Assessing Anatomy as a Basic Medical Science

31

Dujeepa D. Samarasekera, Ponnampalam Gopalakrishnakone, and Matthew C.E. Gwee

Introduction

> "Assessment Drives Student Learning."
>
> Author unknown

The scientific basis of medicine has been convincingly reaffirmed in several official documents published by leading medical authorities (see Tomorrow's Doctors, GMC, U.K., 1993, 2003; joint publication of the Howard Hughes Medical Institute and AAMC, USA, 2009 [1] and the AFMC, Canada, 2010). Together, these documents serve to underscore the importance of the basic medical sciences (BMSc) in the undergraduate medical curriculum. Disciplines in the BMSc (e.g., anatomy, biochemistry, pharmacology, and microbiology) are *highly content-based* disciplines. Moreover, students were simply *passive* listeners and recipients of abundant *factual* content knowledge delivered through numerous lectures.

Dujeepa D. Samarasekera, FAcadMEd(UK), FAMS, MHPE, MBBS (✉)
Medical Education Unit, Yong Loo Lin School of Medicine, National University of Singapore, National University Health System, Singapore
e-mail: dujeepa_samarasekera@nuhs.edu.sg; meddds@nus.edu.sg

Ponnampalam Gopalakrishnakone, FAMS, DSc, PhD, MBBS
Department of Anatomy, Yong Loo Lin School of Medicine, National University of Singapore, National University Health System, Singapore

Matthew C.E. Gwee, PhD, MHPEd, BPharm(Hons)
Medical Education Unit, Dean's Office, Yong Loo Lin School of Medicine, National University of Singapore, National University Health System, Singapore

Focus of Traditional Assessment Strategies

In such a traditional learning environment, assessment strategies commonly designed by the disciplines focused more on testing *factual content* that imposed on students the need to "*memorize, recall, and regurgitate*" in exams. Thus, assessments tested mainly the acquisition of lower-order learning outcomes through mere rote memorization and recall. Moreover, assessment scores were used primarily as final (summative) tests without considering their potential as a learning tool. A major concern of such assessment practices is the *negative steering effect* of assessment on student learning, i.e., students tended to undertake *superficial* or *rote learning* rather than *deep* learning with enhanced conceptual understanding. Thus, students are likely to become *knowledge-rich* but *application-poor* with respect to medical problem-solving and critical reasoning. Myers and Jones [2] have clearly expressed that "*What matters…is not what students know but what they can do with what they know. What's at stake is the capacity to perform, to put what one knows into practice*."

L.K. Chan and W. Pawlina (eds.), *Teaching Anatomy: A Practical Guide*,
DOI 10.1007/978-3-319-08930-0_31,

Assessment Strategies for Twenty-First-Century Medical Education

Today, however, there is clear recognition that assessments have excellent potential as learning tools since they will drive how students learn: "…assessment has a major impact on students' learning behaviour [and it is]…for the test maker to capitalise on this behaviour. …the driving influence of assessment is a powerful tool to ensure that students learn what, and how, teachers want them to learn" [3].

Consequently, there is now a major paradigm shift from the assessment *of* learning to assessment for learning. The latter is best achieved through providing students with feedback to explain the "rights" and "wrongs" of test items, preferably, on an individual basis in a reflective dialogue. Feedback is further enhanced through the use of formative assessments as diagnostic tools rather than as test instruments, which is consistent with the paradigm shift to drive student learning behavior through assessments [4].

Thus, today, assessments should not focus predominantly on student acquisition of *lower- order* learning outcomes based primarily on *rote memorization* of factual content knowledge. Instead, assessments should now focus more on the acquisition of higher-order learning outcomes to help students develop their intellectual skills, i.e., their ability to analyze, integrate, evaluate, and apply the foundational knowledge they have acquired. Myers and Jones [2] have stated that "Students learn not by just absorbing content (taking copious notes and studying for exams), but by critically analysing, discussing and using content in meaningful ways."

Thus, major paradigm shifts in assessment practices have made it imperative to strategize the design of test instruments used in the assessment of students in health professional education. Instead of using instruments to test a student's ability to remember a large amount of factual content, testing should now be aimed at how students can *apply* the knowledge and skills which they have acquired to their next level of learning, or to problem-solve in future professional practice [5, 6], including the relevant *domain independent* skills such as communication, teamwork, and professionalism.

Assessment in Anatomy

> "Students can, with difficulty, escape from the effects of poor teaching, they cannot escape the effects of poor assessment."
>
> Boud [7]

Anatomy is the study of the structure of the human body and associated body functions and, therefore, provides fundamental core knowledge in medical, nursing, and other health professional training programs. Thus, students need to acquire a solid *foundational knowledge* of anatomy as a BMSc which would subsequently enable them to apply the knowledge they acquired to what they will learn about associated structural abnormalities occurring in disease processes and also to patient management. When the students graduate, their knowledge of clinical anatomy will become important in their practice as effective healthcare professionals.

However, assessments have many other useful functions, including:

- Motivating students to study
- Diagnosing a student's strengths, limitations, and difficulties
- Measuring student improvement over time and readiness to proceed to the next level of training
- Providing students with feedback about their learning
- Evaluating the effectiveness of teaching and the educational program
- Making decisions on student understanding of subject matter and competency in skills
- Predicting a student's likely success in future learning or exams
- Ensuring students meet the qualification/certification standards

Since anatomy is a highly content-based BMSc discipline, the following guidelines are recommended:

- Closely align assessment strategies to the specific learning outcomes identified for the anatomy course.
- Design assessment strategies which test beyond just factual recall of knowledge and incorporate testing of higher-order learning outcomes.

- Include *formative* assessments as a potential and powerful tool to drive student learning behavior.

Recommended Best Practices for Anatomy Assessment

> "Everything that can be counted does not necessarily count; everything that counts cannot necessarily be counted."
>
> Albert Einstein

Constructive Alignment in Anatomy Assessment

For any assessment process to be successful, there is a need to ensure that test items are closely aligned to the expected course and program (learning) outcomes [8]. Anatomy is taught mainly in the early part of the medical course. The assessment modalities used in anatomy exams must, therefore, be constructively aligned to the expected course outcomes in anatomy, i.e., to the graduate outcomes for learning to be effective [9]. Figure 31.1 outlines how best to achieve this close alignment through proper planning in the development of an assessment process.

The close alignment of assessment to learning outcomes (i.e., to requisite future practice competencies) will have a strong *positive* steering effect on students' learning behaviors [10]. The best practice in constructive alignment is to develop an *assessment blueprint* (see Table 31.1 below) which allows assessment developers to have a bird's-eye view of the extent to which the assessment covers the discipline-specific outcomes, as well as the relevant program outcomes (end

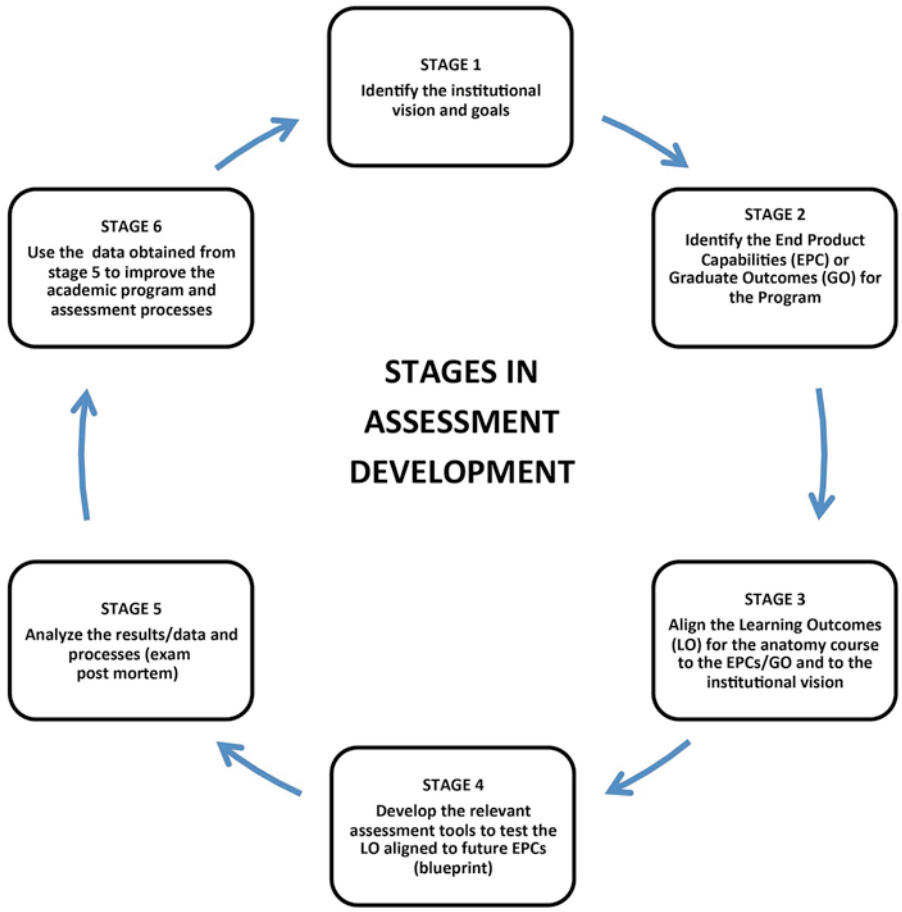

Fig. 31.1 Stages in the development of an assessment process

Table 31.1 Example of a final assessment blueprint for an undergraduate course in anatomy

Outcome capabilities modules	Knowledge				Skills			
	Recall/ comprehension	Application	Analysis	Evaluate	(basic clinical skills/spots)	Professionalism	Teamwork	Communication
Musculoskeletal	MCQ	MCQ		MEQ	MSE		MSE	
Respiratory		MCQ	MCQ/MEQ					
Cardiovascular		MCQ	MEQ		MSE			MSE
GIT	MCQ	MCQ/MEQ	MCQ		MSE	MSE		
Neuro/CNS	MCQ	MCQ	MCQ	MEQ	MSE			
Genitourinary		MCQ	MCQ					
Reproductive		MCQ						
Special senses	MCQ	MCQ			MSE			

MCQ: multiple-choice questions, *MEQ*: modified essay questions, *MSE*: multistation examination (e.g., objective structured practical/clinical examination (OSP/CE))

product capabilities). It also serves to inform test developers, teachers, and students how different assessment instruments test the overall learning outcomes. An assessment blueprint also helps to conceptualize the *level* of testing which is especially useful in identifying the level of testing of intellectual skills in the cognitive (knowledge) domain using Bloom's taxonomy. A simple assessment blueprint will facilitate the development of items that test *higher-order* cognitive skills and, consequently, steering students away from *rote learning* (i.e., through testing mainly *recall of factual content knowledge*) [11].

Utility Index of an Assessment System

The *utility index*, described by Cees van der Vleuten in 1996, provides a good foundation to choose and incorporate different assessment instruments when assessing students [12]. The formula is more conceptual than mathematical, highlighting the fact that no single assessment instrument can achieve high utility if used alone. This is because each instrument has its own limitations and strengths with regard to validity, reliability, educational impact, cost (and feasibility), and acceptability by the stakeholders. However, if the assessment planning is done carefully and incorporates multiple instruments strategically into the curriculum at different time points of student learning, the data gathered from the

Assessment Utility Index

U = Utility index
R = Reliability
V = Validity
E = Educational impact
A = Acceptability
C = Cost
W = Weight

$$U = {}_{w_r}R \times {}_{w_v}V \times {}_{w_e}E \times {}_{w_a}A \times {}_{w_c}C$$

instruments will provide useful information about the learner. This will assist to build a profile of the student's abilities and for examiners to decide on whether the learner is ready to progress to the next level of learning [13]. The utility index therefore can also be used simply to demonstrate the compromise necessary in assessment and given the context, one has to optimize the assessment.

Validity

Validity is defined as the extent to which the instrument measures the competency which it is supposed to measure [14]. To achieve high validity, it is best to first define the domain for the competency to be tested and the relevant assessment instrument to be used to test a student's achievement of the competency.

Validity can be affected by several factors, including: poor alignment of test instruments to the competencies and domains assessed, confusing instructions to assessors and students, poor test development leading to low-quality test items/questions, developed test items/questions too difficult or too easy for the students, and poor sampling of testable content/domain areas.

The *validity* of an assessment can be described under different categories or types, including:

- Content Validity: This refers to whether the assessment actually tests the intended subject area. Thus, if the assessment is to test knowledge of gross anatomical structures and their functions in the human body, the questions should be focused on *human* anatomy and its associated functions.
- Construct Validity: This refers to the extent to which the assessment instrument measures a particular behavior or trait. Thus, the 360° evaluation instrument can measure traits such as *teamwork* and *professionalism*. However, it should be noted that achievement of competency in *behavioral constructs,* such as teamwork or professionalism, cannot be measured using only one instrument.
- Predictive Validity: This refers to whether student performance on an assessment instrument could predict his/her future performance. For example, the *predictive validity* of an anatomy multiple-choice question (MCQ) paper in year 1 of a medical program should predict similar performance in the surgery MCQ exam in year 4. The higher the correlation of scores in both tests, the higher the predictive validity.
- Concurrent Validity: This refers to how well the performance in one test correlates with other validated tests administered at the same time. If the student scores obtained for the anatomy MCQ test correlate positively with the scores of the physiology MCQ test in year 1 exams, then the concurrent validity of the anatomy MCQ paper is high.
- Face Validity: This refers to stakeholders' views and perceptions on whether the test measures what it is supposed to measure and the fairness of the assessment. This is also an important aspect to consider since if the students and teachers feel that the assessment is fair and aligned to future practice or assists in further learning, students will be motivated to learn and teachers will engage in depth in teaching–learning activities.

Reliability

Reliability refers to the consistency or reproducibility of student performance. For an assessment to be reliable, a student can be expected to attain the same score if the test, or the same or similar test with the same degree of difficulty, is administered at a later date or over multiple times. If the assessment is unreliable, then the student's score may vary widely, depending on whether other factors are also aligned to the intent of the exam. The reliability can also be affected by several factors such as efficiency of the test administration, inadequate sampling of the content/domain areas tested, and lack of objectivity of scoring by assessors as a consequence of a lack of or poor assessor training.

Reliability also consists of a few categories as outlined below:

- Test–Retest Reliability: This refers to whether a student's performance in an exam will be similar if the test is administered repeatedly at different times. Test–retest reliability measures the consistency of an examination: for example, will the student scores be similar when the final anatomy modified essay exam is administered at the end of year 1 and again at the beginning of year 2?
- Inter-rater Reliability: This refers to the concurrence of scoring among assessors on a single performance. *Inter-rater reliability* is considered good if two assessors independently give similar scores to the same assessment, i.e., either in an anatomy modified essay question (MEQ) or in an OSCE station.
- Split-Half Reliability: This refers to the internal consistency of an assessment. The process involves dividing the exam questions/items into two halves which assess the same knowledge or domain skills. The examination is then administered and the scores for each half are obtained and correlated. The *split-half correlation* is considered reliable if the scores of the two sets correlate positively.

Educational Impact

One of the most important aspects of assessment is its *relevance* to and its *impact* on student learning. Assessments could be used in a strategic way to drive students to learn what is important in the curriculum and for their future learning and practice. This alignment of assessment to students' next level of learning or to future practice is known as the educational impact and is gaining prominence in the design of new assessments. Furthermore, such a close alignment will assist teachers to focus on relevant teaching–learning activities. Incorporating this constructive alignment of learning relevance and future practice relevance to assessment is also called the consequential validity of assessment [15].

If the test item in the anatomy MEQ examination uses contextualized future practice clinical scenarios, then the students, during their learning, will focus on applying knowledge of the gross anatomical structures and the associated functions which they learned to applied clinical anatomy. The teachers can also be expected to emphasize knowledge application to students during their teaching–learning activities.

Acceptability

Acceptability and the perceived fairness by the stakeholders, i.e., teachers, students, administrators, professional and employing bodies, patients, and the communities, regarding the robustness of the assessment and its processes are also important. This forms the basis of the trust placed by these stakeholders with regard to the graduate's effectiveness as a practitioner.

Cost/Feasibility

Another factor that needs careful attention when developing an assessment instrument or a process is the cost of development or the overall feasibility of employing the instrument or the process. The time it takes to develop and operationalize the assessment, the number of test developers and administrators involved, interpreting the scores, and provision of feedback to learners and to the teachers will have an impact on the assessment. If these areas are not carefully considered, as discussed before, the reliability as well as the educational impact will also be affected.

To elaborate this further, if an anatomy multistation examination involves ten stations (at which students spend six minutes each) assessing the practical skill sets of 300 medical students, you need to either set up many stations testing the same skill or develop the assessment over few days with similar skills stations to run smoothly. For this to be operational and practical, the school needs to allocate substantial financial as well as human resources, and planning becomes very critical. If

not, the exercise will become laborious and ineffective leading to substandard assessment.

Effective Use of Formative and Summative Assessments

As already discussed, assessment should drive student learning behavior, not only to develop higher-order discipline-specific cognitive skills but also to prepare students to acquire competencies for future professional practice. However, assessing the achievement of competences in all the relevant domains cannot be achieved using a single assessment instrument. The students need to be closely supervised and guided on how to develop their skills and gain knowledge and also to be given feedback on how best to further improve/develop as they progress through the learning program. This is best achieved using formative assessments (FAs). The focus of FA is to provide relevant, specific, and immediate feedback to students on how well they are achieving the necessary knowledge, skills, and attitudes. The student's performance obtained at these assessments will not be used for final grading or pass–fail decisions. Thus, FA allow students to refine, enhance, and optimize their performance before the scheduled summative assessments.

Summative assessments (SAs) are used to evaluate whether students have reached the required level of competency, so that they will be effective learners at the next level of learning, safe to patients, and contributing to team care when applying their acquired skills in practice as healthcare professionals. The FAs are planned usually at the end of a particular period of learning or skills training, so that one can judge the students' level of competence.

Effective Use of Feedback in Assessment

Feedback is the cornerstone of assessment, and in contemporary health professional education, it is becoming one of the critical components of learning and assessment. Providing relevant, focused, and immediate feedback on student performance assists learners to identify areas for improvement and enhances their individual areas of strength.

A typical feedback setting involves a face-to-face verbal discourse between the teacher and student. Students usually get feedback on their skills assessments, whereas only a score is often provided for written assessments. However, increasingly more elaborate feedback is also provided on written assessments, especially for MCQs through the use of technology.

The computer-based online feedback can provide more granular feedback regarding the students' performance by comparing the individual's score relative to the cohort as well as for each content area tested. It can further benchmark the student performance to local or international cohorts taking the same or similar exams.

For face-to-face feedback, the most commonly used format is the modified *Pendleton's* feedback model. The model not only identifies areas of strength and areas for improvement but also improves students' *metacognition* and *empowers* learners to be more self-directed in their learning [16].

Modified Pendleton's Feedback Model [16]

- Ask the learner *what went well* and *why*.
- The teacher says what went well and why.
- Ask the learner *what can be done better* and *how*.
- The teacher says what can be done better and how.
  - Summarize strengths and *list three things to concentrate on.*
  - How you (or your colleagues) could help the learner in achieving the above.

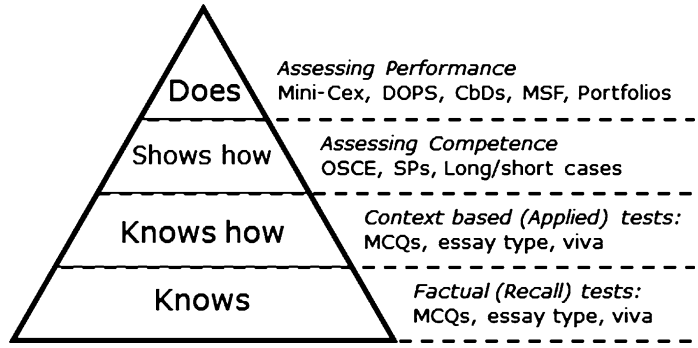

Fig. 31.2 An assessment framework using Miller's pyramid of competence [17]

Assessment Formats in Anatomy

Written Assessment

Written assessments are used mainly to assess knowledge acquisition: the lower-order outcomes of cognitive or knowledge-processing skills of students, relating to the "knows" and "knows how" levels in the Miller's pyramid of competence shown in Fig. 31.2 [17].

Multiple-Choice Question Formats

MCQs are the most common formats of assessment used in medical and health professional programs for anatomy [18]. If the test items are developed properly (see section on various developmental stages) and constructed using blueprinting, MCQ assessments can achieve high psychometric qualities such as high test–retest reliability, content validity, and educational impact [19].

MCQs which are context-rich with a single best response are widely used today. The other formats (e.g., true/false) have now been largely phased out, whereas the extended matching items are increasingly gaining popularity.

Modified Essay Questions

Modified (or structured) essays are commonly used to assess the students' *knows how* level of learning. In most instances, these are best used to assess in-depth knowledge of a particular content area in anatomy. The disadvantage of this format is that it cannot sample widely, and hence, the reliability is generally low.

Long Essays

Long essay questions are no longer used widely today, due to the expected low reliability when testing a subject area like anatomy. The long essays are now limited for use only in special circumstances where students need to develop or synthesize a response to a question relating to a given situation or condition.

Assessing Skills

Multistation Examinations (MSE, e.g., Objective Structured Practical/Clinical Examination, OSPE/OSCE)

The most common format used to assess skills is the MSE such as the OSPE/OSCE. The use of MSE to assess relevant skills acquired using task trainers and simulated participants (SPs) with checklist-based scoring is a popular strategy used in most medical and health professional training programs. The advantage of this format is that it allows assessment of multiple domains in a single test. The domains could be applied knowledge—radiological or anatomical specimen,

clinical case vignettes, audiovisual specimen, communication skills station using SPs, etc. See blueprint shown in Table 31.1. Therefore, the MSE, if developed properly and based according to the overall assessment blueprint, can have high validity and educational impact. However, the main disadvantages are that it requires a large amount of resources to develop as well as to conduct and poor reliability due to low sampling of subject content area.

Spot Tests

Anatomy *spot* tests are used for quick identification of anatomical structures. The students are presented with several prosected specimens bearing anatomical structures such as muscles, nerves, blood vessels, etc., which are highlighted for students to identify. The questions may lead to further testing of associated function, or limitation/s due to any dysfunction. The major disadvantage of this assessment strategy is that it promotes a negative steering effect (i.e., *rote memorization* and *recall* of knowledge), and it has largely been replaced by MSE.

Oral Examination

Oral (or viva voce) examinations were used extensively in the past in medical and health professional training programs. Oral examinations (OEs) have been phased out in many assessment settings due to their apparently poor reliability. However, recent studies show that if the OE is well structured (using a blueprint for specific item development), examiners are briefed and trained, and scoring is based on a marking template, the OE can achieve good reliability as well as identify the learner's abilities in higher-order cognitive skills [20].

Portfolio Assessment

The use of *portfolios* to assess student learning during the phase in the BMSc is now gaining in popularity. The portfolio not only assesses different *cognitive* skills such as reflective learning but also evaluates some areas of affective skills [21]. The students need to show achievement of required outcomes in these domains and provide evidence of their learning at multiple time points of the program. Feedback regarding their performance is given by faculty who have been trained to evaluate portfolios. The main advantage of using this modality of assessment in the early part (i.e., BMSc phase) of the curriculum is that the faculty can evaluate, over different phases, how the learners are developing in a discipline and, consequently, can provide immediate feedback for remediation whenever required. Portfolios are also used as summative assessments, especially to make value judgments and high-stakes decisions on a learner's ability to move to the next level of learning or to certify that a learner is fit for practice.

Standard Setting in Assessment

Deciding on whether a learner has achieved the required competency, based on his/her performance in tests, is an important process in deciding whether or not a learner is ready for the next level of learning or professional practice. The score deciding whether a student can pass or fail should be based on relevant and context-specific educational rationale [22]. According to Norcini *"A standard is a special score that serves as a boundary between those who perform well enough and those who do not."* Broadly speaking, there are two types of assessment standards: relative (also known as norm referenced) and absolute (fixed or criterion referenced) [23]. In relative standards, a student's performance is compared with the performance of other students in the cohort, and the pass/fail mark is then set accordingly. For example, this method is used widely on entry tests to programs where the numbers of student vacancies are limited. The advantage is that you can limit the number of passes and failures based on the requirements of the school's

program. However, the main disadvantage is that setting pass/fail standard scores will differ from one cohort to the other, since it is based mainly on student performance.

When deciding on a pass/fail score using absolute standards, the passing standard is set before the exam is conducted. The standard set will not, therefore, differ from cohort to cohort. Consequently, an entire cohort can pass or fail the assessment, depending on whether the minimum passing score based on the standard has or has not been achieved. The score is set using a *criterion-based* standard setting process in which experts first define the acceptable minimum passing standard before students sit for the assessment. The major advantage of this method is that the passing score is based on a minimum competency standard and will not vary according to the performance of the cohort. It has also been shown that criterion-based standards promote teamwork and collaborative learning as the passing grade is not set based on cohort performance. The main disadvantage is that it is a time-consuming process as well as being resource intensive. There are few well-established methods of criterion-based standard setting, such as Angoff, Ebel, and Nedelsky [24–26].

References

1. Association of American Medical Colleges-Howard Hughes Medical Institute Committee. 2009. Scientific Foundations for Future Physicians. Report of the AAMC-HHMI Committee. https://www.aamc.org/download/271072/data/scientificfoundationsforfuturephysicians.pdf
2. Myers C, Jones TB. Promoting active learning: strategies for the college classroom. San Francisco: Jossey-Bass; 1993.
3. Dijkstra J, van der Vleuten CP, Schuwirth LW. A new framework for designing programmes of assessment. Adv Health Sci Educ Theory Pract. 2010;15:379–93.
4. Van der Vleuten C, Schuwirth L, Driessen E, Dijkstra J, Tigelaar D, Baartman L, Van Tartwijk J. A model for programmatic assessment fit for purpose. Med Teach. 2012;34:205–14.
5. Woods NN, Brooks LR, Norman GR. The role of biomedical knowledge in diagnosis of difficult clinical cases. Adv Health Sci Educ Theory Pract. 2007;12:417–26.
6. Larsen DP, Butler AC, Roediger III HL. Test-enhanced learning in medical education. Med Educ. 2008;42:959–66.
7. Boud D. Assessment and learning: contradictory or complementary? In: Knight P, editor. Assessment for learning in higher education. London: Kogan Page; 1995.
8. Biggs J. Teaching for quality learning at university: what the student does. SRHE and Open University Press imprint; 1999.
9. Ramsden P. Learning to teach in higher education. Routledge Falmer; 1992.
10. Brown SA, Knight P. Assessing learners in higher education. London: Kogan Page; 1994.
11. Bloom BS, Englehart MB, Furst EJ, Hill WH, Krathwohl DR. Taxonomy of educational objectives, the classification of educational goals – handbook I: cognitive domain. New York: McKay; 1956.
12. Van der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. Adv Health Sci Educ. 1996;1:41–67.
13. Van der Vleuten CP, Schuwirth LW. Assessing professional competence: from methods to programmes. Med Educ. 2005;39:309–17.
14. Moskal BM, Leydens JA. Scoring rubric development: validity and reliability. Pract Assess Res Evaluat. 2000;7(10).
15. Paul E, Elder L. Consequential validity: using assessment to drive instruction. 2007. http://www.criticalthinking.org/files/White%20PaperAssessmentSept2007.pdf
16. Pendleton D, Schofield T, Tate P, Havelock P. The consultation: an approach to learning and teaching. Oxford: Oxford University Press; 1984.
17. Miller GE. The assessment of clinical skills/competence/performance. Acad Med. 1990;65:S63–7.
18. Craig S, Tait N, Boers D, Mcandrew D. Review of anatomy education in Australian and New Zealand medical schools. ANZ J Surg. 2010;80:212–6.
19. Case SM, Swanson DB. Constructing written test questions for basic and clinical sciences. 3 ed. NBME; 2002.
20. Kearney RA, Puchalski SA, Yang HY, Skakun EN. The inter-rater and intra-rater reliability of a new Canadian oral examination format in anesthesia is fair to good. Can J Anaesth. 2002;49:232–6.
21. Driessen E, Van Tartwijk J, Vermunt JD, van der Vleuten CP. Use of portfolios in early undergraduate medical training. Med Teach. 2003;25:18–23.

22. Wass V, van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. Lancet. 2001; 357:945–9.
23. Norcini JJ. Setting standards on educational tests. Med Educ. 2003;37:464–9.
24. Ebel RL. Essentials of educational measurement. 3rd ed. Upper Saddle River, NJ: Prentice Hall Inc. 1979.
25. Angoff WH. Scales, norms and equivalent scores. In: Thorndike RL, editor. Educational measurement. 2 ed. Washington DC: American Council on Education; 1971.
26. Nedelsky L. Absolute grading standards for objective tests. Educ Psychol Measur. 1954;14:3–19.