2011; 33: 206–214 MEDICAL TEACHER

Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 Conference

JOHN NORCINI[1], BROWNELL ANDERSON[2], VALDES BOLLELA[3], VANESSA BURCH[4], MANUEL JOÃO COSTA[5], ROBBERT DUVIVIER[6], ROBERT GALBRAITH[7], RICHARD HAYS[8], ATHOL KENT[9], VANESSA PERROTT[10] & TRUDIE ROBERTS[11]

[1]FAIMER, USA, [2]AAMC, USA, [3]Universidade Cidade de São Paulo, Brazil, [4]University of Cape Town and Groote Schuur Hospital, South Africa, [5]University of Minho, Portugal, [6]Maastricht University, The Netherlands, [7]National Board of Medical Examiners, USA, [8]Keele University, UK, [9]University of Cape Town, South Africa, [10]University of Cape Town, South Africa, [11]University of Leeds, UK

Abstract

In this article, we outline criteria for good assessment that include: (1) validity or coherence, (2) reproducibility or consistency, (3) equivalence, (4) feasibility, (5) educational effect, (6) catalytic effect, and (7) acceptability. Many of the criteria have been described before and we continue to support their importance here. However, we place particular emphasis on the catalytic effect of the assessment, which is whether the assessment provides results and feedback in a fashion that creates, enhances, and supports education. These criteria do not apply equally well to all situations. Consequently, we discuss how the purpose of the test (summative versus formative) and the perspectives of stakeholders (examinees, patients, teachers-educational institutions, healthcare system, and regulators) influence the importance of the criteria. Finally, we offer a series of practice points as well as next steps that should be taken with the criteria. Specifically, we recommend that the criteria be expanded or modified to take account of: (1) the perspectives of patients and the public, (2) the intimate relationship between assessment, feedback, and continued learning, (3) systems of assessment, and (4) accreditation systems.

Context

Definitions

Assessment involves testing, measuring, collecting, and combining information, and providing feedback.

Criteria provide the basis and the framework for judgments or decisions.

It is clear that assessment has played and continues to play a central role in medical education. The importance given to the characteristics of a good assessment varies, depending on whether you are being assessed, doing the assessment, or relying on the results. In each case, meeting established criteria for good assessment is critical to both value and credibility for all stakeholders.

Assessment in medical education is multifaceted. It drives and stimulates learning, provides information on educational efficacy to institutions and teachers, and protects patients. For example, examinees need to know what is expected of them and they also need to receive feedback that helps them improve. Those who assess – often teachers and teaching institutions – must ensure that learners are making progress, guarantee that programs are consistent with their mission, and meet the requirements of society and accrediting bodies. Ultimately, patients and society place strong emphasis on summative testing and on assessment programs because

Practice points

The criteria for good assessment outlined above are intended to act as a set of overarching principles. From them, a series of practice points can be derived that might provide useful guidance to various stakeholders. Some of these practice points follow.

Examinees

- Examinees should know the purpose of the assessments they take.
- Examinees should be assured of the quality of assessments they take.
- Examinees should receive feedback that fosters ongoing learning.
- Examinees should participate actively in receiving and acting on feedback.
- Examinees should be informed in a timely fashion about the scoring and standard-setting process.
- . . .

Patients

- Patients should be included as assessors when that role is consistent with their expertise (e.g., communication skills).
- Patients should contribute to improving understanding of facets of competence and performance.

Correspondence: J. Norcini, FAIMER, 3624 Market Street, 4th Floor, Philadelphia, PA 19104, USA. Tel: 215-823-2170; email: jnorcini@faimer.org

ISSN 0142–159X print/ISSN 1466–187X online/11/030206–9 © 2011 Informa UK Ltd.
DOI: 10.3109/0142159X.2011.551559

- Patients should be assured of the quality of assessments trainees take.
- Patients should be included as educators when, within the scope of their expertise, they can contribute to the educational effects of assessments.
- ...

Teachers
- Teachers should design their assessments in ways that maximize examinee learning.
- Teachers should address learning objectives in their teaching.
- Teachers should use assessment results to improve the quality of future learning.
- ...

Educational institutions
- Educational institutions should provide training in assessment for faculty.
- Educational institutions should allocate resources (clinical staff) to ensure assessment is done well.
- Educational institutions should analyze the quality of their assessments as part of processes for monitoring the quality of their teaching.
- Educational institutions should ensure that their curricula are consistent with their assessments.
- ...

Healthcare systems
- Healthcare systems should offer opportunities for ongoing formative assessment.
- Healthcare systems should facilitate a culture of encouraging response to formative assessment.
- Healthcare systems should promote research in assessment in workplace settings.
- ...

Regulators
- Regulators should take account of the educational effects of their assessments.
- Regulators should offer assessments which ensure ongoing competence.
- Regulators should recognize the catalytic effects of assessment on the education and healthcare systems.
- ...

they provide assurance that graduates have met minimum standards and are "fit for purpose". Assessment criteria are necessary to ensure that the results generated are of sufficient quality to meet the needs of each of these and other stakeholders.

No matter the perspective, the dictionary definition carries two distinct meanings to the verb "to test" (Crossley et al. 2002). One is to discover the worth of something by trial, with the purpose of obtaining more information about the object of assessment. The other is to improve the quality of something by trial (i.e., the impact of assessment). These two meanings are central to understanding the importance of assessment, its applications, and to identifying the criteria for good assessment.

In the remainder of this section we provide a historical perspective and argue for the importance of defining criteria for good assessment. In the sections that follow, we identify the current issues, present a set of criteria, make recommendations for how to proceed, and offer a series of practice points.

Historical perspective

Assessment has been part of various societies for more than 2000 years (Gipps 1999). Measurement of knowledge and/or performance for the purposes of selection has been its most pervasive role throughout time. The earliest records of assessment date back to the Han dynasty in China (206 BC to 220 AD) where candidates were selected for government service. The practice of medicine in medieval Islam required competence testing and by the seventeenth century Jesuit priests were using competitive examination for entry into their schools, possibly influenced by the missionaries who had traveled to China.

With regard to medical education, the first step toward the development of formal assessments was the introduction of examinations during an internship in Viennese and French medical schools. From 1788, entry to these internships in Paris was decreed to be by competition in the form of written and oral examinations (Lesky 1970; Poynter 1970). Exit level examinations for medical students were subsequently introduced in Britain in the 1850s at Oxford and Cambridge universities. By 1861, such examinations became a statutory national requirement stipulated by the General Medical Council established in Britain in 1858. This practice rapidly spread throughout medical schools in Europe in the latter part of the nineteenth century.

Across the Atlantic, in the USA the situation was quite different. During the 1800s there had been a proliferation of "medical colleges" both privately and publicly funded, in which the standards of teaching, training, and assessment varied widely as described in the report authored by Abraham Flexner (Flexner 1910). This report subsequently revolutionized medical education in the USA and by 1912, a group of licensing boards formed the Federation of State Medical Boards which agreed to base their practice on academic standards (criteria) as determined by the American Medical Association's Council on Medical Education (Kassebaum 1992). (Flexner (1912) also authored a less influential report published in 1912 about medical education in Europe, England, and Scotland.) By the 1930s, medical training in the USA had been standardized and colleges offered laboratory-based and hospital-based training with exit examinations (Starr 1982).

Over the past 50 years, there have been at least four major developments relevant to the assessment of undergraduate medical students and postgraduate trainees worldwide. These are the:

- development of a wide range of assessment tools, directed to different dimensions of medical competency,
- development and application of new teaching and learning approaches,

- increased sophistication of psychometrics and its application to individual assessment tools and results, and
- growing role of the computer as an integral part of assessments (Norcini 2005)

Until the middle of the twentieth century, medical school examinations relied heavily on the use of essays and oral examinations and the standards for passing were subjective. Recognition of the arbitrary nature of such examinations and their poor reliability led to the development of a large array of psychometrically robust assessment tools over the past 50 years. These include multiple choice questions (best option or extended matching item formats) and a range of modalities assessing performance both in an examination setting (objective structured clinical examination; OSCE, directly observed clinical encounter examination) as well as in the workplace (mini-CEX, clinical encounter cards) (Case & Swanson 1996; Norcini & Burch 2007; Kogan et al. 2009).

These developments have been driven by a few criteria:

- the assessments need to be reproducible (reliable), valid, feasible, fair, and beneficial to learning (van der Vleuten 1996),
- the content and form of assessments need to be aligned with their purpose and desired outcomes,
- broad sampling is needed to achieve an accurate representation of ability since examinee performance is case or content specific (multiple biopsies),
- systematically derived pass–fail scores and the overall reliability of an assessment are important, and
- assessments need to be constructed according to clearly defined standards and derived using systematic and credible methods.

The importance of defining criteria for good assessment

Stakeholders

A number of different stakeholders are involved with or affected by assessments and their results. Stakeholders include the patients, general public, healthcare employers, professional and regulatory bodies, universities, medical schools, training organizations, individual teachers, and, finally and equally important, the examinees themselves (Amin et al. 2006). The stakeholders make different uses of even the same assessments and, not surprisingly, have somewhat different priorities when it comes to the importance of various criteria against which those assessments should be judged.

Students come from a specific socio-cultural context, which affects their learning, and they have their development shaped by assessment (Vygotsky 1978). If successful with these ongoing assessments, the student gradually adopts new roles within society such as healer, counselor, or scientist (Downie & Calman 1987; Rees & Jolly 1998). Further, Boud (2000), has proposed that assessment is a key feature of lifelong learning. Rushton (2005) supports this perspective, stating "(it) equips students with the preparation required to continue independent assessment of their future learning experiences".

The various teaching and learning institutions have a slightly different perspective, from students, on assessment. The vision of the institution – for example its commitment to community-based education – can be supported and grown through assessment and feedback from the students (which is simply another form of assessment). At the same time, assessment both focuses the learner's attention on what is considered core knowledge and influences the content of the undergraduate curriculum. Skills are assessed and attitudes formulated by the feedback assessment provides. The assessment process must be carried out in such a way that only competent and skilled health practitioners emerge.

Regulatory bodies have a critical role in ensuring good assessment since they serve as gatekeepers for patients, the general public, and employers. Assessment in this context is closely linked with the maintenance of professional standards and with accountability – both to the individual and to society – which reinforces the need to have clear criteria for good assessment. In the end, the public entrusts itself to individual doctors based on the belief that the assessment process has been carried out in such a way that all are competent and skilled health practitioners.

Good criteria for assessment are important not only to improve quality but also to avoid unintended effects. Newble (1998) described how a mismatch between assessment and curriculum reform resulted in undesirable effects on student behavior. As part of curricular reform, he describes how didactic teaching was replaced with ward-based teaching. However, as the year progressed students were seldom seen on the wards, didactic teaching was increasingly requested and more time was devoted to book learning. The reason for this was that the assessment methods did not match the curricular reform but favoured the former style of didactic learning. Thus, at an institutional level, the assessment methodology was undermining the institutional mission and the goal of the educational program (Trigwell 2001). This example highlights the importance of aligning the assessment with educational practice.

Learning and teaching

Many well-known adages emphasize the central role of assessment in the educational process such as "Assessment is the tail that wags the dog" or Miller's (1990) assertion that "Assessment drives learning" and Ben-David's (2000) view that "Assessment expands professional horizons". These fundamental tenets are central to understanding the role of assessment and its application to teaching and learning. As Gipps (1999) points out, it is inadequate to conceive of assessment as measurement alone. In order for it to achieve its two goals – that of discovering worth as well as improving quality – the assessment of learning is critical (Arnold 2002). Institutions and educators have moved from viewing assessment as only a tool for *accountability* to viewing it as a method for *improvement* as well (Colliver 2002; Cottrell 2006). The emphasis is on the need for the robust assessment of learning and the development of a theory to support it. This is still a work in progress; as Norman and Schmidt (1999) note: "When educators do make reference to theory, it is more

frequently used the same way as a drunkard uses a light post – more for support than for illumination".

It would be a mistake to recognize the importance of assessment and yet not to connect it with the scholarship of teaching and learning (Shepard 2000). Simply stated, it implies that those who are responsible for assessment can improve it by both taking account of the research literature and conducting research when needed. The scholarship of teaching is not a new concept, but was highlighted as one out of four types of scholarship by The Boyer Commission (Boyer 1990). Trigwell's model demonstrates the growth from excellent teachers to scholars of education by the application of the scholarship of teaching and learning (Trigwell et al. 2000). Unfortunately, along with the scholarship of integration, the scholarship of teaching and learning is still not as highly valued (in financial and other terms) as the well-recognized scholarships of discovery and clinical practice (Curry 2002). Scholarship would also be a way to combat the tendency, in some institutions, to base the practice of assessment and teaching on intuition rather than evidence.

Current issues in criteria for good assessment

The state of the art of assessment may be organized into three categories:

- Areas where practice is consistent with the evidence: Assessment situations where there is evidence that informs practice and where practice is generally consistent with that evidence.
- Areas where practice is not yet consistent with the evidence: Assessment situations where there is evidence but it is generally ignored in practice (e.g., where there are issues of feasibility).
- Areas where there is a lack of evidence: Assessment situations that are not informed by the evidence (i.e., research is needed).

Aspects of any particular assessment fall into one of these three categories and no assessment falls exclusively into only one. Despite the fact that there is a mix, criteria for certain assessments are further developed than others.

Category 1: Practice is consistent with the evidence

Written examinations. The assessment of knowledge, synthesis, and judgment through multiple choice questions, essays, and similar formats falls predominantly into the first category. The criteria for the assessments in this category are generally well established and accepted. There is a sizeable evidence base and, where reasonable resources are available, their application in high stakes (local, national, and regional examinations) and low stakes settings, is typically consistent with the evidence. There remain areas where practice is inconsistent with the evidence (category 2) primarily due to feasibility issues such as test security, test development, and test/item review. And, there is a continued need to develop the evidence base in areas such as score aggregation and standard setting (category 3).

Objective structured clinical examination. Assessment of clinical skills using the OSCE is included in this category. Over the past 30 years, an extensive body of research about the reliability, feasibility, and validity of the OSCE and the use of standardized patients has been developed. The OSCE format has been applied in a variety of high and low stakes situations in a fashion consistent with the evidence (category 1). Issues that require resolution for the application of OSCEs to be more consistent with evidence from the research include case development, standardized patient training, and security of the assessment (category 2). Additional research is needed to improve the evidence around scoring and standard setting (category 3).

Category 2: Practice is not yet consistent with the evidence

Simulation. Advances in technology have led to the development of simulations that recreate, with varying degrees of fidelity, aspects of the practice of medicine. Research done over the past few decades is very supportive of the use of this technology in assessment and broad guidance is available for its successful deployment in a variety of different situations (category 1). The main impediment to the general application of simulation relates to its feasibility. Specifically, the devices are expensive, they may require the creation of a dedicated facility (simulation center), and the development of good testing material can be resource intense (category 2). In addition to these issues, research is needed to provide guidance on a variety of issues including scoring and assessment situations that profit from high fidelity simulation (category 3).

Workplace-based assessment that supports clinical training. In recent years, there has been an increasing emphasis on directly observed formative assessment that supports clinical training. Preliminary research is generally supportive and the literature provides broad guidance on issues such as the number of assessors and encounters needed for various purposes. Feasibility (category 2) is the major obstacle to its implementation and, in particular, it is difficult for clinical faculty to find time to perform a sufficient number of assessments. Additional research (category 3) is needed as well, especially to develop guidance and training for faculty on how to effectively score the encounters and provide feedback.

Category 3: Lack of evidence

Assessment of work. With the growing public interest in doctor accountability and the implementation of continuous quality improvement (CQI) processes in the healthcare system, there is a need to assess the actual, unobserved performance of doctors at work. Included in any assessment of practice performance are both patient outcomes (e.g., mortality, morbidity, patient satisfaction) and the process of care (immunizations, monitoring HbA1c in diabetics).

Considerable research is needed to determine which aspects of patient care are most appropriate (i.e., those for which the doctor is directly responsible), the number of

patients needed to produce reliable results, and means for adjusting the outcomes and processes for case mix and patient complexity (category 3). Feasibility and acceptability are major issues for most of the available measures since they require continuous access to accurate patient records (category 2). Finally, there are a few measures, such as patient satisfaction measures, for which there is good evidence and that are feasible (category 1).

Assessment of newer competencies. The recent shift of focus from the process of education to the required outcomes, along with changes in societies' expectations of doctors, has led to an increased emphasis on a range of newer competencies. There are several schemes for describing the major domains of proficiency (e.g., Accreditation Council for Graduate Medical Education; ACGME, Good Medical Practice, CANMEDs) and, for example, the ACGME competencies are medical knowledge, patient care, communication skills, professionalism, systems-based practice, and practice-based learning and improvement. Each competency is defined as follows:

- Medical knowledge: Demonstrate knowledge of established and evolving biomedical, clinical, epidemiological, and social-behavioral sciences, as well as the application of this knowledge to patient care.
- Patient care: The ability to provide patient care that is compassionate, appropriate, and effective for the treatment of health problems and the promotion of health.
- Practice-based learning and improvement: The ability to investigate and evaluate the care of patients, to appraise and assimilate scientific evidence, and to continuously improve patient care based on constant self-evaluation and lifelong learning.
- Interpersonal and communication skills: Demonstrate interpersonal and communication skills that result in the effective exchange of information and collaboration with patients, their families, and health professionals.
- Professionalism: Demonstrate a commitment to professional responsibilities and an adherence to ethical principles. Demonstrate:
  - compassion, integrity, and respect for others;
  - responsiveness to patient needs that supersedes self-interest;
  - respect for patient privacy and autonomy; and
  - accountability to patients, society, and the profession.
- Systems-based practice: Demonstrate an awareness of and responsiveness to the larger context and system of healthcare, as well as the ability to call effectively on other resources in the system to provide optimal healthcare.

These competencies embody the concepts of patient-centeredness, attitudes, values, teamwork, interprofessional collaboration, etc., and they can be thought of as a three-dimensional framework for structuring an assessment system. Along the first dimension are the competencies that need to be assessed, along the second is the level of assessment required, and along the third is the trainee's stage of development (Dreyfus & Dreyfus 1980; Miller 1990; Norcini et al. 2008).

Of these competencies, there is a substantial literature on the assessment of medical knowledge, patient care, and communication skills (category 1) and a growing literature in the assessment of professionalism (category 2), while practice-based learning and improvement and systems-based practice are relatively new and considerable research is needed to determine the criteria for good assessment of these competencies (category 3).(Arnold 2002; Driessen et al. 2005; Cruess et al. 2006; Epstein 2007; Lurie et al. 2009; Varkey et al. 2009). Methods such as portfolios have been proposed for practice-based learning and improvement. Issues of feasibility, security of data, and their application throughout the continuum of medical education require further research (category 3) (Burch & Seggie 2008).

Draft consensus criteria for good assessment

No single set of criteria for good assessment apply equally well to all situations. In fact, the same criteria should be expected to have different importance depending on the purpose and context of assessment. For example, a good summative examination designed to meet the need for accountability for the knowledge of medical graduates (e.g., a medical licensing examination) cannot be expected to, at the same time, produce detailed feedback that would guide future learning or curricular reform.

Similarly, the criteria are not of equal weight for all stakeholders even given the same assessment. For example, the validity or coherence of a licensing examination may be of more importance to patients than how much it costs the doctors who take it or the government that finances it. The importance of the criteria will vary with the perspective of the stakeholder.

To respond to these issues, we have listed a set of criteria for good assessment with short definitions of each. We then include sections on purpose (summative versus formative) and stakeholders (a limited set: examinees, patients, teachers-educational institutions, healthcare system, and regulators). In these, we discuss how the perspective of the stakeholder influences the importance of the criteria.

Criteria for good assessment

The criteria for good assessment follow and are applicable to a single assessment or a system of assessment focused around one purpose. Many of these criteria have been described before and we continue to support their importance here. However, we place particular emphasis on the catalytic effect of assessment.

(1) Validity or coherence. There is a body of evidence that is coherent ("hangs together") and that supports the use of the results of an assessment for a particular purpose.
(2) Reproducibility 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 creates, enhances, and supports education; it drives future learning forward.
(7) Acceptability. Stakeholders find the assessment process and results to be credible.

The criteria and assessment purpose

Formative assessment. Effective formative assessment is typically low stakes, often informal and opportunistic in nature, and is intended to stimulate learning. By definition, the criterion that stands out to characterize it is "catalytic effect". It works best when it (1) is embedded in the instructional process and/or work flow, (2) provides specific and actionable feedback, (3) is ongoing, and (4) is timely. Consequently, the importance of criteria such as equivalence and reproducibility-consistency diminishes to some degree. Validity-coherence remains central while educational effect and educational quality become paramount. Feasibility also increases in importance in response to the fact that formative assessment is more effective if it is ongoing, timely, and tailored to examinees' individual difficulties. Likewise acceptability, both for faculty and students, is especially important if they are to commit to the process, give credibility to the feedback they receive, and ensure that it has a significant effect.

Summative assessment. Effective summative assessment is typically medium or high stakes and is primarily intended to respond to the need for accountability. It often requires coherent, high-quality test material, significant content expertise, a systematic standard-setting process, and secure administration. Consequently, criteria such as validity-coherence, reproducibility-consistency, and equivalence are paramount. Feasibility, acceptability, and educational effect are also important, but not to the same degree as the psychometric criteria, which will to a great extent determine credibility in the scores and the underlying implications. A catalytic effect is desirable but is less emphasized in this setting. However, by not providing useful feedback, we miss the opportunity to support the learners in their continuing education.

The criteria and stakeholders

Examinees. Examinees have a vested interest in both formative and summative assessment and they must be actively involved in seeking information that supports their learning. For formative assessment, educational effects, catalytic effects, and acceptability are likely to be of most concern to examinees since they are the drivers of learning. Examinees may take validity-coherence for granted and feasibility will be an issue in terms of cost and convenience. Equivalence and reliability-consistency are less immediate.

For summative assessment, issues related to perceived fairness will be most salient for examinees. Hence, criteria such as validity-coherence, reproducibility-consistency, equivalence, and acceptability will be most important. The catalytic effect will support remediation, especially for the unsuccessful examinees. When successful examinees are not provided feedback or do not use it, it misses the opportunity to support ongoing learning.

Teachers-educational institutions. These stakeholders have interests in every facet of the assessment of students to fulfill their dual roles in education and accountability. Consistent with what was outlined above, the criteria apply differently to these two purposes.

For both teachers and institutions, student assessment information serves an important secondary purpose. These data speak to the outcomes of the educational process. In other words, students' summative assessments, appropriately aggregated, often serve as formative assessment for teachers and institutions. When combined for this purpose, criteria such as equivalence and reproducibility-consistency are a bit less important while educational effect and educational effect are a bit more important. Validity-coherence is important but should be addressed as part of good student assessment, while feasibility should be straightforward since the data are already available.

Beyond repurposing student assessment, institutions engage in the assessment of individual teachers and programs. These assessment applications can be broadly classified as either formative or summative and the criteria apply as noted above.

Patients. For patients, it is most important that their providers have good communication skills, appropriate qualifications, and the ability to offer safe and effective care. While patients certainly support the use of formative assessment, summative assessment is a more immediate concern. Consequently, criteria such as validity-coherence, reproducibility-consistency, and equivalence are of the most importance. Feasibility, acceptability, educational effect, and catalytic effect are of less concern to this group. In the long term, however, formative assessment that supports continuous improvement will be of equal or greater importance.

Healthcare system and regulators. The most pressing need of the healthcare system and the regulators is to determine which providers are competent and safe enough to enter the workforce. This need implies correct decisions based on summative assessment, so validity-coherence, reproducibility-consistency, and equivalence are paramount. Feasibility is also important.

It is growing more common for health systems to engage in some form of CQI. These systems are often embedded in the work flow and they provide ongoing, specific feedback to healthcare workers about their activities and outcomes. Validity-coherence is central, along with educational and catalytic effects, feasibility, and acceptability.

Likewise, many regulators are beginning to time limit the validity of their registration-licensure-certification decisions. This is often accompanied by the addition of a CQI component to the revalidation process. As with the healthcare system, such a component would need to emphasize validity-coherence, educational effect, educational quality, feasibility, and acceptability with less stress on equivalence and reproducibility-consistency.

Recommendations for future work

(1) Criteria must recognize the legitimacy and incorporate the perspectives of patients and the public. As the recipient of care the patient has a central role to play in the development and implementation of the criteria for assessment. Utilizing their experiences, we should strive to derive the hitherto difficult but critical facets of the doctor–patient relationship.

(2) Criteria must recognize the growing awareness of the intimate relationship between assessment, feedback, and continued learning. To maximize CQI, relevant and useful feedback must be provided in a way that encourages and supports the examinees' progress. Ideally, this feedback would be adaptive to the individual, his/her place in the developmental continuum, and the broader system of assessment.

(3) Criteria need to be developed for systems of assessment. The focus of the document to this point has been single purpose assessment processes, but systems of assessment require consideration as well. Such systems integrate a series of different individual measures that are developmental and cover the continuum of assessment. Good assessments within a system are designed to take account of the content and results of former and future assessments.

(4) Criteria need to be developed for accreditation processes. The implementation of accreditation processes for educational programs is growing rapidly and internationally. As part of such processes, educational programs are evaluated against a set of standards. There is no published data about whether such processes improve quality and what the criteria are for judging actual performance against the standards being promulgated. At the end of the day, criteria for good assessment must apply equally to institutions and individuals.

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the article.

Notes on contributors

JOHN NORCINI, President and CEO, Foundation for Advancement of International Medical Education and Research, USA.

BROWNELL ANDERSON, Senior Director, Educational Affairs, Association of American Medical Colleges, USA.

VALDES BOLLELA, Divisão de Moléstias Infecciosas e Tropicais do Departamento de Clínica Médica da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo (FMRP-USP), Brazil.

VANESSA BURCH, Professor and Chair of Clinical Medicine, Department of Medicine, University of Cape Town and Groote Schuur Hospital, South Africa.

MANUEL JOÃO COSTA, Life and Health Sciences Research Institute (ICVS), School of Health, Sciences, University of Minho, Portugal.

ROBBERT DUVIVIER, Maastricht University, The Netherlands.

ROBERT GALBRAITH, Co-Executive Director, Center for Innovation, National Board of Medical Examiners, Philadelphia, USA.

RICHARD HAYS, Dean, Faculty of Health Sciences and Medicine, Pro-Vice Chancellor Quality, Teaching & Learning, Bond University, Australia.

ATHOL KENT, Department of Obstetrics and Gynecology, University of Cape Town, South Africa.

VANESSA PERROTT, University of Cape Town, South Africa.

TRUDIE ROBERTS, Prof. Medical Education and Director, Leeds Institute of Medical Education, University of Leeds, UK.

References

Amin Z, Seng CY, Eng KH. 2006. Practical guide to medical student assessment. Singapore: World Scientific Publishing Co. Ltd. Pte.

Arnold L. 2002. Assessing professional behavior, yesterday, today, and tomorrow. Acad Med 77(6):502–515.

Ben-David MF. 2000. The role of assessment in expanding professional horizons. Med Teach 22(5):472–477.

Boud D. 2000. Sustainable assessment: Rethinking assessment for the learning society. Stud Cont Educ 22:151–167.

Boyer EL. 1990. Scholarship reconsidered: Priorities of the professoriate. Princeton, NJ: Princeton University Press, The Carnegie Foundation for the Advancement of Teaching.

Burch VC, Seggie JL. 2008. Use of a structured interview to assess portfolio-based learning. Med Educ 42(9):894–900.

Case M, Swanson DB. 1996. Constructing written test items for basic and clinical sciences. Philadelphia, PA: National Board of Medical Examiners.

Colliver JA. 2002. Educational theory and medical education practice: A cautionary note for medical school faculty. Acad Med 77 (12, Part 1):1217–1220.

Cottrell S. 2006. A matter of explanation: Assessment, scholarship of teaching and their disconnect with theoretical development. Med Teach 28(4):305.

Crossley J, Humphries G, Jolly B. 2002. Assessing health professionals. Med Educ 36:800–804.

Cruess R, McIlroy JH, Cruess S, Ginsburg S, Steinert Y. 2006. The professionalism mini-evaluation exercise: A preliminary investigation. Acad Med 81(10): S74–S78.

Curry L. 2002. Achieving large-scale change in medical education. In: Norman G, van der Vleuten C, Newble D, editors. International handbook of research in medical education: Part 2. 1st ed. Dordrecht, The Netherlands: Kluwer Academic Publishers. pp 1039–1084.

Downie RS, Calman KC. 1987. Healthy respect – Ethics in health care. 1st ed. London: Faber and Faber.

Dreyfus H, Dreyfus S. 1980. A five–stage model of the mental activities involved in directed skill acquisition. Berkely, CA: Operations Research Center, University of California.

Driessen E, van der Vleuten C, Schuwirth L, van Tartwijk J, Vermunt J. 2005. The use of qualitative research criteria for portfolio assessment as an alternative to reliability evaluation: A case study. Med Educ 39(2):214–220.

Epstein RM. 2007. Assessment in medical education. N Engl J Med 356:387–396.

Flexner A. 1910. Medical education in the United States and Canada. New York, NY: Carnegie Foundation for the Advancement of Teaching.

Flexner A. 1912. Medical education in Europe. Bulletin six. New York, NY: Carnegie Foundation for the Advancement of Teaching.

Gipps C. 1999. Sociocultural aspects of assessment. Rev Res Educ 24:355–392.
Kassebaum DK. 1992. Origin of the LCME, the AAMC–AMA partnership for accreditation. Acad Med 67(2):85–87.
Kogan JR, Holmboe ES, Hauer KE. 2009. Tools for direct observation and assessment of clinical skills in medical trainees: A systematic review. J Am Med Assoc 302:13–16.
Lesky E. 1970. Medical Education in England since 1600. In C.D. O'Malley (Ed.) The history of medical education. P 235-250.Los Angeles: University of California Press.
Lurie SJ, Mooney CJ, Lyness JM. 2009. Measurement of the general competencies of the accreditation council for graduate medical education: A systematic review. Acad Med 84(3):301–309.
Miller G. 1990. The assessment of clinical skills/competence/performance. Acad Med 65(9):S63–S67.
Newble D. 1998. Assessment. In: Jolly B, Rees L, editors. Medical education in the millennium. 1st ed. Oxford: Oxford University Press. pp 131–142.
Norcini JJ. 2005. Current perspectives in assessment: The assessment of performance at work. Med Educ 39:880–889.
Norcini J, Burch V. 2007. Workplace-based assessment as an educational tool. Med Teach 29:855–871.
Norcini JJ, Holmboe ES, Hawkins RE. 2008. Evaluation challenges in the era of outcomes-based education. In: Holmboe ES, Hawkins RE, editors. Practical guide to the evaluation of clinical competence. Philadelphia, PA: Elsevier Health Sciences. pp 1–9.
Norman GR, Schmidt HG. 1999. Of what practical use is a baby? Perspectives on educational research as a scientific enterprise. Prof Educ Researcher Quarterly 20:1–5.
Poynter F. 1970. The development of bedsire teaching at the Vienna Medical School from scholastic times to special clinic. In C.D. O'Malley (Ed.) The history of medical education. Los Angeles: University of California Press, pp 217–234.
Rees L, Jolly B. 1998. Medical education into the next century. In: Jolly B, Rees L, editors. Medical education in the millennium. 1st ed. Oxford: Oxford University Press. pp 245–260.
Rushton A. 2005. Formative assessment: A key to deep learning? Med Teach 27(6):509–513.
Shepard LA. 2000. The role of assessment in a learning culture. Educ Res 29(7):4–14.
Starr P. 1982. The social transformation of American medicine. New York, NY: Basic Books.
Trigwell K, Martin B, Benjamin J, Prosser M. 2000. Scholarship of teaching: A model. High Educ Res Dev 19:155–168.
Trigwell K. 2001. Judging university teaching. Int J Acad Dev 6(1):65–73.
van der Vleuten C. 1996. The assessment of professional competence: Developments, research and practical implications. Adv Health Sci Educ 1:41–67.
Varkey P, Karlapudi S, Rose S, Nelson R, Warner M. 2009. A systems approach for implementing practice-based learning and improvement and systems-based practice in graduate medical education. Acad Med 84(3):335–339.
Vygotsky LS. 1978. Mind in society: The development of higher psychological processes. Cambridge, MA: Harvard University Press.

Bibliography

(1) Gronlund NE. 1998. Assessment of student achievement. 6th ed. Needham Heights, MA: Allyn and Bacon. pp. 17–22. This is a really good book which provides very sound advice and guidelines for non-experts who need to develop or administer assessments without specialized training. I would recommend it for the list of references/readings at the end of the document. I have the 6th edition and there is a more recent version, but essentially the list of guidelines for effective assessment remains unchanged. They are (verbatim):

- A clear conception of all intended learning outcomes is required.
- A variety of assessment procedures should be used.
- Instructional relevance of the assessment procedures need to be considered.
- An adequate sample of student performance should be obtained.
- The assessment procedure should be fair to everyone.
- The specifications of criteria for judging performance should be clear.
- Students should get feedback that emphasizes strengths of performance and weaknesses to be corrected.
- A comprehensive grading and reporting system needs to be in place.

(2) Shumway JM, Harden RM. 2003. AMEE Guide no. 25: The assessment of learning outcomes for the competent and reflective physician. Med Teach 25:569–584.

- Another list of recommendations set out by the UK GMC regarding assessment. Very similar to the list of Gronlund.
- A comprehensive list of assessment methods in use.
- For each method, a description of the method, its strengths and weaknesses and impact on learning.
- An annotation of Miller's pyramid providing suitable assessment methods for each level of the pyramid.
- A demonstration of how to match learning outcomes to assessment methods using the Dundee curriculum learning outcomes.

(3) Kogan JR, Holmboe ES, Hauer KE. 2009. Tools for direct observation and assessment of clinical skills of medical trainees: A systematic review. J Am Med Assoc 302:1316–1326

- A very nice review with a comprehensive table detailing all assessments known to man!
- A comprehensive bibliography.

(4) Swanson DB, Norman GR, Linn RL. 1995. Performance-based assessment: Lessons from the health professions. Educ Researcher 24:5–11.

- This is an old paper, but I consider it a "classic". The lessons outlined in the paper still hold true and are valuable for all working in the performance assessment field.
- The fact that examinees are tested in realistic performance situations does not make the test design and domain sampling simple and straightforward.
- No matter how realistic a performance-based assessment is, it is still a simulation and examinees do not behave in the same way they would in real life.
- While high-fidelity performance-based assessment methods often yield rich and interesting examinee behavior, scoring that behavior can be problematic.

- Regardless of the assessment method used, performance in one context does not predict performance in other contexts very well.
- Correlational studies of the relationship between performance-based test scores and other assessment methods targeting different skills typically produce variable and uninterpretable results.
- Because performance-based assessment methods are often complex to administer, multiple test forms and test administrations are required.
- All high-stakes assessments, regardless of the method used, have an impact on teaching and learning.

(5) Wass W, van der Vleuten C, Shatzer J, Jones R. 2001. Assessment of clinical competence. The Lancet 357:945–949.

- Again, an old paper but very useful for the average clinician educator.
- The paper explains the levels on Miller's pyramid and provides examples of assessment methods for the different levels.
- Concept of "blueprinting" of assessments.

(6) Schuwirth LWT, van der Vleuten CPM. 2004. Changing education, changing assessment, changing research. Med Educ 38:805–812.

- Brief update on validity, reliability, and educational impact of assessment following on his earlier paper in 1996 where the criteria were originally proposed.
- A few important points:
  - Measure competence of roles or tasks, not single traits.
  - Authenticity and integration should ensure that there is optimal congruence between assessments on the one hand and educational goals and the demands of future practice on the other.
  - Assessment is an issue of educational design rather than a measurement problem.
  - Rather than adopting a method that has been successful in a certain situation, one should adopt its underlying concepts and translate them to fit the unique demands of the local situation.

(7) Van der Vleuten CPM, Schuwirth LWT. 2005. Assessing professional competence: From methods to programmes. Med Educ 39:309–317.

- Also an older paper but again the principles in the paper are very valuable.
- Important concepts:
  - Think about assessment programs rather than individual assessment methods.
  - Use appropriate, adequate sampling.
  - Move assessment back to the real world of the workplace.
  - Global and holistic assessment rather than breaking down competency into small units.
  - Need to use multiple assessment methods.
  - Rely more on the professional judgment as a basis for decision making.
  - Assessment is inextricably woven together with all other aspects of a training program.
  - A good assessment program will incorporate several competency elements.
  - Use multiple sources of information.
  - Use multiple occasions to test.
  - Use credible standards.