INTRODUCTION

Clinical skills remain fundamental to the practice of medicine and form a core component of the professional identity of the physician. However, evidence exists to suggest that the practice of some clinical skills is declining, particularly in the United States. A decline in practice of any skill can lead to a decline in its teaching and assessment, with further decline in practice as a result. Consequently, assessment not only drives learning of clinical skills, but their practice. This article summarizes contemporary approaches to clinical skills assessment that, if more widely adopted, could support the maintenance and reinvigoration of bedside clinical skills.

WHAT ARE CLINICAL SKILLS?

Clinical skills are typically regarded as the combination of:

- The gathering of clinical information by talk and touch (the history and physical examination)
- The interpretation and application of information gathered by these processes (diagnostic reasoning and clinical thinking)
The acquisition of clinical skills depends on learning how to perform certain motor skills (procedural knowledge), understanding why one should perform these skills (factual knowledge of basic medical sciences), and applying reasoning to interpret the findings from skills (conditional knowledge). This is shown schematically in Fig. 1.

In clinical practice, and in teaching and assessment, it is often difficult to dissociate one clinical skill from others. As such, the teaching and assessment of each often overlaps with that of others, and indeed must do so if validity is to be maximized and professional competence, rather than mere performance of an individual skill, measured. Because the teaching and assessment of practical procedures is usually regarded separately in the educational literature, it is not covered in this article.

WHAT ARE WE TRYING TO ASSESS?

The acquisition of competence, variously defined as the ability to undertake a specific task, or a component of activity that a professional must undertake within a task, is the aim of teaching and learning in medicine. Different assessment methods are suggested for different components of competence, as illustrated in Fig. 2.

In this model, assessing “does” is regarded as more authentic, and thus important, than assessing “shows how.” This has led to an emphasis on workplace assessments conducted while the learner actually “does” the tasks their role requires them to perform.

---

**Fig. 1.** Clinical skills: what they are and the knowledge that underpins their acquisition. *(Data from Michels ME, Evans DE, Blok GA. What is a clinical skill? Searching for order in chaos through a modified Delphi process. Med Teach 2012;34(8):e573–81.)*
Current thinking promotes Entrustable Professional Activities as a practicable means of ensuring the professional competence of learners. In this method, multiple informing competencies are combined into broader areas of professional activity, and assessment based at the level of the activity rather than the competence. For example, an expected professional activity for an internal medicine resident is the assessment of a new referral in an out-patient clinic. Successful delivery of this activity depends on many competencies, including the ability to take a history, undertake a physical examination, and communicate information to the patient. In basing assessment at the level of the activity, it is assumed that successful performance of the activity means that each informing competency has also been successfully attained.

This model may therefore be best suited to those nearing certification. At other stages of education, greater focus may be required on individual informing competencies; for example, the method of physical examination should be specifically assessed in the prelicensing medical student. Arguably, it should not be assumed that these specific informing competencies will be maintained into graduate medical education (GME); the Entrustable Professional Activities method may incorrectly assume that they are. Suggested differences in focus at different stages of medical education are shown in Fig. 3.

**WHAT IS THE PURPOSE OF ASSESSMENT?**

Assessment has a variety of functions in medical education, some of which are summarized in Box 1. The purpose of assessment should dictate the type of assessment chosen.

**WHAT FACTORS INFORM THE CHOICE OF ASSESSMENT USED?**

A key first step in the design of any assessment is the consideration of its purpose; its form follows its function. That is, an assessment that must primarily be developmental takes one form, and one that must set an objective standard to be attained takes another.
Other considerations must be applied when considering the format of an assessment. Reliability, or reproducibility, is traditionally believed to be more important in summative than formative settings. Validity, which may be regarded as how closely the assessment measures what it is intended to measure, is said to be easier to achieve in workplace-based settings. Practicality is an additional factor; assessments using real patients may not be feasible for large cohorts of learners. Other logistic factors also influence choice, such as “What space is available for the assessment? and “How many faculty are willing and able to be involved?” The overall utility of an assessment has been expressed in conceptual form, as shown in Fig. 4.

In addition to psychometric and practical characteristics of each assessment method, the educational impact, and, in the opinion of this author, the public message that the assessment provides, should be considered. Specifically, high stakes assessments of clinical skills send a clear signal to the public that these skills are taken seriously by the profession.

**Box 1**

**Some purposes of assessment in medical education**

- Ensure minimum standards of knowledge or skills
- Rank or grade trainees for selection or other purposes
- Inform licensing, certification, or other key progression decisions
- Provide a forum for direct observation of clinical skills
- Provide feedback to the learner on performance of clinical skills
- Provide feedback to the teacher on the progression of the learner
- Provide feedback to the teacher on the impact of a teaching program

---

**Fig. 3.** From novice to expert: shifting the focus of assessment of clinical skills over time. PE, physical examination.
Assessments for learning, or formative assessments, are developmental, with the primary intention of providing feedback to the learner, thus aiding the acquisition of new skills or knowledge. Assessments of learning, or summative assessments, are designed to ensure attainment of a specific standard at a specific point in time. The relative advantages and disadvantages of formative workplace-based and high stakes assessments of clinical skills are summarized in Table 1.

Development of the learner and the acquisition of competence are critically dependent on direct observation and feedback delivered as close to the learning event as is possible. This, coupled with the view that authenticity of assessment is greatest when the assessment is structured around the performance of a real clinical activity, has given emphasis to workplace assessment in recent years.

However, it has proven difficult to ensure that faculty has the time and expertise to deliver workplace assessments to a sufficient standard and frequency. Some observers also point to the lack of independence of workplace assessors and the
increasing complexity of systems that prove challenging for faculty to deliver. Furthermore, although learners appreciate observation and feedback, they may find the experience intimidating if it is to be used to inform their progression. In UK GME, workplace assessments have been retitled “supervised learning events,” to minimize the sense that they are in any way summative. In addition, it is always necessary to make progression decisions by some method, and it has proven difficult to translate ratings derived from global workplace assessments of clinical skills into reliable pass/fail classifications or grades.

Given these issues, most GME systems around the world continue to use summative high stakes clinical skills examinations to inform key progression decisions in medicine, such as those relating to licensing and certification, in combination with the information derived from workplace-based assessments. No specific combination of formative and summative methods has been found to be more successful in ensuring the attainment of any specific level of clinical skill. In most educational programs, a variety of assessment methods are used, to maximize the information known about each learner, and triangulate assessment between multiple assessors and complementary methods, within the overall concept of “programmatic assessment.”

### THE UNITED STATES APPROACH

US GME assessment in internal medicine is unusual in global educational terms, in that assessment of clinical skills after the point of licensing is wholly based on assessments of fundamentally formative style, delivered in the learner’s workplace. Summative assessments the Boards are restricted to assessments of knowledge,
which can assess aspects of clinical thinking, and theoretic knowledge underpinning clinical skills, but clearly cannot assess the actual practice of these skills. This approach contrasts starkly with other GME systems around the world, including in Australia, Canada, Ireland, and the United Kingdom, that require success in a high stakes summative assessment of clinical skills as a mandatory component of progression, and thus certification, in addition to completion of workplace-based assessments.

Critics of high stakes clinical skills examinations point to their potential cost, low validity (particularly if simulation is used), the difficulties of ensuring sufficient interrater reliability to support a pass/fail decision, and learner stress as justifications for their exclusion from assessment systems. However, the educational impact of these examinations on learners and faculty is substantial, and the consequences of their absence from an assessment system must be taken into as much consideration as would the absence of high stakes assessments of knowledge.

**SOME SPECIFIC WORKPLACE ASSESSMENTS OF CLINICAL SKILLS**

**Mini-CEX**

The mini-CEX evolved from the CEX, which was promoted as an alternative to a summative clinical skills examination when the American Board of Internal Medicine dropped an oral clinical skills examination in 1972. The CEX was time consuming, and suffered from the same problems as similar assessments delivered in a summative setting, with case specificity and assessor inconsistency foremost. The mini-CEX is shorter (15–20 minutes) and theoretically permits sampling of a wider range of skills across a range of encounters, by different faculty members. Time for feedback is increased, construct and criterion validity high, and derived judgements more reliable.

**Case-Based Discussion**

This is a form of chart-stimulated oral discussion, which is used to assess diagnostic reasoning and management planning, but does not typically involve direct observation of the learner interacting with the patient. Broader professional bedside attributes, such as behavior in teams and professional attitude, may be assessed by peer or colleague review in “Multi-source Feedback” but are beyond the scope of this article.

**SUMMATIVE ASSESSMENTS OF CLINICAL SKILLS**

**Traditional Models**

Assessments of bedside clinical skills should ideally assess information gathering, synthesis, analysis and application of gathered information, and its communication. Traditional summative clinical skills examinations attempted to cover these domains in the “short case–long case–viva” model.

**The short cases**

In this format, candidates are asked to examine one part of the patient (eg, the hands or facies) to describe what they see or find and suggest the cause, directly observed and in discussion with one or two examiners. A short introductory statement is provided, typically in the form of a one line history: “This patient has painful hands.” No communication with the patient is typically permitted.

The ability to elicit and interpret signs, diagnostic reasoning, and clinical thinking is assessed and, given the brevity of each encounter, sampling across different organ systems or conditions can occur. However, some of what is assessed is more
consistently assessed with photographic or video material. Validity is clearly impaired if the candidate is not allowed to take a history, and detailed guidance may be necessary to standardize examiner-candidate interaction, and questions asked. Despite these limitations, this model is still used in many assessment settings.

**The long case**
The candidate spends up to an hour with a patient, typically unobserved. They then present the case to one or more examiners, and may discuss aspects of differential diagnosis, and clinical thinking relating to investigation and management. The candidate may be taken back to the bedside to demonstrate physical signs. This format has good face validity and contains elements of the “case-based discussion” or chart stimulated recall formats now used in workplace assessments. However, critics of its use in a summative setting point to the fact that style of communication, attitudes and behaviors with the patient, and method of physical examination are not directly observed and cannot be assessed. Standardization of the examiner-candidate interaction is also seen as a limitation, as is standardization of the clinical content, because different patients may present markedly different challenges.

The Objective Structured Long Examination Record was introduced as a means of attempting to standardize the long case assessment and improve its objectivity, validity, and reliability. The 10-point checklist helped examiners to structure the areas of assessment and was conceptually important, but gained little practical support and is not now widely used.16

**The viva**
The traditional viva or oral examination is a means of assessing applied clinical knowledge, particularly diagnostic reasoning, knowledge of methods of investigation, or guidelines for treatment. Critics cite inconsistency in the examiner-candidate interaction, the long testing time required to produce adequate reliability, and the fact that much of what is assessed is assessed more effectively in written examination settings, as justifications for exclusion from assessment programs. Despite this, viva examinations persist in many international medical assessment settings.

**Newer Models**
Newer models of summative clinical skills assessment attempt to find a balance between validity, reliability, and practicality.12 No model is perfect, and inclusion of a summative skills examination in an assessment system is inherently associated with acceptance of a compromise between these different elements.

Reliability is enhanced by standardization of content and interaction between candidate and examiner. Emphasis is placed on sampling across and between skills and across multiple diseases or presentations in a series of encounters. Direct observation of candidate-patient interactions, usually by an expert clinician examiner, is regarded as important. Most formats retain some examiner-candidate interaction, but in some, most notably the USMLE Clinical Skills Step 2 examination, where the simulated patient also takes the role of examiner, there is no interaction other than that demanded by the simulated clinical encounter.

Assessment theory also suggests that interactions between candidate and patient should be as standardized as possible. In examinations with many candidates the consequence of this thinking is that standardized or simulated patients are used.
Although this may not affect the assessment of method of physical examination, it greatly limits the assessment of the ability to elicit abnormal physical signs and is believed to reduce validity.

**The objective structured clinical examination**

The Objective Structured Clinical Examination\(^{17}\) was developed as an alternative to traditional models of clinical examinations and typically takes the form of a circuit of stations, each of one or more clinical encounter or linked clinical task (e.g., writing a prescription for a patient assessed in the previous encounter), some of which is observed by at least one examiner, who is usually a more senior clinician.

Uncertainty and debate persist about the optimal number of stations, the length of the stations, the number of examiners, whether all tasks are observed, and relative benefits of real and simulated patients.\(^{18}\) In practice, these psychometric and academic considerations must be balanced against the resources available (i.e., the availability of time, space, faculty to act as examiners, and patients to participate).

Key considerations in the construction of a summative clinical skills examination are summarized in **Box 2**.

---

**Box 2**

**Some factors to consider in the design of a summative assessment of clinical skills**

- Will it use real or simulated patients/mannequins, or both?
- Will the candidate-patient interaction be observed or unobserved?
- If observed, will the examiner be a clinician or a trained lay person, or both, and how many examiners will there be?
- What interaction between candidates and examiners will be permitted?
- Will marking be based on an itemized checklist or global rating scale?
- What domains of performance will be defined and assessed?
- Will the pass standard permit compensation between different domains, or must each be passed separately?
- Will clinical challenges be based on a system (e.g., cardiovascular) or a symptom or presentation?
- What level of interaction between the patient and candidate is permitted (physical examination stations)?
- How many stations/encounters can be accommodated, and are necessary?
- How long will each encounter be?
- Will some encounters include data or image interpretation, with or without a patient present?
- How will the degree of difficulty of each encounter be graded and/or compared and/or standardized?
- Will there be any sequential or “linked” encounters, with tasks that follow from an interaction at a previous encounter?
- Will some encounters assess specific clinical skills only (e.g., physical examination) or will all attempt to integrate all clinical skills?
- How will the pass standard be set?
- How will feedback be provided to candidates? (pass/fail classification only: Score and Rank: Examiner comments)
SPECIFIC SUMMATIVE CLINICAL SKILLS EXAMINATIONS OF INTEREST

USMLE Step 2 Clinical Skills

The USMLE Step 2 CS examination is the largest graduation or licensing level examination in the world, with more than 30,000 candidates per annum examined in five centers. It is highly standardized, and of good published reliability, but critics point to the fact that interactions are not observed, that assessment of the interaction is by a trained simulated patient rather than a clinician expert, that no real patients and thus real physical signs can be included, that feedback is limited, and that costs for candidates are high. A noncompensatory three-domain marking structure is used, in which performance is assessed in the Integrated Clinical Encounter, Spoken English Proficiency, and Communication and Interpersonal Skills, with a separate pass standard required in each.

Running in parallel with final examinations in US medical schools, its prime purpose is to ensure, with a high degree of reliability, that a minimum standard of clinical practice is attained by US and international medical graduates. In achieving that aim, it remains the prime example of a successful, high-volume, summative clinical skills examination.

It is noteworthy that this is the last summative assessment of bedside clinical skills that most practitioners of internal medicine in the United States undergo. For some, it may even be the last time that the basic clinical skills of physical examination and history taking are ever directly observed, albeit by a trained simulated patient, rather than a faculty member or other expert clinician examiner.

The MRCP(UK) Practical Assessment of Clinical Examination Skills Examination

The Practical Assessment of Clinical Examination Skills (PACES) is the largest summative international GME examination of clinical skills in the world, sat by around 5000 doctors annually. It is the third part of the MRCP(UK) Diploma examination, the first two parts being knowledge-based examinations in single best answer, best of five options format, with content and standard like that of the American Board of Internal Medicine Certification examinations. Successful completion is mandatory for certification in internal medicine.

PACES provides a portable objective structured clinical examination–based model in which candidates rotate around five stations each of 20 minutes, each station comprised of one to two patient encounters (Fig. 5).

A total of eight patients are seen in these encounters, of which a minimum of four are real patients with real physical findings. All candidate-patient interactions are directly observed by two examiners, such that a total of 10 examiners have observed and assessed the candidates over the duration of the examination. Examiners independently mark seven defined domains of clinical skills (Table 2).

Candidates are assessed using a structured domain-based marking system, which is a compromise between an itemized checklist and global rating scale, these two methods having attracted considerable academic interest. Limited interaction between one of the two examiners and the candidate occurs at each encounter, but most of the assessment is based on observation of the candidates. Because real patients participate, a specific method of minimizing variation between examiner assessment is used, entitled calibration. In this process, pairs of examiners personally evaluate the clinical findings of each patient before the examination starts, to establish the findings; grade their difficulty; and agree what methods, findings, and applied knowledge are required to justify the award of a satisfactory grade for those domains.
The examination is resource and time intensive; only five candidates are assessed by 10 examiners, using up to six real and four simulated patients, over 125 minutes. Reliability, which is difficult to calculate in an examination with eight separate pass standards, meets the UK medical regulator, the General Medical Council, requirement for a high stakes examination. PACES includes encounters that integrate history and physical examination and in doing so with real patients enhances validity, but at the expense of standardization and hence reliability.

**The Place of Simulation**

Practical considerations relating to the number of candidates requiring assessment have led to widespread use of surrogates and actors as simulated patients, typically performing to a strictly defined script, in many examinations. Simulation of this sort improves standardization of content, but may compromise validity, particularly if mannequins or computer-based avatars are used as representations of patients. This increase of the use of simulation, to aid the logistics of delivery and improve standardization, should not be interpreted as overall evidence of superiority over examinations in which clinical content is based on real patients, and it remains to be seen
whether the apparent benefits of simulation translate into improved core bedside clinical skills, and overall clinical competence, when practiced on real patients.25

**SUMMARY AND FUTURE CONSIDERATIONS**

The quality of assessment of clinical skills influences the quality of their practice, and vice versa. If practice is to be sustained or reinvigorated, frequent, structured assessment must occur, particularly in GME. Although psychometric considerations relating to reliability and standardization cannot be entirely overlooked, they should not become an absolute barrier to the occurrence of some form of assessment, because the educational impact of the absence of any assessment of clinical skills is likely to be significant. At the GME level, direct observation by faculty, with clinical encounters that include real patients with real physical findings, is highly preferable.26 Whether the assessment is primarily summative or formative is of secondary importance, and hybrid models should be considered. In short, the most important thing about a clinical skills assessment is that it happens—everything else is detail.

**REFERENCES**