to be addressed through anonymity, clear criteria, and faculty oversight.

Catalytic Effect

Combining peer assessments with other assessment methods like faculty observations or OSCEs can strengthen validity by providing multiple perspectives on performance. Regular feedback from learners and faculty members on the effectiveness of the peer assessment will allow for necessary adjustments to enhance the evaluation over time. By thoughtfully incorporating peer assessment into medical education, programs and institutions can leverage its unique benefits to foster the development of well-rounded, competent physicians who are prepared to excel in all aspects of patient care and professional practice.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

References

Brierley, Clarissa, Leila Ellis, and Emily Roisin Reid. 2022. “Peer-Assisted Learning in Medical Education: A Systematic Review and Meta-Analysis.” *Medical Education.* 56 (4): 365-373. https://pubmed.ncbi.nlm.nih.gov/34595769/.

Lerchenfeldt, Sarah, Misa Mi, and Marty Eng. 2019. “The Utilization of Peer Feedback During Collaborative Learning in Undergraduate Medical Education: A Systematic Review.” *BMC Medical Education* 19 (1): 321. https://pubmed.ncbi.nlm.nih.gov/31443705/.

Norcini John J. 2003. “Peer Assessment of Competence.” *Medical Education* 37 (6): 539-543. https://pubmed.ncbi.nlm.nih.gov/12787377/.

Speyer, Renée, Walmari Pilz, Jolien Van Der Kruis, and Jan Wouter Brunings. 2011. “Reliability and Validity of Student Peer Assessment in Medical Education: A Systematic Review.” *Medical Teacher* 33 (11): e572-585. https://pubmed.ncbi.nlm.nih.gov/22022910/.

Stenberg, Marie, Elisabeth Mangrio, Mariette Bengtsson, and Elisabeth Carlson. 2021. “Formative Peer Assessment in Higher Healthcare Education Programmes: A Scoping Review.” *BMJ Open* 11 (2): e045345. https://pmc.ncbi.nlm.nih.gov/articles/PMC7875268/.

Topping, K. 1998. “Peer Assessment Between Students in Colleges and Universities.” *Review of Educational Research* 68 (3): 249–276. **https://psycnet.apa.org/record/1999-08081-001**.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Portfolio

Description

A portfolio is a collection of evidence intended to demonstrate an individual's learning journey over time. In medical education, a portfolio can include documents such as self-assessments, ILPs, reflective essays, and assessments that reflect a learner's professional development and can be used for both formative and summative assessment. Portfolios can serve multiple purposes. They can be designed to contain mandated records of achievement, with specified levels of performance used for selection or promotion (i.e., dossier portfolio); mandated collections of acquired skills and competencies in a fixed format (i.e., training portfolio); purposeful collections of evidence for personal growth and development (i.e., reflective portfolio); or as a personal development portfolio, containing guided self-assessments of progress in time as well as documenting and enabling ILPs. Regardless of purpose, a critical requirement for portfolio use is active participation by the learner.

As interest in portfolio use has grown in the medical education community, the concept of the "comprehensive portfolio" has emerged. Comprehensive portfolios include content agreed upon by the learner and the program/institution that is standardized and defensible, allowing for summative decision making when required, and documenting that a graduate has attained the desired competence to practice unsupervised, safe, and effective patient care or to enter residency.

Key characteristics of a comprehensive portfolio include:

- a multifaceted approach to assessment;
- assessment based on "triangulation," i.e., assessing multiple domains of competence and utilize multiple assessors;
- longitudinal and iterative content established through the interaction of the learner and faculty assessor(s);
- learner self-assessment and reflection;
- evidence of meaningful learner engagement demonstrating professional growth; and
- portfolio development and use that is transparent to the learners who should have full access to content and a sense of portfolio "ownership."

Competencies

Depending on the design and desired outcome of a portfolio, virtually every Core Competency can be assessed using a portfolio. Portfolios can be particularly useful in documenting growth in the Practice-Based Learning and Improvement and Professionalism, both of which can be difficult to document using traditional assessments.

Validity

Given that portfolios will contain a mix of quantitative and qualitative content, standard approaches to determining the validity and reliability of portfolios can be challenging. In general, the assessor of a portfolio must be able to determine that the portfolio content demonstrates the learner has achieved the desired outcomes established by the program/institution. Such assessment is dependent on the quality of the evidence provided in the portfolio and the process used by the portfolio assessors (e.g., faculty members).

Feasibility

Portfolio use requires significant preparation and can be time-consuming for learners to prepare and faculty members to review. The specific purpose of the portfolio must be clearly defined. Learners and faculty members will need to be specifically educated in the use and purpose of the portfolio. Additionally, programs and institutions will need to determine the time commitment associated with required or desired portfolio activities for both the learner and the faculty members who will be monitoring portfolio activities. If these requirements are adequately addressed, portfolio use is feasible.

Acceptability

Acceptability of a portfolio can vary based upon the portfolio's intended purpose and the preparation of faculty members and learners for using a portfolio. Potential barriers to successful portfolio use can include variable learner and faculty member engagement using a portfolio, time constraints completing required portfolio activities, and inexperience with portfolio use and monitoring. Acceptability can be increased by addressing each of these potential barriers and by clearly defining the intended purpose of the portfolio. For instance, will portfolio content be used for formative and/or summative assessment, and have legal issues, such as patient and learner confidentiality, been clearly addressed?

Catalytic Effect

Portfolio assessment enables learners to reflect on their real performance, identify areas of weakness and strength, and document development of competence. Portfolios also encourage learners to take responsibility for their own learning and can foster master adaptive learning. Additionally, portfolios can be a valuable resource for feedback faculty members and mentors.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

References

Driessen, Erik. 2009. “Portfolio Critics: Do They Have a Point?” *Medical Teacher* (4): 279-281. https://pubmed.ncbi.nlm.nih.gov/19404889/.

Friedman Ben David M., M.H. Davies, R.M. Harden, P.W. Howie, J. Kier, M.J. Pippard. 2001. “AMEE Educational Guide 24; Portfolios as a Method of Student Assessment.” *Medical Teacher* 23(6): 535-551. https://pubmed.ncbi.nlm.nih.gov/12098472/.

Heeneman, Sylvia., and Erik W. Driessen. 2017. “The Use of a Portfolio in Postgraduate Medical Education–Reflect, Assess and Account, One for Each or All in One?” *GMS Journal for Medical Education* 34(5): 1-12. https://pubmed.ncbi.nlm.nih.gov/29226225/.

Mueller, Paul S. 2015. “Teaching and Assessing Professionalism in Medical Learners and Practicing Physicians.” *Rambam Maimonides Medical Journal* 6 (2): e0011. https://dx.doi.org/10.5041%2FRMMJ.10195.

O’Sullivan, Patricia. S., Carol Carraccio, and Eric S. Holmboe. 2017. “Portfolios.” In *Practical Guide to the Evaluation of Clinical Competence* by Eric S. Holmboe, Steven J. Durning, and Richard E. Hawkins. Elsevier.

Van Tartwijk, Jan, and Erik Driessen. 2009. “Portfolios for Assessment and Learning: AMEE Guide No. 45.” *Medical Teacher* 31 (9): 790-801. https://pubmed.ncbi.nlm.nih.gov/19811183/.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

Section III

Implementation

Developing a robust system of programmatic assessment using the tools described above is essential to the valid and reliable assessment of learners in medical education. While each of the listed assessment methods or tools has strengths, the perfect assessment does not exist. As a result, programmatic assessment should use a combination of tools and methods to assess learners in each of the six Core Competencies over time and in various contexts. The table below provides specific assessment tools and methods for each competency.

Assessment Tool/Method	Targeted Competency
Faculty member global assessment	Multiple competencies
Direct observation	• Patient Care and Procedural Skills • Interpersonal and Communication Skills • Medical Knowledge (“in vivo”) • Professionalism
Multisource feedback	• Professionalism • Interpersonal and Communication Skills • Systems-Based Practice • Medical Knowledge
Audit and performance data (clinical and patient safety indicators)	• Practice-Based Learning and Improvement • Systems-Based Practice • Medical Knowledge
Simulation (if available)	• Patient Care and Procedural Skills • Interpersonal and Communication Skills • Medical Knowledge
Written and In-training exam (if available)	• Medical Knowledge
Case or procedure logs	• Patient Care and Procedural Skills • Practice-Based Learning and Improvement

Programs and institutions must know where and how frequently assessments are being used. Developing a tracking table can help ensure that learners are assessed appropriately during learning activities. The table below provides an example of how such tracking can be completed.

Assessment Method/Tool	Core Competency(ies) Targeted for Assessment	Assessment of Tool’s Effectiveness (High/Medium/Low)*	Rotation or Location of Application	Frequency of Assessment

©2026 Accreditation Council for Graduate Medical Education (ACGME)

*The effectiveness of an assessment tool can be determined using a framework like the Ottawa Framework for Good Assessment (Norcini, et al. 2018). This framework lists attributes of an assessment tool and asks the user to determine how effectively the tool achieves those attributes. Attributes include reliability, validity, reproducibility, feasibility, educational effect (of and for learning), and acceptability. This framework informs judgement regarding the likelihood that a specific method or tool will generate good assessment. If a tool lacks any of these attributes, its effectiveness will be substantially diminished. For instance, a tool such as direct observation that is not accepted by faculty because it might take too long or is deemed too expensive because it decreases Relative Value Unit (RVU) generation may be unlikely to be successfully implemented (and therefore has low effectiveness). If an assessment method/tool is assessed to have low effectiveness, the program/institution may consider faculty development activities that could enhance the tool's effectiveness.

Miller's pyramid (Miller 1990), mentioned elsewhere in this Guidebook, is worth revisiting as programmatic assessment systems are developed. Programmatic assessment should sample appropriately across all learning venues and at expected levels of learning. While the emphasis of assessment at the GME level should focus on work-based assessment (i.e., the "Does" of Miller's pyramid), programmatic assessment should also investigate learning as appropriate across the full continuum of "Knows" to "Does." Assessment in UME will start with "Knows" and "Knows How" before progressing to "Shows How." Tracking levels of learning being assessed and where, how, and how frequently assessments are being completed will ensure that robust assessment is completed across all necessary competency domains throughout the educational program.

As discussed throughout this Guidebook, the quality of data generated by assessment tools and methods is dependent on the abilities of the individuals using them. Effective programmatic assessment requires that assessors understand the goals and objectives of the educational program and that they have a shared understanding, or mental model, of how the assessment program defines and documents the developmental progression of learners. These goals and desired outcomes also need to be shared with the learners to ensure they understand the goals of assessment and the importance of using assessment data to develop and implement ILPs.

Finally, assessment methods and tools must be "fit for their intended purpose." If an assessment is elegantly designed and deployed but does not generate data that allows for effective judgements of a learner's developmental progression, it is insufficient. Likewise, if assessment data do not directly inform decisions about the achievement of desired educational outcomes, those assessments and data are "not fit for purpose" and should be either redesigned to address educational outcomes or discontinued. Hauer et al. (2018) have identified six principles of programmatic assessment that can help avoid this situation and should be used by all programs/institutions as they implement programmatic assessment.

Programmatic Success Principles
Centrally coordinated plan for assessment aligns with and support curricular vision.
Multiple assessment tools used longitudinally generate multiple data points.
Learners require ready access to information-rich feedback to promote reflection and informed self-assessment.
Coaching is essential to facilitate effective data use for reflection and learning planning.
The program of assessment fosters self-regulated learning behaviors.
Expert groups make summative decisions about grades and readiness for advancement.

References

Hauer, Karen E., Patricia S. O'Sullivan, Kristen Fitzhenry, and Christy Boscardin. 2018. "Translating Theory into Practice: Implementing a Program of Assessment." *Academic Medicine* 93 (3): 444-450. https://pubmed.ncbi.nlm.nih.gov/29116979/.

Norcini, John, M. Brownell Anderson, Valdes Bollela, et al. 2018. "2018 Consensus Framework for Good Assessment." *Medical Teacher* 40 (11): 1102-1109. https://pubmed.ncbi.nlm.nih.gov/30299187/.

Schut, Suzanne., Lauren A. Maggio, Sylvia Heeneman, Jan van Tartwijk, Cees van der Vleuten, and Erik Driessen. 2021. "Where the Rubber Meets the Road — An Integrative Review of Programmatic Assessment in Health Care Professions Education." *Perspectives in Medical Education*. 10 (1): 6-13. https://pubmed.ncbi.nlm.nih.gov/33085060/.

The Role of the Milestones and Entrustable Professional Activities in Programmatic Assessment in GME

Building effective programmatic assessment requires creation of an organized combination of assessment methods and tools for determining a learner's developmental progression in each of the six Core Competencies. The Milestones serve as the framework for tracking that developmental progression. The Milestones were **not** designed for use as stand-alone faculty evaluation forms in GME. Rather, the Milestones provide a roadmap for the interpretation of rotation-based assessment data (especially work-based assessments) that guides the synthetic judgements of the Clinical Competency Committee. As such, graduate medical educators must recognize that the Milestones are not graduation requirements. Rather, they are targets to guide the development of GME learners. If a learner's trajectories are consistently missing expected targets in any area of general competency growth (i.e., progression in the Milestones), GME programs should critically assess whether the learner requires additional support while simultaneously reviewing curriculum content and its delivery and assessment methods/tools to ensure the educational program is providing the appropriate learning environment. Through this process, programs can support individual learning needs while also identifying and removing or improving ineffective learning and assessment activities as part of programmatic quality improvement.

The Milestones are also dynamic and subject to revision. In 2016, the ACGME launched the Milestones 2.0 effort to review, refine, and revise all the initial Milestones sets. Milestones 2.0 addresses the substantial variability in content and developmental progression in the initial subspecialty milestones and simplifies and standardizes language used to describe developmental progression. The ongoing Milestones 2.0 initiative has identified a set of standardized (or "harmonized") subcompetencies in the four non-Patient Care and Medical Knowledge competency domains. The evolving Milestones should guide programs as they review and update their assessment programs to ensure they continue to meet desired educational outcomes.

As the ACGME's accreditation model has matured over time, interest in EPAs has also grown. While use of EPAs is not required for ACGME accreditation, EPAs have gained support as a strategy for structuring clinical assessment. EPAs were introduced by Olle ten Cate (2005) as a framework to define and assess essential clinical activities required of the profession. EPAs describe the essential work of the profession, whereas Milestones and competencies describe attributes of the learner and provide a framework for defending decisions about a learner's trustworthiness and readiness to progress professionally. As interest in the use of EPAs as an assessment strategy has grown, many specialty societies have developed specialty-specific EPAs. While such EPAs are valuable, programs can also develop customized EPAs to document achievement of desired outcomes for specific rotations.

The Role of the Milestones and Entrustable Professional Activities in Programmatic Assessment in UME

The Foundational Competencies for Undergraduate Medical Education, released in December 2024 through a collaboration among the AAMC, ACGME, and AACOM, provide a shared national framework and broader conceptualization of the outcomes expected of medical school graduates. Organized across six domains (Professionalism; Patient Care and Procedural Skills; Medical Knowledge; Practice-Based Learning and Improvement; Interpersonal and Communication Skills; and Systems-Based Practice), the foundational competencies establish a common language for learner development across the educational continuum. These competencies are intended to serve as key drivers of programmatic assessment in UME, helping to align curricular design, assessment systems, coaching, and transition-to-residency expectations around meaningful developmental outcomes rather than isolated measures of performance.

Within this framework, milestones and Entrustable Professional Activities (EPAs) play complementary roles in operationalizing competency-based assessment. Milestones describe developmental progression or steps within competency domains, providing observable behavioral indicators that help educators and learners understand growth over time. They are intended to support formative feedback, individualized coaching, learning plans, and the aggregation of evidence across multiple settings and assessors. By articulating developmental expectations, milestones help create a shared understanding of competence and identify areas requiring additional support or enrichment. Milestones for the Foundational Competencies are currently under joint development by the sponsoring organizations.

EPAs focus on the authentic professional work performed in clinical settings. EPAs integrate multiple competencies into meaningful units of practice and emphasize the degree of supervision a learner requires to carry out professional responsibilities safely and effectively. In UME, EPAs for entering residency help bridge educational outcomes with expectations for residency readiness by translating competencies into workplace activities that can be observed and entrusted in practice. The AAMC CoreEPAs developed in 2014 and AACOM Considerations published in 2016 will be revised in the near future by the sponsoring organizations.

Within a programmatic assessment approach, Milestones and EPAs are strengthened through the use of multiple complementary assessment methods collected longitudinally and across contexts. This Guide includes many examples, and each contributes pieces of evidence that, when synthesized, provide a richer and more reliable picture of learner progression. Rather than relying on single high-stakes assessments, programmatic assessment emphasizes continuous evidence-gathering, reflection, coaching, and informed decision-making to support learner development and readiness for the transition to residency.

©2026 Accreditation Council for Graduate Medical Education (ACGME)

References

Association of American Medical Colleges. 2014. *Core Entrustable Professional Activities for Entering Residency: Curriculum Developers' Guide*. Washington, DC.

American Association of Colleges of Osteopathic Medicine. 2016. *Osteopathic Considerations for Core Entrustable Professional Activities (EPAs) for Entering Residency*. Bethesda, MD.

Appendix: Van Melle's Core Components Framework for CBME

Table. Van Melle's Core Components Framework for Implementation of Competency-Based Medical Education Programs[1]

Component	Description
An Outcomes-Based Competency Framework	• Desired outcomes of training are identified based on societal needs. • Outcomes are paramount so that the graduate functions as an effective health professional.
Progressive Sequencing of Competencies	• In CBME, competencies and their developmental markers must be explicitly sequenced to support learner progression from novice to master clinician. • Sequencing must consider that some competencies form building blocks for the development of further competence. • Progression is not always a smooth, predictable curve.
Learning Experiences Tailored to Competencies In CBME	• Time is a resource, not a driver or criterion. • Learning experiences should be sequenced in a way that supports the progression of competence. • There must be flexibility to accommodate variation in individual learner progression. • Learning experiences should resemble the practice environment. • Learning experiences should be carefully selected to enable the acquisition of one or many abilities. • Most learning experiences should be tied to an essential graduate ability.
Teaching Tailored to Competencies	• Clinical teaching emphasizes learning through experience and application, not just knowledge acquisition. • Teachers use coaching techniques to diagnose a learner in clinical situations and give actionable feedback. • Teaching is responsive to individual learner needs. • Learners are actively engaged in determining their learning needs • Teachers and learners co-produce learning.
Programmatic Assessment (i.e., Program of Assessment)	• There are multiple points and methods for data collection. • Methods for data collection match the quality of the competency being assessed. • Emphasis is on workplace-based assessment. • Emphasis is on providing personalized, timely, meaningful feedback. • Progression is based on entrustment. • There is a robust system for decision-making. • Good assessment requires attention to issues of implicit and explicit bias. that can adversely affect the assessment process.

[1]Van Melle, Elaine, Jason R. Frank, Eric S. Holmboe, Damon Dagnone, Denise Stockley, Jonathan Sherbino, et al. 2019. "A Core Components Framework for Evaluating Implementation of Competency-Based Medical Education Programs." *Academic Medicine* 94 (7): 1002-1009. https://pubmed.ncbi.nlm.nih.gov/30973365/.

©2026 Accreditation Council for Graduate Medical Education (ACGME)