Assessment Guidebook

Eric S. Holmboe, MD
William F. Iobst, MD

Acknowledgment

The original Toolbox of Assessment Methods was released in September 2000. We are indebted to this work led by Drs. Susan Swing and Philip Bashook. While much has changed, some of the original text has stood the test of time and is in included in this 2020 first version of the new ACGME Assessment Guidebook. We thank all the authors who contributed to the original Toolbox.

We also wish to thank Daniel Dent, MD; Stephen Durning, MD, PhD; Laura Edgar, EdD; Greg Ogrinc, MD; and Liana Puscas, MD for their review of the guidebook.

General Disclaimer

This ACGME Assessment Guidebook provides general guidance and descriptions of assessment methods and approaches that can be used for assessing residents and fellows. The guidebook does not include all the tools that can or may be used by a residency or fellowship program for evaluating residents and fellows, or by a program director in verifying that a resident or fellow has demonstrated sufficient professional ability to practice without supervision. The ACGME shall not be liable in any way for results obtained in applying these assessment methods. The user, and not the ACGME, shall be solely 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.

Table of Contents

Preface

The ACGME Outcome Project was formally launched in July 2001. A major goal of the Outcome Project was to "enhance residency education through outcomes assessment." To help programs prepare for the Outcome Project, the ACGME, in collaboration with the American Board of Medical Specialties (ABMS), released the *Toolbox of Assessment Methods* in September 2000. The original toolkit, co-lead by Drs. Susan Swing and Philip Bashook, provided a useful and practical compendium of assessment methods for the six general competencies. The authors of this guidebook are indebted to their and others' contributions at the beginnings of the outcomes-based approach in graduate medical education.

Twenty years later substantial progress has occurred but many of the same challenges remain in assessing the six ACGME Core Competencies. The biggest change since 2001 has been the introduction of the Milestones, along with the requirement that programs utilize Clinical Competency Committees (CCCs) to assess learners' longitudinal professional development using programs of assessment.

This Assessment Guidebook builds on this robust history. 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 for assessing residents and fellows. Each method or tool is now described based on key characteristics for "good assessment" (van der Vleuten, 1996; Norcini 2018); Section III provides guidance on mapping assessments to the Core 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 available resources on the Milestones section of the ACGME website (available at https://www.acgme.org/What-We-Do/Accreditation/Milestones/Resources):

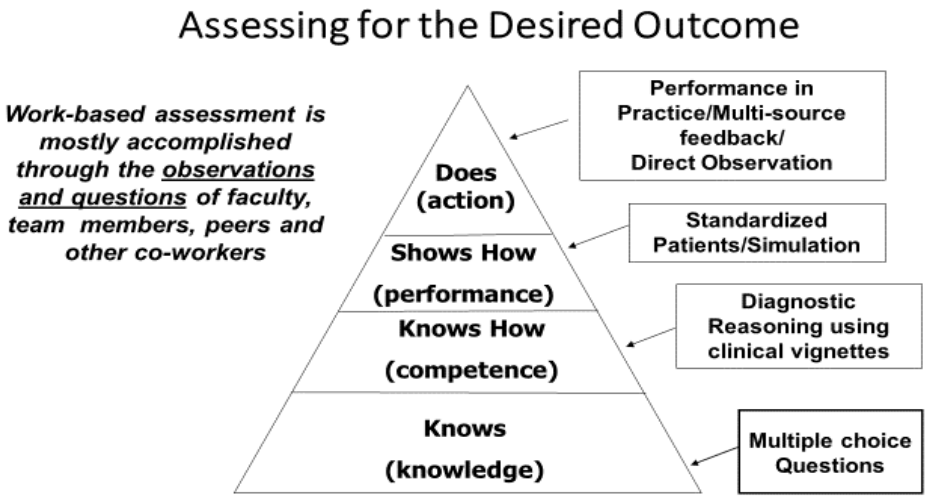
1. Clinical Competency Committee Guidebook, 3rd Edition
2. Milestones Guidebook, 2nd Edition
3. Milestones Guidebook for Residents and Fellows, 2nd Edition
4. Milestones Implementation Guidebook

I. General Principles of Assessment

Importance of Assessment

Assessment is vital to effective professional development, yet is still often neglected, avoided, or performed poorly in graduate medical education (GME). A large body of research has repeatedly confirmed that accurate, robust assessment is essential for effective feedback, coaching, self-regulated learning, and professional growth. Assessment culture in GME has been slow to change despite the introduction of the ACGME Core Competencies and the push toward educational outcomes. The entire GME system needs to accelerate cultural change to make assessment a valued activity. Valid assessment is a social responsibility between learner, patient, faculty members, educational programs, and society. When assessment is done well, learners can more effectively and quickly address gaps to improve and grow. When assessment is done poorly, learners can graduate insufficiently prepared for unsupervised practice, and in a worst-case scenario it leads to patient harm. Poor assessment can also be damaging to learners due to incorrectly identifying strengths or weaknesses, focusing on the wrong abilities or needs, or assessment based on implicit and explicit bias.

Miller's assessment pyramid (Miller 1990) remains a helpful framework to guide programs in building assessment systems:



While the first three levels, *Knows, Knows How, and Shows How*, are important assessment approaches, residency and fellowship programs should place their emphasis on the top of the pyramid: the *Does* level. The *Does* level requires attention to a robust combination of work-based assessments. It is also critical to recognize that the majority of all assessment is based on two primary activities: asking questions and observing. How programs and individuals perform these activities varies from assessment method to assessment method.

Utility and Good Assessment

Van der Vleuten in 1996 introduced the concept of the Utility Index for evaluating the quality of an assessment and helping educators make appropriate tradeoffs when choosing an assessment method or tool (van der Vleuten 1996). His original index was:

$$\text{Utility} = R_w \text{ X } V_w \text{ X } A_w \text{ X } EI_w \text{ X } C_w$$

where R = reliability, V = validity, A = acceptability, EI = educational impact, and C = cost. The *w* refers to the weight a program might place on each element. However, 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.

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

Framework for Good Assessment: Single 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 and improves overall program quality.
7. Acceptability. Stakeholders find the assessment process and results to be credible.

As is clear to see, many of the elements of the Utility Index are included in this list. Two changes are important to note. First, cost has been included in the 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 Milestones.

Primer on Validity

There is a lot of work over the past decades on the validity of assessments. For the purposes of this guidebook, the Messick framework is helpful and practical. Cook and colleagues produced a practical article that describes the Messick framework for medical education (Cook 2006). Examples using a direct observation of clinical care assessment are provided for each element.

1. Content: do items on the assessment completely and properly represent the construct of interest?
    a. Example: does a tool for direct observation capture the proper domains or components of clinical skills, such as medical interviewing, physical examination, etc.?
2. Response process: the relationship between the intended construct and the thought processes of subjects or observers
    a. Example: Have the members of the faculty been trained to use the direct observation method? Do the members of the faculty have a shared mental model of what they are being asked to rate?
3. Internal structure: acceptable reliability and factor structure
    a. Example: 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 viewed the same encounter, would the two faculty members rate the encounter the same (inter-rater reliability)?
4. Relations to other variables: correlation with scores from another instrument assessing the same construct
    a. Example: Would direct observation ratings by faculty members correlate with the quality of care delivered to the patient?
5. Consequences: do scores really make a difference?
    a. Example: Are the direct observation ratings used by the CCC in its Milestones deliberations? Did the learner receive feedback?

Programs of Assessment (Programmatic Assessment)

It is essential to conceptualize GME programs through a systems lens. Simply defined, a system consists of two or more interdependent parts that work together to accomplish a shared aim. GME 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 ultimately physicians highly prepared for 21st century practice. To design a competency-based medical education (CBME) program using systems thinking requires a better understanding of the core components of a CBME-designed program. Van Melle and associates 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 (Appendix 1; van Melle 2018). Effective programmatic assessment is essential to ensuring that 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 that work together on a regular basis to perform appropriate assessments that enable valid determinations of learner progression towards competence. This program of assessment shares agreed-upon goals and outcomes, linked individual learner assessment and program evaluation processes, information about learner's performance (both feedback and feed-forward mechanisms), and the desire to produce a physician fully prepared to enter a subspecialty fellowship program or 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 the quality improvement activities of the program.

No single assessment tool or method is sufficient to judge the overall abilities and readiness for unsupervised practice of physician learners. GME programs will need to choose a combination of assessments to effectively cover the Competencies and the Milestones in their discipline. In addition to using the above criteria for good assessment and the utility index, programs need to create an assessment and curriculum map to ensure their assessment program is robust. Norcini and colleagues also provided guidance for assessment systems (Norcini et al 2018).

Framework for Good Assessment: Systems of Assessment

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

II. Recommended Assessments Methods by Competency

Patient Care and Procedural Skills

Direct observation is the primary and overarching method to assess patient care and procedural skills. 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 (the "Shows how" level of Miller pyramid). A mastery-based approach is strongly recommended when using any form of simulation to teach and assess skills, especially procedural skills (McGaghie 2020). Research on mastery-based approaches for some procedural skills training has demonstrated this type of training with robust assessment translates into better outcomes for patients.

Direct observation of patient encounters, procedures, family meetings, and other clinical activities is essential, and many tools exist that programs can use to capture and document these observations (the "Does" level of Miller pyramid). These tools range from single encounter tools (e.g., mini-clinical evaluation exercise (mini-CEX), objective structured assessment of technical skills (OSATS)) 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, 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 medical records, discharge summaries, operative notes, and others 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 the competency of systems-based practice.

Medical Knowledge and Clinical Reasoning

Multiple choice tests have long been used to assess the capability of learners at the "Knows" and "Knows how" levels. Some standardized knowledge assessments target the "Shows how" level. The most common standardized assessment is the in-training examination (ITE) now available in most specialties. These ITEs have predictive validity for performance on certification examinations.

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. Clinical reasoning is a complex process that is affected by numerous contextual factors. Assessing clinical reasoning requires faculty members and others to be expert questioners using evidence-based theory. Structured assessment tools, such as chart stimulated recall (CSR) and the assessment of reasoning tool (ART), directly target clinical reasoning. Additionally, clinical reasoning can be assessed through observation, faculty evaluation forms, and interactive methods, such as the "think aloud" or SNAPPS presentation model.

Professionalism
Effective assessment of professionalism requires, at a minimum, a multisource feedback (MSF) approach. A complete MSF assessment approach must include patient experience surveys, and when appropriate, family experience surveys. While physician faculty members can certainly assess what they observe with regards to professional behaviors, and several tools exist for this purpose, physician faculty members in general are poorer observers and assessors of professionalism unless critical incidents or deficiencies are displayed in their presence. Multiple studies have found that physician faculty members and nursing assessments of learners often do not correlate.

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

Practice-based Learning and Improvement
Practice-based learning and improvement now focuses on two core subcompetencies: evidence-based practice (EBP) and reflective practice (RP). For EBP, programs can use assessment of articles using research quality criteria to judge a learner's ability to effectively review and apply literature for patient care. A number of tools also exist to judge a learner's ability to perform an EBP review. EBP case logs, using the patient-intervention-comparator-outcome (PICO) format are an effective tool for capturing application of EBP skills in patient care ("Does").

For RP, informed self-assessment and creation of individualized learning plans (ILPs) are essential. Informed self-assessment involves residents or fellows using their assessment and clinical performance data for individual continuous quality improvement. ILPs enable residents and fellows 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. Use of patient safety and quality of care performance data is essential for assessing this competency. It is simply not possible to effectively assess systems-based practice without clinical performance data. While attribution of clinical performance data to any single resident or fellow can be difficult, programs can use the lens of contribution to review the role the resident or fellow played in the performance measure results. Research has also recently shown that programs can use a systematic approach to choose

quality and safety measures that possess a strong connection to the learner's actions (Schumacher 2019).

For system navigation for patient-centered care, an MSF approach is essential. Tools are currently being developed for assessing knowledge in systems, but programs may want to consider using educational programs such as the Institute for Healthcare Improvement Open School (www.ihi.org), or one of their specialty society's resources.

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 Core Competencies. The table below provides a basic menu of assessment tools/methods appropriate for use in each Core Competency domain.

By Competency:

Competency	Competency-Based Assessment Options
Medical Knowledge	• In-training exam • Faculty work-based assessments • Chart stimulated recall, Assessment of Reasoning Tool, others
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
Professionalism	• Informed self-assessment • Multi-source feedback, such as a 360-degree evaluation • Patient experience surveys
Interpersonal and Communication Skills	• Patient reported feedback and experience surveys • Multi-source feedback, such as a 360-degree evaluation, especially regarding interprofessional care
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
Systems-based Practice	• Feedback from multiple faculty evaluations regarding ability to practice in a complex health care system • Multi-source feedback, such as a 360-degree evaluation, especially regarding interprofessional care • Assessment of cost-conscious care

By Basic Assessment Method:

Assessment Tool/Method	Targeted Competency
Faculty assessment (can be interprofessional)	Multiple competencies
Direct observation	Patient Care and Procedural Skills, Interpersonal and Communication Skills, Medical Knowledge (“in vivo”), Professionalism
Multi-source 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, and Medical Knowledge
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

III. Assessment Methods and Tools

Assessment of Medical Knowledge and Clinical Reasoning

Description
When assessing medical knowledge, distinguishing between the acquisition of knowledge and its application 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, and clinical reasoning. These abilities are collectively described as clinical reasoning, the process wherein clinicians observe, collect, and interpret data to diagnose and manage patients.

Components of Clinical Reasoning:
1. Information gathering
2. Hypothesis generation
3. Problem representation
4. Differential diagnosis
5. Leading or working diagnosis
6. Diagnostic justification
7. Management

Acquisition of medical knowledge is traditionally assessed using standardized testing with multiple-choice questions (MCQs). Many GME programs rely on an ITE for this assessment. The literature supporting this approach to assessment is well established and is both valid and reliable. While not commonly used in ACGME-accredited programs, a program of progress testing, such as is used at Maastricht University, can also be used to assess medical knowledge. The Maastricht progress testing program consists of a series of four MCQ exams that are administered on an annual basis to assess a learner's progress in achieving program learning objectives.

Assessment of acquired knowledge is foundational to any assessment program but while essential, it alone, is not sufficient for determining competence in the medical knowledge competency domain. Learners must also be able to apply that knowledge in the clinical setting.

In the clinical setting, medical knowledge (both acquired and applied) can also be simultaneously investigated using clinical questioning. Such questioning can be accomplished with multiple formats. Daniels and associates have categorized the assessment of clinical reasoning into non-workplace-based assessment, assessment in simulated clinical environments, and workplace-based assessment (Daniels 2019). Methods of assessment appropriate to each of these categories are described by Daniels and associates in the provided reference, and the reader is encouraged to review the Daniels article for full descriptions of these methods. 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 written form. A classic questioning strategy in this setting is the one-minute preceptor. 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.

One-Minute Preceptor

1. Get a commitment from the learner. Ask, "What is the likely diagnosis in the case being presented?"
2. Probe for supporting evidence/underlying reasoning. Ask, "What supports/contradicts this diagnosis?"
3. Teach general rules relevant to the topic.
4. Reinforce what was done right by the learner. Provide positive feedback.
5. Correct mistakes with suggestions on how to approach a similar situation next time.

Clinical reasoning can also be assessed using methods such as the "think aloud" or SNAPPS. 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.

Competencies

Assessments of medical knowledge and clinical reasoning interrogate the Core Competencies of medical knowledge and patient care.

Validity

Assessment of medical knowledge acquisition through standardized testing has been extensively studied and is both valid and reliable. For instance, performance on ITEs have been shown to correlate with subsequent certification board passage rates and in general provide a more accurate global assessment of medical knowledge than do faculty ratings. However, an ITE is usually a one-time assessment event and mostly functions at the lower levels of the Miller pyramid. Assessment of medical knowledge application is a more complex process with variable validity and reliability. The implementation of such assessments should be carefully planned and address where and how the assessment will be used by the program.

Feasibility

Standardized testing for medical knowledge is time friendly, places low demand on faculty members and is predictive of ultimate board certification exam performance. It allows for feedback to the learner and the curriculum and accesses a foundational attribute of the learner – medical knowledge. While these benefits make use of such exams quite feasible, programs should recognize that such testing can shift emphasis away from the importance of actual patient care, is associated with cost, and is usually given once a year.

Assessment of medical knowledge application and clinical reasoning is essential to achieving the outcome of a future physician workforce capable of providing safe and effective patient care. For that reason, programs must develop assessments of this critical clinical skill. With careful planning and appropriate communication to faculty members and learners, an assessment program that interrogates all appropriate levels of Miller's pyramid, from "knows" to "does," is possible and feasible.

Acceptability

Tools such as in-service examinations and other standardized MCQ tests are highly acceptable to both faculty members and learners. The other assessment methods referenced in this section require that faculty members and learners fully understand how the assessment will be completed and its purpose. With the appropriate preparation and communication, any of these assessment methods can be a credible part of programmatic assessment.

Catalytic Effect

When data from these assessments is used for documentation of both what has been learned (of learning) and what is needed to achieve competence (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's curriculum and assessment program.

SUGGESTED REFERENCES

Daniel, M., J. Rencic, S.J. Durning, et al. 2019. “Clinical Reasoning Assessment Methods: A Scoping Review and Practical Guidance.” *Academic Medicine* 94(6): 902-912. https://doi.org/10.1097/ACM.0000000000002618

Maastricht University. Progress Testing. https://www.maastrichtuniversity.nl/meta/368652/progress-test-examination-year-1

Neher, J., K. Gordon, B. Meyer, and N. Stevens. 1992. “A Five-Step ‘Microskills’ Model of Clinical Teaching.” *Journal of American Board of Family Practice* 5: 419-424.

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

Wolpaw, T., Klara K. Papp, and G. Bordage. 2009. “Using SNAPPS to Facilitate the Expression of Clinical Reasoning and Uncertainties: A Randomized Comparison Group Trial.” *Academic Medicine* 84(4): 517-524. https://doi.org/10.1097/ACM.0b013e31819a8cbf

Chart Stimulated Recall and the Assessment of Clinical Reasoning in the Workplace (CSR/ART)

Description

Chart-stimulated recall (CSR) is a work-based assessment tool that complements chart audit by combining a chart review of a patient encounter with a structured oral examination. CSR provides a structured approach to the use of clinical questioning and can assess all of the Core Competencies. With appropriate faculty development, a faculty member uses a set of predetermined questions to probe a learner's thought process in areas including the clinical assessment (medical knowledge and 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, and interpersonal and communications skills). Using reflective questioning and prompted self-assessment, CSR can also assess some aspects of professionalism and practice-based learning and improvement. An example of a structured CSR tool is provided in Appendix 1.

The Assessment of Reasoning Tool (ART) by the Society to Improve Clinical Reasoning 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 CSR 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 that can be accessed at www.improvediagnosis.org/art.

When conducting a CSR 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

CSR and ART predominantly assess medical knowledge and patient care and procedural skills. CSR can address other competencies based on the structure and desired outcome of the CSR tool/exercise.

Validity

When structured appropriately, CSR can provide valid assessments of clinical reasoning. As early as 1982, the American Board of Emergency Medicine demonstrated that well-designed

CSR exercises were valid and reliable enough to inform board certification decisions. In that setting, the physician examiners completing the CSR required specific training on how to question examinees and evaluate and score their responses to questions. This degree of structure and preparation is not easily reproduced in GME programs. GME programs may therefore prefer to use CSR as a formative assessment tool to guide ILPs and formative feedback to learners regarding clinical reasoning skills.

Feasibility

Successful use of CSR requires that the specifics of how, when, and why are clearly understood by everyone participating in the exercise. Programmatic use of CSR requires advanced planning and is not an assessment that can be used "on the fly." ART 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. Regarding CSR, programs should determine the rotation(s)/learning venues that will use CSR, the location, the learners who will be assessed, the faculty members who will conduct CSR, and the desired outcome of the exercise. The specific questions used in the exercise and the frequency CSR is completed should also be standardized. Finally, adequate time (30 to 60 minutes) must also be allotted to complete the CSR exercise. Given the complexities of CSR, care must be taken to ensure its value is fully defined and explained.

Acceptability

With appropriate preparation and communication, CSR can be developed as an important work-based assessment tool. Given the time commitment and preparation required to conduct CSR, its role in the program's assessment system must be clearly understood by both learners and faculty members. If this is clearly communicated and understood, CSR can be developed as an acceptable and essential part of a program's assessment portfolio. ART is more portable and can be completed as part of daily educational activities.

Catalytic Effect

Both CSR and ART 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 CSR 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.

SUGGESTED REFERENCES

Ilgen, J.S., A.J. Humbert, G. Kuhn, M.L. Hansen, G.R. Norman, K.W. Eva, B. Charlin, and J. Sherbino. 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

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

Sinnott, C., 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 Serv Res* 17: 583. https://doi.org/10.1186/s12913-017-2539-y

Thammasitboon, S., J.J. Rencic, R.L. Trowbridge, A.P.J. Olson, M. Sur, and G. Dhaliwal. 2018. "The Assessment of Reasoning Tool (ART): Structuring the Conversation Between Teachers and Learners." *Diagnosis (Berl)* 5(4): 197-203. https://doi.org/10.1515/dx-2018-0052

Faculty Global Assessment Forms

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 scheduled rotations or educational experiences with learners typically complete an assessment of each learner. These are usually global assessments that include a rating scale and section for written comments in each of the six Core Competencies. These forms typically assess a learner's knowledge, skills, and attitudes using specific frameworks for defining the anchors for what is being assessed. Common frameworks include analytic assessments of specific and individual components of a task (physical exam or counseling); developmental frameworks, such as the Dreyfus and Dreyfus model (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 (a.k.a. "co-activity scales") or the amount of supervision required.

Competencies

Faculty assessment forms can be used to assess all six Core Competencies.

Validity

Multiple studies have demonstrated major intra- and inter-rater reliability issues with faculty assessment. When using faculty assessment forms, programs must understand that the major source of variability is the faculty member, not the form. The horn or halo effect, leniency (dove), or severity (hawk) error, straight-line and central tendency errors can all reduce the utility of faculty assessment. Such variability occurs when faculty members do not know or apply assessment criteria accurately and highlights the importance of faculty development to grow consensus on what is being assessed and a "shared mental model" of program 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. Programs are encouraged to focus time and energy into 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 still struggle to get assessment forms completed and returned in a timely fashion. Mobile tools and apps may help. Faculty members should also be encouraged to capture assessment notes and observations during their time with the learner. This "aide-de-memoir" approach can help faculty members remember key details when the time comes to complete the assessment form.

Acceptability
Faculty assessment forms are accepted as routine and necessary in virtually all GME programs. 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 and 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 for the rotation. 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 when critical feedback is provided.

Catalytic Effect
The catalytic effect of faculty 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 faculty assessment may be diminished through lack of continuity with the faculty member who initiated this commitment to change.

SUGGESTED REFERENCES

Crossley, J., G. Johnson, J. Booth, and W. Wade. 2011. "Good Questions, Good Answers: Construct Alignment Improves the Performance of Workplace-Based Assessment Scales." *Medical Education* 45: 560–569. https://doi.org/10.1111/j.1365-2923.2010.03913.x

Rekman, J., W. Gofton, N. Dudek, T. Gofton, and S. J. Hamstra. 2016. "Entrustability Scales: Outlining Their Usefulness for Competency-Based Clinical Assessment." *Academic Medicine* 91(2): 186-90.

Rodrigues, R.G. and L.N. Pangaro. 2012. "AM Last Page: Mapping the ACGME Competencies to the RIME Framework." *Academic Medicine* 87(12): 1781. https://doi.org/10.1097/acm.0b013e318271eb61

Silber, C.G., Thomas J. Nasca, L. Paskin, Glenn Eiger, Mary Robeson, and Jon J. Veloski. "Do Global Rating Forms Enable Program Directors to Assess the ACGME Competencies?" 2004. *Academic Medicine* 79 (6): 549-556. DOI: 10.1097/00001888-200406000-00010.

Tolsgaard, G., H. Arendrup, B.O. Lindhard, J.G. Hillingsø, M. Stoltenberg, and C. Ringsted. 2012. "Construct Validity of the Reporter-Interpreter-Manager-Educator Structure for Assessing Students' Patient Encounter Skills." *Academic Medicine* 87: 799-806.

Williams R.G., D.A. Klamen and W.C. McGaghie. 2003. "Cognitive, Social and Environmental Sources of Bias in Clinical Performance Ratings." *Teaching and Learning in Medicine* 15(4): 270-292. https://doi.org/10.1207/S15328015TLM1504_11

Procedure or Operative Case Logs

Description

Procedure, operative, or case logs document each patient encounter by medical conditions seen, surgical operation, or procedures performed. The logs may or may not include counts of cases, operations, or procedures. Patient case logs currently in use involve recording of some number of consecutive cases in a designated time frame. Operative logs in current use vary; some entail comprehensive recording of operative data by CPT code, while others require recording of operations or procedures for a small number of defined categories. However, volume of procedures is a crude proxy for competence. Procedure and case logs should now be accompanied by assessment of the level of competence in that procedure. Any complications that may have occurred during the procedure can also be documented. With the advancement of information technology, completing case logs can be accomplished by mobile apps or over an Internet connection to a learning management system. 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 experience. Regular review of logs can be used to help residents and fellows 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 the patient care and procedural skills competency. Depending on how the case log tool is constructed, they could be useful for practice-based learning and improvement and systems-based practice (if data and assessment on patient safety issues and events, complications, etc. are included in the assessment). Informed consent is an essential communication skill insufficiently taught and assessed and 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 residents'/fellows' recording. Unless defined by CPT or other codes, cases or procedures counted for a given category may vary across residents/fellows and programs. Minimum numbers of procedures required for accreditation and certification have not been rigorously validated against the actual quality of performance of an operation or patient outcomes. Assessing actual performance along a growth curve should now be the standard for assessing procedural and operative skills. A mastery-based approach to teaching and assessing procedural skills should become the standard. Research for some procedures (e.g., central lines, colonoscopy) have demonstrated validity for some patient outcomes.

Feasibility/Practicality
Electronic recording devices and systems can facilitate the collection and summarization of patient cases or procedures performed. Several surgical 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 will need enter the data into their electronic system. Data entry of manual records typically can be performed by a clerk but is time consuming depending on the number of residents or fellows in the program and log reporting requirements.

Smart phone apps and/or web-based tools can substantially enhance the feasibility of gathering case log data. These tools also can provide instant access for residents and fellows 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 residents or fellows and programs is related to some learners' tendency to only log minimum requirements.

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

SUGGESTED REFERENCES

Abdelstattar J.M., Y.N. Aljamal, R.K. Ruparel, P.G. Rowse, S.F. Heller, and D.R. Farley. 2018. "Correlation of Objective Data with General Surgery Resident In-Training Evaluations and Operative Volumes." *Journal of Surgical Education* 75: 1430-6.

Ghaderi, I., F. Manji, Y.S. Park, D. Juul, M. Ott, I. Harris, and T.M. Farrell. 2015. "Technical Skills Assessment Toolbox. A Review Using the Unitary Framework of Validity." *Annals of Surgery* 261: 251-62.

Naik, N.D., E.F. Abbott, J.M. Aho, T.K. Pandian, C.A. Thiels, S.F. Heller, and D.R. Farley. 2017. "The ACGME Case Log System May Not Accurately Represent Operative Experience Among General Surgery Interns." *Journal of Surgical Education* 74(6): e106-110 doi: 10.1016/j.jsurg.2017.09.032

Stulberg, J.J., R. Huang, L. Kreutzer et al. 2020. "Association between surgeon technical skills and patient outcomes." JAMA Surgery. 2020; published online: doi: 10.1001/jamasurg.2020.3007

Williams R.G., M.J. Kim, and G.L. Dunnington. 2016. "Practice Guidelines for Operative Performance Assessments." *Annals of Surgery* 264: 934-48.

Clinical Performance (Record) Review

Description

Critique of the clinical (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: 1) an archive of important patient medical information for use by other health care providers and patients; 2) 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. Residents and fellows 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 and systems-based practice. Clinical records can also serve as the content for CSR to explore clinical reasoning (for the patient care and procedural skills and medical knowledge competencies).

Validity

An extensive body of literature now exists on the validity and reliability of a host of clinical performance measures. Both the Agency for Healthcare Research and Quality (AHRQ: https://www.ahrq.gov/patient-safety/quality-measures/index.html) and the National Committee for Quality Assurance (NCQA: https://www.ncqa.org/hedis/measures/) house a large warehouse of clinical performance measures with supporting evidence as two sources available for programs interested in the science behind performance and safety measures. Many specialty societies also publish and support clinical performance and safety measures within their specialty. Programs are encouraged to check with their respective specialty societies to see what is available.

The major challenge for residency and fellowship programs is attribution of performance measure results to a single resident or fellow. It is not necessary to attribute 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 contribution of this learner to the performance on this set of measures?). However, 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 residents and fellows within a specific clinical setting. Examining performance at the 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 future practice performance of graduates.

Finally, there is some recent literature on processes that can be used to choose performance measures where the resident or fellow provides the majority of the "contribution" to a quality or safety measure.

Feasibility

Feasibility depends heavily on the method used to conduct the review and its purpose. Almost all teaching hospitals and clinics now use an electronic health record (EHR). Quality and safety measures are commonly embedded into EHRs. Programs should check with their quality department and information technology (IT) department about how and what quality and safety measures can be pulled from the EHR for the residents/fellows and the program. One major challenge for most systems is properly identifying or attributing a patient to a resident/fellow. Mechanisms exist to do this, though, and programs are encouraged to work with their IT departments to choose an appropriate clinical setting and process to pull data for individual learners.

While the allure of using easily extracted EHR data is appealing, manual medical record audits can be an extremely meaningful educational as well as assessment experience. These manual audits can be performed by the individual resident or fellow. This approach supports learning about how a quality or safety measure is constructed, enables immediate 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 should use a standardized collection form to guide the data extraction. There is a wealth of information garnered from the EHR review other than the simple performance outcome measure that adds to the learning and reflection. Depending on the focus and number of performance measures and setting 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 GME. Factors that diminish acceptability include a) receiving performance data for patients not primarily cared for by the resident/fellow (i.e., resident's/fellow's contribution to the outcome measure is missing); b) lack of understanding of the measure; c) lack of efficacy in how to improve performance on the measure; d) training in dysfunctional clinical settings or systems where improving performance is challenging; e) lack of meaningful feedback or coaching; and f) no 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 residents and fellows 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.

SUGGESTED REFERENCES

Lynn, L.A., B. J. Hess, W. Weng, R.S. Lipner, and E.S. Holmboe. 2012. "Gaps in Quality of Diabetes Care in Internal Medicine Residency Clinics Suggest Better Training is Needed in Ambulatory Settings." *Health Affairs (Millwood)* 31(1): 150-8. https://doi.org/10.1377/hlthaff.2011.0907

Myers, J.S. and B.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. doi: 10.1136/bmjqs-2018-008305

Schumacher, D.J., E.S. Holmboe, C. van der Vleuten, J.O. Busari, and C. Carraccio. 2018. "Developing Resident Sensitive Quality Measures: A Model from Pediatric Emergency Medicine." *Academic Medicine* 93(7): 1071-1078. doi: 10.1097/ACM.0000000000002093

Schumacher, D.J., D.T.Y. Wu, K. Meganathan, L. Li, B. Kinnear, D.R. Sall, E.S. Holmboe, C. Carraccio, C. van der Vleuten, J. Busari, M. Kelleher, D. Schauer, and E. Warm. 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. doi: 10.1097/ACM.0000000000002748

Schumacher, D.J., A. Martini, E.S. Holmboe, C. Carraccio, C. van der Vleuten, B. Sobolewski, J. Busari, and T.L. Byczkowski. 2020. "Initial Implementation of Resident-Sensitive Quality Measures in the Pediatric Emergency Department: A Wide Range of Performance." *Academic Medicine* 95(8): 1248-55. doi: 10.1097/ACM.0000000000003147

Wong, B.M., K.D. Baum, L.A. Headrick, E.S. Holmboe, F. Moss, G. Ogrinc, K.G. Shojania, E. Vaux, E.J. Warm, and J.R. Frank. 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. doi: 10.1097/ACM.0000000000002937

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 (such as venipuncture arms and intubation mannequin heads) to dynamic computer-enhanced systems that can respond to user actions (such as full-body anesthesia 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 examinations – 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; they allow examinees to reason through a clinical problem with little or no cueing; they permit examinees to make life-threatening errors without hurting a real patient; they provide instant feedback so examinees can correct a mistaken action; and they 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 (VR) 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. The VR 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 the Miller pyramid. 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 is growing in interest and can assess capability in the interpersonal and communication skills competencies. A collection of simulation exercises has potential to better discriminate between resident performance than clinical evaluations.

Validity
Substantial progress has been made in validity research 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 OSCEs discussed in another section. Mastery-based learning using various forms of simulation has found some positive correlations between performance in the simulations and quality of care for patients. VR 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 and the reader is encouraged to access the references provided below.

Feasibility
It is strongly recommended that programs interested in using simulation contact experts at a simulation center. If an institution does not possess a simulation center or specific simulation tools and materials, the authors recommend reaching out to one in the local or regional area. Simulation centers are increasingly creating consortia to share expertise and provide services to outside programs. The biggest feasibility challenges in simulation, depending on the specific method or approach used, is cost and access to the requisite expertise, materials, and setting.

Acceptability
Acceptability with 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 are rare but serious. This appears to be especially true when mastery-based learning principles are incorporated into the simulation design and experience.

SUGGESTED REFERENCES

Abdelstattar J.M., Y.N. Aljamal, R.K. Ruparel, P.G. Rowse, S.F. Heller, and D.R. Farley. 2018. “Correlation of Objective Data with General Surgery Resident In-Training Evaluations and Operative Volumes.” *Journal of Surgical Education* 75: 1430-6.

McGaghie, W.C., J.H. Barsuk, and D.B. Wayne. 2020. *Comprehensive Healthcare Simulation: Mastery Learning in Health Professions Education.* Springer: New York.

Motola, I., L.A. Devine, H.S. Chung, J.E. Sullivan, and S.B. Issenberg. 2013. “Simulation in Healthcare Education: A Best Evidence Practical Guide. AMEE Guide No. 82.” *Medical Teacher* 35(10): e1511-30. DOI: 10.3109/0142159X.2013.818632

Scalese R. 2018. “Simulation-Based Assessment.” In *Practical Guide to the Evaluation of Clinical Competence*. Elsevier: Philadelphia.

Standardized (Simulated) Patients (OSCE)

Description

Direct observation that occurs via simulation of a patient by lay individuals (i.e., “actors”) uses what is 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 for specific competencies. Individuals who undergo more extensive training in order to portray and score a scenario with a high degree of reliability are referred to *standardized* patients. SPs can also be trained to document and report back resident/fellow actions and behaviors, teach residents/fellows via role-play and repeated practice, rate interpersonal and communication skills, and also provide detailed and timely feedback. SPs are an excellent way to provide first exposure—combined with assessment, feedback, and coaching—for 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.

Objective structured clinical examinations (OSCEs) are a formal, standardized use of SPs in series of stations to assess and rate the clinical skills of trainees. OSCEs can be higher or lower 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.

Competencies

SPs operate at the “Shows how” level and can be a very useful supplement to direct observation in 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 from the past four decades on SPs regarding the validity elements of content, response process, and reliability. Until recently, OSCEs were part of the USMLE Step 2 clinical skills examination. 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 examination using SPs involves identification of the specific competencies to be tested, training of SPs, developing checklists or rating forms, and setting criteria. Development time can be considerable but can be made more time efficient by sharing of SPs in a collaboration of multiple residency programs or in a single academic medical center. A new SP can learn to stimulate a new clinical problem in eight to 10 hours; and an experienced SP can learn a new problem in six to eight hours. About twice the training time is needed for SPs to learn to use checklists to evaluate resident/fellow performance. Facilities needed for the examination include an examining room for each SP station and space for residents/fellows 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 standardized patients.

Catalytic Effect
SPs can provide real-time excellent feedback when use mostly for teaching and feedback. Score reports from more formal SP encounters (OSCEs, unannounced SPs) can also be valuable provided sufficient, specific information about performance is provided. Providing learners with just numeric ratings or scores will not be very useful for guiding individual learning plans.

SUGGESTED REFERENCES

Harden, R.M., P. Lilley, and M. Patricio. 2015. *The Definitive Guide to the OSCE.* Elsevier: Philadelphia.

Motola, I., L.A. Devine, H.S. Chung, J.E. Sullivan, S.B. Issenberg. 2013. "Simulation in Healthcare Education: A Best Evidence Practical Guide. AMEE Guide No. 82." *Medical Teacher* 35(10): e1511-30. doi: 10.3109/0142159X.2013.818632

Pell, G., R. Fuller, M. Homer, and T. Roberts of the International Association for Medical Education. 2010. "How to Measure the Quality of the OSCE: A Review of Metrics - AMEE Guide No. 49." *Medical Teacher* 32(10): 802-11. doi: 10.3109/0142159X.2010.507716

Weiner, S.J., and A. Schwartz. 2014. "Directly Observed Care: Can Unannounced Standardized Patients Address a Performance Gap in Performance Measurement?" *Journal of General Internal Medicine* 29(8): 1183-7. doi: 10.1007/s11606-014-2860-7

Direct Observation of Clinical Skills

Description

A discussion on direct observation of clinical skills refers to observing a resident or fellow interacting with a patient taking a medical history, doing a physical exam, informed consent, or shared decision making (i.e., counseling) for the purpose of assessing the learner.

Workplace-based assessment is defined as the assessment of day to day practice in the authentic clinical environment. As such, direct observation of clinical skills is a work-based assessment strategy that sits at the top of the Miller assessment pyramid because it captures what a learner "does" with patients (despite the possibility the act of observation may change the level of performance, known as the Hawthorne effect).

Competencies

Direct observation is essential for assessing the patient care and interpersonal and communication skills competencies. Direct observation can also be combined with questioning to judge clinical reasoning and medical knowledge given the importance of the medical interview and physical examination in making a proper diagnosis and clinical treatment plan.

Validity

Substantial literature exists regarding the reliability of various direct observation tools such as the mini-clinical evaluation exercise and its variants. In general, high levels of reliability can be achieved when a sufficient number of assessments by multiple observers are performed. Evidence for the other validity components is more mixed. One reason is the lack of shared mental models among faculty members and lack of faculty development. Faculty members can lack sufficient levels of skill in the clinical skills they are judging. While faculty development cannot fix all challenges in direct observation, training in both the clinical skills of interest and in good assessment practice can help.

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. New smartphone assessment apps can help faculty members complete assessments more efficiently. Examples of snapshots are provided here:

Interview	Physical Examination	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 program culture and frequency of observation. Program cultures that discourage direct observation, make each observation high stakes, lack faculty member buy-in, and provide infrequent or poor feedback all undermine acceptability of direct observation. Infrequent performance of direct observation also leads to each encounter feeling high stakes to the learner. Programs that make direct observation a habit, train and support faculty members in this assessment skill, and provide a safe learning environment for learners to ask and seek direct observation achieve greater acceptability.

Catalytic Effect
Direct observation of clinical skills is essential in enabling effective feedback and coaching. Too often faculty members judge clinical skills through proxies, such as presenting a patient at morning report or as part of rounds in the hallway or preceptor room. When done well, direct observation can be very impactful in helping residents and fellows improve clinical skills.

SUGGESTED ASSESSMENT TOOL AVAILABLE FROM ACGME
The **Direct Observation of Clinical Care (DOCC) app** is a tool for faculty members and other evaluators to do on-the-spot or scheduled direct observation assessments of residents and fellows performing five clinical activities in which they are expected to achieve competence: performing a history and physical exam; effective clinical reasoning; informed decision making; breaking bad news; and safe hand-offs. The DOCC app is designed as an open access tool that Sponsoring Institutions and programs can implement locally through an integration with a residency management system or other database.

You can access information about **DOCC** here: https://dl.acgme.org/pages/assessment

SUGGESTED REFERENCES
Hauer K.E., E.S. Holmboe, and J.R. Kogan. 2011. “12 Tips for Implementing Tools for Direct Observation of Medical Trainees’ Clinical Skills During Patient Encounters.” *Medical Teacher* 33(1): 27-33. doi: 10.3109/0142159X.2010.507710

Kogan J.R., E.S. Holmboe, and K.R. Hauer. 2009. “Tools for Direct Observation and Assessment of Clinical Skills of Medical Trainees: A Systematic Review.” *JAMA* 302: 1316-26**.** doi: 10.1001/jama.2009.1365

Kogan, J.R., R. Hatala, K.E. Hauer, and E.S. Holmboe. 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. doi: 10.1007/s40037-017-0376-7

Direct Observation of Procedural Skills

Description

Discussing direct observation of procedural skills, refers to observing a resident or fellow performing a range of procedures, from lumbar puncture to central venous catheters to endoscopy to complex surgical procedures with patients. Each specialty has its own core set of procedures considered essential to that specialty. These core procedures are codified by the ACGME Review Committees and the various certification boards.

As for clinical skills, direct observation is essential to the assessment of procedural skills. Faculty 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 the Competencies of patient care and procedural skills and interpersonal and communication skills. Performing procedures, especially in the context of surgery, almost always involves a team. Direct observation by faculty members and others can also inform clinical reasoning and interprofessional teamwork competencies.

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, and other tools. Similar to assessment tools for clinical skills, high levels of reliability can be achieved when a sufficient number of assessments by multiple observers are performed. This cannot be overemphasized. 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 affect 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 judging. While faculty development cannot fix all challenges, training in both the procedural skills of interest and assessment can help.

Feasibility

Procedural skills can enable opportunities for direct observation as faculty members are often participating in the procedure in some manner (i.e., supervising, assisting). New smartphone assessment apps can help faculty members complete assessments more efficiently.

Acceptability

Acceptability among faculty members and learners depends heavily on program culture and frequency of observation. Program cultures that discourage direct observation, make each observation high stakes, lack faculty member buy-in, and 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 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 ask and seek direct observation of procedural skills achieve greater acceptability.

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 presenting a patient at morning report or morbidity and mortality conferences. When done well, direct observation can be very impactful in helping residents and fellows improve procedural skills and outcomes for patients.

SUGGESTED REFERENCES

Abdelsattar, J.M., Y.N. Aljamal, R.K. Ruparel, P.G. Rowse, S.F. Heller, and D.R. Farley. 2018. "Correlation of Objective Data with General Surgery Resident In-Training Evaluations and Operative Volumes." *Journal of Surgical Education* 75: 1430-6. doi: 10.1016/j.jsurg.2018.04.016

George, B.C., E.N. Teitelbaum, S.L. Meyerson, M.C. Schuller, D.A. DaRosa, E.R. Petrusa, L.C. Petito, and J.P. Fryer. 2014. "Reliability, Validity, and Feasibility of the Zwisch Scale for the Assessment of Intraoperative Performance." *Journal of Surgical Education* 71(6): e90-6. doi: 10.1016/j.jsurg.2014.06.018

George, B.C., J.D. Bohnen, R.C. Williams, et al. of the Procedural Learning and Safety Collaborative. 2017. "Readiness of US General Surgery Residents for Independent Practice." *Annals of Surgery* 266(4):582-594. doi: 10.1097/SLA.0000000000002414

George, B.C., J.D. Bohnen, M.C. Schuller, and J.P. Fryer. 2020. "Using Smartphones for Trainee Performance Assessment: A SIMPL Case Study." *Surgery* 167(6): 903-906. doi: 10.1016/j.surg.2019.09.011

Hatala, R., D.A. Cook, R. Brydges, and R. 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. doi: 10.1007/s10459-015-9593-1

Husk KE, Learman LA, Field C, Connolly A. 2020. "Implementation and Initial Construct Validity Evidence of a Tool, myTIPreport, for Interactive Workplace Feedback on ACGME Milestones". *Journal of Surgical Education*. Published online June 13, 2020. doi: 10.1016/j.jsurg.2020.05.002

Thanawala, R.M., J.L. Jesneck, and N.E. Seymour. 2019. "Education Management Platform Enables Delivery and Comparison of Multiple Evaluation Types." *Journal of Surgical Education* 76(6): e209-216. doi: 10.1016/j.jsurg.2019.08.017

Multisource Feedback (360˚ Feedback)

Description

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

Competencies

MSF assessments must be a core component of any program of assessment. It is essential for assessing the Competencies of 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 MSF in medicine and medical education. Studies of practicing physicians have found associations between MSF ratings and patient complaints and malpractice claims. Research supports the use of MSF for mostly formative purposes; when MSF 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 MSF. Data from MSF assessments can and should be used by the CCC to make judgments on the professionalism, interpersonal and communication skills, and systems-based practice Milestones, depending on who completes the MSF and what specific MSF tool is used. Three systematic reviews on MSF are provided in the references below.

Feasibility

MSF requires some effort and coordination to be used effectively. Programs 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 professions can assess professionalism and interpersonal teamwork skills. Ideally a program should try and use an existing tool with validity evidence, but if programs develop their own MSF instruments, they should evaluate the quality of the instrument using the Utility Index or Ottawa Criteria for good assessment.

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

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

SUGGESTED ASSESSMENT TOOL AVAILABLE FROM ACGME
The ACGME has launched the Teamwork Effectiveness Assessment Module (TEAM). The TEAM module is meant for use by individual residents and fellows to gather and interpret feedback from their 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 in fellows in the assessment of key competencies, such as interpersonal skills and communication and professionalism, and the milestones.

For example, the TEAM assessment module is particularly well suited for assessing the milestones of *Interprofessional and Team Communication* subcompetency, one of the subcompetencies for the general competency of interpersonal skills and communication. TEAM will also be helpful in assessing and providing feedback for *Professional Behavior and Ethical Principles* (general competency of professionalism); *Accountability and Conscientiousness* (general competency of professionalism); and *Physician Role on Healthcare Systems* (general competency of systems-based practice).

You can access information about the **TEAM** here: https://dl.acgme.org/pages/assessment

SUGGESTED REFERENCES

Chesluk B.J., S. Reddy, B. Hess, E. Bernabeo, L. Lynn, and E.S. Holmboe. 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. doi: 10.1002/chp.21267

Donnon, T., A. Al Ansari, S. Al Alawi, and C. Violato. 2014. "The Reliability, Validity, and Feasibility of Multisource Feedback Physician Assessment: A Systematic Review." *Academic Medicine* 89(3): 511-6. doi: 10.1097/ACM.0000000000000147

Frost, J.S., D.P. Hammer, L.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. doi: 10.1080/13561820.2018.1515733

Laggo, J., W.R. Berry, K. Miller, et al. 2019. "Multisource Evaluation of Surgeon Behavior is Associated with Malpractice Claims." *Annals of Surgery* 270: 84-90. doi: 10.1097/SLA.0000000000002742

Lockyer, J. 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

Stevens, S., J. Read, R. Baines, A. Chatterjee, and J. Archer. 2018. "Validation of Multisource Feedback in Assessing Medical Performance: A Systematic Review." *Journal of Continuing Education in the Health Professions.* 38(4): 262-268. doi: 10.1097/CEH.0000000000000219

Wilkerson, L. and V. Sigalov. N.D. "Multi-Source Feedback Tool for Interprofessional Collaborative Practice." UCLA David Geffen School of Medicine. https://apps.medsch.ucla.edu/ipe/docs/10A_multi-source_feedback_guide_FINAL.pdf

Patient Experience Surveys

Description

Surveys of patients to assess their experience with care during hospitalization, clinic or outpatient visits, and telehealth visits are now widely available. Patient experience surveys are essential if a program truly wants to assess 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. Survey questions often assess specific aspects of patients' interactions with the health care system, such as whether their questions were satisfactorily answered and whether they 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 physician's explanations, listening skills, and provision of information about examination findings, treatment steps, and drug side effects. A typical patient survey asks patients to rate their agreement with statements describing the care (e.g., "The doctor kept me waiting," --Yes, always; Yes, sometimes; or No, never or hardly ever) or to rate the physician on a numeric scale. These types of scales are becoming the most common. 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 (CAHPS)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 Competencies of patient care (aspects of data gathering, treatment, and management; counseling, and education; preventive care); interpersonal and communication skills; professionalism; and aspects of systems-based practice (patient advocacy; coordination of care). If survey items about specific physician behaviors are included, the results can be used for formative assessment and performance improvement. While patient survey results can be used for summative assessment, the numbers of responses needed for a reliable score are substantial and attribution issues can also be challenging, especially in the hospital setting (the physician may be held responsible for the entire experience when aspects not under the physician's control are being measured). Regardless of the reliability challenges, some type of patient experience surveys is essential for programmatic assessment.

Validity

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 CAHPS, 45 surveys are needed for a high degree of reliability. These numbers are very difficult to achieve for individual residents and fellows, and thus studies of reliability estimates for residents and fellows are lower. An older study using the

short American Board of Internal Medicine Patient Satisfaction Questionnaire reported that 20-40 patient responses were needed to obtain a reliability of 0.70 to 0.82 on individual resident's ratings. Despite the challenge in collecting enough surveys for higher-stakes purposes, patient experience surveys are essential for formative (feedback) purposes. Evidence has accrued that shows patient experience surveys correlate with various patient outcomes (a strong measure of validity) and graduates must learn how to effectively use patient experience data for professional development and practice improvement.

Feasibility

A variety of patient experience surveys are available from multiple sources. Programs 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 CAHPS can be obtained from the Agency for Healthcare Research and Quality (AHRQ); programs can access and download an extensive portfolio of surveys for different settings (https://www.ahrq.gov/cahps/index.html). 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., smart phone app, iPad, or easily accessible computer terminal) is 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 enough per-resident surveys to provide reproducible results; (3) the resources required to collect, aggregate, and report survey responses; and (4) assessment of the resident's contribution to a patient's care separate from that of the health care team. Because of these concerns, patient experience surveys are often conducted by the institution or by one or more clinical sites and reports specific to the residency/fellowship program may or may not be prepared.

Acceptability

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 MSF where only a few patients may be sampled. Some studies have found that patient feedback surveys do provide valuable information to programs. Programs 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.

Catalytic Effect

Patient experience surveys can definitely produce and support improvement and professional growth While not a typical psychometric approach, use of a narrative, clinimetric 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.

SUGGESTED REFERENCES

Agency for Healthcare Quality and Research. 2020. "About CAHPS." Page last reviewed March 2020. Accessed at https://www.ahrq.gov/cahps/about-cahps/index.html

Agency for Healthcare Quality and Research. 2017. "Fielding the CAHPS Clinician and Group Surveys." Document updated June 2017. Accessed at https://www.ahrq.gov/sites/default/files/wysiwyg/cahps/surveys-guidance/cg/survey3.0/fielding-the-survey-cg30-2033.pdf

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.

Hess, B.J., L.A. Lynn, L.N. Conforti, and E.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: 909-915. https://doi.org/10.1111/j.1532-5415.2011.03370.x

Price R.A., M.N. Elliott, A.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. doi: 10.1177/1077558714541480

Portfolio

Description
A portfolio is a collection of evidence intended to demonstrate an individual's learning journey over time. In GME, a portfolio can include documents including 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 (dossier portfolio); mandated collections of acquired skills and competencies, in a fixed format (training portfolio); purposeful collections of evidence for personal growth and development (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 GME community, the concept of the "comprehensive portfolio" has emerged. Comprehensive portfolios include content agreed upon by the resident or fellow and the program that is standardized and defensible, allowing for summative decision making when required, and documenting that a program graduate has attained the desired competence to practice unsupervised, safe, and effective patient care.

Key characteristics of a comprehensive portfolio include:

1) A multifaceted approach to assessment
2) Assessment based on "triangulation" –assessing multiple domains of competence and utilize multiple assessors
3) Longitudinal and iterative content established through the interaction of the learner and faculty assessor
4) Learner self-assessment and reflection
5) Evidence of meaningful learner engagement demonstrating professional growth
6) And portfolio development and use that is transparent to the learner – learners 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 practice-based learning and improvement and professionalism competencies, both of which are 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 portfolio content 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. Such assessment is dependent on the quality of the evidence provided in the portfolio and by the process used by the portfolio assessors (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 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 resident/fellow 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 responsibilities for their own learning.

SUGGESTED REFERENCES

Driessen, E. 2009. "Portfolio critics: do they have a point?" *Medical Teacher* (4): 279-281. doi: 10.1080/01421590902803104

Friedman, Ben, M. David, M.H. Davies, R.M. Harden et al. 2001. "AMEE Educational Guide 24; Portfolios as a Method of Student Assessment." *Medical Teacher* 23(6): 535-551.

Heeneman, S., and E.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. doi: 10.3205/zma001134

Mueller, P.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, P. S., C. Carraccio, and E.S. Holmboe. 2018. "Portfolios." In E.S. Holmboe, S.J. Durning, and R.E. Hawkins (Eds.), *Practical Guide to the Evaluation of Clinical Competence, Second Edition*. (pp 270-287). Philadelphia: Mosby Elsevier.

Van Tartwijk, J., and E. Driessen. 2009. "Portfolios for Assessment and Learning: AMEE Guide No. 45." *Medical Teacher* 31(9): 790-801. doi: 10.1080/01421590903139201

Implementation

Developing a robust system of programmatic assessment using the tools listed above is essential to the valid and reliable assessment of GME learners. 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. The table below provides specific assessment tools and methods that can be used to assess learners in each of the six Core Competencies.

Assessment Tool/Method	Targeted Competency
Faculty evaluation	Multiple competencies
Direct observation	Patient Care and Procedural Skills, Interpersonal and Communication Skills, and Medical Knowledge (“in vivo”)
Multisource feedback	Professionalism, Interpersonal and Communication Skills, and Systems-based Practice
Audit and performance data	Practice-based Learning and Improvement and Systems-based Practice
Simulation (if available)	Patient Care and Procedural Skills and Interpersonal and Communication Skills
IT exam (if available)	Medical Knowledge

Programs must know where and how frequently assessments are being used. Developing a tracking table can help ensure that residents/fellows 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

**The effectiveness of an assessment tool can be determined using a framework like the Ottawa Framework for Good Assessment (Norcini J 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 (takes too long) or is deemed too expensive (decreased Relative Value Unit (RVU) generation) may be unlikely to be successfully implemented (low effectiveness). Provide a judgment on the effectiveness of the proposed assessment tools. If the assessment tool scores low using this framework, identify potential faculty development activities that may enhance the tool’s effectiveness.*

Miller’s pyramid (Miller 1990) was mentioned earlier in this guidebook and 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 (the "Does" of Miller's pyramid), programmatic assessment should also investigate learning as appropriate across the full continuum of "Knows" to "Does." 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 faculty members 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 be shared with the learners undergoing assessment 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 by the CCC and program director, it is insufficient. Likewise, if assessment data do not directly inform decisions about the achievement of desired program outcomes, those assessments and data are "not fit for purpose" and should be either redesigned to address program outcomes or discontinued. Hauer and associates have identified six principles of programmatic assessment (table) that can help avoid this outcome and should be used by all programs 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 K.E., P.S. O'Sullivan, K. Fitzhenry, and C. Boscardin. 2018. "Translating Theory into Practice: Implementing a Program of Assessment." *Academic Medicine*. 93(3): 444-450. https://doi.org/10.1097/acm.0000000000001995

Schut, S., L.A. Maggio, S. Heeneman, J. van Tartwijk, C. van der Vleuten, and E. Driessen. 2020. "Where the Rubber Meets the Road — An Integrative Review of Programmatic Assessment in Health Care Professions Education." *Perspectives in Medical Education*. Published online October 21, 2020: 1-8. https://doi.org/10.1007/s40037-020-00625-w

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

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. Rather, they provide a roadmap for the interpretation of rotation-based assessment data (especially work-based assessments) that guides the synthetic judgements of the CCC. 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 (Milestones progression), 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 educational programs to ensure they continue to meet desired educational outcomes.

As the ACGME's current accreditation model (formerly referred to as "the Next Accreditation System" or "NAS") has matured, interest in entrustable professional activities (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 ten Cate 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.

SUGGESTED REFERENCES

Edgar, L., S. McLean, S.O. Hogan, S. Hamstra, and E.S. Holmboe. *The Milestones Guidebook*. *Version 2020*. Accessed July 27, 2020 at https://www.acgme.org/Portals/0/MilestonesGuidebook.pdf.

Edgar, L., S. Roberts, and E.S. Holmboe. 2018. "Milestones 2.0: A Step Forward." *Journal of Graduate Medical Education* 10(3):367-369. https://doi.org/10.4300/jgme-d-18-00372.1

Holmboe, E.S., K. Yamazaki, T.J. Nasca, S.J. Hamstra. 2020. "Longitudinal Milestones Data and Learning Analytics to Facilitate the Professional Development of Residents: Early Lessons from Three Specialties." *Academic Medicine* 95(1):97-103. https://doi.org/10.1097/acm.0000000000002899

ten Cate O. 2005. "Entrustability of Professional Activities and Competency-Based Training." *Medical Education* 39(12):1176-1177. https://doi.org/10.1111/j.1365-2929.2005.02341.x

GENERAL REFERENCES

Frank, J. R., L.S. Snell, O. ten Cate, E.S. Holmboe, C. Carraccio, S.R. Swing, and D. Dath. 2010. "Competency-Based Medical Education: Theory to Practice." *Medical Teacher* 32(8): 638–645. https://doi.org/10.3109/0142159x.2010.501190

Hauer, K.E., P.S. O'Sullivan, K. Fitzhenry, and C. Boscardin. 2018. "Translating Theory into Practice: Implementing a Program of Assessment." *Academic Medicine* 93(3): 444-450. https://doi.org/10.1097/acm.0000000000001995

McGaghie, W.C. 2015. "Mastery Learning: It Is Time for Medical Education to Join the 21st Century." *Academic Medicine* 90: 1438-1441. https://doi.org/10.1097/acm.0000000000000911

Miller, G.E. 1990. "The Assessment of Clinical Skills/Competence/Performance." *Academic Medicine* 65(9): S63-67. https://doi.org/10.1097/00001888-199009000-00045

Norcini, J.M., B. Anderson, V. Bollela, V. Burch, M. João Costa, R. Duvivier, R. Hays, M.F.P. Mackay, T. Roberts, and D. Swanson. 2018. "Consensus framework for good assessment." *Medical Teacher*. 40:11: 1102-1109. https://doi.org/10.1080/0142159x.2018.1500016

ten Cate O. 2005. "Entrustability of Professional Activities and Competency-Based Training." *Medical Education* 39(12): 1176-1177. https://doi.org/10.1111/j.1365-2929.2005.02341.x

van der Vleuten C.P., L.W. Schuwirth, E.W. Driessen, et al. 2012. "A Model for Programmatic Assessment Fit for Purpose." *Medical Teacher* 34(3): 205–214. https://doi.org/10.3109/0142159X.2012.652239

van der Vleuten, C.P.M. 1996. "The Assessment of Professional Competence: Developments, Research and Practical Implications." *Advances in Health Sciences Education* 1: 41-67. https://doi.org/10.1007/bf00596229

van Melle, E., J.R. Frank, E.S. Holmboe, D. Dagnone, D. Stockley, J. Sherbino for the International Competency-Based Medical Education Collaborators. 2019. "A Core Components Framework for Evaluating Implementation of Competency-Based Medical Education Programs." *Academic Medicine* 94(7): 1002-1009. https://doi.org/10.1097/acm.0000000000002743

GLOSSARY

Construct: The meaning of construct, in the context of medical education and measurement, is a tool or concept used to help define and facilitate understanding of human performance. For example, does an assessment tool used for direct observation include and target the proper domains or components of clinical skills such as history taking and physical examination?

Generalizability: Whether measurements (scores) derived from an assessment tool can be shown to apply to more than the sample of cases, clinical encounters, or test questions used in a specific assessment.

Reliability/Reproducibility: When measurements (scores) are repeated and the new assessment results are consistent with the first scores for the same assessment tool on the same or similar individuals for the same competencies measured. There are essentially three types of reliability:

- Consistency over assessors (inter-rater)
- Consistency over time (test-retest and intra-rater)
- Consistency over items (internal consistency, aka Cronbach's alpha)

Reliability is measured as a correlation with 1.0 being perfect reliability and below 0.50 considered as unreliable. Measurement reliabilities above 0.65 and preferably near or above 0.85 are recommended, especially when the assessment involves higher stakes.

Validity: A process of accumulating evidence about how well an assessment measures represent or predict a resident's ability or behavior. Validity refers to the specific measurements made with assessment tools in a specific situation with a specific group of individuals. It is the scores not the type of assessment tool that are valid. Validity is best viewed as an ongoing argument and process to continually gather the evidence across multiple aspects of validity. See the Messick model described earlier.

Formative Assessment: Assessment in which findings are accumulated from a variety of relevant assessments designed primarily for catalytic educational effects and personal improvement. Formative assessment is intended to provide specific, accurate assessment information and date to support constructive feedback and coaching to individual residents during their training.

Summative Assessment: Assessment in which findings and recommendations are designed to accumulate all relevant assessments for a high-stakes ("go/no-go"; "pass/fail") decisions. Summative assessment is used to decide whether the resident or fellow qualifies to continue to the next year in the educational program, should be dismissed from the program, or at the completion of the residency can be judged as ready for unsupervised practice and recommended for board certification. Of note, a clear distinction, or dichotomy, between formative and summative assessment is unhelpful. In reality, in programs of assessments the assessments and judgments will exist across a spectrum of stakes depending on the purpose of the assessment and developmental stage of the resident or fellow.

Appendix: Van Melle CBME Components Framework

Table 2. Van Melle Framework for Competency-Based Medical Education[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 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, E., J.R. Frank, E.S. Holmboe, D. Dagnone, D. Stockley, and J. Sherbino for the International Competency-Based Medical Education Collaborators. 2019. "A Core Components Framework for Evaluating Implementation of Competency-Based Medical Education Programs." *Academic Medicine* 94(7): 1002-1009. https://doi.org/10.1097/acm.0000000000002743