Update in
Anaesthesia

The Practice of Assessment

Daniel Dubois
Correspondence email: dadubois@toh.ca
doi: 10.1029/WFSA-D-21-00005

Abstract

Assessment is a central feature of teaching and the curriculum. The most important part of assessment is the correct interpretation and use of the information for its intended purpose. There should be a strong emphasis on using frequent and timely formative assessment to optimize individual progress, as inevitably assessments will influence students' learning strategies. There are a multitude of different assessment approaches that will need to be aligned with the desired learning objectives as a useful way to encourage trainees to attend to the most important outcomes. Consider the complementarity of different methods at your disposal to leverage the strengths and compensate for the individual weaknesses which will influence validity in your own setting. Despite the educators' best intentions, problems can develop when they attempt to assess trainees and challenges should be anticipated. A properly constructed system of assessment can, over time and using multiple methods and judges, provide greater validity and coverage of a curriculum.

Key words: educational assessment; competence; performance evaluation; programmatic assessment

INTRODUCTION

Assessment plays a major role in how students learn, their motivation to learn, and how teachers instruct. It may seem easy for clinician-teachers to determine whether a trainee has met the criteria to complete an educational experience ("I know it when I see it"). After all, we have been exposed to assessment since the age of childhood and are implicitly expected to understand the basics of assessment by the time we take on teaching responsibilities. The reality is that many clinician-teachers and the trainees themselves may soon realize their understanding of assessment is insufficient, especially when it comes to understanding the overall purpose and underlying principles of assessment. The following article will provide a broad overview of the fundamentals of assessment which will be relevant to both the trainee and faculty given that both effective receivership and delivery is integral to achieving the overall goals of assessment.

Principles of Assessment

Assessment can be defined as "the process of collecting, synthesizing and interpreting information to aid decision-making"[1]. Assessment defined in this way appears to be a simple process, until you are faced with the challenge of making it work in practice. There are three important principles everyone should know before entering a conversation about assessment.

Purposeful

Firstly, in choosing or designing assessment tools, it is critical to articulate the purpose of the assessment. Though the terms are often used interchangeably, 'assessment' can refer to either formative or summative depending on the intended purpose. Classically, formative assessment occurs during an educational experience, and summative occurs at the end of an educational experience.

Table 1: Purposes of Assessment

Assessment for learning	Formative, low stakes, often informal and opportunistic by nature and is intended to stimulate learning
Assessment of learning	Summative, medium or high stakes and intended to respond to the need for accountability

Dr. Daniel Dubois
Department of Anesthesiology and Pain Medicine,
University of Ottawa
1053 Carling Ave
Room B311
Ottawa
Ontario
K1Y 4E9
CANADA

Assessment of learning is the traditional summative assessment which is familiar to all of us. This may take the form of a grade or formal report card which sums up the learners' attainment of the objectives of the curriculum. It often requires coherent, high quality test material, a systematic standard-setting process, and secure administration. On the other hand, assessment for learning is formative and informs both the learner and teacher about the learners' progress towards attaining the objectives of the curriculum and provides insights to guide further learning. It is important to recognize that both summative and formative assessment indicate the purpose of assessment, not the method. A distinction should be made between assessments that are suitable only for formative use and those that have sufficient psychometric rigor (validity-coherence, reproducibility-consistency) for summative use. This distinction is especially important when developing high-stakes assessments (i.e., licensing and certification examinations).

Goal Oriented

Achieving competence (e.g., independent professional practice) is a longitudinal process that requires a sampling of knowledge, skills, and attitudes across all the domains required of professional practice. One of the most broadly used frameworks for the assessment of competence is Miller's Pyramid. Miller's model provides a framework for understanding the hierarchical progression from "knows" to "knows how" to "shows how" to "does". Miller's ideas strive to define education by its outputs and not by its inputs. It argues that to truly know whether our learners are achieving what we want them to achieve we should assess them in the setting that we expect them to be delivered[2]. It is important to appreciate that the assessment arising from each individual domain helps generate a small window into the overall understanding of how a learner is progressing from "knows" to "does" over the course of a training program.

Validated

All assessments should aim to facilitate acceptable and defensible decisions about the individual being assessed. To make these decisions, evidence needs to be evaluated to understand the strengths and weaknesses of the assessment in question. There are many ways to judge the quality of an assessment. Historically, there was emphasis on the measurement properties of the test alone (reliability and validity). Reliability is a measure of the reproducibility of the scores of an assessment, so that the outcome is the same if the assessment is repeated over time. Validity is not an inherent property of the test itself, but rather refers to the use of a test for a particular purpose. Messick and later Kane have proposed the most widely cited frameworks on validity. They evaluate the fundamental claims, assumptions, and inferences linking assessment scores with their intended interpretations and uses[3,4]. Cees van der Vleuten expanded the list of qualities, pushing beyond the traditional measurement characteristics to include issues related to the test's effect, acceptability, feasibility, and impact on future learning[5]. More recently Cook has proposed a more practical approach to validation of Kane's argument which serves as a desirable concise review of modern validity[6]. All of the aforementioned validity criteria were reaffirmed and added to an international consensus statement of the 2010[7], and 2018 Ottawa Conference[8] which resulted in the following criteria and framework for good assessment outlined in table 2.

Table 2: International Consensus for Good Assessment

Criteria for Individual Assessment	Framework for a System 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.	**1. Coherent** The system of assessment is composed of multiple, coordinated individual assessments and independent performances that are orderly and aligned around the same purposes.
2. Reproducibility or consistency The results of the assessment would be the same if repeated under similar circumstances.	**2. Continuous** The system of assessment is ongoing and individual results contribute cumulatively to the system purposes.
3. Equivalence The same assessment yields equivalent scores or decisions when administered across different institutions or cycles of testing.	**3. Comprehensive** The system of assessment is inclusive and effective, consisting of components that are formative, diagnostic, and/or summative as appropriate to its purposes.
4. Feasibility The assessment is practical, realistic and sensible, given the circumstances and context.	**4. Feasible** The system of assessment and its components are practical, realistic, efficient, and sensible, given the purposes, stakeholders, and context.
5. Educational effect The assessment motivates those who take it to prepare in a fashion that has educational benefit.	**5. Purposes driven** The assessment system supports the purposes for which it was created.
6. Catalytic effect The assessment provides results and feedback in a fashion that creates, enhances, and supports education; it drives future learning forward.	**6. Acceptable** Stakeholders in the system find the assessment process and results to be credible and evidence-based.
7. Acceptability Stakeholders find the assessment process and results to be credible.	**7. Transparent and free from bias** Stakeholders understand the workings of the system and its unintended consequences are minimized.

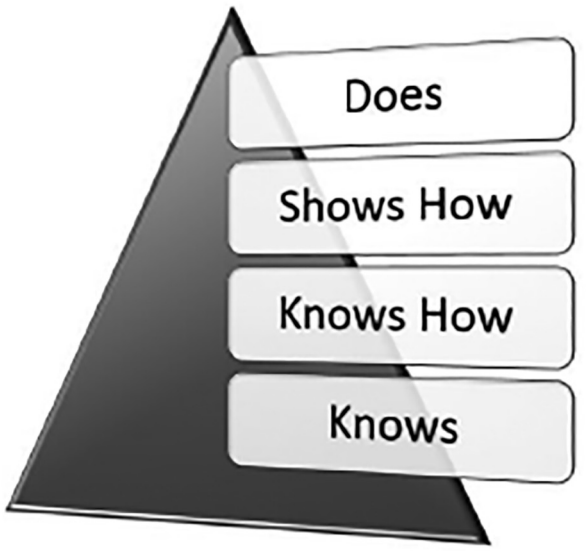

Figure 1: Miller's model of competence (adapted from Miller: Academic Medicine 65(9 suppl): S63–7, 1990.).

Types of Assessment

We will now look at the different methods of assessment available to assess clinical skills and behaviors in academic or workplace settings. We will discuss how and why they are used (e.g., formative versus summative), and some of the practical aspects (e.g. environmental and resource constraints) to be onsidered for educators wishing to make use of them. In the process of selecting or designing an assessment approach, instructors should consider the following questions.

1. What are the learning objectives that the assessment seeks to evaluate?
2. What are the skills and abilities that students need to do well?
3. Have students and faculty been adequately prepared to meet assessment expectations?
4. How will this assessment be utilized to enhance the student learning process?

Written Examinations

Multiple-choice questions (MCQs) are commonly used for assessment because they can provide a large sampling of examination items and from a testing perspective can be efficient since selected-response items take relatively little time to correct. Having a group of experts contribute to test creation provides different perspectives allowing broad representation, the elimination of non-contributory questions and the ability to have the final questions validated by the group[9]. Modified essay and short answer questions have the advantage of assessing clinical reasoning but are less common given the disadvantages of question marking and standardization. Though time consuming, a blueprint for item construct can minimize gaps in assessment content through appropriate sampling of all objectives. This blueprint should be shared well in advance of the examination with anyone being assessed. If using a "pass-fail" summative approach valuable assessment data is discarded along the way; including the information about the answers not chosen by the learner, the specific questions that were answered correctly versus those answered incorrectly, and even percentage correct. Progress testing for formative use is a method wherein you can sample the content of your curriculum in an ongoing continuous fashion by administering progressive tests prepared from a single large item bank itemized according to the blueprint of content areas. Students are given the results of the test to help identify knowledge gaps, and on repeat testing over the course of their training will hopefully demonstrate continued progress in their overall knowledge over time.

Structured Clinical Exams

The oral exam is a "knows how" traditional form of assessment in which one or more examiners deliver questions to a candidate to attempt to assess the candidate's knowledge of a subject, depth of understanding and to test clinical reasoning skills. They have been criticized for lacking structure and standardization, having poor inter-rater reliability, and potential examiner bias[10]. The structured oral examination (SOE) and objective structured clinical exam (OSCE) now exists to remove some of the traditional bias through use of standardized scoring rubrics and multiple scenarios or examiners[11]. To improve the validity a minimum of 10 stations is necessary to achieve a reasonable reliability for summative examinations[12]. The observing faculty member uses either a checklist of specific behaviors or a global rating form to evaluate the student's performance. To limit any subjectivity in this regard the criteria for answers provided in a scoring rubric will ensure clear guidelines on what is and not an acceptable answer. Faculty development for examiners on the appropriate usage of these rubrics should be provided to develop a shared mental model. The advantages of using this assessment approach is the sampling of competencies or procedures which are normally difficult to assess under conditions with high fidelity and patient safety. Important practical points to consider when administering and setting up an OSCE are that it's both time and resource heavy. Costs do vary significantly and can be mitigated with employing lower fidelity approaches, and volunteerism.

Workplace Based Assessment

The ability to "show how" can be accomplished through simulation but can also be accomplished with limited resources using work-based assessments (WBA). WBA allow for demonstration and observation of performance in the workplace. Faculty are asked to record their assessment of students on a checklist or rating scale. It is especially helpful if the rater includes narrative comments with their ratings. Direct Observation of Procedural Skills (DOPS), involves the direct observation and scoring of performance on a rating scale. Physicians-in-training are given the list of procedures for which they will need to be assessed in advance. For a resident anesthesiologist, typical procedures might include endotracheal intubation or arterial cannulation. The sequence of WBA is outlined in table 3 and typically involves a 15-minute, direct observation and a 5-minute, structured feedback session. The mini-clinical examination (mini-CEX) is another version of work-based assessment for faculty to assess physicians-in-training as they interact with patients. Research in the use of WBA suggests mini-CEX assessments are similar to simulation-based assessments but with higher fidelity and lower cost[13].

Table 3: Process for any work-based assessment (WBA)

1. Direct Observation
2. Learner Self-Evaluation
3. Structured, faculty-driven written and verbal feedback
4. Development of an action plan for improvement
5. Follow-up with frequent WBA [with different faculty assessors]

Direct observation and timely communication can be powerful sources of feedback to students. If the ratings are used for formative purposes the feedback generated can be used to improve performance on subsequent attempts. If the ratings are used solely for summative purposes the student may be encouraged to hide their weaknesses and limit any benefits gathered through the feedback process. The other major issue in the assessment of students by faculty is the lack of reliability of the faculty assessor. The scores may be biased by the different standards and understanding by individuals completing the ratings[14].

Portfolios

A portfolio might serve to organize assessment data into readily accessible fashion for review at regular periodic coaching meetings. Multiple assessment data points may be used to build a developmental portfolio in which students and their coaches can follow the student's progress from novice to expert. It is important to specify what to include in portfolios as doctors will naturally present their best work, and the evaluation of it will not be useful for quality assurance. In addition, if there is a desire to compare doctors or to provide them with feedback about their relative performance, then all portfolios must contain the same data collected in a similar fashion[15]. Otherwise, there is no basis for legitimate comparison or benchmarking.

Programmatic Assessment

A program of assessment is used to collect and combine information from various assessment sources to inform about the strengths and weaknesses of each individual learner. This helps mitigate limitations in a single assessment as the weaknesses or deficiencies of some instruments can be compensated by the strengths of other instruments. Multiple sampling through various assessments leads to a diverse spectrum of complementary measurement tools to better understand competence as a whole[16]. When reviewing a student using programmatic assessment, individual data points, garnered from individual assessments, are maximized for learning and feedback value. Whereas high-stakes decisions on a learner's competency are based on the aggregation of many data points. Thus, no high-stakes decisions are made without a detailed collection of information that is supported by thorough measures to ensure their reliability. Programmatic assessment considers assessment to be as important as the curriculum itself, thus requiring intense planning and review.

Challenges in Assessment

Despite the educators' best intentions, problems can develop when they attempt to assess trainees. It is generally acknowledged that assessment drives learning; however, assessment can have unintended consequences for both the trainee and the program. Some of the common problems you may encounter are discussed below and outlined in Table 4.

1. Incoherent Approach

If the assessment process is not integrated into the curriculum, it may produce data that are not meaningful or that inappropriately skew the direction of the curriculum. Devising and operating an assessment system using a comprehensive blueprint that maps learning objectives to multiple assessment tools ensures the program is robust and there are no gaps in the assessment process. A key principle of the approach is that individual data points are maximized for learning and feedback value whereas high stake decisions are based on an aggregation of many data points. Thus, each assessment point is optimized for learning using meaningful feedback but the key decisions about progress on the program are never taken on single assessment points but only on an aggregation of points. Data collected early in the curriculum and at the midpoint can provide feedback about learners in trouble to both educators and the learners in question. A formal system for addressing failures can prevent the creation of ad hoc solutions, along with their potential challenges.

2. Insufficient Data

An assessment system should generate a meaningful amount of data. To aid busy faculty and learners, data acquisition processes should be simple and automated. Learners should be encouraged to seek out feedback, and mandatory activities that require assessment data should be implemented. Many assessment programs pursue objectivity over subjectivity as it is easier to summarize and compare objective information but choosing to ignore the details of well-gathered subjective evaluations discards the great value of this subjective information. Using quantitative and qualitative data in combination can bring greater meaning to learner assessment. Do not assume that quantitative data are more reliable, valid, or useful than qualitative data.

3. Biased Assessment

This is perhaps one of the most challenging pitfalls of assessment. The problem is complex and related to the amount of time required to provide precise data; the inherent biases in our assessment, and the reluctance of supervisors to provide necessary "negative" assessments because of the emotional, personal, and professional repercussions that assessments may have. All tools are only as good as the people who are using them. The best way to accurately assess learners is not found in a holy grail assessment tool, but rather to have a wide range of faculty members use multiple efficient tools to assess a learners' performance in several different context. This approach, in combination with faculty development sessions and the use of well-defined scoring rubrics with criterion referencing, may provide a foundation for improving the reliability and validity of assessment data.

4. Evaluation Fatigue

The administrative setup required to manage a robust and defensible system must be planned and supported, with due regard for costs. The issue of having sound educationally based systems and high quality needs to be balanced against time and financial costs. Not everything

Table 4: Challenges in Assessment

1. Incoherent Approach
2. Insufficient Data
3. Biased Assessment
4. Overburdened Practices

that can be measured, needs to be measured. Many assessments are highly predictive of each other and of subsequent similar assessments. Consequently, designing the system of assessment with the aim to limit redundancies or assessments of low educational yield should reduce the resources needed to run them and make assessment more feasible.

CONCLUSION

Assessment is a central feature of teaching and when done well is a powerful catalyst for learning. Ideally, any assessment should enhance a student's capacity for learning and engagement with the curriculum. Assessment should therefore aim to reinforce students' intrinsic motivation to learn and to inspire them to set higher standards for themselves. However, this doesn't always happen in practice as it is heavily dependent on the form of assessment and whether timely comprehensive feedback is given. These elements are in sharp contradistinction from established practice where assessment measures are often applied in isolation or at least in an uncoordinated fashion. These uncoordinated measures are often combined to reach an overall decision based on weights dictated by tradition. A system of assessment explicitly blends single assessments to achieve the different purposes (e.g., formative versus summative; high vs. low stake) for a variety of stakeholders (e.g., students, faculty, patients, regulatory bodies). Those involved in educating, training, and certifying anesthesiologists should build in a systematic approach to assessment with evidence to support the validity and reliability of their approach. The advanced skills of a specialist can often be difficult to evaluate, finding effective assessment methods, especially systems of assessment that can eventually lead to more capable practitioners is necessary to effect long-term improvements in practice.

REFERENCES

1. Airasian P. Classroom Assessment. 3rd ed. New York (NY): McGraw-Hill, 1997.
2. Miller G. The assessment of clinical skills/competence/performance. *Acad Med*. 1990; **65(9 Suppl)**: S63-7.
3. Messick S. Standards of validity and the validity of standards in performance assessment. E*ducational measurement: Issues and practice.* 1995 Dec; **14(4)**: 5 8.
4. Kane M. The argument-based approach to validation. *School Psychology Review*. 2013 Dec1; **42(4)** :448-57.
5. Van Der Vleuten CP. The assessment of professional competence: developments, research and practical implications. *Advances in Health Sciences Education*. 1996 Jan 1; **1(1)**: 41-67.
6. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to K ane's framework. *Medical education*. 2015 Jun; **49(6)**: 560-75.
7. Norcini J, Anderson B, Bollela V, Burch V, Costa MJ, Duvivier R, Galbraith R, Hays R, Kent A, Perrott V, Roberts T. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. Medical teacher. 2011 Mar 1; **33(3)**: 206-14.
8. Norcini J, Anderson MB, Bollela V, Burch V, Costa MJ, Duvivier R, Hays R, Palacios Mackay MF, Roberts T, Swanson D. 2018 Consensus framework for good assessment. *Medical teacher*. 2018 Nov 2; **40(11)**: 1102-9.
9. Bandaranayake RC. Setting and maintaining standards in multiple choice examinations: AMEE Guide No. 37. *Medical Teacher*. 2008 Jan 1; **30(9-10)**: 836 45.
10. Maxim BR, Dielman TE. Dimensionality, internal consistency, and interrater reliability ofclinical performance ratings. ***Med Educ*** 1987; **21**: 130-137.
11. Anastakis, D.J., Cohen, R. and Reznick, R.K., 1991. The structured oral examination as a method for assessing surgical residents. *The American journal of Surgery*. 162(1), pp.67-70.
12. Khan KZ, Gaunt K, Ramachandran S, Pushkar P. The objective structured clinical examination (OSCE): AMEE guide no. 81. Part II: organisation & administration. *Medical teacher*. 2013 Sep 1; **35(9)**: 1447-63.
13. Norcini J, Burch V. Workplace-based assessment as an educational tool: AMEE Guide No. 31. *Medical teacher*. 2007 Jan 1; **29(9-10)**: 855-71.
14. Noel GL, Herbers JE Jr., Caplow MP, Cooper GS, Pangaro LN, Harvey J. How well do internal medicine faculty members evaluate the clinical skills of residents. *Ann Intern Med*. 1992; **117**: 757–65.
15. Van Tartwijk J, Driessen EW. Portfolios for assessment and learning: AMEE Guide no. 45. *Medical teacher.* 2009 Jan 1; **31(9)**: 790-801.
16. Vleuten CP, Schuwirth L, Driessen EW, Govaerts MJ, Heeneman S. Twelve tips for programmatic assessment. *Med Teacher*. 2015; **37(7)**:641-6.