Chapter 11
Assessing Student Performance

Brian Mavis

Assessment bridges the gap between teaching and learning. Perhaps second only to teaching, assessing student performance is a fundamental role in the life of a teacher. Assessment is important because it provides students with feedback about their performance; this information reinforces their areas of strength and highlights areas of weakness. Using this feedback, students can direct their study strategies and seek additional resources to improve their performance.

From the perspective of the teacher, another equally important function of student assessment is providing evidence necessary for decisions about student progress. The various student assessments within a class define the types and levels of achievement expected of students. As part of a course of study, student assessments describe a developmental process of increasing competency across a range of domains deemed necessary for graduation.

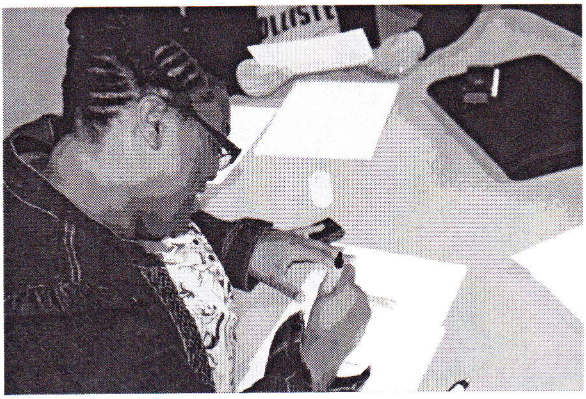
Any thoughtful teacher realizes the important role that student assessments play in their lives as teachers as well as in the lives of their students. Less obvious are the principles of educational measurement underlying sound student assessment practices. The purpose of this chapter is to provide an overview of some of the key features to consider when choosing among various student assessment strategies. This chapter will also provide information on how to create fair student assessments, that is, assessments that are both reliable and valid.

Reasons for Assessing Student Performance

As stated above, the assessment of student performance provides feedback to students about what they have or have not learned, and provides information that teachers can use in student progress decisions. However, these are only two of many possible goals that can influence your selection of student assessment strategies. As you can see by the list below, the goals that can drive the selection of student performance measures are many and far reaching:

B. Mavis (✉)
College of Human Medicine, Michigan State University, East Lansing, MI, USA

W.B. Jeffries, K.N. Huggett (eds.), *An Introduction to Medical Teaching*,
DOI 10.1007/978-90-481-3641-4_11, © Springer Science+Business Media B.V. 2010



- Providing feedback to students about their mastery of course content
- Grading or ranking students for progress and promotion decisions
- Offering encouragement and support to students (or teachers)
- Measuring changes in knowledge, skills or attitudes over time
- Diagnosing weaknesses in student performance
- Establishing performance expectations for students
- Identifying areas for improving instruction
- Documenting instructional outcomes for faculty promotion
- Evaluating the extent that educational objectives are realized
- Encouraging the development of new curriculum
- Demonstrating quality standards for the public, institution or profession
- Articulating the values and priorities of the educational institution
- Informing the allocation of educational resources

Clearly, many of these goals are related directly to the interaction between teacher and student. Nonetheless, this same information can be used by other stakeholders in the educational process for a variety of other important decisions. From a practical perspective, it is unlikely that any single assessment strategy can provide information to support more than a few of these goals. The likelihood for misinterpretation or inaccuracy increases when student assessment data are used for purposes other than those originally intended. The sheer breadth of the list of goals above also demonstrates the importance of considering the use of multiple strategies to assess student learning.

In the curriculum development cycle presented in Chapter 10, student assessment follows instruction and is the impetus for reflection, evaluation and curricular improvement. While this cycle might reflect how many teachers approach teaching, from the student perspective, it is the assessment phase of the curriculum that drives learning. This is most often manifest when students ask, "do we have to know this for the test?" For many students their motivation is survival, and in an educational setting, one key element of survival is passing the test, whatever form student assessments might take. This is particularly true when results from assessments will be used for student progress decisions.

Learning the Language of Assessment: A Few Definitions

Before getting much deeper into a discussion of student assessment, it is important that we clarify a few definitions of key terms and their meaning in this context.

Assessment Versus Evaluation

Both assessment and evaluation refer to processes of gathering information for the purposes of decision-making. In medical education, assessment most often refers to the measurement of individual student performance, while evaluation refers to the measurement of outcomes for courses, educational programs or institutions. Practically speaking, students are assessed while educational programs are evaluated. However, it is often the case that aggregated student assessments serve as an important information source when evaluating educational programs.

Formative Versus Summative Assessment

Formative assessments are used to give students feedback about their learning. Practice test questions or problem sets, in-class peer-graded assignments and reviews of video recorded simulated patient encounters are examples of frequently used formative assessment strategies. Formative assessments are most valuable when they are separated from summative assessments, so that they are perceived to be low threat performance experiences. For conscientious students, this represents an opportunity to document both strengths and weaknesses. However, some students might dismiss formative assessments for their lack of consequences, and not put their best effort forward to use these experiences to maximal advantage.

Summative assessments are used to gather information to judge student achievement and to make student progress decisions. These assessments are very familiar to students and teachers, and for students often provoke anxiety. A substantial component of this anxiety comes from the student progress decisions that are predicated on performance. However, to the extent that uncertainty about the summative assessment strategy itself is a source of anxiety, teachers can take steps to reduce student anxiety. This includes providing information about the types of assessments to be used, their timing within a course, how they are scored and how each contributes to the final grade or progress decision. Students often become anxious in an unfamiliar assessment situation, such as a standardized patient encounter or new computer-based testing software. Sample interactions, in-class demonstrations or opportunities for non-graded practice can help students anticipate what to expect under these circumstances, which might help reduce their anxiety.

Competence

It is increasingly common in medical education for discussions of student assessment to lead to discussions of competence. One current definition of competence provided by Epstein and Hundert (2002) gives us a sense of the tip of the iceberg implied by these discussions. They wrote that "professional competence is the habitual and judicious use of communication, knowledge, technical skills, clinical reasoning, emotions, values and reflection in daily practice for the benefit of the individual and community being served." From the practical perspective of an educator, competence requires us to set expectations of satisfactory performance appropriate for the students' progress within the curriculum.

This approach to thinking about competence has been articulated by George Miller (1990), who has described a developmental model that is helpful when thinking about appropriate forms of learner assessment. For novice learners, competence is determined by what students *know*, i.e., their mastery of factual knowledge. At the next level, competence is defined by an assessment requiring students to demonstrate what they *know how* to do, such as how to use and apply knowledge to solve new problems, or demonstrate the clinical skills necessary to gather clinical data. At the next level, *showing how* in an assessment setting would be required to demonstrate competence. Assessments at this stage would require students to actually demonstrate their ability to acquire, interpret and translate knowledge. At the highest levels of competence, students would demonstrate competence by *doing*, which would be assessments of how they perform in an encounter with a patient in a real world setting.

Key Features of Student Assessment Methods

There are a number of factors to consider when choosing a method of student assessment. Attention to these factors at the planning stage will go a long way to helping you create a high quality student assessment. Five factors to consider are:

1. *Reliability*
2. *Validity*
3. *Feasibility*
4. *Acceptability*
5. *Educational Impact*.

Reliability and validity are characteristics that generally refer to the development of an assessment. Feasibility, acceptability and educational impact more often reflect contextual features of the assessment, which are related to when and how an assessment method is implemented.

Reliability

When talking about the reliability of an assessment method, we are referring to the consistency or repeatability of measurement. In practice a reliable assessment should yield the same result when given to the same student at two different times or by two different examiners. One of the advantages of tests comprised of multiple-choice questions is that they are highly reliable: the results of the test are unlikely to be influenced by when the test is administered, when the test is scored or by who does the scoring. Hence the term "objective" is often used when referring to these kinds of assessments. On the other hand, reliability is an important concern when grading essay questions, rating clinical skills or scoring other assessments requiring judgment or interpretation. In these situations, clear scoring criteria are needed to attain a high level of reliability, regardless of whether one or multiple people will be involved in grading the responses. Writing clear test questions and test instructions are important strategies for improving the reliability of an assessment by reducing the likelihood that test questions are ambiguous to the reader and open to multiple interpretations. Writing clear test questions also increases the likelihood that the assessment is testing desired knowledge, skills or attitudes rather than reading proficiency or verbal reasoning skills.

Internal consistency is another form of reliability that is more often used to describe assessments based on multiple-choice questions. This term refers to the coherence of the test items, or the extent to which the test questions are interrelated. The primary difference between this and other estimates of reliability is that calculations of internal consistency involve only one administration of the test. For example, a set of multiple choice questions focused on assessing students' knowledge of childhood immunizations should have high internal consistency. When questions testing other knowledge or abilities are added, the internal consistency is lowered. Internal consistency estimates are frequently provided as part of the output for machine-scored multiple choice tests. The concept of internal consistency can be applied to other methods of assessment; this is best done in consultation with a measurement specialist.

Validity

Validity is the extent to which an assessment measures what it is intended to measure. Validity is related to reliability, insofar as a test that has low reliability will have limited validity. A test with low reliability is subject to biases in interpretation and scoring, and when these biases are unrelated to specific content or student performance, the validity of the assessment is diminished.

Among the many types of validity discussed in education and social science research, content validity is the approach most commonly used to assure quality in student assessment. Essentially, an assessment is valid when it samples representatively from the course content. A common method for assuring that the assessment content is representative of the course content is to develop a table of specifications, often referred to as a test blueprint. The blueprint organizes course content by course objectives, such as students' ability to recall factual information, understand concepts or apply knowledge to new problems. Another approach to organizing content could be based on patient cases (well child visit, asthma, developmental delay, etc.), or by disciplines (pathology, physiology, pharmacology, nutrition, etc.). In reality, any organizing structure that reflects the logic of the course content can be used as the basis of the blueprint (Table 11.1).

Table 11.1 Sample blueprint for a clinical competence examination

Clinical competency	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
Communication skills	x	x	x	x		x
History-taking		x	x	x		x
Physical examination		x	x	x		x
Data interpretation			x	x	x	
Assessment/diagnosis			x	x	x	x
Patient education	x					
Written record	x		x	x	x	x

Ideally, the course content was initially designed around a blueprint based on the course objectives. In this way, the organization of the course informs the organization of the assessment content, creating a valid assessment. In situations where there are no preexisting course objectives, it might be necessary to derive them from the content and reverse engineer a blueprint prior to creating the assessment.

Content based on a blueprint approach works well for assessments that focus on recall and application of knowledge. This approach also can be extended to assessments of clinical skills. The specific steps used to teach communication, history-taking and physical exam skills define the content, and a rating form can be developed to judge student performance of these skills. Another approach frequently used for the assessment of skills is expert judgment. A number of practitioners or "experts" can be polled to determine what they would identify as the key components of a specific skill, and this consensus is the basis for the checklist developed to rate student performance.

In general, sampling assessment content from the same blueprint that was used to define instructional content can enhance the validity of an assessment. It might be necessary to use multiple methods of assessment when complex performance is the focus of assessment. Further, some assessment methods are more appropriate for some types of performance than others, so choosing appropriate methods can increase validity.

Feasibility

The feasibility of an assessment method is a judgment of the resources needed to implement it in light of the information to be gained. The development of a multiple choice test requires significant time in the development phase for question writing, but requires relatively little effort for administration or scoring. Conversely, a comprehensive assessment of clinical skills might require as many resources for scoring as for development and implementation, though the types of resources required might be different for each phase. Essentially, any assessment requires time for development, implementation and scoring. Additional resources might include examiners and/or simulated patients, as well as training time for each. Proctors, computers, biological samples, clinical case information, curricular time and building space are among other possible resources that can facilitate or constrain an assessment.

Acceptability

Consider a teacher weighing the use of three assessment strategies: weekly tests, a more traditional mid-course and final exam, or a single cumulative examination at the end. Each approach has its merits as well as limitations, depending on the purpose of the assessment. The acceptability of an assessment is based on the responsiveness of faculty and students to the assessment. If the assessment requires too much time from faculty and staff or requires too many resources to implement, the long-term survival of the assessment is jeopardized. Similarly, an assessment approach might be aversive to students because of the timing, length, content or other features. When this occurs and students do not prepare as expected for the assessment or do not see it as valuable to their education, the validity of the assessment can be jeopardized.

Educational Impact

The impact of an assessment is the sum of many influences. The intent of the assessment relative to the course objectives is a consideration. The thoughtful use of both formative and summative assessments can positively affect student learning and subsequent student performance. Educational impact also reflects the appropriateness of the match between the content and the assessment method; a mismatch reduces the educational impact. Since content and assessment method are linked, the use of multiple assessment methods can enhance the impact. Relying on a single method tends to focus assessment on the content most amenable to the method. The method can also influence how students prepare for an assessment, such as the differences in preparation for a multiple choice exam versus a standardized patient encounter.

Choosing an Assessment Method

Consider this example. As part of the neuroscience curriculum, second-year medical students were asked to identify the location of a lesion based on a written case included as part of their final examination, constructed from multiple-choice questions. The correct response was chosen by 88% of the class. The next day when the same question was presented in a different format within the context of a simulated patient encounter only 35% of the class got the answer correct. The conclusions about student performance based on each assessment would be very different. This example reflects the impact of assessment method in terms of both format and cognitive demands on the student.

When choosing an assessment method, there are several factors to consider. One of the first considerations is the type of performance to be assessed: is the assessment focusing on knowledge, skills or attitudes? Another factor that might influence the choice of assessment method is whether it is being used for formative or summative assessment. A related concern is the reliability and validity of the method. Some methods are more practical than others when considering the resources required to achieve reliable and valid results. Another consideration is whether more than one assessment approach should be used. When choosing an assessment method, it is important to remember that no single method "does it all." For this reason, a multiple-methods approach will probably provide a more accurate picture of student performance or achievement than relying on a single approach. As educators this idea makes intuitive sense; in practice we tend to stick to what is familiar.

The chart below summarizes the relative strengths of various assessment methods when measuring different types of performance. The chart provides guidance in choosing assessment methods, but is not intended to be absolute in matching methods and performance (Table 11.2).

Methods of Student Assessment

There are a wide range of methods available when developing your approach to student assessment. Presented below are assessment methods most common to medical education and while not exhaustive, this list represents a wide range of options available to faculty. The methods described in this section are organized into four broad categories: assessments based on written exercises, assessments derived from faculty ratings of performance, simulation-based assessments, and methods of global performance assessment. Each of the methods is described in terms of (a) strengths, (b) limitations, (c) reliability and validity and (d) construction tips.

Written Exercises

- Multiple choice questions (MCQs)
- Extended matching questions
- Short answer questions
- Essays and structured essay questions

Faculty/Preceptor Assessments

- Faculty global ratings
- Faculty checklist ratings
- Oral examinations

Table 11.2 Strengths of various assessment methods

	Types of student performance										
Assessment methods	Knowledge recall	Knowledge application	Communication skills	History-taking skills	Physical examination skills	Procedural skills	Professionalism	Team work	Written documentation	Attitudes	Self-reflection
Written exercises											
Multiple choice questions	X	X									
Extended match questions	X	X									
Short answer questions	X	X									
Essay questions		X							X	X	X
Faculty/preceptor assessments											
Faculty global ratings			X	X	X	X	X	X	X		
Faculty checklist ratings			X	X	X	X	X	X	X		
Oral exams	X	X								X	X
Simulated clinical encounters											
OSCE/standardized patients			X	X	X	X	X	X	X	X	
Technology-based simulations					X	X		X			
Other global assessments											
Peer assessment		X	X				X	X		X	
Self-assessment							X	X		X	X
Portfolios		X	X						X	X	X

Simulated Clinical Encounters

- Standardized patients and OSCEs
- Technology-based simulations

Other Global Assessments

- Peer assessments
- Self-assessments
- Portfolios

Multiple Choice Questions (MCQ)

Assessments based on multiple choice questions (MCQ) are one of the most common approaches to measuring student performance. Typically, a multiple choice question consists of two parts: the question (stem) and the possible answers (response options). Most MCQs include four or five response options and the student is asked to choose the best response. The stem also can make reference to tables, graphs or other information sources that the student must use in order to determine the correct response.

> Which of the following best describes the path of air into the lungs of humans?
>
> (a) Alveoli, trachea, bronchi, bronchioles
> (b) Trachea, bronchi, bronchioles, alveoli
> (c) Bronchi, bronchioles, trachea, alveoli
> (d) Trachea, bronchioles, bronchi, alveoli
> (e) Alveoli, bronchioles, bronchi, trachea

Strengths

- Multiple-choice questions are familiar to most students, given their common usage throughout most levels of education.
- MCQs provide broad coverage of content. It is relatively easy to build an examination using MCQs that covers a wide range of course content.
- MCQs can be simply written to test for recall of factual knowledge, or can make reference to graphs, tables or illustrations to test cognitive skills. MCQs also can be posed within the context of a science problem or clinical case to test knowledge application and problem solving.
- Scoring of MCQs is highly reliable and objective.
- Scoring of MCQs can be automated, making scoring efficient and reducing the turnaround time for feedback to students. Automated scoring also facilitates the calculation of psychometric properties of each MCQ.

- MCQs are more flexible than True and False questions, which require absolute statements. MCQs are more flexible in terms of absolutes since the student is asked to choose the best answer from several possible options.

Limitations

- Good MCQs are challenging to write, especially for applications beyond knowledge recall such as knowledge application and problem solving. The time saved in scoring a MCQ examination is usually required up front in the preparation of the questions, which requires time to construct to avoid cueing students about the correct response option.
- MCQs frequently focus on recall of factual information and rely on students' recognition of the correct answer from among the options provided.
- Guessing can be a successful test-taking strategy for those questions where the student can rule out a number of the response options.
- MCQs are limited as a means of providing instructive feedback to students since usually the only information provided is the correct response option.
- The ease of use and economy of scoring associated with MCQs can lead to their overuse in situations when other types of assessment would be more appropriate.

Reliability and Validity

- Reliability tends not to be an issue for MCQs, which are the most frequently used objective test format. MCQ scoring is highly reliable in terms of consistency from one time to the next as might occur if the exams are scored over several sessions.

Scoring is consistent between examiners; scoring of MCQs is not dependent on who is scoring the exams.

- Validity of an exam based on MCQs is enhanced through the use of a test blueprint, which assures that the distribution and coverage of examination content matches the instructional objectives and major content areas.

Construction Tips

- The response choices should be relatively brief, with the major content elements of the question included in the stem. The content can include graphs, images, clinical scenarios, research findings or other complex information that requires interpretation.
- Write each response option so that it matches the grammar of the stem.
- Equally distribute the position of the correct response across a series of questions. For example, the correct answer should not always be the third response option. A strategy to avoid such bias is to always order your response options alphabetically or numerically. Knowing this, the students cannot expect position bias.
- Do not use "all of the above" or "none of the above" as a response option.
- Avoid questions worded with negatives or double negatives.
- The correct and incorrect response options should be about the same length.
- Avoid the use of response options that are irrelevant or silly. This increases the likelihood of guessing the correct response.
- One MCQ can be viewed as a cluster of true/false questions, where one response choice is true and all of the others are false. When using true and false questions, the number of true and false questions should be roughly equal, and care should be used to construct questions of approximately the same length. A limitation of true and false questions is that for questions where the correct choice is false, you do not have assurance that the students in fact know the true answer. Another limitation of true and false questions is that they require statements that are absolutely true or absolutely false. These types of questions are best used for assessing factual recall.

Extended-Matching Questions

The extended-matching question format was developed to address some of the limitations of the MCQ format. The major advantage over MCQs is that the larger number of response options reduces the likelihood that the question will cue the student to the correct answer; students also are less likely to recognize the correct answer. In many ways, the strengths and limitations of extended-matching questions are similar to those of multiple-choice questions.

Extended matching questions are organized around themes, and include multiple response options, instructions and a series of stems. Here is an example:

Theme: Endocrine glands and hormones

Options
A. Luteinizing hormone
B. Vasopressin
C. Calcitonin
D. Glucagon
E. Estrogen
F. Insulin
G. Testosterone
H. Melatonin
I. Norepinephrine
J. Prolactin
K. Oxytocin
L. Progesterone

Instructions For each statement below, select one hormone that best fits the description.

Stems
1. Secreted by the thyroid gland.
2. Stimulates ovulation and corpus luteum formation.
3. Lowers blood sugar.
4. Secreted by pineal gland.

Strengths

- This question format can be used to construct an exam covering a wide range of content.
- This question format can be used to test knowledge recall as well as knowledge application.
- Scoring is highly reliable and objective, and like MCQs can easily be automated for efficiency.
- This question format is often used to test recall of factual information; there is less of a chance of students recognizing or guessing the right answer compared to MCQs. They can also be used to test problem-solving skills such as clinical diagnosis or patient management.

Limitations

- As with MCQs, time and practice are needed to write good questions that take advantage of the strengths of this format but do not cue the respondent.
- Similar to MCQs, these questions provide only minimal feedback to students to enhance their learning. Some examination software applications allow additional feedback either during or following the examination.

Reliability and Validity

- Like MCQs, this item format has high reliability because of the consistency of scoring over time or across examiners.
- The validity of an exam using extended matching questions is based on the representativeness of the test content compared to the instructional content. Like MCQs, questions derived from a test blueprint can assure a fair test in terms of content.

Construction Tips

- Extended matching questions are usually written around a theme. When the theme is based on a clinical scenario, research abstract or an image, the questions can require students to recall knowledge, interpret findings or suggest possible diagnoses.
- The response choices should be relatively brief, with the major content elements included in the questions.
- Avoid questions worded with negatives or double negatives.
- Avoid the use of response options that are irrelevant or silly.

There is an excellent resource for extended-matching questions available at no charge from the National Board of Medical Examiners website (www.nbme.org). Under publications, look for "Item Writing Manual: Constructing Written Test Questions for the Basic and Clinical Sciences" by Susan Case and David Swanson.

Essays and Modified Essay Questions

These types of questions are characterized by the requirement that the student constructs a response rather than choose a correct response from one or more options provided. Essay and modified essay questions provide an opportunity to assess student's ability to apply knowledge to solve problems, organize ideas or information, and synthesize information. A sample essay question might be,

> You are treating Sandy, a 57 year old woman who was diagnosed six months ago with Stage 2 adenocarcinoma of the right lung. Until a few days ago, her pain has been well-controlled. You have reevaluated the pain control and decided to initiate treatment with sustained release oral morphine. Sandy's brother is coming to the next appointment; he has concerns that his sister will become addicted to the pain medication. What will you say to Sandy's brother?

Modified essay questions are an assessment format that addresses some of the limitations of the essay question. A modified essay question is made up of one or more short answer questions. The student is provided with basic science or clinical information and then asked to write brief responses to one or more questions. When a series of questions is presented, additional information about the original problem can be provided at each subsequent step, guiding the students through an analytical process.

David is a 26 year old computer programmer, who lives alone with his dog Max. He has come to your office complaining of a persistent cough.

1. What are three likely diagnoses?
   a.
   b.
   c.
2. List five specific questions that would help you distinguish among these possibilities.
   a.
   b.
   c.
   d.
   e.

David tells you that the cough started about 5 days ago, and that many people in his office have called in sick lately. He has felt feverish and had some chills yesterday evening. He has been coughing up a small amount of thick green sputum.

3. List two diagnostic tests appropriate for work-up of this case.
   a.
   b.

Strengths

- This question format can be used to test written communication skills.
- Essay questions can focus on content related to knowledge or attitudes.
- Essay questions are best used to assess depth of knowledge within a limited area of content.
- This question format is familiar to students, and fairly straightforward for faculty to construct.

- Essay and modified essay questions are well-suited for formative feedback, since students can be provided with a model answer to help them understand their performance and prepare for future assessments.
- Modified essay questions require less time to score than traditional essay questions. Because student responses tend to be shorter and more succinct, these types of questions are less subject to scoring bias and can provide broader content coverage, both of which increase the reliability and validity of the assessment. While essay questions can be used to assess higher levels of student cognitive ability, modified essay questions are ideal for testing knowledge recall that is not based on recognizing the correct answer as in MCQs.

Limitations

- Reliability is a major concern and there is a need to assure consistency of scoring over time and when multiple individuals are involved in scoring. Scoring of the written responses are more subject to general subjective biases of the scorer, often referred to as "halo effects." In practice halo effects occur when there is a possibility of giving some students the benefit of the doubt more often than other students while scoring written responses. An example might be students who are known to be strong performers, or students who have done well on other parts of the test might be given the benefit of the doubt more often compared to weaker performers. The possibility of halo effects is more likely when student identities are known to scorers or when a single scorer is used.
- More time is required for scoring these responses than other formats.
- Tests based on essay-type questions are more limited in their content coverage because of the length of time required for students to respond to the questions as well as the length of time required for scoring.
- Essay questions require at least minimal written communication skills, and if communication skills are not the focus of the assessment, a lack of communication skills might limit a student's ability to achieve a high score even if they know the content.

Reliability and Validity

- Reliability is a major concern with these types of questions. The individuals scoring written responses might need to make some inferences about what the respondent meant because of poor written communication skills including organization, grammar and vocabulary, or due to vague wording. The opportunity for inference tends to reduce reliability.
- Reliability can be increased by having a clear scoring scheme developed prior to grading the questions. One approach would be to create a model answer and then allocate points to specific features of the answer, such as mentioning specific key content, presentation of a logical argument, recognition of a counter-argument or

alternative explanations, or whatever else is appropriate to the question. When possible, all of the answers to one question should be scored at the same time, by the same person. If multiple people are scoring the exam, then each should grade all of the responses to a single question. Each essay question or set of modified essay questions should be graded independently of the other questions, and when possible the identity of the students should be unknown to the person grading the question to reduce the likelihood of bias.

- To create a test with high validity, it is important to make sure that the essay questions address important content as indicated by the course objectives and overall course plan. Having several content experts review the model answer to each question can strengthen the validity of the assessment. This is particularly true when the question asks students to integrate concepts from across several domains, which might not have been taught by the same instructor.

Construction Tips

- Write questions that outline a specific task for the students. Asking students to discuss a content area is not as clear or helpful as asking students to compare and contrast, describe, provide a justification or explain.
- To improve reliability and the sampling of course content, it is more effective to use a large number of modified essay questions requiring short answers than to use a more limited number of essay questions requiring long written answers.
- Prepare a model answer after constructing the test question. This helps to increase scoring consistency by assuring that the answer you expect is reasonable given the question, and clarifies how points are assigned to content and presentation of the answer.
- To reduce bias and improve consistency, score only one essay question or set of modified essay questions at a time, and if feasible have separate scorers for each essay question. When this is not possible, rescoring a small set of answers can help maintain consistency. The subset of rescored answers should be sampled from throughout the set of examinations to make sure that the application of the scoring criteria did not change over time.
- When possible, score the answers to the questions with the identity of the students anonymous to the scorer.
- When used for formative assessment, student learning can be enhanced by providing students with a model answer as well as feedback about common errors observed when scoring.

Short Answer Questions

Short answer questions require students to provide brief answers to questions. The responses usually require only one or two words or a brief phrase. Short answer questions are often presented as fill-in-the-blank questions.

1. A middle-aged financial planner presents with a several month history of stomach discomfort. He has found limited relief with over-the-counter antacids, although these are now less effective than before. His discomfort is aggravated by caffeine, alcohol and late night snacking. What is the likely differential diagnosis for this patient?
2. In planning the diagnostic work-up for this patient, list two tests you would definitely include to aid in your diagnosis.

Like essay questions, short answer questions require students to provide a response rather than choose (recognize) a response from list of possibilities provided. However, because short answer questions require briefer responses, more questions can be included within an exam, achieving greater content coverage than with essay questions.

Strengths

- High content coverage is possible.
- This question format has high reliability and validity.
- Faculty find it relatively easy to construct short answer questions.
- This question format is familiar to students.

Limitations

- These types of questions tend to focus on knowledge, and are used to test knowledge recall and comprehension rather than higher level abilities.
- Scored questions indicating the correct answers provide limited feedback to students to improve learning.
- Scoring cannot easily be automated: this question format requires more time to score than MCQs but less time than essay or modified essay questions.

Reliability and Validity

- Reliability can be achieved by writing questions that are clear to the student, as well as writing clear model answers to each question. The distribution of points for the responses should be clearly specified. While bias is less likely to apply to scoring short answer questions, the likelihood of halo effects can be minimized by the same procedures described for essay and modified essay questions.

- Using a blueprint to create short answer questions that representatively sample from the course objectives and content is important. As mentioned previously, having several content experts review the model answers can strengthen the validity of the assessment. This is particularly true when there might be multiple possible correct answers for a question.

Construction Tips

- Write questions that are clear and specific.
- Prepare the short answer questions and the model answers at the same time. Afterwards, reread the questions and answers again to assure that the expected answer is reasonable given the question.
- To reduce bias and improve consistency, score the all the answers to a single set of questions at the same time. Rescoring a small subset of answers can help maintain consistency throughout the scoring process.
- Score the answers to the questions with the identity of the students anonymous if possible.

Faculty Global Ratings

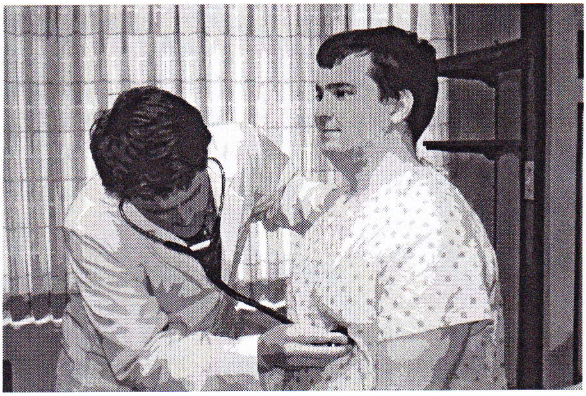
Faculty global ratings can be used to summarize impressions of overall performance within a defined situation such as a clinical encounter, or aggregated over time to represent typical behavior in key situations, such as a clinical rotation or small group setting. Global ratings are based on a number of key domains of performance or behavior, with a judgment of the extent to which desired performance was observed.

1. Rate the student's ability to obtain information about the presenting problem in this simulated clinical encounter.

    a. Obtained little or no information
    b. Obtained some information, but with major omissions or errors
    c. Satisfactory performance with most information obtained
    d. Very thorough exploration of patient's presenting problem

2. Rate the student's participation in your small group discussion section with regards to his/her ability to disagree with or question others without conveying disrespect.

    a. Never
    b. Seldom
    c. Usually
    d. Always



Strengths

- Global ratings by faculty tend to have high validity when they are based on the observation of behaviors of interest within a real or simulated context.
- This format can be used to assess general categories of performance such as clinical skills, problem-solving, teamwork, presentation skills and professionalism; participation and preparation in a small group learning context such as problem-based learning; or achievement of course objectives when used to rate group presentations.
- This format is useful in formative assessment to provide learner feedback.

Limitations

- Reliability can be variable when there are differences in expectations and standards among faculty providing the ratings.
- Faculty need to set aside time to observe students prior to completing the global ratings.
- Halo effects and other subjective biases are common.

Reliability and Validity

- Specific training to achieve consensus among faculty raters as to the rating categories and expectations for the ratings within a category is needed to achieve standardization. The rating form should have specific descriptions of the expected behaviors for each of the rating points within a category. To limit the impact of halo effects and other subjective biases, the use of multiple raters and multiple rating sessions is recommended.

- Validity is evident from the behaviors sampled on the rating form. Expert faculty can be used to determine those behaviors that represent the target performance in the learning context (e.g., PBL group, clerkship, etc.) where the form will be used.

Construction Tips

- The rating form should only include observable behaviors; these behaviors should occur at a frequency that makes them readily observable.
- Some people are better raters than others, depending on the rating task. Over time, it might be evident that some raters are more reliable than others. Using individuals who are reliable will improve the usefulness of the data obtained from the rating form for feedback and student performance decisions.
- The rating form should have specific descriptions of the expected behaviors for each of the rating points within a category.
- Halo effects and other subjective biases can be reduced through the use of multiple raters and multiple rating sessions.
- Student learning can be enhanced by providing students with copies of the rating forms to be used for assessment, indicating the categories of performance and the expected performance within each category.

Faculty Checklist Ratings

Global ratings as discussed above are frequently based on assigning scores on a multi-point rating scale for each of a variety of behaviors. In contrast, checklist ratings of direct observations are based on checklists that indicate the presence or absence of a specific behavior or a component of a behavior.

Please rate this student's performance of the following portions of the neurologic exam.

	Done correctly	Done incorrectly	Not done
1. Motor			
a. strength of arms	1	2	3
b. arms outstretched, eyes closed	1	2	3
c. strength of legs	1	2	3
2. Reflexes			
a. biceps (inside of elbow)	1	2	3
b. triceps (back of arm at elbow)	1	2	3
c. brachioradialis (wrist/forearm)	1	2	3
d. Patellar (knee)	1	2	3
e. Achilles (ankle)	1	2	3

Strengths

- Checklist ratings tend to have high validity because they are based on direct observation of specific behaviors of interest. The checklist represents a list of the specific skills expected of students.
- Checklist ratings can be used to assess specific skill sets such as clinical skills related to communication, history-taking, physical examination, or presentation skills related to a class project or clinical case.
- Checklist ratings provide specific feedback to learners about the elements of their performance judged to be present or absent.

Limitations

- Rater factors such as poor standardization, inconsistent expectations, and halo effects can reduce the reliability of the assessments.
- A limitation of rating forms is that for rating purposes, target skills are broken down into essential key elements. While this approach is appropriate and helpful for students learning a new skill, it is less appropriate for assessing the performance of more experienced students or practitioners.

Reliability and Validity

- Reliability can be improved by having clearly defined checklist items, and raters familiar with the skills to be rated. The less inference required of raters when completing the checklist, the greater the likelihood of reliable ratings. This also reduces the likelihood of halo effects.
- Validity is high for this approach because the rating forms are based on specific target behaviors, often broken down into key elements. For students learning new skills, this can provide feedback about specific components that were omitted or incorrectly executed. The items on the rating form can be based on the list of steps used to teach the skill.
- More advanced students and practitioners, with practice, will move beyond step-by-step performance of skills as they were initially learned to more integrated performance. More advanced learners are less likely to repeat the key elements in a rote fashion while still effectively performing the desired task. For this reason, when assessing the skills of advanced learners, global ratings might be more appropriate than checklists.

Construction Tips

- Reliability is improved when the number of options for each checklist item is limited. Frequently checklists are constructed with two or three options per item on the checklist, such as (1) done and (2) not done, or (1) satisfactory or

(2) unsatisfactory. Sometimes a third option might be included indicating that a step was attempted: (1) done correctly, (2) done incorrectly (3) not done. This format can be used when the student attempts a skill unsuccessfully, and there is reason to distinguish this from skills not attempted, such as when the rating form is used for formative feedback to students.

- If the checklist ratings are to be performed from memory, such as might occur when a standardized patient completes the checklist after a simulated encounter with a student, the number of total items on the checklist should be limited to what can reasonably be remembered by the rater.

Oral Examinations

An oral examination requires students to answer a series of preselected questions; these are typically based on standard stimulus information such as a patient case. Based on the patient information provided, the examiner can ask questions about differential diagnoses, pertinent missing data, additional testing, patient management as well as reasoning and interpretation of data underlying the student's responses. The length of time per case can vary depending on whether breadth or depth of understanding is desirable, as well as whether the exam is being used for formative or summative assessment.

> David is a 26 year old computer programmer, who lives alone with his dog Max. He has come to your office complaining of a persistent cough.
>
> 1. List three diagnoses that you would include in your differential diagnosis.
> 2. List five specific questions that would help you distinguish among these possibilities.
> 3. List two diagnostic tests appropriate for work-up of this case.
>
>   a. What is the rationale for each?

Strengths

- Oral exams can be used to assess knowledge and attitudes.
- This assessment format can be used to assess higher order clinical problem-solving such as application and synthesis of knowledge, ability to prioritize features of a patient case and evaluate treatment options.
- Oral exams provide insights into students' organizational and verbal skills.
- When used in formative settings, oral exams can be used to provide students with immediate feedback and provide instructors with information about students' approaches to problem-solving and reasoning.

Limitations

- Reliability can be problematic as a result of rater factors such as poor standardization, inconsistent expectations, and halo effects.
- Like essay exams, oral exams provide limited coverage of content and cases, which can limit the validity of the assessment.
- Oral exams require verbal and language skills, which can limit students' ability to communicate their content knowledge.
- This assessment format is not familiar to many students, which increases their anxiety.
- Time is required for scoring the results of an oral exam, particularly when a large number of examiners are involved.

Reliability and Validity

- Significant training of the examiners is required for reliability to be achieved. The training must address performance expectations and standards, as well as the use of structured rating forms to record student performance. The use of multiple examiners is recommended to reduce halo effects and other rater biases.
- Because oral examinations are limited in the amount of content that can be covered, longer exams are more valid than shorter exams. It is also important that the exam is standardized in terms of the content to be covered and the specific rating forms for scoring each examinee.

Construction Tips

- An effective strategy to improve reliability is the use of paired or tripled examiners for each question. Thus each student will have a different group of raters for each oral exam question. Each examiner should grade or rate the examinees independently.
- To improve the validity of this exam, the selection of the cases to be covered should focus on important content; longer exams are more valid than shorter exams because of the increase in content coverage.
- When cases are used as the stimulus for the oral exam, the same cases and questions should be used for all examinees to maintain standardization. However the order of questions can be varied across examinees.
- As with essay exams, model answers and explicit grading criteria for each question should be developed prior to the oral exam. All raters should be familiar with the grading criteria and rating form.

Standardized Patients and OSCEs

Standardized patients are actual patients or laypeople trained to portray a patient for teaching and/or assessment purposes. The standardized patient can provide a test of the student's skills related to communications, history-taking or physical examination. As the term suggests, standardized patients are used in assessment to provide a standard clinical encounter against which to judge student performance.

Standardized patients are typically used as part of an objective standardized clinical examination (OSCE). An OSCE provides an opportunity to assess student knowledge and skills that are not easily assessed using more traditional paper and pencil-based examinations. This assessment format involves students moving through a series of stations, with each station requiring specific tasks. Skills related to communications, history-taking, physical examination and written records are typically a part of an OSCE. A typical OSCE might be made up of about eight stations, each involving a 15-min encounter with a standardized patient, with 10 min afterwards to complete a written record or answer specific knowledge or interpretation questions about the case. In each of these situations, student performance is judged using the methods described above under checklists, rating forms, essay and oral exams. The Medical Council of Canada has used an OSCE as part of their licensure process for over ten years; an OSCE was introduced as part of the United States Medical Licensing Examination (USMLE) in 2004.

Strengths

- Simulated encounters provide a realistic yet safe context for assessing student performance of basic clinical skills as well as more integrated performance required in complex clinical encounters, such as deriving differential diagnoses, treatment planning and documenting clinical findings. The complexity of the encounters can be varied to accommodate the experience of the learners.
- Simulated encounters can provide students with immediate feedback about their performance. Alternatively they can be recorded for later review and critique. This review can involve students and faculty reviewing the recordings together, or can be completed by the student independently as a self-assessment.
- Simulated encounters can be customized to focus on educational goals and values important to the institution.

Limitations

- Despite the high fidelity of the approach, it does require some suspension of belief on the part of students.

- Simulated encounters are resource-intensive in terms of time, case development, raters and standardized patients.
- This format is unfamiliar to students and initially can cause anxiety, especially when used for summative assessment.

Reliability and Validity

- Reliability and validity increase with the length of the OSCE. Adding more cases or stations increases content coverage and improves validity. To accomplish this, the length of each station might be reduced to shorten the overall testing time per student. However, as the time per encounter decreases, the fidelity of the encounter might be reduced.
- In rating the performance of students during the encounters, checklists and global rating forms are frequently used. Each has their own strengths and limitations with regards to reliability and validity that must be addressed.

Construction Tips

- The formative use of simulated encounters is a very powerful technique and can provide students with tangible feedback to improve performance. In formative settings students often value this approach to assessment.
- The checklists to record student performance during a simulated encounter should be only as long as necessary. When completed from memory as is often the case when used by standardized patients, the value of long checklists and rating forms is limited by patients' ability to remember the specifics of the encounter.

Technology-Based Simulations

Technology-based simulations for performance assessment provide standardized conditions for studying and assessing clinical performance. Through the use of mannequins, computers, artificial limbs, virtual reality and other tools, simulations can be created for assessment purposes that provide a realistic challenge to students based on a clinical problem. The advantage of this format is that there is no danger to patients, and depending on how they are implemented the simulations can provide instant feedback. Written and computer-based simulations have been used to assess clinical reasoning, diagnostic plans and management. Simulators can be used alone or in conjunction with standardized patients. Additional information about this format is found in Chapter 7.

Strengths

- Technology-based simulations are particularly well-suited for assessing procedural skills, critical care decision-making and teamwork.
- When observed by faculty, this approach can be a powerful tool for improving instruction and providing feedback to students.
- This type of simulation provides important skill-focused training in a context that does not jeopardize patient safety.
- Technology-based simulations are useful for both formative and summative assessment.

Limitations

- Technology-based simulations are less realistic than standardized patient encounters, but can provide opportunities to demonstrate skills that might be impractical, uncomfortable or embarrassing for standardized patients.
- Simulator technology often is expensive.

Reliability and Validity

- Technology-based simulations create highly standardized test situations for students. Some simulators collect response data and provide quantitative feedback. This tends to be reliable and valid. To the extent that performance is judged on the basis of checklists or rating scales, these approaches each have strengths and weaknesses, which have previously been discussed.

Construction Tips

- A complete discussion of the effective use of simulators can be found in Chapter 7

Peer Assessments

Peer assessment is usually implemented based on global ratings forms; respondents are asked to rate the student's performance or to indicate the relative frequency of specific target behaviors of interest. Peer assessments are useful in that they provide feedback from multiple sources about an individual's performance, usually aggregated across a variety of situations or encounters.

For each of the attributes listed below, please rate the performance of this student compared to other students in the program this year.

	Well below average		Average		Well above average
1. Is professional in appearance	1	2	3	4	5
2. Reliable and responsible	1	2	3	4	5
3. Carries fair share of workload	1	2	3	4	5
4. Adheres to ethical behavior	1	2	3	4	5
5. Interacts appropriately with patients	1	2	3	4	5
6. Responds appropriately to advice	1	2	3	4	5

Strengths

- This format can be used to assess knowledge, skills and attitudes.
- Peer assessments can provide insights into professional behavior and teamwork, which are often difficult to assess using other methods.
- Peer assessments provide a credible source of performance information related to daily observable behaviors. This is especially useful for formative feedback when provided in a timely and confidential manner.
- Participation in peer assessments provides students with valuable experience in giving and receiving feedback. It also provides students with an opportunity to systematically compare their performance with the performance of others within a similar context.

Limitations

- Peer assessments can be provided through the use of rating checklists, global ratings or written narratives. Each of these methods has inherent limitations that have already been described.
- A general lack of familiarity with this approach is threatening to students.

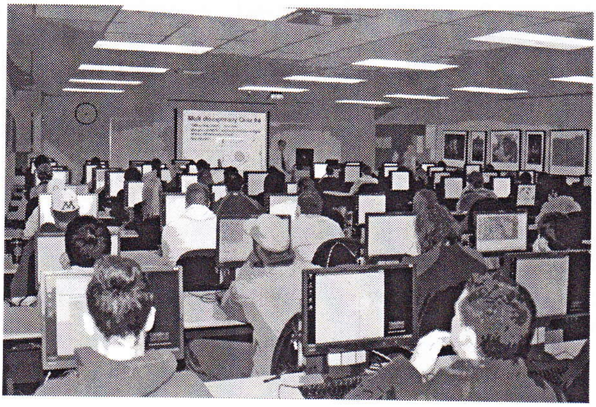
- Some evaluators, especially peers, might be reluctant to provide negative feedback to their fellow students.
- Data collection, analysis and feedback can be cumbersome when used for assessing many individuals.
- A supportive learning environment is essential. When confidentiality and trust are not safeguarded, the validity of the data collected and the value of the feedback to students is significantly diminished. In the worst case scenario, peer assessments can be experienced by students as critical and hurtful.

Reliability and Validity

- A large number of respondents are required to obtain reliable ratings: nurses have been found to be reliable raters while patients and faculty are less reliable, which requires more ratings. This is particularly important when this approach is used for high stakes outcomes such as recertification. Individuals chosen to provide ratings should have multiple opportunities to observe the behavior of the student in question.
- Validity is a function of the process used to develop the rating form, the individuals from whom ratings are obtained, and the length of time over which raters have observed the target behaviors.

Construction Tips

- Students should know the rating categories used in the peer assessment in advance so the process is transparent and less threatening.
- It is helpful to provide guidelines and examples about giving feedback to others so that the feedback is constructive and appropriate for the expected level of performance.



- The quality of the feedback will improve over time with practice; this is especially true for students, who frequently have little experience with this form of assessment.
- Peer assessments can be implemented as part of a 360 degree assessment, which involves ratings-based assessment of an individual's behavior. The ratings are completed by a wide range of others who have contact with the individual. In a clinical setting, this frequently includes peers, supervisors, instructors, nursing staff and allied health personnel; in some cases patients might be included.
- Peer assessment can focus on skills related to interpersonal and written communication, professionalism, teamwork and leadership.
- Although most frequently used in clinical settings to evaluate student performance, this type of assessment can be used in small group instruction, clinical skills training and other settings.

Self-Assessments

Self-assessment is often an informal process for students as they progress through their education. There are relatively few opportunities for students to use structured self-assessment for formative assessment; for a variety of reasons, some of which are obvious, the use of self-assessment for summative progress decisions is even rarer. The actual format of self-assessments can be written or based on rating forms, focusing on global attributes. Self-assessments are probably most valuable when used in conjunction with similar assessments from other sources such as peers or teachers.

Strengths

- Self-assessment provides students with a valuable opportunity to become more critical of their own performance as well as to develop insight and responsibility for their performance.
- Self-assessment provides a setting for reflection and creation of a self-initiated plan for personal and professional development.
- This approach can be used to self-assess strengths and limitations related to knowledge, clinical skills and personal attitudes.
- Because the assessment is self-generated, it provides a unique perspective on students' abilities, particularly when used with other information to provide formative feedback.
- The structure and content of the self-assessment form can be used to direct the scope of the self-assessment.

Limitations

- The most significant limitation of self-assessment is our own difficulty of seeing ourselves as others see us. This makes self-assessment a challenging task for many to do well, and even with experience, our inherent limitations in self-monitoring restrict the application of this approach.
- Since rating scales are typically the format used for self-assessment, the limitations associated with that technique apply to self-assessment.
- Students' lack of familiarity with a systematic approach to self-assessment makes this form of assessment threatening for many students.
- This approach is best used within the context of a supportive learning environment, where students feel safe to reveal their own limitations and confidentiality is assured.

Reliability and Validity

- Reliability can be increased by clearly specifying the self-assessment rating task in terms of the behaviors to be rated, the time period covered by the assessment, and well-defined criteria and standards to guide the assessment.
- The assessment should be structured using explicit criteria that are acknowledged and endorsed by students and faculty. This will enhance the validity of the self-assessment ratings.

Construction Tips

- The rating form used to guide student self-assessment should focus on specific behaviors and outcomes, e.g., what was tried and/or accomplished.
- To promote broad support and endorsement of the self-assessment rating scheme, open discussions involving students and faculty can be used to delineate the criteria for judging performance as well as elicit possible standards for each criterion related to satisfactory and unsatisfactory performance.
- This form of assessment can be combined with peer and faculty assessments to provide multisource feedback. This approach also helps balance unrealistic self-appraisals.

Portfolios

A portfolio is a collection of evidence organized around specific themes as a means of assessing knowledge skills and attitudes. The key components of a portfolio include a statement of purpose for the portfolio, examples of evidence selected by the student to document performance, as well as a reflective statement by the student regarding the portfolio content.

Strengths

- The task of selecting representative evidence of achievement provides an opportunity for reflection and self-appraisal.
- A wide range of evidence can be included in a portfolio including written documents and projects, letters of appreciation or recognition, presentations, digital media and resources, citations, logbooks of patient encounters, and survey results.
- The assembled evidence provides insight into the learner's ability to apply their knowledge and skills in integrative tasks, as well as the growth of their knowledge and abilities over time.
- A key component of most portfolios is a reflective essay that provides insight into higher level cognitive abilities as well as the learner's own ability to self-assess their achievements and what has been learned.
- Portfolios are frequently used for formative assessment, and can be an important source of information when combined with faculty mentorship.
- Portfolios can also be used for summative assessment such as faculty promotion decisions.

Limitations

- The task of selecting, organizing and interpreting the representative evidence of achievement is time consuming.
- Presumably the examples selected for inclusion in the portfolio are the best evidence the learner has of their performance, and therefore only a selective sample of performance is presented.
- This format is unfamiliar to many students and faculty, both in terms of putting a portfolio together and making judgments from a portfolio.

Reliability and Validity

- As with other forms of assessment, clear specification of the purpose and content of the portfolio is important to assure validity. The relationship of the portfolio to course objectives or promotion criteria enhances validity and helps define the types of evidence appropriate for inclusion, the number of examples to include, and the content of the reflective essay.

- Reliability is achieved through the use of multiple ratings of the portfolio content, as well as the use of multiple forms of evidence included in the portfolio to demonstrate specific educational outcomes or performance.
- Students and raters must understand the criteria by which the portfolio will be judged as well as the rating form that is derived from these criteria.
- Because the assessment of the portfolio is ultimately made through the use of rating forms, the issues associated with the reliability and validity of rating forms also have bearing here.

Construction Tips

- Providing learners with the responsibility to meaningfully choose the evidence to include in the portfolio enhances their ownership of the portfolio. Another way of promoting ownership is involving students in the discussion of how the portfolios will be evaluated: the criteria and standards to be used.
- Portfolios can include a wide range of evidence such as: an abstract or brief description of research or educational projects; publications or presentations; case studies; self- or peer- assessments; awards or letters of recognition/appreciation documenting professional achievements; conference proceedings or reviewer lists indicating contributions to professional organizations; awards; materials from websites or digital media that have been developed; personal reflections on specific achievements, activities, ethical dilemmas, challenging patients, etc. Almost any type of evidence might have value in a portfolio depending on its purpose.
- Software tools have been developed to assist in the compilation of evidence into an electronic portfolio. These tools range from blogs, wikis, online learning management systems to specific portfolio systems. See Chapter 9 for more information.
- Criteria for judging the content of a portfolio often focuses on the student's reflective essay regarding the achievements represented by the assembled evidence. The evidence can also be interpreted in terms of the breadth and depth of content, comparison of different types of content; areas of strength, weakness or achievement not represented within the portfolio. Another use of the portfolio is as a stimulus for discussion between students and instructors or mentors.
- An assessment derived from a portfolio can focus on the skills, knowledge or attitudes in judgments of the technical achievements represented by the evidence, as well as the application of theory or the ethics and values inherent in the content.
- There are circumstances when standard criteria for assessing portfolios might not be desirable, such as when the portfolio is implemented as a means of documenting the achievements and progress made by individuals as part of an individualized educational plan for independent study or remediation.

Reporting and Feedback

Feedback to Students

As mentioned earlier in this chapter, an important consequence of assessment is that students receive feedback about their learning. Many of the assessment methods described in this chapter will be used for summative assessment, providing reliable and valid information from which student progress decisions can be made. Some of these assessments are also well-suited for providing students with detailed information about their strengths as well as areas for improvement. It is this level of detail in the feedback to students that provides them with the greatest opportunities for learning from their assessment experiences and building confidence in their abilities. Formative feedback is also important for building confidence and reducing anxiety when students are confronted with forms of assessment that are unfamiliar to them.

Depending on the assessment, feedback can take the form of detailed model performance such as model answers for essay and oral exams, videos of expected skill performance, sample portfolios and the like. Other forms of feedback include summaries of the most common errors made by students during an assessment, and information about why a specific response choice was right or wrong. Of course, written comments related to the students' specific responses are very helpful but can be very time consuming. Another strategy is to have students self-assess their performance as a means of comparison with instructor feedback. To optimize the value of assessments as feedback experience for students:

- use clear criteria for grading performance
- provide feedback in a timely manner
- include both positive and negative feedback when practical
- make feedback as specific as possible

Feedback to Faculty

It is important that aggregated student performance information be available to the medical school committees responsible for oversight of the curriculum. Aggregate performance information can be used to provide evidence of the success of new programs, curricula or modes of instruction. Another important use of aggregated student performance data is to provide valid evidence for decision-making and supplement the perceptions of students or faculty. It provides a systematic approach to data collection that can be used to answer specific questions about effectiveness and outcomes, and perhaps give rise to further questions about the curriculum. Such evidence can be crucial in the face of personal testimonials or opinions derived from one person's experience with a specific student. This information can be part of an on-going effort to monitor an educational program or diagnose curricular problems as part of a systematic program review. See Chapter 12 for more information.

References

Case S, Swanson D, Constructing written test questions for the basic and clinical sciences. National Board of Medical Examiners. Available via www.nbme.org. This document can be downloaded free-of-charge from the website. Under publications, look for "Item Writing Manual."

A helpful resource is Cashin's paper, Improving essay tests. Available via: http://www.theideacenter.org/sites/default/files/Idea_Paper_17.pdf

For Further Reading

Amin Z, Seng CY, Eng KH (2006) Practical guide to medical student assessment. World Scientific Publishing Company, Singapore.

Anderson L, Krathwohl D, (Eds.) (2001) A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. Longman, New York.

Challis M (1999) AMEE medical education guide no. 11 (revised): Portfolio-based learning and assessment in medical education. Medical Teacher 21: 370–386.

Epstein R, Hundert E (2002) Defining and assessing professional competence. Journal of American Medical Association 287(2): 226–235.

Fitzgerald J, White C, Gruppen L (2003) A longitudinal study of self-assessment accuracy. Medical Education 37(7): 645–649.

Gray J (1996) Global rating scales in residency education. Academic Medicine 71: S55–S63.

Hardin R, Gleeson F (1979) Assessment of medical competence using an objective structured clinical examination (OSCE). Medical Education 13(1): 41–54.

Holmboe E, Hawkins R, Huot S (2004) Effects of training in direct observation of medical residents' clinical competence: A randomized trial. Annals of Internal Medicine 140(11): 874–881.

Mancall EL, Bashook PG (eds) (1995) Assessing clinical reasoning: the oral examination and alternative methods. American Board of Medical Specialties, Evanston, IL.

Mehrens WA, Lehmann IJ (1991) Measurement and evaluation in education and psychology, 4th edn. Holt, Rinehart and Winston, Fort Worth, TX.

Miller AH, Imrie BW, Cox K (1998) Student assessment in higher education. In: A handbook for assessing performance, Kogan Page, London.

Miller G (1990) The assessment of clinical skills/competence/performance. Academic Medicine 65(9): S63–S67.

Raksha J, Ling F, Jaeger J (2004) Assessment of a 360-degree instrument to evaluate residents' competency in interpersonal and communication skills. Academic Medicine 79(5): 458–463.

Shumway J, Harden R (2003) AMEE Guide No. 25: The assessment of learning outcomes for the competent and reflective physician. Medical Teacher 25(6): 569–584.

Rademaker J, ten Cate T, Bar P (2005) Progress testing with short answer questions. Medical Teacher 27(7): 578–582.

Tekian A, McGuire C, McGaghie W (Eds.) (1999). Innovative simulations for assessing professional competence. University of Illinois Press, Chicago, IL.

Wallace P (2007) Coaching standardized patients. Springer, New York.