References

Assessment Strategies For Le Cordon Bleu

1. How to Write Better Tests – Handbook for Improving Test Construction Skills

2. Evaluation and Grading of Students – Pangaro and McGaghie (online at www.treysystems.com/projects/)

3. Developing Better Multiple Choice Questions (www.treysystems.com/projects/)

April 28, 2008

Michael Simonson

HOW TO WRITE BETTER TESTS

A Handbook for Improving Test Construction Skills

Introduction

This handbook is designed to help instructors write better tests—better in that they more closely assess instructional objectives and assess them more accurately. A number of problems keep classroom tests from being accurate measures of students' achievement.
Some of these problems are:

1. Tests include too many questions measuring only knowledge of facts. One of the most common complaints from students is that the test content did not reflect the material discussed in class or what the professor seemed to indicate was most important. This may happen because knowledge questions are the easiest to write.
2. Too little feedback is provided. If a test is to be a learning experience, students must be provided with prompt feedback about which of their answers were correct and which were incorrect.
3. The questions are often ambiguous and unclear. According to Milton (1978), ambiguous questions constitute the major weakness in college tests. Ambiguous questions often result when instructors put off writing test questions until the last minute. Careful editing and an independent review of the test items can help to minimize this problem.
4. The tests are too short to provide an adequate sample of the body of content to be covered. Short tests introduce undue error and are not fair to students.
5. The number of exams is insufficient to provide a good sample to students' attainment of the knowledge and skills the course is trying to develop. The more samples of student achievement obtained, the more confidence instructors have in the accuracy of their course grades.

PLANNING THE TEST

A taxonomy of teaching objectives (Bloom, 1956) lists several cognitive outcomes typically sought in college instruction. These outcomes are listed hierarchically in Table1 and include Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation. If these are desired outcomes of instruction, then classroom tests must include assessment of these objectives.

Table 1. Examples of Bloom's Cognitive Levels

Bloom's Cognitive Level	Student Activity	Words to Use in Item Stems
Knowledge	Remembering facts, terms, concepts, definitions, principles	Define, list, state, identify, label, name, who? when? where? what?
Comprehension	Explaining/interpreting the meaning of material	Explain, predict, interpret, infer, summarize, convert, translate, give example, account for, paraphrase
Application	Using a concept or principle to solve a problem	Apply, solve, show, make use of, modify, demonstrate, compute
Synthesis	Producing something new or original from component parts	Design, construct, develop, formulate, imagine, create, change, write a poem or short story
Evaluation	Making a judgment based on a pre-established set of criteria	Appraise, evaluate, justify, judge, critique, recommend, which would be better?

The easiest way to ensure a representative sample of content and cognitive objectives on the test is to prepare a table of specifications. This table is simply a two-way chart listing the content topics on one dimension and the cognitive skills on the other. We want to include content and skills in the same proportion as they were stressed during instruction. Table 2 shows a simple table of specifications; it is intended to be illustrative, not comprehensive.

Table 2. Table of Specifications for a Chemistry Unit Test on Oxygen

Content (%)	Knowledge	Comprehension	Application	Total (%)
Physical Properties	8	6	6	20
Chemical Properties	12	9	9	30
Preparation	4	3	3	10
Uses	16	12	12	40
Total	40	30	30	100

This table indicates the content topics and the objectives to be covered and the proportion of the test that will be devoted to each. Evidently, more class time was spent on the uses of oxygen because 40 percent of the test questions deal with uses compared with only 10 percent on preparation. The column totals indicate that 40$ of the items will be written at the knowledge level with the remaining divided equally between comprehension and application. Using the percentages assigned to each cell, one writes the appropriate number of items. For example, because 20% of the test is to cover physical properties and 30% is to be application, then 6% of the total test would measure the ability to apply knowledge about oxygen's physical properties to new situations.

Coordinating test content with instruction content ensures content validity of the test. Using a table of specifications also helps an instructor avoid one of the most common mistakes in classroom tests, namely writing all the items at the knowledge level.

THE TEST FORMAT

After planning the content and cognitive objectives for the test, instructors must decide on the best way to measure them; that is, they decide on the test format. The format refers to whether the test will be objective (multiple choice, true false, matching, etc.) or essay. What factors do faculty consider when deciding on the format of the test?

1. What is to be Measured?

We should choose the format that is most appropriate for measuring the cognitive objectives on the test. If instructors want students to contrast A and B, take a position on an issue and defend it, create a plan, and perform other similar tasks, then they would most likely use an essay format. For example, if an instructor wants students to explain the role of the press in the coming of the Civil War, he/she would probably choose an essay item. But if the objective is to identify the authors of selected writings about the coming of the War, then the instructor could use an objective type format.

Many times instructors have a choice. Objective-type items can be used quite effectively to measure high level cognitive objectives. A common myth depicts objective items as measuring simple factual recall and essays as evaluating higher-order thinking. But multiple choice items, for example, can be written to measure reasoning, comprehension, application, analysis, and other complex thinking processes. What other factors might influence the decision about format?

2. The Size of the Class

Class size is often an important factor influencing the decision about test formal. It is very difficult to give essay tests when there are 400 students in the class because the scoring time is prohibitive. A survey of 1100 professors from across the country (Cross, 1990) showed that class size is the factor that professors consider most important when they decide what test format to use. Two-thirds of the faculty surveyed said they preferred the essay format but could not use it because of the size of their classes. They used essay tests only in small classes.

3. Time Available to Prepare and Score Test

It takes a long time to score an essay test. By contrast, it takes a long time to construct a multiple-choice test. Instructors must consider whether they will have more time available when preparing or when scoring the test. If instructors are short of time when a test must be prepared, then they might choose an essay test, if class size permits. We are not implying that good essay questions are easy to write; essay tests are easier to prepare only because fewer questions have to be written.

ESSAY ITEMS

Let us look at the relative strengths and weaknesses of the essay format.

Strengths of Essay Items

1. Essay items are an effective way to measure higher-level cognitive objectives. They are unique in measuring students' ability to select content, organize and integrate it, and present it in logical prose.
2. They are less time-consuming to construct.
3. They have a good effect on students' learning. Students do not memorize facts, but try to get a broad understanding of complex ideas, to see relationships, etc.
4. They present a more realistic task to the student. In real life, questions will not be presented in a multiple-choice format, but will require students to organize and communicate their thoughts.

Limitations of Essay Items

1. Because of the time required to answer each question, essay items sample less of the content.
2. They require a long time to read and score.
3. They are difficult to score objectively and reliably. Research shows that a number of factors can bias the scoring:

   A) Different scores may be assigned by different readers or by the same reader at different times.

B) A context effect may operate; an essay preceded by a top quality essay receives lower marks than when preceded by a poor quality essay.
C) The higher the essay is in the stack of papers, the higher the score assigned.
D) Papers that have strong answers to items appearing early in the test and weaker answers later will fare better than papers with the weaker answers appearing first.
E) Scores are influenced by the expectations that the reader has for the student's performance. If the reader has high expectations, a higher score is assigned than if the reader has low expectations. If we have a good impression of the student, we tend to give him/her the benefit of the doubt.
F) Scores are influenced by quality of handwriting, neatness, spelling, grammar, vocabulary, etc.

Writing Good Essay Items

1. Formulate the question so that the task is clearly defined for the student. Use words that "aim" the student to the approach you want them to take. Words like discuss and explain can be ambiguous. If you use "discuss", then give specific instructions as to what points should be discussed.

   Poor: Discuss Karl Marx's philosophy.

   Better: Compare Marx and Nietzsche in their analysis of the underlying problems of their day in 19th century European society.

   Clearly stated questions not only make essay tests easier for students to answer, but Also make them easier for instructors to score.

2. In order to obtain a broader sampling of course content, use a relatively large number of questions requiring shorter answers (one-half page) rather than just a few questions involving long answers (2-3 pages).

3. Avoid the use of optional questions on an essay test. When students answer different questions, they are actually taking different tests. If there are five essay questions and students are told to answer any three of them, then there are ten different tests possible. It makes if difficult to discriminate between the student who could respond correctly to all five, and the student who could answer only three. Use of optional questions also affects the reliability of the scoring. If we are going to compare students for scoring purposes, then all students should perform the same tasks. Another problem is that students may not study all the course material if they know they will have a choice among the questions.

4. Indicate for each question the number of points to be earned for a correct response. If time is running short, students may have to choose which questions to answer. They will want to work on the questions that are worth the most points.

5. Avoid writing essay items that only require students to demonstrate certain factual knowledge. Factual knowledge can be measured more efficiently with objective-type items.

Writing Essay Items at Different Levels of Bloom's Taxonomy

The goal is to write essay items that measure higher cognitive processes. The question should represent a problem situation that tests the student's ability to use knowledge in order to analyze, justify, explain, contrast, evaluate, and so on. Try to use verbs that elicit the kind of thinking you want them to demonstrate. Instructors often have to use their best judgment about what cognitive skill each question is measuring. You might ask a colleague to read your questions and classify them according to Bloom's taxonomy.

Another point that should be emphasized when writing items that measure higher cognitive processes is that these processes build on and thus include the lower levels of knowledge and comprehension. Before a student can write an essay requiring analysis, for example, he/she must have knowledge and a basic understanding of the problem. If the lower level processes are deficient, then the higher-level ones won't operate at the maximum level. The following are examples of essay items that appear to measure at different levels:

Knowledge: Identify the "wage fund doctrine".

Comprehension: Explain the following: Aquinas was to Aristotle as Marx was to Ricardo.

Application: Use the "wage fund doctrine" to explain wage rate in the writing of J.S. Mill.

Analysis: Compare and contrast the attitudes toward male and female sex roles in the work of Ibsen and Huysmans.

Synthesis: Write an essay contrasting Nietzsche's approach to the question of "truth" with that of Comte. development.

Evaluation: Using the five criteria discussed in class, critically evaluate Adam Smith's theory of economic

Scoring Essay Tests

The major task in scoring essay tests is to maintain consistency, to make sure that answers of equal quality are given the same number of points. There are two approaches to scoring essay items: (1) analytic or point method and (2) holistic or rating method.

1. Analytic: Before scoring, one prepares an ideal answer in which the major components are defined and assigned point values. One reads and compares the student's answer with the model answer. If all the necessary elements are present, the student receives the maximum number of points. Partial credit is given based on the elements included in the answer. In order to arrive at the overall exam score, the instructor adds the points earned on the separate questions.

2. Holistic: This method involves considering the student's answer as a whole and judging the total quality of the answer relative to other student responses or the total quality of the answer based on certain criteria that you develop.

   As an instructor reads the answers to a particular question, he/she sorts the papers into stacks based on the overall quality. The best answers go into the first stack, the average go into the second stack, and the poorest into the third stack. After further examination of the answers in each stack, one may want to divide some of these stacks to make additional ones. Then points are written on each paper appropriate to the stack it is in.

Suggestions for Scoring Essays

1. Grade the papers anonymously. This will help control the influence of our expectations about the student on the evaluation of the answer.

2. Read and score the answers to one question before going on to the next question. In other words, score all the students' responses to Question 1 before looking at Question 2. This helps to keep one frame of reference and one set of criteria in mind through all the papers, which results in more consistent grading. It also prevents an impression that we form in reading one question from carrying over to our reading of the student's next answer. If a student has not done a good job on say the first question; we could let this impression influence our evaluation of the student's second answer. But if other students' papers come in between, we are less likely to be influenced by the original impression.

3. If possible, also try to grade all the answers to one particular question without interruption. Our standards might vary from morning to night, or one day to the next.

4. Shuffle the papers after each item is cored throughout all the papers. Changing the order reduces the context effect and the possibility that a student's score is the result of the **location** of the paper in relationship to other papers. If Mary's B work always followed John's A work, then it might look more like C work and her grade would be lower than if her paper were somewhere else in the stack.

5. Decide in advance how you are going to handle extraneous factors and be consistent in applying the rule. Students should be informed about how you treat such things as misspelled words, neatness, handwriting, grammar, and so on.

6. Be on the alert for bluffing. Some students who do not know the answer may write a well-organized coherent essay but one containing material irrelevant to the question. Decide how to treat irrelevant or inaccurate information contained in students' answers. We should not give credit for irrelevant material. It is not fair to other students who may also have preferred to write on another topic, but instead wrote on the required question.

7. Write comments on the students' answers. Teacher comments make essay tests a good learning experience for students. They also serve to refresh your memory of your evaluation should the student question the grade.

Preparing Students to Take Essay Exams

Essay tests are valid measures of student achievement only if students know how to take them. Many college freshmen do not know how to take an essay exam, because they haven't been required to learn this skill in high school. You may need to take some class time to tell students how to prepare for and how to take an essay exam. You might use some of your old exam questions, and let students see what an A answer looks like and how it differs from a C answer.

MULTIPLE-CHOICE ITEMS

Many users regard the multiple-choice item as the most flexible and probably the most effective of the objective item types. A multiple-choice item consists of two parts: (1) the stem, which presents a specific problem to the test taker and (2) a list of possible solutions or answers called distractors. The stem may be written either as a question or as an incomplete statement. There should be only one correct or best answer while the other three or four options serve as distractors.

Strengths of Multiple-Choice Items

1. Versatility in measuring all levels of cognitive skills.
2. Permit a wide sampling of content and objectives.
3. Provide highly reliable test scores.
4. Can be machine-scored quickly and accurately.
5. Reduced guessing factor compared with true-false items.

Limitations of Multiple-Choice Items

1. Difficult and time-consuming to construct.
2. Depend on student's reading skills and instructor's writing ability.
3. Ease of writing low-level knowledge items leads instructors to neglect writing items to test higher-level thinking.
4. May encourage guessing (but less than true-false).

Writing Multiple-Choice Items

The challenge is to write questions that test a significant concept, that are unambiguous, and that don't give test-wise students an advantage.

1. The stem should fully state the problem and all qualifications. To make sure that the stem presents a problem, always include a <u>verb</u> in the statement.

2. Concentrate on writing items that measure students' ability to comprehend, apply, analyze, and evaluate as well as recall.

3. Include words in the stem that would otherwise be repeated in each option. Following this guideline not only saves time for the typist but also saves reading time for the student.

   Poor: Sociobiology can be defined as
   a. the scientific study of humans and their relationships within the environment.
   b. the scientific study of animal societies and communication.
   c. the scientific study of plants and their reproductive processes.
   d. the scientific study of the number of species in existence.

   Better: Sociobiology can be defined as the scientific study of
   a. humans and their relationships within the environment.
   b. animal societies and communication.
   c. plants and their reproductive processes.
   d. the number of species in existence.

4. Eliminate excessive wording and irrelevant information in the stem.

5. Make sure there is only one correct or best response.

   Poor: The function of the hypothesis in a research study is to provide

   a. tentative explanation of phenomena.
   b. proven explanation of phenomena.
   c. framework for interpretation of the findings.
   d. direction for the research.

   There is no single or best answer, options a, c, and d are correct. The options need to be reworded so that only one is clearly best or correct. Or one could change the stem to read: According to the lecture (or the text), the **most important** function of the hypothesis is…

6. Provide a minimum of three, but not more than five, plausible and attractive options for each item. A good procedure is to think of errors that students are likely to make and use these as distractors.

   Poor: The recent (1989) research suggesting that controlled nuclear-fusion could be effected in a laboratory experiment at room temperature was conducted by

   a. Watson and Crick.
   b. Pons and Fleischmann.
   c. Koch and Jenner.
   d. Fermi and Bohr.

   While the first two options are plausible, the last two are not. The latter should be replaced by the names of contemporary scientists.

7. Make all the options for an item approximately homogeneous in content, form, and grammatical structure. Increasing the homogeneity of the content among the options can increase the difficulty of an item. (Difficulty of a test should not be based on inclusion of obscure content.)

8. Avoid the use of the all-of-the-above and none-of-the-above options. The problem with "all of the above" as an option is that it makes the item too easy. If students can recognize at least one incorrect option, they can eliminate "all of the above" as a viable option. On the other hand, if they can recognize at least two correct options, then they know that "all of the above" is the correct answer. Furthermore, research shows that when "all of the above" is used as a distractor, it is too often the correct response. Students are quick to pick up on this clue.

   "None of the above" should be used only when absolute standards of correctness can be applied, such as in math, grammar, spelling, geography, historical dates, and so on. Otherwise, students can often argue about the correctness of one of the other options.

9. Avoid verbal associations between the stem and the correct option, e.g., the same reference word should not appear in the stem and an option. Also make sure that the options are grammatically consistent with the stem.

   Poor: The correlation coefficient found by correlating students' scores on a classroom math test with their scores on a standardized math test is called a
   a. validity coefficient.
   b. index of reliability.
   c. equivalence coefficient.
   d. internal consistency coefficient.

   Option (a) is the only one that is grammatically consistent with the stem. It could be correctly selected without knowing anything about the content. One should change the "a" in the stem to "a(n)".

10. Avoid making the correct answer markedly longer or shorter than the other options.

11. If there is a logical sequence in which the alternatives can be arranged (alphabetical if a single word, in order of magnitude if numerals, in temporal sequence, or by length of response), use that sequence.

12. Use negatively stated stems sparingly. When used, call attention to the negative word by underlining and/or capitalizing.

13. Randomly distribute the correct response among the alternative positions throughout the test. That is, have approximately the same proportion of A's, B's, C's, D's, and E's as the correct response.

14. Watch for specific determiners such as "all", "always", "never" which are more likely to be in incorrect options. Others like "usually" and "sometimes" are more likely to be in the keyed response.

15. Multiple-choice items should be independent. That is, an answer to one question should not depend on the answer to another question.

16. Avoid the use of language that your students won't understand. For example (unless it's a French test), use "cause" instead of "raison d' etre" in the question.

17. State items so there can be only one interpretation of their meaning.

    Poor: Which one of the following is the best source of heat for home use?
    a. Gas b. Electricity c. Oil d. Geo-thermal

    The answer would depend on how one interprets the question. Are we talking about the best source economically, in terms of cleanness, in terms of efficiency, or just what? Also the correct answer might depend on what part of the world we're asking about.

    Better: The most economical source of heat in the Midwestern U.S. is
    a. gas b. electricity c. oil d. geo-thermal

Suggestions for Writing Multiple-Choice Items Which Measure Higher Objectives

It is difficult and time-consuming to write multiple-choice items that measure the higher thinking skills. The item writer has to be creative in order to develop challenging questions. The following suggestions may provide some ideas for writing these kinds of questions.

1. Present practical or real-world situations to the students. These problems may use short paragraphs describing a problem in a practical situation. Items can be written which call for the application of principles to the solution of these practical problems, or the evaluation of several alternative procedures.

2. Present the student with a diagram of equipment and ask for application, analysis, or evaluations, e.g., "What happens at point A if …?," "How is A related to B?"

3. Present actual quotations taken from newspapers or other published sources or contrived quotations that could have come from such sources. Ask for the interpretation or evaluation of these quotations.

4. Use pictorial materials that require students to apply principles and concepts.

5. Use charts, tables or figures that require interpretation.

Table 3 shows multiple-choice items that measure at different levels.

Table 3. Multiple-Choice Items That Measure at Various Levels.

1. Knowledge
   Which of the following are the raw materials for photosynthesis?
   a. Water, heat, sunlight
   b. Carbon dioxide, sunlight, oxygen
   c. Water, carbon dioxide, sunlight
   d. Sunlight, oxygen, carbohydrates
   e. Water, carbon dioxide, carbohydrates

2. Comprehension
   If living cells similar to those found on earth were found on another planet where there was no molecular oxygen, which cell part would most likely be absent?
   a. cell membrane
   b. nucleus
   c. mitochondria
   d. ribosome
   e. chromosomes

3. Application
   Phenylketonuria (PKU) is an autosomal recessive condition. About one in every fifty individuals is heterozygous for the gene but shows no symptoms of the disorder. If you select a symptom-free male and a symptom-free female at random, what is the probability that they could have a child afflicted with PKU?
   a. (.02)(.02)(.25) = 0.0001+0.01%, or about 1/10,000
   b. (.02)(.02) = 0.0004 = 0.04%, or about 1 /2,500
   c. (1)(50)(2) = 100% = all
   d. (1)(50)(0) = 0 = none
   e. 1/50 = 2%, or 2/100

4. Analysis
   Mitochondria are called the powerhouses of the cell because they make energy available for cellular metabolism. Which of the following observations is *most* cogent in supporting this concept of mitochondrial function?
   a. ATP occurs in the mitochondria.
   b. Mitochondria have a double membrane.
   c. The enzymes of the Krebs cycle, and molecules required for terminal respiration, are found in mitochondria.
   d. Mitochondria are found in almost all kinds of plant and animal cells.
   e. Mitochondria abound in muscle tissue.

5. Evaluation
   Disregarding the relative feasibility of the following procedures, which of these lines of research is likely to provide us with the most valid and direct evidence as to evolutionary relations among different species?
   a. Analysis of the chemistry of stored food in female gametes.
   b. Analysis of the enzymes of the Krebs cycle.
   c. Observations of the form and arrangement of the endoplasmic reticulum.
   d. Comparison of details of the molecular structure of DNA.
   e. Determination of the total percent protein in the cells.

Note: The writers are indebted to Dr. Michael Tansey of the Biology Department of Indiana University, Bloomington, for these items.

TRUE-FALSE ITEMS

The true-false item typically present a declarative statement that the student must mark as either true or false. Instructors generally use true-false items to measure the recall of factual knowledge such as names, events, dates, definitions, etc. But this format has the potential to measure higher levels of cognitive ability, such as comprehension of significant ideas and their application in solving problems.

T F 1. Jupiter is the largest planet in the solar system.
T F 2. If Triangle ABC is isosceles and angle A measures 100 degrees, then angle B is 100 degrees.
T F 3. If a distribution of scores has a few extremely low scores, then the median will be numerically larger than the mean.
T F 4. The larger the number of scores in a distribution, the larger the standard deviation of the score must be.

The first example above measures recall of a specific fact. The other examples, however, show how a true-false item can be written to measure comprehension and application.

Strengths of True-False Items

1. They are relatively easy to write and can be answered quickly by students. Students can answer 50 true-false items in the time it takes to answer 30 multiple-choice items.
2. They provide the widest sampling of content per unit of time.

Limitations of True-False Items

1. The problem of guessing is the major weakness. Students have a fifty-percent chance of correctly answering an item without any knowledge of the content.
2. Items are often ambiguous because of the difficulty of writing statements that are unequivocally true or false.

Writing True-False Items

1. Test significant content and avoid trivial statements.
2. Write items that can be classified unequivocally as either true or false.
3. Avoid taking statements verbatim from textbooks.

T F Poor: The square of the hypotenuse of a right triangle equals the sum of the squares of the other two sides.
T F Better: If the hypotenuse of an isosceles right triangle is 7 inches, each of the two equal sides must be more than 5 inches.

4. Include only a single major point in each item.
5. Avoid trick questions.

T F Poor: "The Raven" was written by Edgar Allen Poe.
T F Better: "The Raven" was written by Edgar Allan Poe.

The intent of the question should be to determine if students know that Poe write "The Raven", not to see if they notice the incorrect spelling of his middle name.

6. Try to avoid using words like "always," "all", or "never which tend to make the statement false; words like "usually," "often," "many" usually make the statement true.
7. Avoid using negatively worded statements.

T F Poor: *Silas Marner* was not written by Thomas Hardy.
T F Better: *Silas Marner* was written by Thomas Hardy.

8. Put the items in a random order so as to avoid response patterns that could serve as clues (such as T,T,F,T,T,F)
9. Try to avoid long drawn-out statements or complex sentences with many qualifiers.
10. Avoid making items that are true consistently longer than those that are false.
11. Use slightly more false items than true items. False items tend to discriminate more highly among students than do true items. Research shows that when students guess they are more inclined to respond with a true than with a false. We can compensate for this "acquiescent response set" by having a few more false statements than true.

Variations of the T-F Format

Changing false statements to make them true:

The student indicates whether the statement is true or false; if false, he/she must change an underlined word to make the statement true.

T F electrons 1. Subatomic particles of negatively charged electricity are called protons.
T F _____ 2. The green coloring matter in plants is called chlorophyll.

Items measuring ability to recognize cause-and-effect:

The item has two parts, both of which are true; the student must decide if the second part explains why the first part is true.

Yes No 1. Leaves are essential *because* they shade the tree trunk.
Yes No 2. Iron rusts *because* oxidation occurs.

MATCHING

A matching exercise typically consists of a list of questions or problems to be answered along with a list of responses. The examinee is required to make an association between each question and a response.

Example:

I	II
1. a substance of low solubility	A. distillation
2. two liquids that do not dissolve in each other	B. miscible
3. a substance that does the dissolving	C. immiscible
4. a method of purifying a substance	D. precipitate
5. the substance being dissolved	E. soluble
	F. solute
	G. solvent

The problems can be in various forms. The most common is to use verbal statements, but other types of material can be used. For example, the problems might be locations on a map, geographic features on a contour map, parts of a diagram of the body or biological specimens or math problems. Similarly, the sponses don't have to be terms or labels, they might be functions of various parts of the body, or methods, principles, or solutions.

Example:

I	II
1. 3/4	A. 0.060
2. 3/5	B. 0.500
3. 5/8	C. 0.600
4. 3/50	D. 0.625
5. 14/28	E. 0.750
	F. 0.875

Previously, it was difficult to use machine scoring for the matching format. However, a ten-choice, machine-scannable answer sheet that makes it possible to use matching exercises with up to 10 possible responses per question can be purchased from BEST.

Because matching items permit one to cover a lot of content in one exercise, they are an efficient way to measure. It is difficult, however, to write matching items that require more than simple recall of factual knowledge.

Guidelines for Constructing Matching Items

1. Use homogeneous material in each list of a matching exercise. Mixing events and dates with events and names of persons, for example, makes the exercise two separate sets of questions and gives students a better chance to guess the correct response. For example, if one stem were "president of U.S. during World War II", the student could ignore all the responses other than names. Using homogeneous materials requires students to distinguish or discriminate among things which makes for a more challenging task.

2. Include directions that clearly state the basis for the matching. Inform students whether or not a response can be used more than once and where are answers are to be written.

3. Put the problems or the stems (typically longer than the responses) in a numbered column at the left, and the response choices in a lettered column at the right. Because the student must scan the list of responses for each problem, one should keep the responses brief. This saves reading time for the student.

4. Always include more responses than questions. If the lists are the same length, the last choice may be determined by elimination rather than knowledge.

5. Arrange the list of responses in alphabetical or numerical order if possible in order to save reading time.

6. All the response choices must be plausible, but make sure that there is only one correct choice for each stem or numbered question.

COMPLETION ITEMS

The completion format requires the student to answer a question or to finish an incomplete statement by filling in a blank with the correct word or phrase. The advantages of completion items are (1) they provide a wide sampling of content; and (2) they minimize guessing compared with multiple-choice and true-false. The limitations are they (1) rarely can be written to measure more than simple recall of information; (2) are more time-consuming to score than other objective types; (3) are difficult to write so there is only one correct answer and no irrelevant clues.

Guidelines for Writing Completion Items

1. Omit only significant words from the statement, but do not omit so many words that the statement becomes ambiguous.

   Poor: The Constitutional Convention met in _____ in _____.
   Better: The Constitutional Convention met in the city of _____ in 1787.

2. Write completion items that have a single correct answer, if possible.

   Poor: Abraham Lincoln was born in _____.
   There are several legitimate answers: Kentucky, 1809, February, a log cabin, etc.

   Better: Abraham Lincoln was born in the state of _____.

3. Use blanks of the same length throughout the test so that the length is not a clue

4. Avoid grammatical clues to the correct response. For example, if the indefinite article is required before a blank, use a(n) so that the student doesn't know if the correct answer begins with a vowel or a consonant.

   ` Poor: A subatomic particle with a negative electric charge is called an _____.
   The student could eliminate proton, neutron, and meson as possible responses.

   Better: A subatomic particle with a negative electric charge is called a(n) _____.

5. If possible, put the blank at the end of a statement rather than at the beginning. Asking for a response before the student understands the intent of the statement can be confusing and may require more reading time.

   Poor:_____ is the measure of central tendency that is most affected by extremely high or low scores.
   Better: The measure of central tendency that is most affected by extremely high or low scores is the _____.

6. Avoid taking statements directly from the text.

Scoring

Scoring completion items is less objective than multiple-choice or true-false because the student supplies his/her own response. It is difficult to write completion items so that there is only one correct answer. When preparing a key, one should list the correct answer and any other acceptable alternatives. Be consistent in using the key; it would not be fair to accept an answer as right on one paper and not accept it on others.

Chapter 6: Evaluation and Grading of Students

Louis N. Pangaro, MD and William C. McGaghie, PhD, Lead Authors

Michael Ainsworth, MD, T. Andrew Albritton, MD, Michael J. Battistone, MD, David Carnahan, MD, Julia Corcoran, MD, Gerald D. Denton, MD, Ruth-Marie E. Fincher, MD, Shiphra Ginsburg, MD, Cyril M. Grum, MD, Paul A. Hemmer, MD, Eric Holmboe, MD, S. Barry Issenberg, MD, Thomas W. Jamieson, MD, Lisa E. Leggio, MD, John A. Poremba, MD, David A. Rogers, MD, Ross J. Scalese, MD, Karen Szauter, MD, Co-authors

Sections

Section 1. General Introduction

William C. McGaghie, PhD

The term evaluation frequently has a negative connotation, especially for medical learners engaged in a program of study. Medical students and residents rarely view their evaluations as opportunities for improvement even though better performance and public accountability are the principal aims of medical education and the evaluation of its outcomes. Instead, evaluations are seen by learners as hurdles grounded in threat. Evaluations are barriers that channel learner thinking and behavior, frequently motivated by fear of failure, with adverse consequences for those who fall short. Such learner perceptions contrast with faculty intent where evaluation is considered a tool needed to boost student competence and protect the public. Nonetheless, learners perceive the stakes to be high and so is their anxiety. Evaluation is a process to which most medical learners grudgingly submit. It is rarely a process they seek and enjoy.

But evaluation in medical education has an upside, especially as learners and teachers acknowledge the goal is to produce superb clinicians. When educational evaluation data are seen and used as a tool, not as a weapon, the outlook becomes improvement and mastery rather than enforcement. This outlook also changes the psychological climate toward constructive progress instead of apprehension. An illustration is when internal medicine residents express enthusiasm about the acquisition and mastery demonstration of ACLS skills in an educational program featuring deliberate practice and rigorous outcome evaluation.[1,2]

This section provides an overview about evaluation in medical education and sets a point of departure for 14 sections that follow. The section has four parts that lay a foundation for subsequent chapter writings: (a) purposes of learner evaluation, (b) evaluation goals, (c) matching evaluation goals and tools, (d) evaluation and learner motivation. Much of this contextual writing amplifies work published elsewhere nearly two decades ago.[3] There are many similarities with the earlier work although the material has been updated to capture new developments.

Purposes of Learner Evaluation

There are at least eight purposes for learner evaluation in medical education. Each of these purposes is addressed in many ways throughout the remaining chapter sections. They are all important but for different reasons. So except for the first, all of the other purposes for learner evaluation should be assigned equal weight.

Accreditation requirement

A program of undergraduate or postgraduate medical education simply cannot operate, or stay in operation, without being accredited. In the U.S., undergraduate medical accreditation is managed by the Liaison Committee on Medical Education (LCME), jointly sponsored by the American Medical Association (AMA) and the Association of American Medical Colleges (AAMC). U.S. graduate medical education is accredited by the Accreditation Council on Graduate Medical Education (ACGME). (See also Chapter 15: The Clerkship Director and the Accreditation Process) Each of the medical accreditation agencies imposes detailed requirements for learner evaluation that medical education programs must fulfill just to stay in business. Cyclic accreditation reviews assure that once met, a medical education program's learner evaluation criteria and standards do not erode.

Assess competence

Assessment of medical student competence is a basic responsibility for all programs of clinical medical education. Such assessments represent accomplishment benchmarks, tangible signs of medical student progress along the educational continuum. They depend, of course, on *a priori* statements of cognitive, procedural, or affective learning goals; high performance standards; and measurement methods that yield reliable data about student achievement. Clerkship directors realize that sound competence assessments provide focused feedback to students and feedback about educational program effectiveness for faculty and administration (see Chapter 6, Section 3). Competence assessment is a cornerstone of quality medical clerkship education.

Competence assessments external to a clerkship are also imposed in the form of board examinations. With very few exceptions, clerkship medical students have successfully passed USMLE Step 1 and are beginning preparation for Step 2, especially its clinical skills component (2 CS) involving standardized patients (SPs). Students need to be aware that the best way to prepare for these high stakes competence assessments is active engagement with the clinical curriculum.[4]

Document learner experience

Most clerkship directors struggle to exercise control over the type or variety of cases seen by medical students in the clinic or hospital. Patients arrive for clinic visits or are admitted to an inpatient service due to concerns about their health, not because the patients want to advance medical education. Individual cases, and the health problems they represent, often present on an uncontrolled, seemingly random basis. Unless patients having different problems are selectively distributed among clinical learners in a controlled way, clinical medical education can be an uneven experience.

Documenting and managing student exposure to a variety of clinical problems is difficult to fulfill. Hand held computers and wireless data entry and manipulation may simplify the task. The growing use of standardized patients (see Chapter 6, Section 8) and other forms of medical simulation (See Chapter 6, Section 7) can complement contact with real patients. This increases the odds that the clinical curriculum can be uniform.

Gauge academic progress

Similar to competence assessment at important clinical milestones, educational evaluations are also used to gauge and monitor student academic progress more frequently. Medical students are expected to advance through the clinical curriculum on a “critical path” achieving successive program goals both within individual clerkships and across the clerkship year. Wide deviations from that path are a source of concern for clerkship directors. Similar to monitoring infant development using the *Denver II* development chart, medical student academic progress should be gauged frequently to insure it is within normal limits.

Predict performance

Today’s educational evaluations are often used to forecast performance on future assessments. The success of educational forecasts usually stems from the similarity of the skills being assessed, congruence of measurement methods, and the time span between the measurements (shorter is better). The conventional wisdom that “the best way to predict future behavior is to rely on one’s current and past overt behavior” is correct.[5] Rigorous evaluations that produce reliable data give teachers and medical learners a snapshot of each student’s performance status and a window to future student performance.

Feedback for improvement

A common complaint among medical students is that they rarely receive concrete information about "how they are doing" clinically or educationally. Medical learners are usually eager to discuss their experiences and are anxious to discuss ways in which they can boost their fund of knowledge or improve their clinical skill. Performance *feedback* is a term that is widely used to describe information that gives learners knowledge of the results of their study and clinical work. Given *specific feedback* about their progress or deficits, medical students can either move to new areas of clinical practice or take steps to improve marginal performance (see Chapter 6, Section 15).

An educational program needs to have three basic features before useful feedback can be given to learners. First, the program needs to have clear goals that represent a graduated set of milestones for medical students. Second, the program needs to have a means to collect, store, and routinely retrieve data that learners and their teachers can use for educational feedback. Third, the program needs faculty who are willing to take time to candidly review the evaluative data with students, tied to clerkship goals. Effective feedback about educational progress cannot occur unless a plan is in place that identifies goals to be accomplished, routinely collects data about student progress, and provides frequent opportunities for trainees and faculty to discuss clinical learning.

Assign grades

Clerkships operate inside a clinical department, within an undergraduate medical curriculum, usually wrapped in a university environment. Clerkships are one of many threads in a broad academic fabric. Academic tradition holds that variation in student achievement is acknowledged by the assignment of low and high grades. One of the toughest everyday responsibilities that clerkship directors face is translating data about student performance into medical school grades (see Chapter 6, Section 13). This is a medical school and university requirement, a practical reality that comes with the clerkship director's job, which cannot be avoided.

Grades assign value to medical student work. Grades can be given in a normative way ("on the curve") to compare students against their peers or in ways that compare all students against a fixed achievement standard. The bottom line is that assigning grades to medical students as a sign of their achievement is part of every clerkship director's job. Fair and impartial grade assignment is a necessary condition of clerkship education.

Judge program effectiveness

Learner evaluation data including board examination scores, results of OSCEs and SP-based clinical exams, conative measures, and tests using medical simulations can be employed in a variety of ways to judge the effectiveness of a medical education program. The clerkship works to the degree that medical students meet or exceed *a priori* expectations about their acquisition of the knowledge, skill, and affective outcomes stated in the program plan. Achievement of clerkship goals is documented by medical student performance data. A clerkship is successful if a high proportion of its medical students measure up to expectations based on tough but fair assessments of their learning. [See Chapter 7: Evaluation of the Clerkship: Clinical Teachers and Program]

Quality Improvement (QI) is another outcome when medical student performance data are used to judge clerkship effectiveness. The clerkship matures and prospers as student performance data accumulate, are studied, and used for program improvement. Medical student

performance data not only tell a story about individual learners but also about the quality of clerkships and curricula that shape student learning.

Evaluation Goals

Medical student evaluation has at least five goals to amplify the eight purposes already stated. The five goals are evaluation of: (a) professional knowledge, (b) technical and procedural skills, (c) professionalism, (d) professional relationships, and (e) physician-patient relationships. Each evaluation goal is addressed by different measures of medical student achievement.

Professional knowledge

Evaluation of professional knowledge has been the mainstay of medical competence evaluation since the formation of the National Board of Medical Examiners in 1915.[6] Today, medical knowledge assessment is done via internal (e.g., course, clerkship) and external (e.g., USMLE Step 1 and 2) evaluations that rely mainly on multiple-choice questions (see Chapter 6: Section 10 [Clerkship Examinations], Section 11 [Use of In-house Examinations] and Section 12 [Writing Multiple Choice Questions]). These evaluations, by intent and format, propel the idea that the acquisition and maintenance of a broad and deep fund of knowledge is essential for medical practice. The primacy of these tests asserts that knowledge acquisition is a basic goal of medical education.

Technical and procedural skill

Assessment of medical student technical and procedural proficiency has grown in frequency and sophistication over the past decade. Measures are now available that permit objective evaluation of such skills as cardiac auscultation,[7] ACLS maneuvers,[1,2] and the female pelvic examination.[8,9] Most of these measures rely on simulation technology embodied in SPs or medical simulators that vary in human fidelity.[10] These technologies are covered elsewhere in this chapter (see Chapter 6: Section 6 [Procedural skills], Section 7 [Simulators], and Section 8 [Standardized patients]).

Professionalism

Professionalism is expressed in each young physician's character, reliability, honesty, ability to keep confidences, and other nonacademic qualities that embody "the good doctor." Professionalism is more than maturity and less than sainthood; it connotes promises of expertise and duty. In medical circles professionalism is usually conspicuous by its absence and taken for granted when present. Measurement and evaluation of medical professionalism has recently been expressed as a key outcome of medical education witnessed by the Medical School Objectives Project of the AAMC[11] and the subsequent Outcomes Project of the ACGME.[12]

Teaching and evaluating student professionalism has become one of the highest priorities of U.S. medical schools. Teaching is done via customary methods including reading, case discussions, study of professional codes of conduct, and especially by faculty example in clinical settings. Assessing professionalism is difficult to do with precision.[13] However, such assessments are essential because, "Unprofessional behavior in medical school is associated with subsequent disciplinary action by a state medical board."[14] Chapter 6, Section 9 [Evaluating Professionalism] presents a detailed discussion about evaluating professionalism.

Professional relationships

A fourth evaluation goal is professional relationships. This goal goes beyond personal integrity to embrace respect for other members of the health care team, administrative staff, and other colleagues. Professional relationships are addressed infrequently in undergraduate medical education while its profile is rising in graduate and continuing medical education (GME, CME). Received clinical wisdom in addition to recent writing about patient safety[15,16] teach that clinical patient care is rarely a solitary activity. Instead, nearly all patient care is now delivered by teams of individual clinicians having different credentials and skills. The emerging educational goal is [how to] turn a team of experts into an expert team.

Professional relationships, individual and team skill acquisition, team member interchangeability, and team cognition are several of the many variables involved in the preparation of expert teams. Another key variable in team effectiveness is dissolution of traditional professional hierarchies that have existed in clinical medical settings. A growing literature on team training and new professional relationships in medical practice is now beginning to affect medical curricula.[17,18]

Physician-patient relationships

The doctor-patient relationship has been a hallmark of effective clinical practice from antiquity through Osler and Halsted to the present day. Fostering these interpersonal skills and sentiments has been a key feature of medical education and sound clinical practice though always threatened by lapses in honesty by either doctor or patient. More recent threats to doctor-patient relationships include time pressures due to the managed care environment, social class differences, ethnic differences, and many others**.** Holmboe (Chapter 6, Section 4) addresses direct observation of physician-patient relationships in depth.

Matching Evaluation Goals and Tools

A persistent problem in evaluation and grading of students on medical clerkships is matching evaluation goals with the right evaluation tools. Many different tools are available ranging from long aptitude tests such as the MCAT to simulations, OSCEs, and short bedside encounters. Some evaluative tools such as board certification examinations like USMLE Steps 1 to 3 are highly quantitative and objective whereas others such as letters of recommendation are qualitative, subjective. Each type of measure has a place in medical learner evaluation. However, the decision to use one of the tools should be based on a clear understanding of one's evaluative purpose and context.

Table 6.1.1 describes 16 common evaluation methods in medical education. The table also contains a short comment about the advantages of each method and a statement about potential problems associated with using each procedure. At least one citation is given for each method to encourage further reading by those who seek more detailed information.

Table 6.1.1. Evaluation methods commonly used in medical education			
Method	**Description**	**Advantages**	**Problems**
1. Descriptive evaluation by teachers[19-21]	Gives a clear portrait of student status and achievements. Highly individualized; underscores the uniqueness of each medical learner in light of clerkship goals	Requires in-depth faculty knowledge of each student. A qualitative, clinical snapshot of student performance on the clerkship. Works best with faculty group consensus.	Reliability is a concern without rater training and checks for rater bias. Are results consistent and reproducible?
2. Records of clinical encounters[22-24]	Case-by-case documentation of (a) clinical problems seen, and (b) decisions made about each problem	Long-run formulation of learner practice profile. Helps identify clinical problems where more experience is needed. Best managed using computer database. Useful for gaining hospital privileges after training.	Requires high degree of learner compliance
3. Formal (external) examinations[25,26]	Long, standardized examinations covering large bodies of medical content; often composed of separate disciplinary subtests.	Usually high quality exams that give a general portrait of an examinee's fund of knowledge	Test content may not match local educational objectives exactly. Not useful to pinpoint specific learning deficits. High monetary costs.
4. Local (internal) examinations[27]	Examinations written by local faculty for use in courses or clerkships	Can be created to closely match local teaching emphases; exams unite instruction and evaluation	Quality can suffer if faculty are disinterested or unschooled in test development. Major cost is faculty time.
5.Simulations[1,2,10,28,29]	Static models, mannequins, computer-based, and virtual reality approximations of clinical encounters with patients. May be used for individual or team evaluation.	Lifelike approach to evaluating learner skills and clinical reasoning. Enjoyed by clinicians; also excellent for instruction	Simulations vary greatly in fidelity to genuine patient care problems. Scoring rules to capture clinical performance are difficult to derive. Generalization of performance across cases needs to be better established.
6. Objective structured clinical examination[30]	Examinees rotate through a series of stations where, in about 5 minutes each, they are questioned, asked to interpret clinical data, perform a procedure, or otherwise show proficiency with clinical materials.	Concrete, realistic approach to evaluating discrete clinical skills among learners. Requires prompt responses to real clinical material. Bluffing is unlikely.	Faculty involvement and cooperation is essential; tight management is needed to operate effectively.

Table 6.1.1. Evaluation methods commonly used in medical education			
Method	**Description**	**Advantages**	**Problems**
7. Checklists[31,32]	Step-by-step "yes-no" or "right-wrong" protocols used to assess either skill at a clinical procedure (e.g., ACLS) or at preparing a clinical product (e.g., a sterile tray)	Useful to evaluate specific procedures and products. Little guesswork once checklist items and their order are agreed on.	Can appear simplistic unless procedures and products are critical. Use may require much faculty time. Rater training is essential.
8. Rating scales[33]	General assessments, often of learner character or noncognitive professional qualities, based on the rater's memory rather than direct observation of specific events.	Allows evaluators to quantify important qualitative factors that underlie good clinical care.	Frequent "halo" effect (leniency) meaning low ratings are rare.
9. Oral examinations[34]	Face-to-face learner-evaluator encounters where learners are questioned about clinical subjects; sometimes used to gauge if learners can withstand stress.	Historically grounded, have been used for medical learner evaluation for over 3,000 years. Encourage student-faculty interaction.	Notoriously unreliable approach to learner evaluation. Unstandardized; subject to capricious evaluator behavior.
10. Anecdotal records[35,36]	Dean's letters, faculty letters of recommendation	Highly personalized approach to description of learner achievement and frequently, learner readiness to pursue more advanced training	"Halo" effect is common. Frequently difficult to interpret as recipients try to "read between the lines."
11. Chart (record) reviews[37,38]	Faculty-learner case discussions based on data contained in patient charts and recent progress notes.	High relevance due to grounding in real clinical work. No or low cost; straightforward, immediate feedback about patient management.	Cases selected should be representative of the learner's experience or practice, not chosen because they are unusual.
12. Standardized patients (SPs)[30,39]	Laypersons are trained and calibrated to present patient health problems uniformly. SPs frequently record data about learner performance.	Very high realism. SPs can record reliable data and give learners excellent feedback. Especially useful to evaluate skills in physical diagnosis and interviewing.	SP training and calibration takes time. Careful management of the evaluation plan is needed.
13. A-V reviews[40,41]	Learner-faculty review and critique of taped encounters involving the learner and patients.	Very high realism; allows mutual assessment of patient management and learner's interpersonal skill and professional qualities.	Can be "highly charged." Some learners need time to "desensitize" from seeing or hearing themselves on tape.

Table 6.1.1. Evaluation methods commonly used in medical education			
Method	**Description**	**Advantages**	**Problems**
14. Educational prescription contracts[42]	Written agreement between learner and evaluator about learner's educational goals for a specified period of time	Clear specification of learner's educational intentions and outcome measures. States educational criteria and standards. Notes what support faculty will provide.	Some learners and faculty are reluctant to express expectations for one another.
15. Portfolios[43-45]	Maintenance of a tangible cumulative record of clinical, scholarly, or professional accomplishments. May contain products like publications, records of sites visited, data on teaching skill, and other material. Should not duplicate educational transcript.	Detailed accounting of student's nonacademic accomplishments using hard evidence from material products.	Requires high degree of learner compliance. Must specify inclusion and exclusion criteria for eligible entries.
16. 360^0 evaluation[46,47]	Evaluation of a medical learner using rating data from a variety of sources, e.g., self, peers, supervisors, nursing staff, patients	Broad array of data sources presents a rich portrait of the learner's perceived competence. Allows normative comparisons using different data sets.	Cumbersome and difficult to manage without a computer or web-based system. Requires high compliance from different data sources.

No single evaluation method is valid for all purposes. Academic physicians need to think hard about their reason for wanting to assess a student's knowledge, procedural skill, self-confidence, dependability, honesty, or any other clinically relevant characteristic. Only after identifying the purpose of the evaluation (e.g., educational diagnosis, technical proficiency, overall performance on a rotation) should the clerkship director select a measurement tool that will produce meaningful data to inform the needed decision.

Evaluation and Learner Motivation

Seasoned medical educators know that examinations shape and drive student behavior. Today's medical students live from test to test, usually viewing each evaluation experience as a sentence rather than an opportunity. For medical students, the evaluations they encounter are an operational definition of the curriculum because no matter what is presented, read, practiced, or discussed passing tests defines life in medical school. This issue was raised 44 years ago in 1961 by George Miller in his famous book, *Teaching and Learning in Medical School*.[48] Not cited by Miller, the identical point was made about British medical education in the 19th century including much faculty grousing about "test driven" students.[49]

Recent research confirms that even small changes in the emphasis or format of evaluation procedures prompt revisions in the way that students prepare for and approach examinations. This holds for learners in general[50] and medical students in particular.[51]

The origins of this behavior are not hard to detect as discussed by Good.[52] She astutely describes the widespread and high-level culture of "evaluation apprehension" in the medical profession. Left unchecked this apprehension can have bad effects like needless competition; reduced student cooperation; defensiveness; attempts at one-upmanship; and reliance on expensive, extracurricular commercial test preparation courses that have no tangible benefits.[4] The challenge to medical educators is to craft and use evaluation and grading methods that truly are tools for student improvement not weapons that intimidate. The following sections of this chapter provide blueprints to fulfill that goal.

Section 2. A Primer of Evaluation: Definition and Important Distinctions in Evaluation

Louis N. Pangaro, M.D.

This section provides a lexicon for key issues in evaluation and assessment through a series of definitions and distinctions. The purpose is to provide clerkship directors with a quick reference to key terms that guide the practical decisions to be made in clerkships. Since terms are sometimes used differently in different contexts, and by different authors, etymologies are provided to root meaning in the embryology of the term. (Etymologies are based principally on *The Compact Edition of the Oxford English Dictionary,* Oxford University, Press, 1971.)

The following definitions and distinctions are included:

- Evaluation vs. Grading vs. Assessment
- Formative vs. Summative Evaluation
- Process versus Product Measurements
- Dichotomous vs. Scalar Grading
- Normative vs. Fixed standard Criterion-based
- Compensatory vs. "Weakest-Link" Models
- Descriptive vs. Quantitative Methods ("Subjective" and "Objective")
- Analytic vs. Synthetic Approaches; Developmental Approaches
- Competence vs. Performance
- Reliability and Validity; Feasibility
- Curriculum vs. Syllabus

Evaluation vs. Grading vs. Assessment

Evaluation, rooted in "value" and derived from the Latin *valeo,* (to be strong), indicates a judgment of how well a student strengths correspond with the "values" of the concerned communities, including the department, school, and the profession. *Grading* implies assignment of a label to the level of performance achieved, and derives from the Latin word *gradus,* or step. Grading within a medical schools is, effectively, an administrative action classifying the level of performance achieved. While *evaluation* implies a description in words of how a student is performing, *grade* implies a concise label that can be expressed with letters, labels or even numbers *(*A, B, C, D, etc.; Honors, High Pass, Pass, Low Pass, Fail, Incomplete, Withdrawal; 96%, 76%) of the level achieved. *Assessment* is sometimes used to embrace the entire process of evaluation and grading. It comes from a Latin term meaning to set a tax. (The term assessor would mean someone who "sat at" a judge's bench). However, it is can also be used to refer to the process of measuring something (a radio-immuno-"assay"), or of acquiring direct observations about a learner ("sitting next to" the student). The term *assessment,* then, combines something of the quantitative and qualitative aspects of gathering data for evaluation.

While there is some flexibility, perhaps even disagreement, on which terms are used for which part of the process, it can be useful to construct a sequence in which, together, the terms establish a rhythm (assessment-evaluation-grading), and constitute three-phase process that corresponds to the familiar rhythm of clinical medicine that, in turn, reflects the classical sequence of observation-reflection-action. In this sequence, grading and administrative action, and feedback would be an educational action. (Cf. Table 6.2.1)

Table 6.2.1. The rhythm of the evaluation process

Educational process	Aristotle	Clinical process	(S.O.A.P.)
Assessment = making observations about learners	**Observation**	**History and Physical**	**(S.O.)**
Evaluation = determining learner's	**Reflection**	**Diagnosis**	**(A.)**
Grading/Feedback = taking an action • administrative/societal • educational intervention	**Action**	**Therapy**	**(P.)**

Practically, decisions about who is asked to evaluate a student, and who gets to "grade", have to be decided in each setting, and teachers' responses often depend on how they see the consequences of their role in this process. Grades are often submitted to the registrar's office as terse summative letters (A, B, C, etc.) or steps (Honors, High, Pass, etc); and, these reductions of performance into a single letter can be seen by teachers and students as categorical judgments on the student as a person. Hence, the grading framework used dictates a choice of terms that can affect what teachers are willing to contribute to grading.[53]

Formative vs. Summative Evaluation

Formative evaluation is done to "form" or shape the subsequent performance of a learner, specifically by generating and providing feedback. It is done during an experience, and can be done by teachers as frequently as time will allow, but it should also be done formally at specified times, for instance, halfway through an experience. *Summative* evaluation is done at the end of a unit of time, typically at the end of the clerkship, and "sums" up the student's performance. Whereas formative evaluation is done primarily for the sake of the student, summative evaluation fulfills our responsibility to society, pronouncing the student ready for the next level of training. Summative evaluation often includes a *grade* as well as narrative description of performance and recommendations for improvement. A grade without comment would provide only minimal guidance to a student and would not help the student improve subsequent performance. Therefore, it is recommended that a grade (label) always be accompanied by and evaluation (description in words).

Process versus Product Measurements; Baseline measurements

This distinction is meant to capture the difference between the curriculum that students experience (process) and their achievements (product, outcomes). The concept is often described as the process-product paradigm.[54] Process measurements could include documentation that students have actually completed clerkship tasks (number of patients seen, number of procedures done), while product measurements include typical, end-of-clerkship assessments (e.g., NBME subject exams). Often, our research tries to document the

relationships between what we "do" to students, and how they are changed by the experience. Since research shows that much of what individual students actually achieve depends as much on their personal characteristics as much as on the formal curriculum, it is useful to document to their "baseline" status, that is, what they bring to the clerkship, by having pre-clerkship measurements such as pre-clerkship GPA, or USMLE step 1 scores.[55]

Dichotomous vs. Scalar Grading

Dichotomous grading (etymologically from Greek, “cuts into two”) divides a group of students into those who pass and those who fail. *Polytomous* (“cutting into many parts”) or *Scalar (scala* = steps in Italian) grading recognizes a broader spectrum of student performance by providing for a series of steps for assigning grades, such as Honors, High Pass, Pass, Low Pass, Fail, or the equivalent letter grades, A, B, C, D, F. Continuous grading would refer to a series of numbers which have small intervals, such as 88%, 87%, 86%, etc. Generally speaking, dichotomous grading fulfills our responsibility to society by determining whether a learner is competent or not. Scalar and continuous grading helps faculty and students compare performances among students, and may also help graduate program directors rank their applicants. For quantitative assessments (such as multiple choice examinations or OSCEs) the conversion from an exam score to a final grade can be straightforward, even if the cut points are arbitrary. However, students and teachers have had an ongoing concern about the lack of clarity in how descriptive assessments from teachers are converted into a step-wise grading system (such as Honors, High Pass, etc.). One simple method of addressing this problem is to categorize teachers’ observations about a student's performance into a step-wise, such as, second-year level, third-year level, fourth-year level, internship level; or, reporter, interpreter, manager/educator.[20, 53] (see Chapter 6, Section 13 [Converting Evaluation into Grades])

Normative vs. Fixed standards (Criterion-based)

Normative grading is “relative” and it assigns grades to students’ performance by comparing them with another group, the “norm”, such as a contemporary peer group. This comparison group could be a national reference such as all students taking a certifying examination, or a local group of students taking a clerkship at the same time of year. Normative grading can be done in a mathematical way, generating a “curve”, with grade rankings based on distance above or below the mean score. Normative grading is often done less formally, with half students in the middle (for example, a grade of High Pass), a quarter receiving Pass, and a quarter receiving Honors. In any case, the essence of normative grading is to compare students to each other.

What is often called c*riterion-based* grading sets mastery standards for each grading level (pass, high pass, etc) and is more “absolute”, less relative, than norm-referenced methods. Basically, “criterion-based” grading is really *fixed-standard grading*, in which experts first decide what the tested domain will be (the criterion, the “what”) and then what will be expected standards of proficiency (fixed standards, the “how much?”). This approach depends upon a prior judgment of what has “content validity” (see below). For example, in the domain of manual skill in suturing, the fixed-standard is the degree of proficiency that must be achieved – adequacy of would closure, the number of sutures used, and the time taken to close the wound. The examiner then decides whether the standard has been met, and how well (*crites* means a “judge” in Greek).

Choice of a criterion-based or fixed-standard system is one of the most difficult choices made in a clerkship, and has powerful consequences upon grading decisions. In a fixed or absolute standard system, a group of three students working with a single teacher could all receive grades of Pass or all grades of Honors, depending on the criteria they met. In a normative system, they are competing against each other.

Another consequence of a fixed-standard grading system is that it would typically yield more grades at the upper end of the grading spectrum at the end of an academic year, when students would typically perform better; whereas, a normative grading system would try to assign the same number of Honors grades at the start as at the end of the year. This highlights the difference between evaluation and grading. At the start of the year, performance as a strong "interpreter" might lead to a grade of Honors, but at the end of the year only to a grade of High Pass.

In practice, most clerkship directors agree that the dichotomous pass-fail decision should be based on criteria, rather than an arbitrary failing of a certain percentage of students in each clerkship for each year. It is the distinction between Honors, High Pass, Pass, etc. that is more problematic. Each institution, or perhaps each clerkship, has to decide which is fairer to patients and society (ranking students based on mastery of certain criteria) or fairer to students (assuring equal distribution of grades, irrespective of the time of year a student takes the clerkship.)

Compensatory vs. "Weakest-Link" Models

A *compensatory* grading system averages aspects of a student's performance using various parameters to yield a final grade. For instance, a high score on a multiple-choice final examination plus a failing clinical evaluation might calculate to a grade of Pass. A *non-compensatory* ("weakest link") approach would conclude that the student is not better than his/her lowest level of competence in a core area of evaluation. For instance, an excellent examination score would not compensate for poor professionalism, or vice versa. Therefore, a student with unacceptable performance in any domain of evaluation could not receive a passing final grade. Generally speaking, clerkships must determine which aspects of performance are so important that deficiencies in any cannot be compensated for by proficiency in others.

Descriptive vs. Quantitative Methods ("Subjective" and "Objective")

Descriptive methods of evaluation describe a student's performance using words. *Quantitative* methods try to measure performance and yield a numerical score. Most summative grades are a combination of the two methods with some consistency in weighting descriptive methods more than quantitative ones. A survey of internal medicine clerkship directors reported that, average, 25% of the clerkship grade was derived from the NBME subject examination,[56] this figure was 33% for surgical clerkships,[57] and 31% for Psychiatry Clerkships.[58]

There is a tendency to refer to quantified examinations as "objective" and narrative evaluations as "subjective". However, these terms can be misleading. In comparison to descriptive evaluations, a multiple-choice examination is dispassionate (not caring, for instance, about how confidently a student speaks), has a single "grader" (the scoring device) and its precision and reliability are more easily calculated. However, we should not confuse objectivity with reliability;

and “objectification”[59] may be a better term for MCQs or OSCEs. In any case, objectivity (or objectification) does not mean that in assessment itself has validity. Each step in creating a multiple choice question, decisions about what to test and wording of the item, involves judgments that reflect the opinions of teachers.[25]

Unspoken assumptions in the process of converting teachers’ evaluations into grades often lead students to regard teachers’ evaluations as subjective and arbitrary. Many students protest a lower-than-desired grade by arguing that a high score on a multiple choice test is “objective” (and therefore, valid) and that the narrative evaluation describing unprofessional behavior is “subjective” (and therefore not valid). Yet, descriptive methods can achieve a level of reliability (see below) and validity that is sufficient for high stakes decisions.[21, 60] Both assessment methods have a role in determining summative grades and one is not inherently more valuable than the other, so the terms “subjective” and “objective” – which undervalue the former - should be avoided if possible.

Analytic vs. Synthetic Approaches; Developmental Approaches

Traditional evaluation theory “analyzes”, or “breaks up” a student’s performance (to analyze in Greek is to “loosen up” or “take apart”) into several components, knowledge, skills and attitudes (or, attitudes, skills, and knowledge, “ASK”). Each component can be assessed by tools appropriate for each domain. For instance, multiple choice tests might be used to assess knowledge, and standardized patients can assess history-taking skills.

A “synthetic” approach “puts things together”, and asks how the student’s abilities in several domains come together to achieve a level of proficiency. The RIME Scheme[20] introduces a vocabulary for synthetic evaluation of students’ clinical skills. This describes development in clinical skills from “Reporter” to “Interpreter” to “Manager/Educator” (RIME) in which each task requires *all three* facets of the analytic model. For instance, a reliable “reporter” must combine skill in physical examination technique with the knowledge of what to look for in the patient at hand, and also with respect for the patient’s privacy; the ability to honestly and accurately communicate findings must be combined with a sense of duty to fulfill responsibilities each day.

The rhythm of RIME corresponds to the same sequence as observation-reflection-action and S.O.-A. – P. While there is a developmental aspect to this, it does not imply that all students go sequentially through stages of development. Rather, the RIME scheme is intended as a "razor" defining a level of performance below which the learner should not fall.

Recently, there has been initiative to apply the ACGME approach of the "six competencies" to medical students. Three of the "competencies" fit the analytic model: professionalism, interpersonal skill, knowledge) and three are synthetic: patient care, system based practice, and practice-based learning an improvement.

Analytic and synthetic approaches are complimentary. For instance, the RIME synthetic vocabulary offers an initial assessment framework for organizing observations about a learner’s development toward independence. A teacher who recognizes that a student is an effective reporter, but not yet an interpreter, should switch to an analytic approach in order to determine what will help the student take the “next step”. For example, if there is a problem moving from reporter to interpreter, does the student need to acquire more knowledge, to practice the skill of

differential diagnosis, or to become more confident? Analytic and synthetic approaches reinforce each other.

The ACGME approach is intended to reach a dichotomous decision about competence at the point when a resident leaves training, and moves into unsupervised practice; therefore, it minimizes the developmental approach. Clerkship students are in the transition from pre-clinical status to internship, and some developmental aspect is usually required in framing the evaluation system.

Progressive refinement of cognitive skills has an ancient pedigree. Plato described the progress from observing facts to observing and identifying the abstractions below them; in other words, the progress from reporter to interpreter. Aristotle was even more explicit in defining the fundamental rhythm of cognitive processes: observation-reflection-action, with further reflection based upon action. This developmental approach has been captured educationally in Bloom's taxonomy[61] for cognitive progress in which, simply, there is progress from the possession of facts, to being able to explain the facts, to apply them to new situations, to synthesize intermediate conclusions, and to reach value judgments. The Dreyfus brothers described six stages of progress from novice, to advanced beginner, to competent, to proficient, too intuitive expert and finally to mastery.[62] While these are generalizations, and difficult for every day teachers to apply to specific students, they do capture the expectation that a student will be able to accept progressively higher levels of responsibility. We have to recognize that students can be more advanced in their level of performance on some patients, that on others. This is the principle of content-based expertise. Nevertheless, clerkships often have to decide what is acceptable performance at the end of each clerkship rotation, and whether there should be different at different times of the year, or if a student is returning to the clerkship in the fourth year in remediation for prior substandard performance.

Competence vs. Performance

These terms have complementary meanings, but their meanings are sometimes used interchangeably, and educators should pay careful attention to how the terms are being used in a specific context. In the more common use of the terms, “competence” is what a student has the ability to do at certain times or under test conditions (in this sense, related to the etymology of the word, to strive with, or to “compete”) and “performance” is what a student does consistently on a daily basis, even when not being watched. This distinction is best reflected in the “Know-Can-Do” description of a levels of accomplishment described in Miller's triangle; that is, the student “knows what to do”, "can apply it”, “can do it successfully under test conditions”, and “*does* do it*”* regularly. Alternatively phrased, the student “knows how”, “shows how” and “does”. So, the distinction between competence and performance also highlights two differences, one in the setting - *in vitro* (a simulation center) and *in vivo* (actual practice), and another in process (whether the person is being observed, or is aware of being observed)

However, these terms can also be used in exactly the reverse senses, in which “performance” refers to a display while being observed (i.e., performing for an audience), as in being “on-stage”, in test conditions, and “competence” denotes all the attributes to function independently. In this less conventional use of the terms, competence can actually never be demonstrated until it is actually achieved in a sustained, independent way in practice.

In practice competence is defined in many ways and embodies many frameworks. In the analytic model, competence is proficiency in tasks in each of the contributing domains (knowledge, skills and attitudes). In a developmental model, competence can be described in relation to the steps above it (intuitive expertise), and below it (proficiency).

In the synthetic model, competence is putting all the necessary characteristics and qualities together for each patient in a sustained way. The definition of competence in a profession, in this model is the ability to give to every situation that a professional might face all that properly belongs to that situation, and no more.[63] This means that a competent person first has to make the decision about what a situation requires. Since the efficiency and judgment needed to exclude unnecessary effort implies a level that is beyond most students, it may not be appropriate to use the term "competence" for students at all. Practically, our concrete expectations for students or interns should require that they consistently do all the important things for their patients (for instance, accurately report all important findings) but reward their having the ability to leave out less important with a higher grade.

Do clerkship directors judge that a learner is "competent" (or has "competence") when proficiency is achieved in each of several "competencies", or must they all be brought to bear, consistently, in the care of individual patients? Actual practice situations are truly *in vivo*, and have the complexity of authentic decision-making. *In vitro* tests, such as clinical skills examinations, focus on clinical "performance" and have often narrowed down the task for the learner. While use of the analytic method to create an assessment method for some single aspect of competence is quite useful at the undergraduate level, it can never be entirely successful for a resident about to begin unsupervised practice.

Clerkship directors therefore will typically use a variety of quantitative methods to assess aspects of competence (written examinations, direct observations of interviewing skill, etc. See sections 6 through 12 of this chapter) and rely on summary observations of teachers to see whether they can put things together (see Chapter 6, Section 3 [Descriptive Methods])

"Competencies"

This term has become popular since the introduction by the ACGME of the six "general competencies" which are to guide the teaching and assessment of those in graduate education.[12] The six items do not together *equal* "competence", but all are part of the characteristics and detailed skills sets expected to be present in a resident ready for independent practice. In a sense the "competencies" do not describe competence, but are a framework with which program directors can assess competence, competency by competency with a toolbox of methods for each.[12] This fits quite well with the intention to facilitate the ACGME's Outcomes Project, which will link process in training to product (outcomes) at the end of training or in subsequent practice. This is a very exciting development which should foster educational measurement and research. The framework of competencies will be seen as a combination the "analytic" model noted above in the first three items, and three "synthetic" items that describe tasks to be mastered.

The "competencies" are intended to benchmark the final level of proficiency achieved by each resident, so they do not contain an explicitly developmental aspect. Clerkship directors have therefore debated their utility for medical students. The question has largely been rendered

moot by the influence strong forces of regulation of the ACGME and the endorsement of the AAMC (see Chapter13: Understanding, Navigating and Leveraging American Medicine). Therefore, clerkship directors must articulate what would be expected of a starting and finishing third-year student, and finishing fourth year students. Similarly, program directors must make expectations clear for interns and PGY2 residents.

There are assessment methods appropriate for each of the competencies (please see detailed Table 6.1.1 in Chapter 6, Section 1). Although this chapter is not organized by the "competencies", there are discussions appropriate to each in the following section in this chapter:

- Medical knowledge: Chapter 6: Section 10; Section 11; Section 13
- Interpersonal and communication skills: Chapter 6: Section 3; Section 4; Section 8; Section 9
- Professionalism: Chapter 6: Section 1, Table 6.1.1.(360^0 evaluation); Section 3; Section 8; Section 9
- Patient care: Chapter 6: Section 3; Section 4; Section 6; Section 7; Section 8
- Systems-based practice: Chapter 6: Section 3; Section 4
- Practice-based learning and improvement: Chapter 6: Section 1, Table 6.1.1 (Portfolios); Section 10

Reliability and Validity; Feasibility; Impact

Reliability is the consistency, replicability, stability, or reproducibility of results (in Latin, to rely on - *religare* - is to trust). Reliability is the amount of the observed variance that is due to the student (true score variance) rather than the test and everything else (error variance), and is usually expressed as a decimal figure between zero and 1.0. High reliability suggests that the "signal" (what we want to measure) is sufficiently greater than the "noise" (problems inherent in the assessment method), so that we can consider the results reproducible, or at least representative. For high stakes decisions, at least 80% of the variance should be true score variance (a reliability figure of 0.8).[64] (for discussion of reliability statistics see Chapter 6, Section 12.)

Validity is confidence that you are measuring what you want to measure, what you "value" (similar in etymology to "evaluation"). There are several terms dealing with validity with which clerkship directors should be familiar.[65] *Content validity* reflects whether assessment reflects enough of the domain you want to assess, and this can be made as a judgment of experts, or by comparison with some external standard, such as from the core curricula available from clerkship groups (CDIM, STFM, etc.). *Face validity* judges whether the assessment method seems to experts to be appropriate for competency in question. For instance, use of a multiple-choice test to assess interpersonal skills would not have face validity. C*onstruct validity* means that results are consistent with reasonable theory (e.g., experts perform better than novices). *Criterion/concurrent validity* is more numerical, and determines whether the results of your assessment method agree with other appropriate measures of students' performance. *Predictive validity* refers to whether results of one assessment measure are verified by subsequent performance, and this, too, is best demonstrated with mathematical methods, such as correlations and linear regression. *Consequential validity* is the term applied to a judgment about whether the effects of an evaluation system, typically social effects, are desirable. For

students, and perhaps for clerkship directors, one consequence of grades might be a student's choice of what GME specialty to apply to. Clerkship directors are referred to the excellent articles by Downing[66-68] on these subjects.

Feasibility deals with whether an evaluation can actually be conducted in your own clerkship setting (from the French, *faire*, "to do"). Time to prepare and conduct the assessment, money to support the development, and space all contributes to feasibility. Feasibility is often the rate-limiting step in deciding how we evaluate our clerkship students. To some extent, *acceptability* to students and faculty is another aspect of feasibility. For students, their acceptance may be contingent upon perceived fairness, or upon cost in time and money. For faculty, simplicity of use and perhaps being distanced from legal implications would be[69] the priorities. Nonetheless, it is preferable to develop reliable and valid tools; then try to make them work. Another factor of assessment has been called the "educational impact" on students, how they change their strategies of studying to match not only the content but the format of assessment.[87]

Curriculum vs. Syllabus

To some extent, what we measure and reward will determine what students learn; in other words, "assessment drives the curriculum". The list of topics or skills that we wish students to master is the syllabus (the term, etymologically, means "list"), and the methods we use to help students master the list, collectively, are "curriculum" (that is, the "horse race" we put students through, from "currere", "to run", as in the word "current"). This distinction has implications for evaluation. If each of a school's third year clerkships has a different list of topics to master, these are typically knowledge-based, and will require an emphasis on multiple-choice tests to establish content mastery. On the other hand, if schools wish to have common goals across clerkships, then these must be process-based, such as skills in interviewing and physical examination, in differential diagnosis, and in rapid mastery of the necessary knowledge to go beyond collecting facts to interpret them. In this approach, "curriculum" for third-year students might be seen as an expectation to move from reporter to interpreter; the basic strategy for clinical teachers would be to have a clear expectation that a student will offer a reasonable opinion.

Most clerkships accept a responsibility to be both discipline-specific (proficiency in the unique syllabus of subjects not taught elsewhere) and interdisciplinary (emphasizing common expectations which will lead to a successful performance in residency). As a consequence, the clerkship's blueprint for evaluation might identify, explicitly, the methods to assess both the discipline-specific and the inter-departmental goals.

Section 3: Descriptive Evaluation

David Carnahan, MD and Paul A. Hemmer, MD, MPH

Introduction

The focus of this section will be the descriptive evaluation of medical students by teachers during clinical clerkships. We will discuss the purpose of descriptive evaluation, its characteristics, strengths and potential deficiencies, as well as offer suggestions on how to improve the quality and credibility of descriptive evaluation. We will complete our discussion with a look at a synthetic

framework for evaluating the performance of students using descriptive evaluation known as R-I-M-E.

The purpose of descriptive evaluations

What is descriptive evaluation? Descriptive evaluation is the term applied to the words instructors use in their assessment of students' demonstrated competency across the domains of knowledge, skills and attitudes, and it is usually based on their observations of students over a given period of time. (see also Chapter 6, Section 2) Their words should provide evidence of students' strengths and weaknesses, give examples of achievement or deficiencies, and serve as the basis for direct, meaningful feedback to the student and for recommending advancement or remediation. Some have described this as "clinical performance appraisal."[70]

Characteristics of descriptive evaluation

Unfortunately, descriptive evaluation is often referred to as "subjective" evaluation.[71, 72] This may have been "encouraged by psychometricians and behavioral scientists who have labeled narrative judgments as unreliable and 'soft', and have urged faculty to focus on methods that yield 'objective' assessments"[73] and reflects the bias toward believing that which is expressed in numbers rather than in words.[74] However, Eisner has asserted that expert judgment is likely the superior approach to evaluating competence in fields in which science and art are mixed.[75] We believe that use of the term "subjective" is detrimental to the evaluation process in that students and faculty often infer that a "subjective" assessment method is inferior to an "objective" method, such as multiple choice examinations. One counterpoint made to this notion by Norman et al. states that "objectivity does not necessarily result from the strategies of objectification (a set of strategies to reduce measurement error), and the application of these strategies may have undesirable consequences."[59] "Descriptive" more accurately defines this type of evaluation—conveying one's ideas, thoughts, observations, and a synthesized judgment with words.

Descriptive evaluation is a component of an overall system of evaluation that also frequently incorporates quantifiable examinations of knowledge and/or skills evaluation.[76-77] Descriptive evaluation is unique because it involves all aspects of the evaluation system, including evaluators, students, content of evaluation, and learning environment.[78] Additionally, it assesses competencies not easily measured by knowledge or skills examinations, such as responsibility, integrity, compassion, maturity, and the application of knowledge in the clinical problem-solving of direct patient care.[78]

Clerkship directors place great emphasis on instructors' comments in determining grades.[79] Studies of required clerkships in the United States and Canada demonstrated that clinical instructors' evaluations account for 40 to 60% (range, 0-100%) of students' final clerkship grade[80-84] Given the reliance on descriptive evaluations in the grading process, clerkship directors must strive for reliable and valid descriptions. The evaluations should be based on as many direct clinical observations of the students as feasible, describe students' performance based on uniform criteria established by the clerkship faculty, and cite specific examples of behavior and performance.[33,54,85-86] Evaluators should make specific, behaviorally based comments that cite strengths and weaknesses, thereby providing meaningful feedback to the students. As a result, the evaluations would help clerkship directors and faculty teaching in the clerkship discern and tailor interventions for those students who are superior, average or marginal, as well as those who are failing.[33, 70, 87]

Studies of instructors' ratings of medical students have shown a remarkable similarity in the elements instructors emphasize. Typically, instructors have emphasized the students' interpersonal skills in dealing with colleagues and patients, their professional attitudes and behaviors, as well as their ability to apply knowledge and solve clinical problems.[33, 77, 88-89] However, instructors at various levels of training and experience may place greater emphasis on different factors. Residents are likely to value a student's procedural skills, work ethic, and motivation to help the team, while attending physicians are likely to place greater value on a student's knowledge and reasoning skills.[90-91]

These studies are based primarily on instructors' annotations on an evaluation form rating scale and not on their narrative comments. These rating scales, which usually address a student's knowledge, skills, and professionalism, are used by most clerkships; although, some clerkships may now be adapting their ratings' form to reflect the ACGME core competencies.[92-93] Regardless of the domains assessed, instructors mark or circle the point on the scale they believe corresponds to the observed level of student performance. The scales are usually numerical (from three to nine options per rated domain) and may be either simply numbers or may contain more detailed written descriptors of student performance (Appendix 1). Ideally, instructors should be trained in the proper use of the forms, understand how their evaluation contributes to grading, and understand the criteria for specific levels of student achievement for each rated category, as well as overall performance (e.g., failing, marginal, satisfactory, outstanding). Further issues concerning rating scales will be discussed in the next sections on, Problems with descriptive evaluation and improving descriptive evaluation.

We believe that that most important role that evaluation forms can play is to clearly and concisely communicate goals to teachers. The forms can be one way to communicate expectations for what teachers should assess, and provide guidance and a common language to create a frame of reference from which to evaluate students.

Despite their limitations, these studies demonstrate that instructors' evaluations of students assess the breadth of competency: knowledge and its application, problem-solving skills, and professional qualities. Many faculty believe that assessing qualities of professionalism may be the most important aspect of evaluating medical students.[94] There may be no better evaluation method to assess professional qualities than faculty and residents who observe performance on a daily basis. In fact, recent studies demonstrate that faculty ratings and comments form the centerpiece of an evaluation process focusing on professionalism,[95] and that such comments made by teachers about students may identify those individuals at risk of future unprofessional conduct.[14]

Problems with Descriptive Evaluation

Despite the acknowledged importance of descriptive evaluation, there are problems with this type of evaluation. The Association of American Medical Colleges (AAMC) conducted a survey in 1983 as part of the Clinical Evaluation Program (CEP) to determine faculty perceptions of clinical evaluation systems.[96] Faculty respondents from 136 U.S. and Canadian medical schools cited concerns related to the evaluation process and quality of evaluations, specifically the reliability, objectivity, uniformity, validity and feasibility of evaluating specific content areas. Subsequently, clerkship directors and instructors at 10 U.S. medical schools were asked to comment on the seriousness of problems with the evaluation of medical students.[74] The results are summarized in Table 6.3.1. There was remarkable overlap between the instructors' and clerkship directors' perceptions of the most serious problems: inadequate guidelines or lack of information regarding how to handle students with problems; unwillingness to document poor performance; lack of training as evaluators; unclear definition of the role of evaluators; and insufficient definition of the

criteria of evaluation. Unfortunately, despite the fact that this study was conducted nearly two decades ago, the problems cited are still likely to ring true to clerkship directors.

Instructors' unwillingness to record negative comments in evaluations does not necessarily mean that instructors are not able or willing to identify "marginal" or failing students.[73, 87] Instructors are often willing to verbally discuss their concerns, but are reluctant to document, on either a rating scale or in written comments, these same concerns.[77, 97-99] Reasons for reluctance include fear legal action, lack of administrative support for unpopular decisions, an unwillingness to be involved in follow through on difficult cases, or "passing the buck" to other evaluators.[78] Also, instructors may feel their role as teacher and mentor may be in conflict with that as an evaluator, or they may have difficulty with delivering "bad news". A national survey regarding grade inflation showed that 82% of respondents believed that faculty were reluctant to give low grades because of students' expectations of higher grades, fear of legal action or student "hassle", belief that students with strong work ethic should not fail, and that assigning higher grades may entice students to their specialty. Of further concern, forty-three percent of the clerkship directors surveyed felt that we are unable to identify incompetent students.[100] These findings are disappointing for several reasons. First, the courts have consistently upheld the judgment of faculty in cases in which students have not met academic or professional standards[95,101] (See Chapter 6, Section 14 [Legal Aspects of Failing Grades]). Second, it would also appear that the "halo effect" continues to strongly influence an instructor's evaluation,[102] and finally, students' expectations, sense of entitlement, or tenacity in challenging grades appears to have undue influence on instructors.[103] Even if only one instructor states or records a negative comment, it likely has substantial merit.[14, 89, 98, 104-105]

Reliability

Studies of instructors' ratings of students' written case reports, as well as ratings of videotaped encounters of trainees interviewing, examining, or presenting a patient have shown low intra-rater and inter-rater reliability.[106-108] Although some of the low reliability may be due to instructors focusing on different aspects of student performance, standardized rating scales only modestly improved reliability.[90-91, 107.]

Other studies suggest that instructors' clinical evaluations can achieve sufficient reliability for "high stakes" academic decisions (usually considered to be a reliability coefficient > 0.8). Carline and colleagues[109] analyzed individual instructors' ratings from a standardized, descriptive clerkship evaluation form and achieved a reliability of 0.8 for assigning clerkship grades when at least 7 observations of student performance were available. More recently, Williams et al. found a reliability of 0.8 was possible when evaluating surgical residents at least 8 times with no improvement in the reliability when more rating scales were added to the evaluation form.[93] Time during the academic year, clerkship site, and academic level of the rater had little effect on the ratings. A study of reliability yielded slightly lower coefficients across clerkships when 8 raters evaluated each student.[110] In this study, the student's score did seem to depend on the instructor to whom they were assigned and the clinical context in which the rating was performed. Use of global rating scales yielded inter-rater reliability of 0.83-0.91 in one study.[111] The authors attributed this high inter-rater reliability to definition of the parameters rated, instructors who had direct, prolonged and close observation of relatively few students, ratings which were assigned after consensus among all supervisors, and training the raters to use the evaluation forms. Another study by MacRae et al.[112] compared physician ratings of 120 videotaped medical student encounters using four cases, they noted similar inter-rater reliability with an average reliability coefficient of 0.85. They also attributed the high level of agreement due to collaboration on the rating scales that were used in the study.

While high reliability coefficients are desirable, lack of agreement among instructors' evaluations is not necessarily undesirable. Different instructors may focus on different aspects of student's performance, but in aggregate, the ratings may provide a more comprehensive picture of a student's performance.[90-91, 93] Limited variability in instructors' ratings may be detrimental if it leads to overemphasis on other measures of student performance, such as written examinations.[113] Ultimately, the clerkship director must decide whether areas of disagreement among instructors are desirable or undesirable.

Validity

The validity of descriptive evaluations has been questioned in studies that have centered on the predictive, concurrent, content, and face validity of descriptive evaluations. (See definitions earlier in Chapter 6, Section 2). One study examining the predictive validity of clerkship evaluations found that overall competence could be predicted better than professional behavior during residency. Students with good communication skills were more likely to receive higher overall competence ratings.[114] Students who had either cognitive or non-cognitive deficiencies identified during an internal medicine clerkship were 13 times more likely to receive low ratings or comments from internship directors than those without deficiencies.[87] As previously noted, a case control study suggests that comments and ratings that identify unprofessional behavior of medical students likely highlight individuals that are at risk of continued unprofessional behavior.[14]

Studies have also raised concern about the concurrent validity of instructors' evaluations, as evidenced by low correlation between instructors' end-of-clerkship evaluations and students' performance on end-of-clerkship knowledge and/or skills examinations and licensing exams.[71, 115-118] However, this low correlation may not be unexpected. In addition to assessment of student's knowledge, instructors' evaluations assess clinical skills and attitudes, thereby, assessing characteristics beyond the scope of knowledge or skills examinations. The different types of evaluations may be measuring different characteristics, reinforcing the need for multiple methods of evaluation. [71, 117-122]

Content and face validity of instructor evaluations have also been questioned. For example, instructors' ratings of videotaped case presentations seemed to depend on the "likeability" of the student and judgments about competency reflected students' communication skills.[108, 119] Assessment of one trait (e.g., knowledge) on an evaluation form correlated with assessment of other traits (clinical skills, personal characteristics) in another study.[121] Several studies have also shown that residents tend to give higher ratings to students than faculty.[115-116, 123] This may be due to a greater amount of time spent with the students, leniency in grading, or the "halo effect".[115] Resident evaluations have shown better internal consistency than faculty evaluations of students and adding resident evaluations to those of faculty improves the dependability of the evaluations.[115-116, 123] Nevertheless, Holmboe demonstrated that rating forms tailored to a specific task, such as the mini-CEX for observation of resident's clinical skills, do have content validity and that faculty can be trained to observe and record accurately their observations of a trainee.[124-125]

Many other factors may affect the reliability and validity of instructors' evaluations. These include a sense of personal failure if a student does not improve; a desire to be liked; evaluations that lack specific, behaviorally based comments; substitution of a grade for comments; differing expectations among instructors; limited student-instructor encounter time; lack of a trusting relationship between teacher and learner; failure to directly observe student performance; the interest of raters in the process of evaluation; the types of interactions (such as attending rounds vs. work rounds); and

differences in the training environment.[104, 109, 120, 126] Reliability and validity of evaluations may also be affected by instructors' failure to use the full rating scale and impatience with completing evaluation forms.[115, 127] Perceived lack of rewards for teaching may also impact instructors' willingness to participate effectively in the evaluation process.[115]

Most studies of reliability and validity have focused on the evaluation form rating scale or instructors' final ratings, not on the instructors' comments. However, the most important aspect of descriptive evaluation is the narrative comment, not the box checked on an evaluation form. It is not clear whether the findings of the studies that used rating scales would apply to the comments which instructors make.

Improving descriptive evaluation

General interventions to improve instructors' evaluations include developing and reinforcing clear performance guidelines, improving communication among faculty members, and faculty development regarding evaluation skills.[87, 97, 116] Reliability may be improved by using additional raters or investigating the sources of disagreement among evaluators.[115, 117] Using a computerized evaluation form may improve the timeliness of evaluations as well as the number and quality of comments.[126, 128-129] Relying on instructors for evaluation but not grading may improve the quality of the instructors' comments.[17, 40] Feedback to instructors on their evaluation or grading patterns may help improve future evaluations.[86, 130, 132]

Many efforts have focused on the evaluation form itself. Adding items to the evaluation form does not improve reliability.[110] However, adding behaviorally based descriptors in each evaluation category for each level of performance enhances the reliability of instructors' evaluations, a benefit that was lost when the descriptors were subsequently withdrawn.[89, 123] Behavioral descriptors on an evaluation form may contribute to instructors making more detailed written comments.[131]

However, trying to improve evaluation by continually refining a form is misplaced effort. Once the fundamentals of the evaluation form are set (see below), one must focus efforts elsewhere. Instructors' training and sense of ownership of the process are more important than the evaluation form in developing a reliable and valid evaluation system.[132] Including items related to knowledge, skills, and attitudes on a rating scale and asking faculty to assign a number rating does not make evaluation easier or the results more valid.[73, 87] In fact, evaluation form rating scales are less sensitive than instructors' comments for detecting deficiencies in competency.[98, 105] A study by Battistone et. al demonstrated that using a descriptive vocabulary to assess students created a more normal distribution and a greater range of ratings than did the use of a global numeric method.[133] Evaluation forms with written descriptors of student performance serve two purposes: (a) communicate clerkship and performance goals, and (b) facilitate evaluation. (See Appendix 1)

A recent review by Williams et al.[33] explores the sources of bias and limitations in clinical performance ratings, which we have referred to as descriptive evaluation. This is an excellent review, one that clerkship directors should read. The authors propose sixteen recommendations to improve clinical performance ratings and these are summarized in Table 6.3.2. These recommendations highlight that descriptive evaluation should not be viewed as simply the distribution and collection of rating forms. For the process of descriptive evaluation to be effective, it takes time and it needs to be an interactive process between the clerkship director (or site director) and the teachers—both housestaff and faculty. In the final sections, we will discuss an evaluation process that provides all teachers with a common frame of reference from which to evaluate clerkship student performance that is combined with regular, face to face meetings with teachers that serve as protected time for evaluation, feedback, and faculty development.

The R-I-M-E Framework

The third year internal medicine clerkship at the Uniformed Services University of the Health Sciences (USUHS) uses an evaluation framework designed to assess and foster a student's progression from "Reporter" to "Interpreter" to "Manager/Educator" (RIME).[21, 55]

Reporter: Students must: (1) accurately gather information about their patients, through an independent history and physical examination, chart review, and from other sources such as family or referring physicians; (2) use appropriate terminology to clearly communicate their findings, both orally and in writing; (3) interact professionally with patients and staff, and (4) consistently and reliably carry out their responsibilities. This stage requires that students have an adequate knowledge base, the basic skills to perform fundamental tasks, and core attributes of honesty, reliability, and commitment. Students who are Reporters can answer the "What" questions about their patients.

Interpreter: Students must: (1) demonstrate ability to identify and prioritize problems independently, (2) offer three *reasonable* explanations for new problems, and (3) generate and defend a differential diagnosis. This step requires a greater knowledge base, increased confidence and skill in selecting and applying clinical facts to a specific patient, and the ability to begin to pose clinical questions. Interpreters organize, prioritize, synthesize, and interpret problems. Students who are Interpreters can answer the "Why" questions about their patients.

Manager: Students must be more "proactive", suggesting diagnostic and therapeutic plans that include reasonable diagnostic options and possible therapies. This step takes even greater knowledge, more confidence, and the skill to select interventions for an individual patient. Managers understand their patients' needs and desires and can enter into "or relationship-centered care".

Educator: Becoming a Manager is intricately tied to being an Educator. Students must identify questions related to their patients that cannot be answered from textbooks, cite evidence that new or alternative therapies or tests are worthwhile, and share their acquired knowledge with other members of the health care team. Desire and ability to educate oneself and others is intrinsic to being a "manager" and reflects a desire not only to teach colleagues but also, and most importantly, to help the patient. A Manager/Educator answers the "How" questions, for themselves, and their patients. It is not simply a matter of "bringing in articles to the team."

In our third-year clerkship, "passing" requires mastery of "reporter" skills and evidence of some transition toward "interpreter". Acquisition of skills as a consistent, *reasonable* "interpreter" constitutes a higher level of performance. Consistently demonstrating skills at the "Manager/Educator" level reflects performance beyond expectations for a third-year clerk (what might be expected of a fourth year student). RIME is "synthetic"—each level encompasses the traditional analytic framework of knowledge, skills and attitudes. It is a criterion-based framework for evaluating the performance of students.

Importantly, there is a Rhythm to RIME that cuts across medical specialties (see also Chapter 6, Section 2). It is a readily understood frame of reference from which all teachers can evaluate student performance.[20, 86] RIME captures what clinicians do when they interact with patients: Observation (Reporter), Reflection (Interpreter), Action (Manager/Educator) and what they write: "Subjective/Symptoms" and "Objective/Observations" (Reporter), "Assessment" (Interpreter), "Plan" (Manager/Educator). Furthermore, RIME also helps teachers understand the minimal level of performance below which a trainee cannot fall. For example, it would be unacceptable for a

student to be able to Interpret data that they are given if they cannot demonstrate that they are able to reliably obtain the information themselves from the patient.

R-I-M-E is readily "portable" and applicable in ambulatory care or inpatient ward settings. It can be incorporated into the student's clerkship evaluation form (see Appendix 1), into the student's clerkship handbook, onto "encounter cards" used in ambulatory or ward settings, during orientations to ward teams and ambulatory attendings, and readily becomes part of the terminology that teachers use. In a study looking at the feasibility and acceptability of R-I-M-E, Battistone et al.[53] found that residents and faculty believed that the new descriptive system was "more valid" than the prior evaluation method and that 80% of students found RIME to be "helpful" to "very helpful" with overall student satisfaction. Battistone also noted that more than half of the students noted they heard the RIME terminology in the feedback from their teachers within the first year of implementation.

Formal Evaluation Sessions

Importantly, we evaluate clerkship students using the RIME framework during formal evaluation sessions.[21, 53, 98, 107] The evaluation sessions are formal, planned meetings that are held every 3 to 4 weeks at each clerkship site. The clerkship director, or the on-site coordinator for the clerkship, moderates each session during which 15 minutes is devoted to discussing each medical student currently on the clerkship. All instructors, including residents and faculty, who are working with the student are asked to attend. Each evaluator is asked to describe and assess the student's strengths and/or weaknesses and is allowed to speak uninterrupted. The moderator may ask for clarifications about, or specific examples of, demonstrated knowledge, skills, and attitudes. The most junior evaluator speaks first, with the attending physician adding comments last in an effort to encourage the house staff to voice their observations uninfluenced by the comments of the attending physician. At the end of the evaluator's comments, the facilitator asks for a recommended grade based on the student's performance and the "next steps" for the student to progress along the R-I-M-E framework. The clerkship director or site directors can also provide feedback to the teachers on their comments. The clerkship director or on-site coordinator meets with each student the following day to provide feedback.

In addition to serving as a forum to evaluate students, the evaluation sessions fulfill other needs including: (1) defining clerkship objectives and how they can be assessed; (2) defining expectations of instructors; (3) facilitating communication among faculty members; and (4) providing faculty development to improve the evaluation of students.[78, 97, 99, 100, 121] Faculty development is accomplished in a non-threatening, interactive, "workshop" format. The evaluation sessions are "protected time" for these activities. They provide a regular, recurring time to provide frame of reference training and performance dimension training to teachers, one that is immediately applicable because the teachers are still working with the students. In addition, the evaluation sessions not only meet clerkship directors' need for timely summative evaluation, but also the students' need for formative evaluation and feedback by identifying and discussing strengths and weaknesses *during* the clerkship.

Perhaps most significantly, the evaluation sessions facilitate the identification of marginally performing students by capitalizing on instructors' willingness to verbally discuss concerns regarding students that they may not be willing to document in writing[77, 98, 105] We have demonstrated the enhanced predictive validity of the evaluation sessions over traditional evaluation methods for identifying students with marginal funds of knowledge, as well as identifying those students who are likely to have problems during their first-postgraduate year of training.[21, 98] Evaluation sessions enhance the quality of behavior-based description of a student's professional

demeanor[135] and significantly improve the detection and description of unprofessional behavior.[105] Finally, the use of the R-I-M-E framework in conjunction with the formal evaluation sessions has achieved an internal consistency of descriptive evaluation of student performance similar to that of quantifiable examinations.[136]

Evaluation sessions (or similar activities) have been implemented at other institutions, and on clerkships other than medicine.[53, 127-139] Residency program directors and local leadership have supported these sessions by providing residency lecture time for the sessions, a clear signal as to the importance of trainee evaluation and also teachers' professional development.[53, 137] Teachers will come to the meetings—in the first year of implementation, Battistone et al.[53] found that 79% of residents and 72% of faculty attended the sessions; Ogburn noted near 100% attendance.[137] We recognize that some clerkships may have students at such a large number of teaching sites that face to face meetings may not be feasible. We believe that what is most important is the interaction that takes place between the clerkship director and the teachers. There is a terrific opportunity for research to address other ways of interacting (email, phone, video teleconferencing) that prove valuable.

The 45 to 60 minutes invested per student during a 12-week clerkship to complete this evaluation process is similar to the time invested in evaluation by clerkship directors who use other evaluation and grading methods. The time and resources to administer the evaluation sessions is commensurate with expectations of clerkship directors.[140-141] Finally, the time requirement for the evaluation sessions pales in comparison to our educational and societal obligations to evaluate the competency of medical students.

Conclusion

Two broad conclusions are apparent. First, credible descriptive evaluation of medical students takes time, both for the clerkship director and for the teachers. Second, improving descriptive evaluation also means clerkship directors need to talk to teachers on a regular basis. Both of these can be addressed but certainly require the support of the medical school department and local teaching site leadership. While it is important to convey clerkship goals and expectations in a variety of ways, including using a concise evaluation form with behavioral descriptors, it is unreasonable to assume that, without training, instructors will be able to improve their evaluation skills or feel the needed support to identify concerns regarding student performance. RIME is a readily understood and applicable common frame of reference. RIME encourages formal evaluation sessions using a planned, longitudinal format for student evaluation and feedback that addresses many of the recommendations for improving this type of evaluation.[33]

Table 6.3.1. Problems Encountered in the Evaluation of Medical Students During the Clinical Years			
Clerkship Directors	**%***	**Instructors**	**%***
1. Faculty members' unwillingness to record negative evaluations	40.4	1. Inadequate guidelines for handling problem students	36.7
2. Lack of early warning system regarding problem students	43.6	2. Lack of information about problems students bring with them into the rotation	34.8
3. Breakdown in transmission of information across rotations and clerkships	43.5	3. Faculty members' unwillingness to record negative evaluations	34.5
		4. Failure to act on negative evaluations	36.6
4. Lack of training of evaluators	35.1	5. Lack of training of evaluators	25.8
5. Tardy submission of required evaluations	27.2	6. Reversal or dilution of negative evaluations	25.1
6. Criteria of evaluation insufficiently defined	29.2	7. Criteria of evaluation insufficiently defined	22.6
		8. Delays in feedback to students	21.9
7. Inadequate guidelines regarding repeaters	32.2	9. Role as evaluator not clearly defined	15.3
		10. Insufficient communication with clerkship or site coordinator	15.3
8. Reversal or dilution of negative evaluations	25.6		
		11. Tardy submission of required evaluations	14.3
9. No follow-up of effectiveness of remediation	23.3	12. Insufficient opportunity to observe students directly	18.2
10. Lack of agreement among evaluators	16.9	13. Excessive reliance on residents for information about students	14.3
11. Failure to act on negative evaluations	26.4		
12. Lack of integration of information about the student from various sources	12.5	14. Inadequate evaluation form	12.9
13. Delays in feedback to students	14.1		
14. Lack of integration of information about the student over time	15.1		
15. Inadequate guidelines regarding temporary or conditional grades	17.2		
16. Lack of correspondence between grades and narrative evaluation	16.5		
17. Inadequate evaluation form	12.8		
18. Inappropriateness of remedy applied to problems identified in students	14.5		
19. Underutilization of existing counseling system	9.4		
20. Excessive reliance on residents for information about students	11.3		
21. Paucity of counseling options	8.8		
22. Insufficient support from administration for the evaluative process	9.6		

*% of those surveyed who rated problem as "serious"

Table adapted from Tonesk X, Buchanan RG. An AAMC pilot study by 10 medical schools of clinical evaluation of students. J Med Educ. 1987;62:707-718.

Table 6.3.2. Recommendations for Improvement in Clinical Performance Assessment (Descriptive Evaluation)*	
Recommendation	**Example**
Broad, Systematic Sampling	Plan multiple observations (may be brief), multiple settings, includes simulations; ideally, 7-10 ratings
Observation by Multiple Raters	Addresses "idiosyncrasies" of single raters
Keep Rating Instruments Short	For progress decisions (grades): 5-10 items plus global rating; when feedback is goal, make form specific to event rated
Separate Appraisal for Teaching, Learning, and Feedback from Appraisal for Promotion	Feedback should be immediate, not saved for written comments on end of rotation rating form
Encourage Prompt Recording	Record observations during the clerkship, as they occur
Supplement Formal Observation with Unobtrusive Observation	Using nurse and patient observations
Consider Making Promotion and Grading Decisions via Group Review	Broadens base of knowledge, perspectives; more likely to make "tough" decisions
Supplement Traditional Clinical Performance Ratings with Standardized Clinical Encounters and Skills Training and Assessment Protocols	Allows all members of group to have clinical skills assessed in standard manner; comparisons to peers and gold standards possible
Educate Raters	Familiarize raters with forms; Provide frame of reference training
Provide Time for Rating	Gather raters together to accomplish ratings (e.g., Evaluation Sessions)
Encourage Raters to Observe and Rate Specific Performances	Use of mini-CEX form (from American Board of Internal Medicine)
Use No More than Seven Quality Rating Categories	Discourage two-level rating (e.g., 1-3 unsatisfactory, 4-6 satisfactory)
Establish the Meaning of Ratings	Use consistent rating form; did forms help identify excellent or poor performers (e.g., those asked to leave program); provide descriptors
Give Raters Feedback about Stringency and Leniency	Let them know how they compare to others
Learn from Other Professions	Aviation, clergy, military; team performance
Acknowledge the Limits of Ratings	Insufficient by themselves to assess clinical competence

Table adapted from: Williams RG, Klamen DA, McGaghie WC. Cognitive, Social, and Environmental Source of Bias in Clinical Performance Ratings. Teach Learn Med. 2003;15(4) 270-292.

MEDICINE CLERKSHIP EVALUATION FORM

Student Name: _____ Dates: From _____ TO: _____

Site: _____

For each area of evaluation, please check the appropriate level of ability. Qualities should be cumulative as rating increases, e.g. an outstanding rating for physical exam skills assumes that major findings are identified in an organized, focused manner AND that subtle findings are elicited. Indicate the level at which the student is consistent.

OUTSTANDING	*ABOVE AVERAGE*	*ACCEPTABLE*	*NEEDS IMPROVEMENT*	*UNACCEPTABLE*
DATA GATHERING				
Initial History/Interviewing Skill		If Not Observed, Check Here o		
o Resourceful, efficient, appreciates subtleties, insightful	o Precise, detailed, appropriate to setting (ward or clinic)	o Obtains basic history. Accurate. Identifies new problems	o Incomplete or unfocused	o Inaccurate, major omissions, inappropriate
Physical Examination Skill		If Not Observed, Check Here o		
o Elicits subtle findings	o Organized, focused, relevant	o Major findings identified	o Incomplete or insen-sitive to patient comfort	o Unreliable
DATA RECORDING/REPORTING				
Written Histories & Physicals		If Not Observed, Check Here o		
o Concise, reflects thorough understanding of disease process & patient situation	o Documents key information, focused, comprehensive	o Accurate, complete; timely	o Often late; poor flow in HPI, lacks supporting detail, labs, or incomplete problem lists	o Inaccurate data or major omissions
Progress Notes/Clinic Notes		If Not Observed, Check Here o		
o Analytical in assessment and plan	o Precise, concise, organized	o Identify on-going problems & documents plan	o Needs organization, omits relevant data	o Not core or inaccurate data
Oral Presentations		If Not Observed, Check Here o		
o Tailored to situation (type of rounds); emphasis and selection of facts teaches others key points	o Fluent, focused; good eye contact; selection of facts shows understanding	o Maintains format, includes all basic information; minimal use of notes	o Major omissions, often includes irrelevant facts, rambling	o Consistently ill-prepared
KNOWLEDGE				
In General		If Not Observed, Check Here o		
o Understands therapeutic interventions, broad-based	o Thorough understanding of diagnostic approach	o Demonstrates understanding of basic pathophysiology	o Marginal understanding of basic concepts	o Major deficiencies in knowledge base
Relating To Own Patients		If Not Observed, Check Here o		
(check as applicable) oBroad textbook mastery oDirected literature search o Educates others	o Expanded differential diagnoses, can discuss minor problems	o Knows basic differential diagnoses of active problems in own patients	o Inconsistent understanding of patient problems	o Lacks knowledge to understand patient problems
DATA INTERPRETATION				
Analysis				
o Understands complex issues, interrelates patient problems	o Consistently offers reasonable interpretation of data	o Constructs problem list, applies reasonable differential diagnosis	o Frequently reports data without analysis; problem lists need improvement	o Cannot interpret basic data
Judgment/Management				
o Insightful approach to management plans	o Diagnostic decisions are consistently reasonable	o Appropriate patient care, aware of own limitations	o Inconsistent prioritization of clinical issues	o Poor judgment, actions affect patient adversely
MANAGEMENT SKILLS				
Patient Care Activities		If Not Observed, Check Here o		
o Functions at senior level, involves and coordinates health care team	o Efficient & effective, often takes initiative in follow-up (clinic or ward)	o Monitors active problems, maintains patient records	o Needs prodding to complete tasks; follow-up is inconsistent	o Unwilling to do expected patient care activities; unreliable
Procedures		If Not Observed, Check Here		
o Unusually proficient and skillful	o Careful, confident, compassionate	o Shows reasonable skill in preparing for and doing procedures	o Awkward, reluctant to try even basic procedures	o No improvement even with coaching, insensitive

PROFESSIONAL ATTITUDES

Reliability/Commitment

o Unusual dedication to education & patient care	o Seeks responsibility	o Fulfills responsibility	o Often unprepared, lackadaisical	o Unexplained absences, unreliable

Response to Instruction

o Continued self-assessment leads to further growth	o Seeks and consistently improves with feedback	o Generally improves with feedback	o Inconsistent improvement	o Lack of improvement

Self-Directed Learning (knowledge and skills)

o Outstanding initiative, consistently educates others	o Sets own goals; reads, prepares in advance when possible	o Reads appropriately	o Frequent prompting required	o Unwilling, lack of introspection

PROFESSIONAL DEMEANOR

Patient Interactions

o Preferred provider	o Gains confidence & trust	o Sympathetic, develops rapport	o Occasionally insensitive, inattentive	o Avoids personal contact, tactless

Response to Stress

o Outstanding poise, constructive solutions	o Flexible, supportive	o Appropriate adjustment	o Inflexible or loses composure easily	o Inappropriate coping

Working Relationships

o Establishes tone of mutual respect & dignity	o Good rapport with other hospital staff	o Cooperative, productive member of own team	o Lack of consideration for others	o Antagonistic or disruptive

COMMENTS: (Written comments are also required. **What's the "next step" for this student?** Thanks.)
Please check <u>each</u> step the student has <u>consistently</u> reached: oReporter oInterpreter oManager oEducator

Recommended Grade: _____ Have you discussed this report with the student? ____

			Intern	Resident
Attending		Preceptor		
Printed Name	Signature	Date	_____	_____ _____

Our System is Based on Performance Criteria Rather Than Percentages. Please Use These to Describe Current Level of Student Work

PASS: (Reporter) Satisfactory performance. Obtains and reports basic information accurately; is beginning to interpret; professional qualities are solid. Distinctive personal qualities should be recognized in descriptive comments.

HIGH PASS: (Interpreter) Clearly more than typical work in most areas of evaluation. Proceeds consistently to interpreting data; good working fund of knowledge; an active participant in care. Consistent preparation for clinics.

HONORS (Manager/Educator): Outstanding ratings in most major areas of evaluation. Fourth-year level of patient care, actively suggesting management options; excellent general fund of knowledge, outstanding (broad/deep) knowledge on own patients. Strong qualities of leadership and excellence in interpersonal relationships.

LOW PASS: Performs acceptably in some areas but clearly needs improvement in others. Has shown evidence of steady progress and should be able to perform satisfactorily as a physician with additional experience in Medicine during Fourth Year without having to repeat the third year clerkship.

FAIL: Overall inadequate performance or unacceptable performance in any major area of evaluation. Little improvement with guidance. A grade of Fail will require repeating the clerkship.

Section 4. Direct Observation of Student's Clinical Skills

Eric Holmboe, MD

Background and Importance

Despite tremendous advances in medical technology, the basic clinical skills of interviewing, physical examination, and counseling remain essential to the successful care of patients. The Association of American Medical Colleges (AAMC) strongly endorses the evaluation of students in these clinical skills.[142] The Institute of Medicine has placed *patient-centered care* at the heart of its five core competencies for all physicians.[143] Faculty observation of students performing a medical interview, physical examination, or counseling is still essential for the reliable and valid assessment of these skills. The development of standardized patients to evaluate clinical skills has been a major advance in the assessment of students.[144-148] However, standardized patients are optimally applied in clinical skills teaching and assessment as a supplement to similar activities in the real clinical setting; they cannot replace the observation of students by physicians on an ongoing basis with actual patients.[149-152]

Therefore, despite the growing availability and acceptance of standardized patients and other simulation technologies, teaching faculty will continue to shoulder the primary responsibility for evaluating student skills through direct observation in real clinical settings. Unfortunately many faculty are not sufficiently prepared to accurately observe and provide effective corrective feedback about these clinical skills. In this chapter we will first explore problems in students' clinical skills and the challenges faced by faculty performing direct observation. We will then outline some practical methods to improve faculty observation skills along with useful tools faculty can use when performing observations.

Reasons for and Challenges of Direct Observation

Numerous studies have documented serious deficiencies in medical interviewing and counseling that have persisted over time and in the views of some, history taking skills may have actually declined.[153-157] More importantly, research has demonstrated positive associations between good communication skills and improved patient outcomes.[158] Errors are also common in physical examination skills.[159-164] For example, deficiencies in auscultatory skills among trainees were noted over forty years ago[161-162] and poor cardiac and pulmonary physical exam skills continue to plague U.S. students and residents today.[163-164]

These findings are relevant because we know that despite advances in technology, accurate data collection during the medical interview and the physical exam remains the most potent diagnostic tool available to physicians.[165-167] Two important studies showed that the medical interview alone produced the correct diagnosis in nearly 80% of patients presenting to an ambulatory care clinic with a previously undiagnosed condition.[165, 167] Bordage recently noted that errors in data collection are one of the principle factors in diagnostic errors committed by physicians.[168] As a result, there has been a significant push to re-emphasize both the training and evaluation of clinical skills.[169-171] Without accurate evaluation of clinical skills, which must be accomplished by direct observation, improvement in the clinical skills of physicians is unlikely.

Lack of Direct Observation by Faculty

Perhaps the biggest problem in the evaluation of clinical skills is simply getting faculty to observe students. One of the most prominent physician-scientists and educators of the twentieth century, the late George Engel, strongly advocated direct observation of the history and physical examination skills of trainees over 30 years ago.[172-173] Dr. George Engel commented in a 1976 editorial,

"Evidently it is not deemed necessary to assay students' (and residents) clinical performance once they have entered the clinical years. Nor do clinical instructors more than occasionally show how they themselves elicit and check the reliability of clinical data. To a degree that is often at variance with their own professed scientific standards, attending staff all too often accept and use as the basis for discussion, if not recommendations, findings reported by students and housestaff without ever evaluating the reporter's mastery of the clinical methods utilized or the reliability of the data obtained."[173]

The AAMC found that among 97 medical schools it visited between 1993 and 1998, faculty rarely observed student interactions with patients, noting that the majority of a student's evaluation was based on faculty and resident recollections of student presentation skills and knowledge.[174]

Quality of faculty observation

Although several studies show that four to seven observations produces sufficient reliability in the evaluation of clinical skills for "pass-fail" determinations, little is known about the validity and accuracy of faculty rating. Noel and Herbers, in two important studies of the American Board of Internal Medicine's (ABIM) traditional "long case" clinical evaluation exercise (CEX), found substantial deficiencies in the accuracy of faculty ratings.[175-176] They demonstrated that faculty failed to detect up to 68% of errors committed by a resident scripted to depict marginal performance on a training videotape. Use of specific checklists prompting faculty to look for certain skills increased accuracy of error detection nearly twofold, but the checklist did not produce more accurate overall ratings of competence. Nearly 70% of faculty still rated a resident depicting marginal performance as satisfactory or superior overall.

Kalet examined the reliability and validity of faculty observation skills using videotapes of student performance on an objective structured clinical examination (OSCE) designed to evaluate interviewing skills.[177] She found that faculty were inconsistent in identifying the use of open-ended questions and empathy, and that the positive predictive value of faculty ratings for "adequate" interviewing skills was only 12%. Another study found that faculty could not reliably evaluate 32% of the physical exam skills assessed, and had the most difficulty with examination of the head, neck and abdomen.[178]

Practical Approaches to Training Faculty

Given the essential role of faculty observation in the evaluation of basic clinical skills, medical schools and residency programs must better prepare faculty for this important task. Recent research in medical education has demonstrated that effective training approaches can improve observation skills. A brief description of each approach and how it applies to faculty development for competency evaluation of medical students is described below.

Behavioral Observation Training (BOT)

Behavioral observation training is focused on improving the detection, perception, and recall of actual performance.[179] There are two main strategies emphasized in BOT. The first is simply to increase the number of observations, or increased sampling of actual performance. This helps to improve recall of performance and provides multiple opportunities for skill practice in observation by the rater, the "practice makes perfect" principle." The second strategy is to provide some form of observational aide that raters can then use to record observations, sometimes referred to as "behavioral diaries." Studies show that even something as simple as a 3 X 5 inch index card used to record observation notes improves the quality of information provided on evaluation forms. As described below, the mini-CEX form and checklists can serve as an immediate "behavioral diary" to record a rating of an observation.[180]

Observation of clinical skills also requires that faculty "prepare" for the observation. First, faculty should determine what are the objectives and/or goals of the observation before entering the patient's room with the student. For example, if you plan to perform an observation of student's physical examination skills, what would be the appropriate components of a physical exam for the patient's chief complaint or medical condition? Positioning is also very important because as faculty you want to minimize interference with the student-patient interaction whenever possible. Figure 1 demonstrates the principle of triangulation that maximizes the ability of the faculty to observe while minimizing interference. Table 6.4.1 lists some important yet simple rules for performing student observation.

Performance Dimension Training (PDT)

This type of training is designed to teach and familiarize the faculty with the appropriate performance dimensions used in their own evaluation system.[181-183] PDT simply starts with a review of the definitions and criteria for each dimension of performance or competency. The goal should be to define all those criteria and student *behaviors* that constitute a superior performance from the perspective of patient outcomes. The next step in PDT is to give faculty the opportunity to "interact" with the definitions using videotapes or actual evaluation examples to improve their understanding of the definitions and criteria. The overarching goal of PDT is to ensure faculty first understands the definitions and criteria for the competency of interest as a group so that some degree of consensus is shared among faculty. Appendix 1 provides a very straight forward and useful proactive PDT exercise that can be done with faculty to facilitate interaction with competency in clinical skills. We recommend performing PDT exercises in small groups and then have the small groups share their results. Inevitably differences occur between the groups. These differences, however, lead to productive discussions on what constitute the core elements and criteria of competency in counseling, or other clinical skills. This type of PDT exercise can be done for two clinical skills over approximately one hour of time. Another approach to PDT is reactive: using actual evaluations or videotapes of clinical skills that faculty can react to when performing the PDT exercise

Frame of Reference Training (FoRT)

This type of training specifically targets accuracy in rating. Table 6.4.2 describes the complete FoRT process. As you can see, FoRT is really an extension of PDT; the main goal of FoRT is establishing the different performance criteria that distinguish *levels* of performance. The main focus of FoRT should be to define four levels of performance: unsatisfactory, marginal, satisfactory and superior. The PDT exercise should first define the criteria and definitions for a superior performance from the perspective of optimal patient outcomes. The second step of the exercise, as shown in the Appendix, is to define the minimal criteria for a satisfactory performance. These criteria for a satisfactory performance serve as an important anchoring

point to define marginal and unsatisfactory performance in step 3. Once the group defines marginal criteria, by default any other type of performance is unsatisfactory.

Direct Observation of Competence (DOC) Training

Direct observation of competence training uses the methods of BOT, PDT, FoRT, and standardized patient training methods to train faculty in observation. There are two versions of DOC training. The "short course" form involves BOT, PDT, and FoRT exercises using small group discussion and videotape encounters. The long course version includes a half day of skill practice with standardized residents and patients.[184] An evaluation course that includes DOC training is available through the American Board of Internal Medicine (www.abim.org) .

Useful Tools to Guide Observation

The Mini Clinical Evaluation Exercise (miniCEX)

The mini-CEX was originally designed to evaluate residents in a setting reflective of day-to-day practice. Faculty observe a resident performing a *focused* history, physical, or counseling session during routine care experiences on the inpatient wards, intensive care units, outpatient clinics and the emergency department. However, the miniCEX has also been used in student clerkships.[185] The mini-CEX facilitates multiple observations over time by different faculty members. This improves both the reliability and validity of the evaluations. This longitudinal nature of the mini-CEX is one of its most important strength as an evaluation tool and method.

In the first large study of the mini-CEX, Norcini et al.[186] reported on the results of 388 miniCEX evaluations for 88 residents at 5 different residency programs. Over half of the encounters occurred in the inpatient setting. In this initial study, most of the participating residents were in the PGY-1 year, and each resident underwent a mean of 4.4 observations (range 2-10). The authors noted that the standard error for just 4 miniCEXs per resident was acceptable enough for pass-fail determinations. Trainees reported high satisfaction ratings for the miniCEX format, and interestingly there was a modest correlation between faculty satisfaction ratings and resident performance. In a study of the miniCEX with students, Kogan and colleagues found that nearly 90% of students on a 12 week medicine clerkship were able to obtain at least 9 miniCEX observations.[185] The reliability coefficient for 8 miniCEXs was 0.77 and the miniCEX was used in both the inpatient and outpatient clerkship settings.[185] Holmboe and colleagues, using scripted videotapes, found that the mini-CEX evaluation form does possess construct validity.[187]

Feedback and the MiniCEX

An essential component of the mini-CEX, as with any evaluation, is feedback. A recent study investigated the feedback generated from the miniCEX observation by audio taping the attending – resident feedback session, with a particular focus on interactive feedback.[188] Interactive feedback was defined as any feedback that provided a recommendation plus self-assessment, allowing the learner to react to the feedback, and development of an action plan. The study showed that 80% of the feedback sessions included at least one recommendation for improvement for the resident, and on average each feedback session contained 2 recommendations. The majority of recommendations, as might be expected, involved the clinical skills of medical interviewing, physical examination, and counseling. However, despite the large number of recommendations, only 8 sessions concluded with a specific action plan from the faculty member on how to carry out the recommendation or improve.[187] This is a very important aspect of feedback – including an action plan to enable the learner to act on the recommendations provided.

Checklists and Structured Clinical Observation

Checklists targeting specific skills are another tool that can improve the quality of faculty observation. However, since the purpose of faculty direct observation is to assess performance of actual clinical practice, it is not feasible to develop highly detailed checklists for every patient encounter. Some degree of faculty interpretation of behavior and skills will be required when working in actual clinical settings. A number of checklists for assessment of interviewing skills have been developed and tested for reliability. Both the SEGUE and Calgary-Cambridge checklists are useful tools to guide the evaluation of process and general content of medical interviewing.[189-190] Structured clinical observation is another observation technique that uses guidelines and observations sheets to systematically assess skills in history-taking, physical examination, and information-giving.[191]

Creating an Observation system

There are three simple steps in creating a faculty observation system. First, determine what your faculty are doing in regards to observation. If no observation is occurring, you will probably have to create a "need" for observation. Highlighting the substantial deficiencies in clinical skills among students provide ample evidence you can use to demonstrate the need to perform observation. Second, start small and get the faculty to perform some form of observation. Usually what happens is that faculty will observe these deficiencies. Once that happens, it becomes very difficult for your faculty to argue they no longer need to observe students, especially from a patient-centered perspective.

The next step is to improve faculty skill in observation, and depending on your educational climate, can be done concurrently with creating the need for observation. We recommend you start with performance dimension and behavioral observation training. This is can be done in a series of brief workshops, evaluation sessions, or at faculty meetings. Once your group feels comfortable with the definitions and criteria for the clinical skills competencies, you can then move on to frame of reference training and direct observation of competence training to improve faculty accuracy and ability to distinguish between levels of competence.

Conclusions

The successful practice of medicine requires the effective application of medical interviewing, physical examination, and counseling skills. Studies continue to document significant deficiencies in all three of these clinical skills areas among students. Direct observation by medical faculty remains an essential method to assess core basic clinical skills with actual patients. Furthermore, faculty are in the best position to assess student's acquisition and refinement of clinical skills longitudinally over time.

Table 6.4.1 Five Simple Rules for Observation	
Rule	**Description**
Correct Positioning	As the rater, try to avoid being in the line of sight of either the patient or trainee, especially when they are communicating. Use the principle of triangulation. However, during physical examinations be sure you can view the trainee's techniques accurately.
Minimize external interruptions	Let your staff know you will be with the resident for 5-10 minutes, avoid taking routine calls, etc.
Avoid intrusions	Don't interject or interrupt if at all possible. Once you interject yourself into the trainee-patient interaction, the visit is permanently altered. However, there will be many times at some point in the visit where you need to interject yourself in order to correct misinformation, etc. from the resident.
Be prepared	1. Know before you enter the room what your goals are for the observation session. For example, if a physical exam, have the trainee present the history first; then you will know what the key elements of the PE should be.
Prepare the trainee *and* the patient	1. Let the trainee know what you plan to do during the observation, including your interaction with the patient. You also need to let the patient know what your role will be and your relationship with the trainee.

Table 6.4.2. Steps for Frame of Reference Training	
Step	**Description of task**
1	Performance dimensions training (PDT). Faculty are given descriptions for each dimension of competence followed by a discussion of what they believe the qualifications are for each dimension
2	Faculty define what constitutes superior (the most effective criteria and behaviors) performance from the perspective of optimal patient outcomes.
3	Next, faculty define and reach consensus on the minimal criteria for satisfactory performance. Once the satisfactory criteria are set, marginal criteria are defined. Everything else by default is unsatisfactory performance
4	Participants are given clinical vignettes describing critical incidents of performance from unsatisfactory to average to outstanding. (Frame of reference). For clinical skills, videotaped encounters are the best method.
5	Participants use the vignettes to provide ratings on a behaviorally anchored rating scale.
6	Session trainer/facilitator provides feedback on what the "true" ratings should be along with an explanation for each rating.
7	Training session wraps up with an important discussion on the discrepancies between the participants' ratings and the "true" ratings.

Appendix

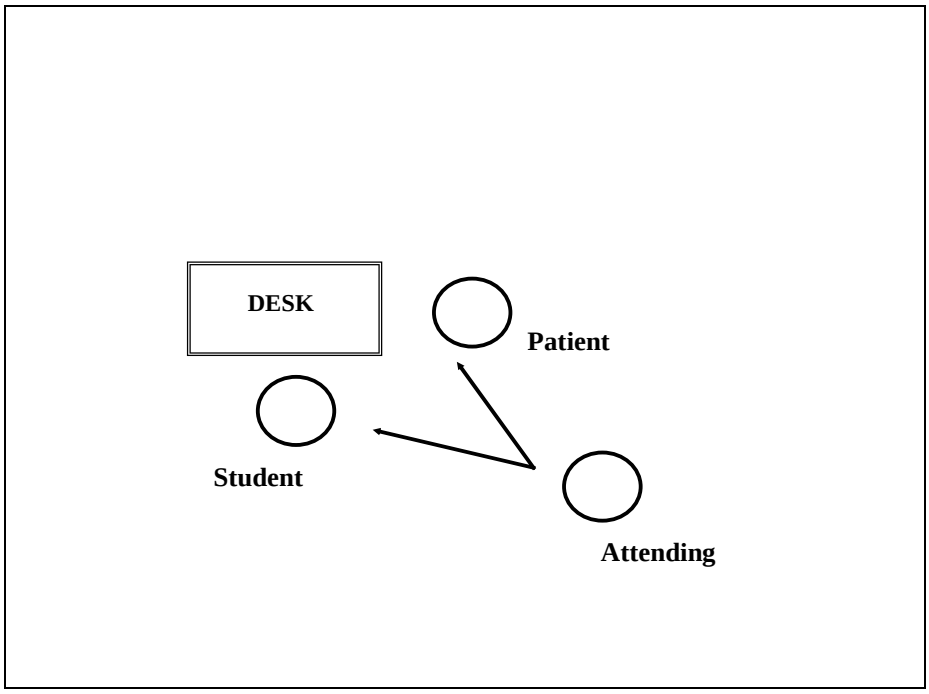
Sample Performance Dimension Training and Frame of Reference Exercise

The purpose of this group exercise is to develop specific criteria for a dimension of clinical competency.

Situation: A student is seeing a patient who has been diagnosed with hypertension and failed a trial of diet and weight loss. The student now needs to start a new medication for this patient. What are the criteria for a superior, highly effective counseling and patient education session? In other words, what criteria will you use to judge the counseling and patient education performance of this student? Once you have defined all the criteria, check off those criteria a student would have to perform in order to receive a ***satisfactory*** rating.

With your group:
Define the components/criteria of *effective* patient counseling and education, based on the Knowledge, Skills, Attitudes (KSA) model. Be sure your criteria are "behavioral" -remember you are developing these elements in the context of faculty observation.

Knowledge	What questions and "content" should the student ask the patient?
Skills	How should the interview be conducted? How should questions be asked?
Attitudes	Define behaviors that would signal to an attending a student was displaying a compassionate, interested, professional attitude.


DESK
Patient
Student
Attending

Section 5. Using pre-clerkship variables to identify high-risk students

John A. Poremba, MD and Gerald D. Denton, MD

Introduction

Pre-clerkship information that may be used to identify students who are at risk for poor clerkship outcome includes pre-matriculation data, basic science performance, incidents of unprofessional behavior, standardized test results (e.g., USMLE Step 1 and NBME subject examination scores), and in-house clerkship pre-tests. This section will briefly review the predictive power of some commonly available measurements, and then discuss the ways this information might be used to identify and help students. It is a truism that "test performance predicts test performance", and prior measures of knowledge may readily predict a student's ability to acquire factual knowledge during clerkships.[192-193] However, competency in clinical skills, professionalism, data analysis and problem solving are also critical to successful clerkship outcomes, but there is much less data supporting the ability of pre-clerkship variables to predict deficits in the skill and attitudinal domains. Therefore, while most of the advice in this chapter is informed by published educational literature, some recommendations rely on the experience and judgment of the authors.

Use of pre-matriculation data

Undergraduate Grade Point Average: A strong positive correlation between undergraduate GPA and subsequent measures of knowledge has consistently been reported by multiple authors.[194-195] Undergraduate GPA, as a measure of knowledge and test taking skills, predicts performance on knowledge assessments in medical school. However, the relationship between undergraduate GPA and clinical skills or professional attitudes has not been well described. Clinical skills deficits and unprofessional behavior do not always track consistently with knowledge deficits. A strong knowledge base does not necessarily "protect" a student from deficits in other areas. Further research exploring this association is needed.

MCAT: Much like undergraduate GPA, performance on the MCAT correlates well with subsequent standardized tests, including licensure examinations and NBME subject examinations.[194] In 1991, the MCAT introduced a writing sample component, as a measure of a student's ability to synthesize and communicate information.[196] The writing sample does not add value to other measurements in predicting USMLE Step 1 or Step 2 performance,[197] but may help predict other clerkship outcomes, such as global clinical competence, data gathering and communication skills, and this correlation persists into residency.[198]

Admissions committee interviews: Narrative comments from medical school admissions interviews are another pre-clerkship source of information. Although the process is hardly standardized,[199] and of low-yield[200] admission committee narratives may predict clinical performance – perhaps even better than undergraduate GPA.[201] Thus, as the earliest form of observation-based evaluation done by a school's own faculty, these narrative remarks have a potential role in identifying students who may subsequently have difficulty in clinical skills and non-cognitive domains. Of course, candidates with adverse comments are less likely to be admitted to a school, and observations about successful applicants are not regularly provided to or used by clerkship directors. It is not at all clear that interventions based on comments from admission committee interviews would be capable of yielding improved clinical performance. This is also a rich area for future research and collaboration.

Use of first and second year (pre-clinical) performance

Basic science GPA: Pre-clinical GPA predicts performance on the end of clerkship examinations,[202] especially if grades are reported on a numerical scale, rather than in letter-graded or pass/fail formats.[203] Non-numerical grading formats may not be as meaningful to clerkship directors because of a lack of predictive validity caused principally by the small breadth of the scales, i.e., 4-point (A-B-C-D) vs. 100-point scales.

USMLE Step I: Several reports have correlated USMLE Step One scores with various clerkship examination scores.[204-205] Failure on the first attempt at USMLE Step One predicts students at risk for poor scores on clerkship final examinations.[194, 202, 206] However, GPA and USMLE measures are not available to approximately 80% of clerkship directors (personal communication, Clerkship Directors in Internal Medicine, October 2003) and are not used by all medical schools. In addition, the NBME does not currently provide a separate score for Step One clinical skills questions, which would be potentially valuable to clerkship directors.[207] Cooperation between the student affairs or dean's office and clerkship directors could result in the availability of excellent baseline prognostic information.

ICM performance: Basic proficiency in fundamental clinical skills learned during pre-clerkship clinical skills courses (ICM) is essential to their successful practice and mastery during clerkships and beyond. While it is intuitive that performance in ICM courses should predict clerkship performance, there is a paucity of literature supporting the predictive validity of ICM courses. Students who perform at a substandard level in an ICM course may be at a much higher risk of a failing Internal Medicine clerkship grade,[208] although multiple-choice testing in an ICM course did not correlate with clerkship outcome.[209] The lack of data in this area is somewhat surprising because ICM courses specifically focus on developing critical skills in preparation for clerkships. Further investigation in this area is warranted.

Use of a clerkship pretest

A test given on the first day of a clerkship (a "pretest") can identify students at risk of poor performance. A clerkship pretest may offer advantages over USMLE Step One and preclinical GPA, such as ready availability, lack of student perception of prejudicial nature, and lack of a need for cooperation from the Dean's office.[210]

Use of professionalism information

Papadakis et al.[14, 96] described an effective system to longitudinally track professionalism deficiencies throughout the four years of medical school. Significant deficiencies in the third and fourth years predicted adverse licensure actions among graduates,[14] while traditional cognitive markers such as grades and test scores did not. This research shows it is feasible to track professionalism issues through the first and second years of medical school and that problems identified in the clinical years were often preceded by deficiencies in the pre-clinical years.[96] Such deficiencies may be amenable to early remediation, but that has not been clearly established. This represents another area of future multi-center investigation to refine generalizability and predictive validity.

Other intangible qualities of character and professionalism probably also predict clerkship performance, but little hard data supports their predictive validity or even how to adequately measure them. Such concepts as punctuality and reliability, work ethic, knowledge of one's

limits, the ability to get along with others, and situational awareness are well-recognized characteristics of successful professionals that would logically lead to enhanced clerkship outcomes. It's unclear how to measure these qualities, and it is even less clear how to design an intervention to improve on deficits that might be identified. Well-designed local or collaborative studies addressing these characteristics should be instituted.

In the physician charter for professionalism, the ABIM spells out a compelling need to reaffirm professionalism as the basis of medicine's societal contract.[211] In an era of renewed emphasis on the basic principles of professional behavior, the need to assess and monitor the professional development of medical students is apparent. Systems of professionalism assessment are optimally managed by a single responsible office, and tracked longitudinally. Clerkship directors should participate in such assessments, not only as a means of improving the early identification of professionalism issues within the context of their individual clerkships, but also as contributors to the longitudinal assessment of students. Wider implementation of longitudinal professionalism assessment and tracking would be useful to clerkship directors, and such systems should be implemented.

Proper use of baseline information

Avoiding bias: Medical school officials and students may be concerned that future teachers and clerkship directors could be unfairly biased knowledge of by prior performance, preemptively profiling students as academically weak. We should temper our enthusiasm for pre-clerkship prognostic measures by acknowledging the paucity of data and lack of predictive validity. Intervening upon some adverse pre-clerkship information has not been proven to improve clerkship performance. Moreover, students who struggle in the pre-clinical years may wish to start with a clean slate on their clinical rotations, so to be fair to students, we should use this information judiciously.

Confidentiality and disclosure: Proper use of baseline information includes protection of student confidentiality. Concerns about confidentiality and fairness may prevent medical school officials from sharing pre-clerkship information like college GPA, MCAT scores, preclinical GPA, and USMLE Step One scores with clerkship directors. If pre-clerkship information is used, disclosure outlining the nature of its use and protections limiting dissemination of prior academic data is important to reassure students that the process respects their privacy and is unbiased.

Protecting future patients, discovering important patterns: In their status as trainees, students participating in clinical clerkships have an important role in patient care as first-line reporters and patient confidants. Thus, any system of evaluation should take into consideration safety and effective patient care. While there may be reluctance to "poison the well" by sharing potentially adverse information about an individual student, the sharing of information between course and clerkship directors is justifiable when defined in the context of patient safety. Deficits in clinical competence should be tracked longitudinally, as isolated clerkship incidents may represent a pattern of repetitive error on the part of the student. In this role, the Office of Student Affairs or similar advisory committees should be engaged in tracking such problem students.

Summary

While not readily available to most clerkship directors, prior assessments of a student's knowledge and prior faculty observations of professional behavior, clinical skills and non-cognitive performance are part of every student's record that would be valuable to clerkship

directors to facilitate early interventions. Use of clerkship pretests can help when prognostic information is not available. However, the literature is not robust, and the utility of the information is unclear, so no specific recommendations for the use of pre-clerkship information can be made. Collaboration across disciplines and further study may lead to the development of useful prediction models to identify at-risk students with adequate sensitivity and specificity to justify early intervention.

Section 6. Evaluating Medical Procedures

David A. Rogers, MD, MHPE

Motivation for Evaluation

There are a number of reasons that support developing a medical student procedural evaluation program. A properly constructed evaluation program provides evidence that an individual medical student has mastered the essential skills that all physicians use. Further, evaluation tools facilitate the provision of feedback that is essential for motor skill acquisition. Finally, group performance evaluation information allows for thoughtful revisions to the curriculum to enhance learning.

The reasons not to develop a procedural evaluation program include a lack of familiarity with the process of instrument development or performance testing and limited resources available for the development of such a program. The goal of this section is to review the basic processes of evaluation instrument development and performance testing. Motor skill performance evaluation can be done with very expensive instruments like those that measure actual motion,[212] but an effective medical student evaluation program can be designed with less costly measurement tools.

Developing Evaluation Instruments

The first step in developing an instrument is to see if one already exists. Procedures relevant to medical student education are sometimes also taught to other health professions students. For example, an individual responsible for teaching phlebotomists has likely already developed and used an evaluation tool designed to assess phlebotomy skill. Nursing or allied health textbooks may also contain these instruments.[213] If an evaluation instrument does not exist, then one must be created. The procedure should be analyzed and divided into the key steps associated with the whole task. These steps should be emphasized in the curriculum and also serve as the framework for the evaluation instrument. An example of this type of analysis of a procedure performed by medical students has been published[214] and could also be done with a group of local procedural experts. Once the key procedural elements have been identified, the next step is deciding whether or not to create a checklist, rating scale or global score.[215] A checklist allows for the evaluator to indicate by a yes or no response whether or not a step has been performance correctly. A rating scale allows a determination of the extent to which a specific step has been performed. For example the evaluator may indicate on an ordinal scale that the performance was superior, acceptable, average, marginal or unacceptable. The final alternative would be of a global rating of the skill that involves an overall assessment of the performance and so does not include an evaluation of the individual steps of the procedure. Each of these options has advantages and disadvantages. Checklist responses would be most appropriate if judging whether each step of the skill can be performed either correctly or incorrectly. For example, medical students performing intravenous cannulation will either recognize a flashback of blood in the hub of needle or they will not recognize this step in the procedure. Rating scales

are appropriate if there are variations in the performance of the steps of a skill. For example, the angle of the needle to the skin in phlebotomy may range from perfect to unacceptable. Both checklists and rating scales are valuable in the provision of feedback because both allow for the identification of steps performed incorrectly and those performed correctly. Global assessment of performance by experts has been shown to produce reliable and valid measures of skill[216] and is efficient but may be less helpful in reminding the evaluator of the specific steps in the procedure that require correction. Therefore, global ratings are most useful for summative evaluation where feedback is not provided to the student. An evaluation instrument may include both rating scale and checklist type responses and this has been done in the creation of the current gold standard of surgical skill evaluation.[217] A similar instrument has been developed for assessing the types of procedures that a medical student might perform.[218]

Evaluation tools have a number of important attributes that that affect their utility.[219] Validity is the most important attribute because it assures that the ability of interest is actually being measured.[220] There are different types of evidence for validity[221] that can be accumulated through the process of instrument development. Reliability is also an important attribute as it is the degree to which an instrument produces reproducible results. There are a number of threats to reliability in performance evaluation and these should be examined in procedural evaluation, as poor reliability will impact validity.[222] Feasibility of the instrument is of significant practical importance because it assures that the instrument is actually used. Feasibility is enhanced through assuring that the evaluation instrument is concise and that instructions are clear. Evidence for validity, reliability and feasibility should be generated before an evaluation instrument is used for student assessment. This is particularly true if the instrument is to be used to generate information that might affect student progress.

Creating an Evaluation Program

A number of options exist in creating a procedural evaluation program.[8] Of these, logbooks, direct, and indirect observation would seem to be most applicable to medical student evaluation. Logbooks are commonly used to allow medical students to record the number and types of procedures but unfortunately this does not provide evidence that actual learning occurred.[223] Direct observation with criteria is generally recognized as the optimal form of performance evaluation. As the name would imply, this method involves the evaluator observing the learner perform the actual task and then evaluating the performance using an instrument that yields psychometrically sound data. This method does have the distinct disadvantage of requiring the physical presence of the evaluator. Faculty may not have enough time to perform this activity and non-physician raters have been used for this purpose with good results.[224-225] Indirect observation with criteria would include a review of a videotaped performance.[226] This method allows for evaluation of the task by multiple raters and fast-forwarding the videotape allows the assessment time to be shortened.[227] This method would not allow for the provision of immediate feedback and so would be most appropriate for summative evaluation. An evolving option is to evaluate the procedural performance in a simulated environment. The standardized patient experience suggests that there is much to recommend in this option. However, the evidence of transfer of the skill from the simulated to the real environment has been mixed. This suggests that for most skills, evaluation of the performance of the procedure on an actual patient is the best evidence that the student has mastered this skill. Adequate practice in a simulated environment with formative evaluation and feedback should enhance learning so that performance on patients is done in a way that minimizes their anxiety and discomfort.

A Real World Scenario

The challenge:

You are presenting your annual report about the clerkship plans and a senior faculty member begins to rant about how medical students aren't learning the procedures that most physicians need to know. You diplomatically acknowledge his concerns and promise to investigate the matter. You become convinced from your search of the literature and discussions with the curriculum committee that your senior colleague may be right. You gain approval to develop a curriculum that is designed to teach medical students how to properly close a simple laceration and plan to develop an evaluation program for this curriculum.

A step-wise solution

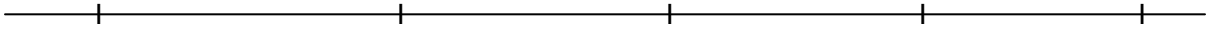
1. You search the health professions literature and textbooks in an attempt to locate an existing evaluation instrument. You find a list of the key steps for this procedure in an emergency medicine textbook but can find no completely developed evaluation instrument for this skill.
2. You do a Google Scholar search of the word "checklists" and are delighted to find a very helpful website dedicated to the development of checklists at http://www.wmich.edu/evalctr/checklists/.

You develop a rating instrument that consists of a checklist of the individual steps of the procedure and an overall global rating of proficiency (Table 6.6.1). Your local evaluation expert advises you to do this so that you can use the form for both formative and summative feedback.

Table 6.6.1 Simple Laceration Repair Rating Instrument

Procedural Step	Performed Yes	Correctly? No
1. Described body fluid precautions		
2. Gently evaluated the depth of the wound		
3. Explained the procedure to the patient and obtained informed consent		
4. Administered the appropriate local anesthetic		
5. Gently debrided and cleaned the wound		
6. Selected the appropriate suture material		
7. Selected the appropriate suture technique (simple interrupted, continuous, etc.)		
8. Performed the suture technique		
9. Applied appropriate dressing		
10. Disposed on needles appropriately.		
11. Provided appropriate wound care instructions.		

Please rate the medical students overall performance on this procedure:



Major Deficiencies Significant problems Proficient Superior Exceptional

Section 7. The Use of Simulators in Assessment

S. Barry Issenberg, MD and Ross J. Scalese, MD

A common challenge for clerkship directors is to determine the most appropriate assessment tool for particular competencies that their students should acquire. In the table below we list several clerkships, clinical competencies, and some of the evaluation solutions traditionally available to educators.

Clerkship	Competency	Solutions
Internal Medicine	Interpret heart murmur	Real patient (RP), *simulator*
OB/GYN	Perform pelvic examination	Standardized Patient (SP), Anesthetized RP, *simulator*
Anesthesia	Perform endotracheal intubation	Cadaver/animal tissue, Anesthetized RP, *simulator*
Surgery	Suture wound	RP, cadaver/animal tissue, *simulator*
Emergency Medicine	Perform ACLS	RP, animal, *simulator*

Many other examples exist, but these illustrate the customary reliance on real patients for assessment of many important skills. More recently, however, ethical considerations and the growing concern for patient safety have appropriately limited the use of real patients as training and assessment "instruments"; it is no longer acceptable as a matter of routine to assess third- and fourth-year medical students' ability to perform critical (e.g., intubation) or sensitive (e.g., pelvic examination) tasks on real (even standardized) patients.

Use of patient substitutes, such as cadaveric or animal tissue models, has its own challenges, not the least of which is maintaining an adequately realistic clinical context. In addition, availability, cost and, very importantly, ethical concerns have limited the use of cadavers and animals for medical skills assessment. Simulators, on the other hand, circumvent most of these obstacles and, thus, recently have come into widespread use for evaluation of learners across the continuum of medical education.

Definition and Characteristics

Now what exactly is a simulator? How is the term defined? Although we could address the use of *simulations* in the broad sense – including any approximation of an authentic clinical situation, such as mass casualty exercises or standardized patient encounters – our discussion here will focus more narrowly on *simulators*, referring to particular devices that aim to imitate real patients, anatomic regions, or clinical tasks. As described previously,[228] when simulators are used for assessment the student must respond to the challenge as he or she would under real-life circumstances. Simulators for evaluation have several common characteristics:

- Students act as they would in the real environment.
- Students see cues and consequences very much like those in the real environment.
- The fidelity (exactness of duplication) of a simulator is never completely identical to "the real thing". Some reasons are obvious: engineering limitations, psychometric requirements, cost and time constraints.

- The complexity of simulated conditions can vary according to the needs of the assessment or level of the learner.

Simulators can take many forms and span the range from low- to high-fidelity. These include: simple three-dimensional but inert anatomic models, such as venipuncture arms and airway trainers; computer-based programs that simulate patient encounters; virtual reality "haptic" systems that provide visual and tactile stimuli for surgical and endoscopic procedures; high-fidelity examination simulators, such as Harvey, the Cardiopulmonary Patient Simulator, that provide very realistic physical findings but are not interactive; and full-body mannequins that provide physical findings and actually respond to user actions.

Assessing Domains and Levels of Competence

Several factors contribute to the increasing role of simulators as assessment tools. These include: improving technologies that allow for more realistic simulations, flexibility in controlling and standardizing the clinical task or scenario and, as mentioned previously, a greater awareness of the problem of medical errors and the resulting emphasis on patient safety. This has led to a recent worldwide shift in focus toward outcomes-based education as a response to public demand for assurance that physicians are competent. The Accreditation Council for Graduate Medical Education (ACGME) describes six domains of clinical competence: 1) patient care, 2) medical knowledge, 3) practice-based learning and improvement, 4) interpersonal and communication skills, 5) professionalism, and 6) systems-based practice.[12] Simulators may be used to evaluate these ACGME domains of competence during, for example, an Internal Medicine clerkship: *patient care* – using a cardiology patient simulator, demonstrate the ability to perform a focused cardiac examination and identify the presence of a third heart sound in a "patient" presenting with dyspnea; *medical knowledge* – using a full-body simulator during a simulated case of ventricular fibrillation, verbalize the correct steps in the algorithm for treatment of ventricular fibrillation; *practice-based learning and improvement* – using an airway mannequin that measures appropriate cricoid pressure, demonstrate the ability to use feedback from the simulator until a defined level of mastery is consistently obtained. (see also Chapter 6, Section 2, [Comptencies])

Within each of these domains, one can assess medical learners at four different levels of competence, according to the pyramid model conceptualized by Miller.[229] These levels are: a) *knows* (knowledge) – recall of basic facts, principles, and theories; b) *knows how* (applied knowledge) – ability to solve problems, make decisions, and describe procedures; c) *shows how* (performance) – demonstration of skills in a controlled setting; and d) *does* (action) – behavior in real practice. The ACGME Toolbox of Assessment Methods[12] suggests that simulators are instruments most appropriate for evaluation of those outcomes that require students to demonstrate or "show how" they are competent to perform a skill.

Now in all clinical assessment there are three variables – the examiner, the patient, and the student.[230] If we standardize the first two variables, we improve the evaluation, such that the student's performance then represents a true measure of his or her clinical competence. Examiner training and the use of reliable evaluation tools allow for standardization of the 'examiner' component. An inherent feature of simulators is the ability to standardize many aspects of the 'patient' variable in the clinical assessment equation, thus offering a uniform, reproducible experience to multiple examinees. Simulators, however, do not comprise the entire assessment *per se*, but rather serve as tools to facilitate standardization and to complement existing evaluation methods. For example, simulators often serve effectively as one of several

tools used in the brief examining stations of an Objective Structured Clinical Examination (OSCE).

Assessing Process and Outcome

Numerous assessment criteria are available to evaluate learners, and clerkship directors must choose whether the competency tested relates to a *process* (such as completing an orderly and thorough "code blue" resuscitation) or an *outcome* (such as the status of the 'patient' after the cardiac arrest).[231] The following summarizes how one can assess processes and outcomes with simulators:

Criteria Type	Example
Measure a Process	A case-specific checklist to record actions during student suturing on a skin wound simulator [see example 1]
Judge a Process	A global rating (with well-defined anchor points) that allows an evaluator to reliably observe and judge the quality of suturing performed by a student on a skin wound simulator [see example 2]
Measure an Outcome	Observing and recording specific indicators of patient (simulator) status (alive, cardiac rhythm, blood pressure) after an ACLS code
Judge an Outcome	A global rating (with well-defined anchor points) that allows an evaluator to reliably observe and judge the quality of the overall patient status after an ACLS code
Combined	Task-specific checklist of cardiac bedside exam and observation and recording of correct identification and interpretation of physical findings

Example 1: Measured Process – Suturing

Process	Not Done or Incorrect	Done Correctly
Held instruments correctly	0	1
Spaced sutures 3-5 mm	0	1
Tied square knots	0	1
Cut suture to correct length	0	1
Apposed skin without excessive tension on sutures	0	1

Example 2: Judged Process – Suturing [232]

1. Time and motion				
1	2	3	4	5
Many unnecessary or repetitive movements.		Efficient time/motion, but unnecessary and repetitive movements.	Clear economy of movement and maximum efficiency.	
2. Instrument Handling				
1	2	3	4	5
Repeatedly makes tentative or awkward moves with instruments through inappropriate use.		Consistent use of instruments, but occasionally appears stiff or awkward.	Fluid movement with instruments.	

Having provided examples of assessment criteria to evaluate a process and/or outcome, Table 6.7.1 provides a list of available simulators that have been used in various clinical clerkships and characteristics that may assist in the decision to include one in a particular program. An important factor to consider is the multi-disciplinary use of simulators, so that their value is distributed across several clerkships.

Table 6.7.1. List of simulators and their characteristics

Simulator	Features	Clerkship	Assessment Criteria	Cost (as of 8/05)	Comments
Human Patient Simulator HPS www.meti.com	Full-sized, high-fidelity mannequin that functionally simulates all organ systems and responds physiologically to procedures and IV medications	Anesthesia Critical Care Emergency Medicine Neurology Internal / Family Medicine	Process – IV access, intubation, ventilation Outcomes – patient status following intervention	XXXX	Numerous studies demonstrate validity of simulator as assessment tool. Best used when assessing multiple competencies and used in a "theater" setting.
Emergency Care Simulator www.meti.com	Full-sized, high-fidelity mannequin that is more portable than the HPS and is programmed for more emergency scenarios. Less sophisticated physiological response to interventions	Critical Care Emergency Medicine	Process – IV access, intubation, ventilation Outcomes – patient status following intervention	XX	Newer simulator that uses much of the same technology as HPS, but is designed to be more portable so that it can be used in numerous environments.
PediaSIM & PediaSIM-ECS www.meti.com	Small child-sized, high-fidelity mannequin that functionally simulates the anatomy and physiology of a child and responds appropriately to interventions	Pediatrics Emergency Medicine	Process – IV access, intubation, ventilation Outcomes – patient status following intervention	XX	Newer simulator that uses much of the same technology as HPS, but is designed to function and react different from an "adult." The ECS version has more sophisticated physiologic responses to interventions.
Pelvic Exam SIM www.meti.com	Female adult-sized, high-fidelity pelvic torso that functionally simulates a variety of gynecological findings and automatically and objectively tracks user examination technique	GYN	Process – pelvic exam technique Outcomes – identification of abnormal findings	XX	Several studies have demonstrated its validity as an assessment tool. Currently being evaluated by NBME for potential use in high-stakes examination settings.
BabySIM www.meti.com	Infant-sized, high-fidelity mannequin that functionally simulates the anatomy and physiology of a 3- to 6-month old infant and responds appropriately to interventions	Pediatrics	Process – IV access, intubation, ventilation Outcomes – patient status following intervention	XX	Available less than a year – uses same technology as HPS.
AirSIM www.limbsandthings.com	Adult-sized, high-fidelity head and neck mannequin that functionally simulates a variety of airway emergencies and responds appropriately to interventions	Anesthesia Critical Care Emergency Medicine	Process – intubation, ventilation Outcomes – patient status following intervention	XX	Several studies demonstrate validity of airway mannequin as assessment tool. Best used when assessing complex airway scenarios.
SimMan www.laerdal.com	Full-sized, high-fidelity mannequin that is portable and provides functionally realistic anatomy for multiple clinical tasks and procedures	Anesthesia Critical Care Emergency Medicine Internal / Family Medicine Surgery	Process – IV access, intubation, ventilation Outcomes – patient status following intervention	XXX	Several studies demonstrate validity and feasibility of simulator as assessment tool. One of the most widely used high-fidelity simulators for a broad range of learner populations and levels.

Simulator	Features	Clerkship	Assessment Criteria	Cost (as of 8/05)	Comments
AirMan www.laerdal.com	Adult-sized upper torso, head and neck that provides functional anatomy and physiological signs of normal and difficult airway conditions	Anesthesia Critical Care Emergency Medicine Internal Medicine	Process – intubation, ventilation Outcomes – patient status following intervention	XX	Uses much of the same hardware and software as SimMan, but focuses on airway management skills.
Resusci Anne www.laerdal.com	Adult-sized, medium-fidelity, portable task trainer that provides functional anatomy for critical lifesaving skills. Optional built-in assessment system that evaluates adequacy of chest compressions	Critical Care Emergency Medicine Internal Medicine	Process – chest compression, intubation, ventilation Outcomes – patient status following intervention	XX	Numerous studies demonstrate reliability, validity and feasibility of device as assessment tool.
CathSim www.immersion.com/medical	High-fidelity task trainer that provides haptic feedback for intravenous access skills. Contains built-in assessment system that evaluates IV access performance	Critical Care Emergency Medicine Internal Medicine Pediatrics OB/GYN	Process – intravenous cannulation technique Outcomes – successful placement of IV	XX	Several studies demonstrate validity of simulator as assessment tool. Advantage is built-in objective assessment system.
Noelle Maternal and Neonatal Birthing Simulator http://www.gaumard.com	Adult- & neonate-sized simulators that provide functional anatomic and medium-fidelity physiological signs to perform complete delivery and post-natal care	OB/GYN Pediatrics	Process – infant delivery technique (normal/forceps, breech), C-section, Outcomes – maternal, fetal status	XX	New simulator that also offers interactive features for more complicated pregnancies and deliveries.
Harvey, the Cardiopulmonary Patient Simulator www.crme.med.miami.edu	Adult-sized high-fidelity mannequin that provides comprehensive cardiac and pulmonary physical findings.	Internal/Family Medicine Pediatrics Critical Care Emergency Medicine Surgery	Process – cardiac and pulmonary exam Outcomes – correct identification of abnormal findings	XXX	Longest continuous high-fidelity simulator with numerous studies demonstrating validity as assessment tool.
Suture Simulator www.limbsandthings.com	Adult-sized arm that provides anatomic functional wound for suturing skills	Surgery Emergency Medicine OB/GYN Family Medicine	Process – suture technique Outcomes – quality of sutures	X	Inexpensive task trainer that provides hundreds of assessment opportunities.
Clinical Female Pelvic Trainer www.limbsandthings.com	Partial adult-sized task trainer that provides functional anatomy of lower abdomen and pelvis, vaginal and rectal findings	OB/GYN Internal / Family Medicine	Process – pelvic exam and pap smear technique Outcomes – correct identification of abnormal findings	X	Inexpensive task trainer for assessing recognition of appropriate landmarks, vaginal and bimanual exam, cervical smear, dry catheterization, and digital rectal exam.
Breast Examination Simulator www.limbsandthings.com	Partial adult-sized task trainer that provides functional anatomy of female upper chest and breast with normal and abnormal findings	OB/GYN Internal / Family Medicine	Process – breast exam technique Outcomes – correct identification of abnormal findings	X	Inexpensive task trainer that provides opportunity to assess breast exam technique (including lower neck, clavicle and both axillae), identification of anatomic landmarks, and pathologic diagnosis.
Episiotomy Suture Simulator www.limbsandthings.com	Partial adult-sized task trainer that provides functional anatomy of vagina and perineum with varying degree lacerations	OB-GYN	Process – episiotomy suture technique Outcomes – quality of episiotomy suture	X	Inexpensive task trainer that provides opportunity to assess episiotomy technique (superficial, subcuticular, deep musculature) and identification of tissue layer.

Simulator	Features	Clerkship	Assessment Criteria	Cost (as of 8/05)	Comments
Central Line Simulator www.kyotokagaku.com	Partial adult-sized task trainer that provides functional anatomy of neck and upper chest including all landmarks for subclavian and internal jugular vein catheterization	Surgery Internal Medicine Critical Care Medicine	Process - central venous catheter insertion technique Outcomes – proper placement of central line (no complications)	X	Newer task trainer that allows assessment of central venous catheter insertion technique (and incorrect technique – pneumothorax, arterial puncture), identification of local anatomy.
Diagnostic Prostate Simulator www.limbsandthings.com	Partial adult-sized task trainer that provides functional anatomy of male rectum, perineum and prostate	Surgery Internal / Family Medicine	Process – prostate exam technique Outcomes – correct identification of normal and abnormal findings	X	Inexpensive task trainer that provides opportunity to assess prostate exam technique and identification of normal, bilateral BPH, unilateral nodule, uni-/bilateral carcinoma
Eye Exam Simulator www.kyotokagaku.com	Partial adult-sized task trainer that provides functional anatomy of external and internal eye with normal and abnormal retinal findings	Internal / Family Medicine	Process – ophthalmologic exam technique Outcomes – correct identification of normal and abnormal findings	X	Inexpensive task trainer that provides opportunity to assess ophthalmologic exam technique and identification of normal and 9 common abnormal retinal findings.
Ear Exam Simulator www.kyotokagaku.com	Partial adult-sized task trainer that provides functional anatomy of external and middle ear with normal and abnormal findings	Internal / Family Medicine	Process – otoscopic exam technique Outcomes – correct identification of normal and abnormal findings	X	Inexpensive task trainer that provides opportunity to assess otoscopic exam technique and identification of 10 normal and abnormal middle ear findings.
Spinal Injection Simulator www.adam-rouilly.co.uk	Partial adult-sized task trainer that provides functional anatomy and landmarks for lumbar puncture and various spinal injections.	Neurology Internal Medicine Anesthesiology	Process – lumbar puncture technique Outcomes – correct placement of spinal needle and collection of spinal fluid	X	Relatively inexpensive task trainer that provides opportunity to assess technique of lumbar puncture and identification of critical anatomic landmarks. Fluid can be added to provide feedback regarding correct placement of needle.
Infant Lumbar Puncture Simulator www.laerdal.com	Partial infant-sized task trainer that provides functional anatomy and landmarks for lumbar puncture technique.	Pediatrics	Process - lumbar puncture technique Outcomes – correct placement of spinal needle and collection of spinal fluid	X	Inexpensive task trainer that provides opportunity to assess technique of lumbar puncture and identification of critical anatomic landmarks. Fluid can be added to provide feedback regarding correct placement of needle.

X: < $1,000 XX: $1,000 - $10,000 XXX: $10,000 - $50,000 XXXX: > $50,000

Many factors influence course directors' choice of evaluation methods for their clerkships. Simulators have assumed an increasing role in such clinical assessments and, ultimately, the decision to use simulators for testing depends on local circumstances, the needs of the particular examination, and the competencies under evaluation. These considerations can guide one in choosing from among simulators that range widely in terms of fidelity, cost, and features. In general, simulators are most appropriate for assessment of competence in performing clinical skills or procedures. Simulators provide standardization of the patient variable in clinical

examinations and contribute to more reliable evaluations of student performance in these domains. Simulators complement other clinical assessment methods, such as the OSCE, and allow one to measure and/or judge the wide range of processes and outcomes encountered in clinical medical education.

Summary Recommendations

- In keeping with principles of curricular alignment, let stated course objectives drive the use of simulators for assessment in your clerkship, rather than the other way around: just because you have a simulator than *can* be used to evaluate certain competencies, does not mean that you *should* use it in this way, if these competencies are outside the scope of the course or inappropriate for your learners' level.

- To satisfactorily complete the course, learners will likely have to master several competencies, not all of which will be amenable to assessment with simulators; in general, demonstration of clinical skills or procedures is most suited to this method of evaluation. Be aware that your clerkship may require several tools to accomplish all the necessary assessments.

- Decide whether the competency to be tested relates to a process or an outcome, and then design an appropriate assessment tool that can be used in conjunction with the simulator: checklists are often employed to evaluate whether learners complete all steps of a process, while global rating scales with well-defined anchor points are frequently used to judge the overall outcome of the learners' performance.

- If establishing a simulation or clinical skills center, or using simulators in your clerkship for the first time, consider acquiring multi-feature devices that can be applied across many clerkships/specialties. Besides the obvious improvement in cost-effectiveness when expenses are distributed among many users, obtaining one such (initially higher priced) simulator offers several potential advantages over buying multiple (less expensive) single-task trainers: presentation/assessment of a wider variety of clinical conditions or scenarios, often greater face validity (realism), and identification in the curriculum and evaluation of those broader core competencies that transcend the boundaries of individual clerkships.

Section 8. Standardized Patient-based Assessment of Clinical Skills in Clerkships

Michael Ainsworth, MD and Karen Szauter, MD

Introduction: Why Use Standardized Patients to Assess Clinical Skills?

Clinical skills assessment presents challenges distinct from those encountered in assessment of medical knowledge. Accurate and reliable assessment of students' medical knowledge depends on use of rigorously designed examinations with blueprint-based topic selection and careful adherence to question construction. Such examinations supplement the less formal assessment of knowledge conducted by teachers in the clinical setting. Likewise, standardized patient (SP)-based assessment of clinical skills provides an objective component that complements the assessment done by physicians who work with students in hospitals and clinics. Analogous to written tests of knowledge, SPs-based assessments also require careful attention to examination blueprints, individual case challenges, and construction of scoring criteria.

Although there are logical overlaps between medical knowledge and clinical skills, evaluation of clinical skills often has limited correlation with measures of medical knowledge and problem solving.[233]

Standardized patients are laypersons trained to accurately and consistently simulate a patient encounter for teaching or evaluation purposes, and are used in virtually all US medical schools for teaching or assessment.[148, 234] Although there are technical distinctions between "standardized patients" and "simulated patients", the term standardized patient is used widely in the literature and will be used consistently throughout this chapter. The value of SPs arises from ability to simulate a range of clinical encounters allowing teaching and assessment of skills that range from basic data-gathering in the medical interview and physical examination, to more complex communication or patient education skills.[235-236] Applications range from instruction in the fundamentals of clinical skills (for example, in Years 1 and 2), to specialized examinations (breast, pelvic, and genital examination), to more comprehensive clerkship-based or interdisciplinary examinations.[237-238]

Use of any SP-based assessment exercise requires a step-wise process of examination design and formatting. For clerkship-based examinations, these decisions should be based, ideally, on the clerkship's goals and objectives, and should reflect the students' learning experiences. For example, inclusion of specific problems and diseases (e.g., abdominal pain from GI pathology) on an SP-based exam should imply that mastery of this problem and recognition of this group of diseases was a stated objective of the course, and that students had opportunity to encounter the problem in their patient care experiences, reading, and/or other learning activities.

Decisions in SP-based Assessment Design

A decision to implement SP-based assessment in a clinical clerkship carries a commitment to decision-making in the following examination design categories.

1. Blueprinting: Design of the exam should parallel the clerkship objectives.
    a. Review the clerkship's content goals and objectives to determine the patient problems (e.g., chest pain) and diseases (e.g., valvular heart disease) that should be emphasized on the examination.
    b. Review the clerkship's skill goals and objectives (e.g., interviewing, physical examination, lifestyle modification counseling) to determine the activities to include in the SP-based exam.
    c. Insure students have sufficient opportunity to learn or practice the desired skills (through patient encounters or supplementary activities) before inclusion in the assessment exercise. Although SP-based activities often focus on assessment, the technique presents powerful opportunities for teaching clinical skills and providing formative feedback as well.[239]

2. Examination construction: Determine, based on educational philosophy and practical constraints, whether to select a format that emphasizes a small number of relatively comprehensive patient encounters, often referred to as a Clinical Examination Exercise (CEX), or a larger number of relatively focused encounters, often referred to as an Objective Structured Clinical Examination (OSCE). This labeling distinction is artificial, but reflects the trade-offs inherent in exam design.
    a. The number of patient encounters should ideally reflect the range of skills being assessed. Adequate sampling is essential to limit the effects of case specificity on examination reliability.[149, 240]

b. The length of each patient encounter should reflect the complexity of the task. For example, a case which directs the student to complete a comprehensive interview and physical exam will require more time than one which is problem-focused. Examination formats can range from as brief as a few minutes ("Perform a history of present illness for this patient's chest pain"), to as long as an hour or more ("Complete a comprehensive interview, physical and neurologic examination on this new patient with hemiplegia, discuss the patient's most likely diagnoses with him, and conduct patient counseling on long-term care and lifestyle modification"). In practice, most SP-based examination encounters in clerkships tend to range from 10-15 minutes for problem-focused encounters to 25-30 minutes for more comprehensive evaluations. It is appropriate to acknowledge that little research has been done in this area, and most decisions regarding encounter length by exam designers are made empirically.
c. While the primary focus of an SP-based examination is the patient encounter itself, the exam format lends itself favorably to linking patient encounters with post-encounter exercises. Such exercises, typically lasting 5-10 minutes and occurring immediately after each patient encounter, most commonly include written notes, oral presentations, or written responses to questions about the preceding case. Since simulation of abnormal physical findings is not always possible by a standardized patient, students may be shown an exhibit (such as a photograph of an optic fundus or a tympanic membrane), which challenges them to identify a specific pathologic finding.
d. The most common method of student assessment in SP-based exercises is checklist-based ratings completed by patients after each encounter. Therefore, checklists must be created that are relevant to the case, of reasonable length, and include items scorable by patients. Long checklists (unless scored by a separate observer) or those that include items requiring medical knowledge to rate, may threaten the reliability and validity of the resulting scores.

3. Use of faculty: Involvement of faculty is a critical component of standardized patient-based assessment in a clerkship. Such involvement can range from case writing and exam development to participation in the assessment exercise itself. Faculty participation in the exercise can include direct observation and grading of the student during the patient encounters, or assessment of oral presentations or responses to questions during post-encounter exercises.

Resource Requirements of Standardized Patient-based Assessment

The clerkship director and involved faculty form the nucleus for successful SP-based assessment, typically by identifying relevant cases, determining performance standards for students, and creating appropriate checklists or criteria for student scoring. Yet, successful integration of SP-based assessment into a clerkship requires investment in resources that often stretch beyond the capacity of individual courses or departments. Valuable (often essential) school-based resources include identification of examination space, whether a dedicated assessment center or access to appropriate shared space, such as use of a clinical facility during non-patient care hours. The most valuable central component of an SP program involves professional staff who can assist with script creation and patient training, as well as assist with the logistic challenges of exam administration.

The Advantages of SP-based Assessment

Traditional clinical skills assessment in the clinical setting relies heavily on random observations or inference based on written records and oral presentations of patient encounters. The use of SPs to achieve systematic and direct observation of patient encounters overcomes four limitations of traditional assessment by clinical faculty.

Table 6.8.1. Assessment Limitations addressed by Standardized patients

Limitations of Assessment in a Clinical Setting	Potential Advantages of SP-based Assessment
Evaluator skill: Faculty who supervise students may not have the motivation, training, or skills for optimal evaluation	SPs can be selected for interest and aptitude, and trained to predefined standards of accuracy
Evaluator time: Faculty who supervise students have other responsibilities which limit their opportunities for direct observation	The sole focus of an SP encounter is clinical teaching and assessment
Evaluator standardization: Faculty observe a range of student behaviors, and have varying standards for student performance	Skills to be assessed and standards for mastery can be defined in advance, and each SP portraying a case can be trained to the same standards
Sampling of Challenges: The random clinical problems encountered by students may not match the problems they need to master	Clerkship directors can select the range of problems an examinee encounters in an SP examination

The Limitations of SP-based Assessment

SP-based assessment cannot stand alone or replace the judgment of clinical faculty in student evaluation. Although SPs can simulate some physical findings (e.g., neurologic abnormalities), evaluation of a student's ability to detect physical abnormalities such as heart murmurs requires SPs with stable pathologic findings, or use of mechanical/electronic simulators to supplement SP encounters.[241]

SPs are ideally suited to scoring of medical interview and physical examination performance on dichotomous scales (done vs. not done) and can also be trained to evaluate technical correctness of many examination and communication skills. However, more subtle distinctions of performance, such as interview organization or sequence, and time management typically require the judgment of a physician. SPs should not be expected to evaluate students' clinical reasoning or problem-solving abilities, which are best accomplished by direct student-faculty contact. For these reasons, face validity of SP-based examinations is highest when SP-based assessment is combined with the expertise of clinical faculty.

SP-based Assessment Costs

SP-based assessment expenses include direct and indirect costs (Table 6.8.1) related to exam development, administration, and infrastructure. Patient training and portrayal costs range from

$15-20/hour for straightforward tasks, $20-30/hour for more complex cases, to $50+/hour for patients who serve as Male Urologic Training Associates and Gynecological Training Associates. Published cost estimates have utilized various methods of cost calculations, and therefore provide little generalizable information for estimating these overall costs.[242]

Table 6.8.2. Estimating Costs for Standardized Patient Based Assessment Exercises

Direct costs	Indirect costs
SP training SP portrayal Training materials ■ Copying costs Consumables ■ Patient gowns and sheets ■ Items used by the students during the examination, such as gloves, tongue blades, disposable tips for otoscopes	Facility use/maintenance SP staff salaries (for recruitment and training of SPs) Faculty time for ■ Case development ■ Examination development, scoring, and interpretation ■ Direct observation during the examination ■ Student orientation and supervision during the exam

Because a typical clerkship-based examination will involve multiple stations, and for larger classes will require more than one SP trained to each case, common estimates of direct (SP) costs for exams range from $50-100/student.[242]

What Has the Research Literature Taught Us About SP-based Assessment?

Patient issues

SP Portrayals: Realism and Consistency

Several studies have demonstrated the ability of well-trained SPs to simulate a clinical encounter so realistically that they cannot be distinguished from genuine patients, even when scheduled into the physician's practice unannounced.[243] Similarly, observation of SP portrayals suggests that an SP can consistently reproduce interview and examination findings more than 90% of the time.

SP Reliability in Checklist Ratings (Inter-rater Reliability)

Several studies have addressed scoring reliability of a single SP subjected to fatigue or long scoring checklists, as well as variation among different SPs simulating the same case. Scoring reliability is influenced by complexity of checklists, but within certain parameters (15-25 checklist items; <4-5 hours of portrayal with breaks), SP reliability appears to match physician raters. Most studies suggest that adding multiple observers to a case adds little to exam reliability.[244]

Examination issues

Case-specificity (Inter-case reliability)

A single patient encounter, whether genuine or SP-based, will not provide a reliable indicator of an examinee's skills when the goal is to provide a comprehensive assessment of clinical ability. Four to six hours (or more) of total testing time is typically required to reach levels of reliability (generalizibility) expected for pass-fail decisions on high-stakes examinations.[245] Exercises designed to assess less complex skills, or which focus on teaching and feedback are typically not as time intensive.

Validity

Assessment of validity of SP-based examinations is a complex issue, influenced by the content and structure of the exam, and by the perceived correlation between SP-based assessment and other performance measures.[246] As mentioned previously, correlations between SP-based assessment and multiple-choice question examinations are only moderate, because these assessments measure skill sets that are only partially overlapping. Most measures of validity attempt to establish a relationship between SP exam performance and faculty clinical evaluations without establishment of a standard for comparison. While SP-based studies have shown that learners at higher levels tend to perform better than their lower-level counterparts (fourth-year students compared to second year students, or residents compared to students, for example), these studies do not establish a precise correlation between experience and performance. Other studies demonstrate the limitations of the checklist-based assessments in SP exams, which may not capture the subtlety of skills demonstrated by higher levels of learners.[247] This is one reason why the decision to have SPs rate students (utilizing discrete checklist items) versus having faculty observers rate students (often utilizing global rating scales) is so challenging. Student perceptions of the scoring system may also influence performance.[248] (for discussion of OSCE performance as a test, see Chapter 6, Section 12)

Incorporating SP-based Assessment into a Comprehensive Clinical Evaluation Program

While the use of SPs elsewhere in the curriculum is not directly under a clerkship director's control, success of a clerkship-based SP program can be substantially influenced by it. Students' acceptance of SP-based testing is enhanced when they have encountered SPs in a teaching mode earlier in the curriculum. Use of SPs by multiple courses also strengthens the practicality of developing stable school-level resources, such as a dedicated SP testing center and permanent professional staff for patient recruitment and training. Ideally, the skills upon which students are tested during clerkships are skills that have been stressed and reinforced through practice and testing in earlier courses.

The July 2004 addition of a standardized patient-based clinical skills component to Step II of the United States Medical Licensing Examination has raised many questions about the optimal use of SPs in the medical curriculum. It is too early to predict which, if any, factors will be helpful in predicting which students are likely to pass or fail this assessment.

A comprehensive listing of articles relating to standardized patients in medical education is available at http://oed.utmb.edu/SP/bibliography.htm

Section 9. Evaluating Professionalism

Shiphra Ginsburg, MD and William McGaghie, PhD

Background

The Introduction to this chapter states, "Professionalism is expressed in each young physician's character, reliability, honesty, ability to keep confidences, and other nonacademic qualities that embody 'the good doctor.' Professionalism is more than maturity and less than sainthood; it connotes promises of expertise and duty. In medical circles professionalism is usually conspicuous by its absence and taken for granted when present."

Medical professionalism is a slippery construct because it can convey many meanings to different people. Some of the meanings are general, like altruism. Others are specific, such as punctuality. Components of medical professionalism are embodied in the three fundamental principles (primacy of patient welfare, patient autonomy, social justice) and 10 professional responsibilities (e.g., professional competence, honesty with patients) in the 2002 *Physician Charter* published jointly by the ABIM Foundation, ACP-ASIM Foundation, and the European Federation of Internal Medicine.[249] Medical professionalism is also a curriculum imperative of the Medical School Objectives Project sponsored by the AAMC.[250] However, despite assertions about its importance, recent scholarship shows that medical professionalism is difficult to measure reliably.[13] Absent reliable measurement data, evaluating the professionalism of medical learners either for formative feedback or summative grading is a real challenge.

Despite these obstacles the Liaison Committee on Medical Education (LCME), the body responsible for medical school accreditation in North America, insists that clerkship directors and other faculty are responsible for evaluating medical student professionalism. The *Standards for Accreditation of Medical Education Programs Leading to the M.D. Degree*[251] assert:

> "The medical school faculty must establish a system for the evaluation of student achievement throughout medical school that employs a variety of measures of knowledge, skills, behaviors, and attitudes." (p. 14)
>
> "A medical school must teach medical ethics and human values, and require its students to exhibit scrupulous ethical principles in caring for patients, and in relating to patients' families and to others involved in patient care." (p. 13)
>
> "'Scrupulous ethical principles' imply characteristics like honesty, integrity, maintenance of confidentiality, and respect for patients, patients' families, other students, and other health professionals. The school's educational objectives may identify additional dimensions of ethical behavior to be exhibited in patient care settings." (p. 13)

In short, North American medical schools and their faculty, clerkship directors included, have no choice about evaluating student professionalism. It simply must be done for a medical school to keep its doors open. Yet there are several other reasons for evaluating medical student professionalism than just maintaining school accreditation. Readers should also know with more specificity why professionalism is so difficult to evaluate. The next two sections treat these two issues briefly.

Why Professionalism is Important to Evaluate

There are at least four reasons why clerkship directors should evaluate student professionalism.

1. Professionalism is a key feature of clinical practice as articulated in the *Physician Charter*[249] and other contemporary writings. This is not a new idea. Nearly identical statements about the need for professionalism in clinical practice, beyond technical competence, are found in writings about medieval and renaissance medicine[252] and medical practice in the ancient world.[253]
2. Organizations that oversee clerkship education in all specialties insist, like the LCME, that student professionalism must be evaluated rigorously.
3. Students who display unprofessional behavior in medical school are much more likely to face disciplinary actions later in practice than students who behave appropriately.[14]
4. Reports of unprofessional medical behavior that appear in the popular press[254] remind the medical education community about its social contract with the public, including the occasional need to remove bad apples.

Why Professionalism is Difficult to Evaluate

Research and experience show that five problems are the main reasons why medical professionalism is difficult to evaluate.

1. Most definitions of the features of professionalism are fuzzy, indistinct. They tend to focus on enduring traits or subjective attributes of people (e.g., Edward is compassionate) rather than on overt, measurable behavior.
2. Medical professionalism is manifested by persons in specific clinical situations, what social psychologists call person X situation interactions.[255] This means that evaluations of individual medical professionalism must be done by assessing a *variety* of behaviors in *many* situations, not just on a one-shot basis.
3. Problems in medical professionalism frequently originate from *values conflicts*. These are circumstances where the student or clinician must strike a balance between competing interests such as honesty vs. patient confidentiality. Appropriate professional behavior is easy to discern in the absence of conflict, yet clinical situations without some conflict are rare.
4. An individual's resolution and judgment is manifest in the behavior s/he displays (e.g., telling the truth, withholding information, breaking confidentiality). The behavior we observe about an individual that is used for evaluation seems more objective than our inferences about the behavior. However, if all we evaluate is the overt behavior we have no way of knowing if the student truly comprehends the situation. Deeper probes are needed.
5. There is a pervasive reluctance among evaluators to transmit bad news. This is the widespread "MUM Effect,"[256] not only observed in medical education evaluation but also in clinical practice and other endeavors. The "MUM Effect" must be acknowledged and addressed to assure honest evaluations of medical student professionalism.

A much more detailed and scholarly review about medical professionalism and its measurement has been published as a monograph recently.[257] Interested readers are urged to obtain and use this resource.

How to Evaluate Professionalism

The Background discussion suggests that evaluating medical student professionalism is similar to Winston Churchill's description of the 1940s Soviet Union: "A riddle, wrapped in a mystery, inside an enigma." But clerkship directors are practical people who are responsible for evaluating and grading medical student professionalism every day, despite the flaws and uncertainties. What shall they do on Monday morning?

Specify Evaluation Criteria

A good place to start is a report of work done by Jon Veloski and colleagues[12] that " . . . analyze[s] the measurement goals and the reliability and validity of the instruments used in studies related to the measurement of professionalism reported in the [medical education] literature over the past two decades" (p. 366). This work is summarized in Table 6.9.1, which is reproduced from the report by Veloski et al.[258]

Table 6.9.1 (Table 1) clearly shows that the attribute of professionalism that has been used most frequently to develop instruments for medical student evaluation over the past 20 years is *Ethics, decision making and moral reasoning.* This is followed, in order of decreasing frequency, by humanism, multiculturalism, and empathy down to self-assessment and mixed attributes. Table 6. distinguishes professionalism as one facet of [medical] competence (reported in nine studies) vs. professionalism as a comprehensive construct (reported in 11 studies). The table also points out other signs of medical professionalism that have been used for evaluation instrument development: abuse and harassment of students, housestaff; patient satisfaction; cheating; attitudes toward [medical] uncertainty; cynicism; and "turfing" (of patients to others).

Clerkship directors who are responsible for evaluating medical student professionalism must translate these broad and fuzzy evaluation criteria into measurable operational definitions. The operational definitions should include a situational component (e.g., truth telling in the outpatient clinic, oncology ward, MICU, etc.) that allow for tailored measurements in real or simulated circumstances. The goal is to isolate and measure facets of student professionalism that coincide with a particular clerkship's goals. Professionalism priorities for pediatrics in Portland are unlikely to share ranks with radiology in Rochester. Variation in clerkship professionalism goals should be acknowledged and prized.

Measuring Professionalism

Once evaluation criteria are specified (evaluation "whats") the responsible clerkship director will choose the best and most practical way to measure the professionalism outcome (evaluation "hows"). Table 6.9.1 identifies 16 measurement methods that can be used to evaluate student professionalism. They range from objective measurement methods including formal examinations, OSCEs, standardized patients, and simulations to more subjective measurement methods that are grounded in an evaluator's perceptions and interpretations. Subjective measures include global and specific clinical ratings by students' supervisors (attending physicians, fellows, residents); assessments by other clinicians (e.g., nurses, EMTs); patient ratings; peer assessments,[259] self-assessments; and many others.

Table 6.9.1

Table 1

Elements of Content Definition Used to Develop Instruments in 134 Studies Related to the Measurement of Professionalism, 1982–2002

Definition	No. of studies
Specific attributes of a professional*	
Ethics, decision making moral reasoning†	48
Humanism	11
Multiculturalism	8
Empathy	4
Values	4
Deception in patient relationships, attitudes toward	3
Indigent, care for	2
Trust	2
Attitudes and communication	1
Confidentiality of patient data	1
Contact with patients, appropriate/inappropriate	1
Emotional intelligence	1
Mental health	1
Self-assessment	1
Mixed attributes	6
Professionalism as one facet of competence	9
Professionalism as a comprehensive construct	11
Other phenomena	
Abuse and harassment of students, housestaff	7
Patient satisfaction	5
Cheating	4
Uncertainty, attitudes toward	2
Cynicism	1
Turfing	1
Total	134

* The list of attributes of a professional is empirical, having been derived from titles of the article, abstracts, purpose, or key terms provided by the reviewers. Each article was assigned to one category. However, the attributes in this list are not necessarily mutually exclusive, and the list is not intended to be exhaustive.

† The heterogeneous category "ethics" includes ethics, ethical decision-making skills, moral reasoning abilities, and related attributes.

(reproduced with permission Veloski et. al. Measuring Professionalism: a review of studies with instruments reported in the literature between 1982 and 2002. Academic Medicine, 80:366-70. 2005)

Each of these methods of measuring student professionalism has strengths and limits depending on one's evaluation goals and common situations of practice. Savvy clerkship directors will work hard to match evaluation tools with evaluation goals, recognizing that the fit is never perfect. They will also use quality control or audit mechanisms to routinely assess the reliability and validity of student evaluation data. High quality evaluation data, whether objective or subjective, is essential for both formative student feedback and summative student grading.

Descriptive evaluation, described in Chapter 6, Section 3, of this chapter, is especially well suited to measure medical student professionalism. In particular, the R-I-M-E Framework (student as reporter, interpreter, manager, and educator) is an excellent approach to capture instances of student professional behavior in a cumulative and consistent way during a clerkship. In reporter mode students must gather data about patients, use appropriate terminology, interact professionally with patients and staff, and fulfill their responsibilities consistently and reliably. As interpreters students identify and rank problems, explain patient problems, and formulate a differential diagnosis. The manager function has students suggesting diagnostic options and possible therapies. The educator role expects students to be self-directed learners and to share knowledge with the health care team. The authors of this section, David Carnahan and Paul Hemmer, report that third year students in their medicine clerkship are expected to master "reporter" skills and show evidence of progress toward the "interpreter" phase in order to advance. The "manager" and "educator" phases are best evaluated among fourth year medical students.

The R-I-M-E Framework provides many opportunities to collect and use descriptive data to evaluate student professionalism. Clerkship directors need to carefully consider the features of professionalism (i.e., criteria) they wish to evaluate and then build a R-I-M-E system that measures their intentions.

Decision Making About Students

Evaluation of medical student professionalism is done for a pragmatic reason, to make decisions about their advancement, promotion, or remediation. Decision-making about individual medical students is one of a clerkship director's most important responsibilities. Judging student readiness to move ahead or keeping them behind is a tough job, frequently layered with emotion and doubt. Having a thoughtful, uniform, and systematic approach to decision-making about student professionalism makes the procedure predictable, manageable, and transparent. Such a system makes student and faculty roles and expectations plain and dispels anxieties about potential hidden agendas, grudges, and favoritism.

Thoughtful decision-making about individual medical students begins with an evaluation plan. The plan involves standardized evaluations so all students are assessed uniformly and fairly. It includes clear-cut evaluation criteria, measurement methods, student achievement standards, and mechanisms of data management and quality control. Such a system will have a uniform definition of professionalism for all students enrolled in a clerkship, a definition that is plain, public, and captured by evaluation tools. This results in no surprises for students and few complaints for faculty.

Section 10. Clerkship Examinations

Cyril Grum, MD

Introduction

End-of-clerkship examinations should be part of the evaluation process for all students in core clinical clerkships. These examinations should supplement clinical evaluations by attending physicians, preceptors, and house officers. Most clerkship directors subscribe to the philosophy that the evaluation of clinical performance should contribute the majority weight to the overall grade. However, end-of-clerkship objective examinations are an important component of a comprehensive evaluation process.

Purposes of Clerkship Examinations

Examinations motivate students to study. The assessment methods and content drive students' learning more than any other single stimulus. Therefore, assessment methods should be matched to learning objectives. If it is desirable for students to memorize and recall answers, test memory and recall; to motivate students in their acquisition and application of knowledge, test knowledge and its application. For example, the National Board of Medical Examiners Subject Test motivates students to read and improve their knowledge base. This examination reinforces the importance of reading. It emphasizes that a sufficient knowledge base and ability to apply it to clinical situations is a critical part of clinical competence. However, multiple-choice tests cannot assess bedside history taking and physical examination skills, communication skills, or procedural skills.

End-of-clerkship evaluations are important to evaluate key features of students' competence. Consider testing specific areas of knowledge, such as electrocardiogram interpretation, chest x-ray interpretation, or recognition of heart sounds, using a set of pass/fail quizzes at the end of the clerkship. Some clerkship directors require passing these quizzes as a course requirement, but do not use the scores in determining the final grade.

Examinations tend to segregate students (i.e., identify the top and bottom performers) to a greater degree than clinical evaluations.

In addition to evaluating students, examinations are helpful for evaluating the clerkship curriculum and the level to which a local curriculum meets national standards.

Assessment Options

National Board of Medical Examiners (NBME) Subject Tests

The NBME Subject Tests are available for all core clinical disciplines and are high-quality, psychometrically sound, clinical-vignette-based, multiple-choice examinations. Greater than 90% of students take the National Board of Medical Examiners' Medicine Subject Exam at the end of the Medicine clerkship.[204] This national, standardized examination should be supplemented by locally generated examinations that are tailored to the goals and objectives of the clerkship.

The NBME Subject Tests are national tests that assess a student's overall knowledge in a clinical specialty. The Medicine Subject test reflects the Internal Medicine portion of USMLE

Step 2. An important advantage of the Subject Tests is that they have superb psychometric properties. The reliability (typically 0.75 to 0.85, Ripkey[260]), validity and standardization of these examinations add to their desirability as an assessment modality.

Questions often arise regarding how the Subject Tests are developed and who writes the questions. Using medicine as an example, a committee composed of clerkship directors and residency program directors writes the questions. That is, educators who have extensive experience with students and residents determine the content and questions. The goal of the Medicine Subject Test Committee is to write questions that a first day intern in internal medicine should know. This is a person who has graduated from medical school and ready for a certain degree of independence, but still has senior residents and attending physicians to guide their practice.

The precise distribution of Subject Test questions must be confidential to preserve the test's integrity. A description of the NBME Subject Examination Program can be found at:

http://www.nbme.org/programs/SrvSubjectExams.asp

A general scheme of the distribution of Subject Exam questions is stated at:

http://www.nbme.org/programs/subjexamsclin.asp

as well as a pdf file with some sample questions. It is clear that the Subject Test template and the learning objectives identified in the Core Medicine Clerkship Curriculum Guide[261] are very similar. [The SGIM/CDIM Core Medicine Clerkship Curriculum Guide can be found at:

http://www.im.org/AAIM/Pubs/Docs/CDIMCurriculumGuide/TableofContents.htm]

This should be no surprise as both the NBME and medical educators nationally feel that students should be tested about common problems that they are likely to see in an undifferentiated medical practice in the United States. The use of a national examination allows clerkship directors to gauge the performance of their students against national standards.[262] The NBME scores the examination and reports scores to the school. The subject exam reports are sent to the chief proctor at each school via the web on a weekly basis. The chief proctors can download a pdf file of the results and send them to the clerkship director. In addition, the Subject Test program provides an item analysis, thereby allowing a clerkship director to identify specific areas where their students are particularly strong or weak. Fees for the 2005-2006 Subject exam administration are $32 per test per examinee. Fee structure is posted at:

http://www.nbme.org/programs/subjexfees.asp

The exam must be proctored. An extensive "Information Guide" as well as a "Test Administration Handbook" is available as a pdf file:

http://www.nbme.org/programs/medsch.asp.

It is also important to recognize what NBME Subject Tests do not address. Specifically, the NBME Subject Exam does not address students' clinical skills, or professional or personal attributes. Although the Subject Test allows comparison with a nationally accepted standard, it does not specifically evaluate whether the local curriculum is effectively delivered.

In essence, a "Subject Exam" is an end-of-third-year examination. Since the Subject Test program was not specifically designed as an end-of-clerkship exam, this may pose some minor problems for clerkships that are given early in the year, but should not be a consideration for clerkships given in the second half of the junior year.[248, 262-264] The Score Interpretation Guide that comes from the NBME with each medicine subject test results indicates that for Internal Medicine the mean score on the subject exams increases only 2.4 points over the course of the year (this data is based on the first time exam takers for the 2001-2002 academic year). The mean score is 72.6 for examinees taking the medicine subject exam in the first quarter of the

year and increases gradually to 75.0 for those students taking it the last quarter of the academic year.

The number of clerkship directors who use the NBME Medicine Subject Test has increased over the past 15 years.[80, 204, 265] Reasons may include the difficulty and time commitment needed to produce a local examination that is valid and reliable, as well as the increasing match between the content of the Subject Exam questions and the core clerkship objectives. The NBME continues to express interest in working with clerkship directors to develop examinations that satisfy clerkship needs[266] Clerkship directors often serve on NBME writing committees. The integral involvement of clerkship directors in the generation of the Subject exam, helps to ensure that questions are relevant and at the appropriate level.

The last survey of clerkship directors by CDIM in the spring of 1999, indicated that 83% of the clerkships are using the NBME Subject Exam, representing greater than 90% of all US third year medical students.[204] Hemmer and colleagues reported in 2002 that over the past decade presiding schools using the NMBE Medicine Subject Exam has increased from 66 to 83% whereas the clerkship using a faculty written examination has decreased from 46% to 27% Since the publication of Hemmer's paper in 2002, these trends have continued. The 1999 survey also indicated that 80% of schools set a passing score, which is required for successful completion of the clerkship. An analysis utilizing 50 clerkship directors by the National Board using a Hofstee method to set passing score indicated that a mean score of 60 for passing, with a standard deviation of 2. The range of passing scores in this analysis was between 53 and 64. This information is included with subject exam results. The exam contributes approximately 25% to the student's final clerkship grade.[204] The CDIM survey indicated that approximately a quarter of the clerkships use written examinations developed by local faculty and a quarter use standardized patient exams. Clerkship directors typically allow students one additional opportunity to retake the subject exam if they fail it on the first time,[204] but most do not allow additional retests after failing the exam for the second time. Students who fail the subject exam retakes are often given a failing grade for the clerkship and are prescribed some type of remediation.

The National Board of Medical Examiner's Medicine Subject Exam is very convenient to use, has extremely strong psychometric properties and is relatively inexpensive to administer. These factors contribute to the fact that the vast majority of clerkships use them and has also led to a decrease in use of faculty-developed examination over the past decades. Clerkship directors have approached the National Board of Medical Examiners to improve the subject exam, especially to have more input in shaping the content of the exam.[266] The National Board has responded by developing a process of flexible blueprinting which is anticipated to be available in future years so that clerkship directors will be able to tailor a subject exam to their local environment.

Faculty-generated Examinations

One-best-answer Multiple Choice Examinations

Faculty-generated written examinations to test knowledge and its application can be developed based on multiple-choice or extended matching formats. Both formats use patient-case vignettes. Attaining acceptable reliability and validity on examinations generated in-house is difficult.

Examinations generated by local faculty have the advantage of testing specific knowledge and clinical skills relevant to that particular clerkship and curriculum. Therefore, local examinations theoretically should have high content validity.

However, there are several potential disadvantages: the psychometric characteristics (especially reliability) are often poor, questions may be included that have not been previously validated, and question quality may be poor. Test construction and grading require considerable faculty time and effort, which is increasingly difficult to acquire. Furthermore, many faculty have not been trained in examination development and question writing and, therefore, cannot be expected to write high quality questions. Poor quality questions that do not reflect the important content often help test-wise students and do not assess their overall knowledge and ability to apply it. Finally, local examinations cannot be used to determine student performance against national standards.

Despite these drawbacks, having an examination process that includes both national standards as well as locally generated examinations is appealing pedagogically. Local examinations have "content validity" and may make students confident they are being examined regarding material their own faculty feel is important. In addition, the use of local examinations reinforces the educational mission of all medical centers and reminds the faculty that evaluation of students is a critical part of the overall teaching process.

If faculty will generate multiple choice examinations to test knowledge locally, an excellent manual exists that will help them improve the quality of these multiple choice questions. This manual entitled "Constructing Written Test Questions for the Basic and Clinical Sciences", 3rd Edition, written by Susan Case and David Swanson is available as a pdf file at the NBME web site: http://www.nbme.org/about/itemwriting.asp. This manual reviews issues related to technical item flaws and issues related to item content. The manual also helps staff to review statistical indices of item quality after test administration. An overview of the standard setting techniques is also provided.

The National Board of Medical Examiners also provides periodically item-writing workshops to help faculty construct better quality multiple choice questions. These are available by special arrangement and usually takes place at the host medical school.

Essay and Other Open-end Question Examinations

Faculty-generated essay questions and other open-ended questions have some pedagogic appeal. However, unless there are only a small group of students per rotation or faculty who are extremely motivated to grade examinations, the time-intensive nature of these examinations makes them infeasible. In addition, it is very difficult to ensure reliability with open-ended responses.[268-269]

Oral Examinations

Faculty-administered oral examinations, once used frequently as end of clerkship assessment in internal medicine are now used rarely because they are not sufficiently reliable and require considerable faculty time to administer.

Oral examinations can assess students' ability to reason and to problem solve, as well as assessing overall factual knowledge. Assessment of factual knowledge using oral examinations is often significantly less precise than using written examinations, whether they are from the National Board of Medical Examiners or locally generated. Furthermore, the logistics of administering an oral examination, as well as determining reliability and validity, can be formidable. Many factors contribute to low reliability. Typical oral examinations allow time for the student to be presented with only one or two clinical scenarios. The student's ability to

perform well in these scenarios may be a reflection of whether the student encountered a patient with a similar problem during a clinical rotation. Although overall performance of a class on oral examinations may be a useful indicator of the class's experience and level of competence, an individual student may score quite high or quite low on the examination, based on the randomness of clinical material rather than an evaluation of their ability. Oral examinations are often administered by many faculty members, leading to variability in examiners' assessment of a particular student. Furthermore, very few faculty examiners have been trained in the process of administering oral examinations and often there is no standardization among the examiners. These factors all contribute to a very low inter-rater reliability. The time-consuming nature of the examination for faculty and students has added to their unpopularity. On the other hand, oral examinations often are a powerful motivating force for students to study and review core material.

If an oral examination is used, it would be wise to limit its scope and standardize it as much as possible. For example, students may be asked to submit a list of patient problems they encountered on a particular rotation. The faculty member limits the oral examination to questions about these specific clinical problems. Oral examinations are currently used much less frequently than NBME Subject Tests or locally generated written examinations. When they are used, they are usually weighted less heavily for determining grades than other examinations.

Objective Structured Clinical Examinations (OSCE)
A clinical practical examination, an objective structured clinical examination (OSCE), assesses students' performance regarding specific clinical skills, usually involving the evaluation of standardized patients. (see Chapter 6, Section 8, [Standardized Patients]) This assessment method is being used increasingly and is a viable method for assessing clinical and interpersonal skills that cannot be assessed using multiple-choice examinations.[149]

Over the last decade the number of clerkships using standardized patient exams has increased from 2% in 1989 to 27.5% in 1999. This mode of examination is increasing exponentially, largely in part to the addition of a clinical skills exam to USMLE Step 2.[204]

OSCEs are becoming increasingly popular to assess specific clinical skills at the end of clerkships because they are perceived as being more objective than evaluations received from attending physicians and house officers.[269-270] The addition of an OSCE component to the USMLE Step 2 Examination, during the academic year 2004-2005, indicates both the popularity of and satisfaction with this examination technique. The presence of a USMLE Step 2 clinical skills exam will further amplify the development of OSCEs at all medical schools. Background research by the National Board of Medical Examiners prior to the implementation of the clinical skills exam indicated that the general public feels quite strongly that clinical skills are necessary for each physician and should be tested. A Harris Interactive survey was conducted by telephone within the US between December 12 and 16, 2002, among a nationwide cross section of 1,023 adults (ages 18+): http://www.usmle.org/news/Step2CSNews/Harris.asp. The survey concluded that an overwhelming majority of Americans consider good clinical and communication skills critical for physicians and believe students should pass an examination that tests these skills before receiving their medical licenses: http://www.usmle.org/news/Step2CSNews/newsrelease2503.htm

Students, faculty, and the general public have the perception that OSCEs test specific clinical skills that are necessary for each physician to possess. As a result, the face validity of the examination is quite high.[271] The OSCE has high face validity because the cases can simulate

experiences the students have encountered during clinical rotations and core competencies that will be necessary for internship and beyond. OSCEs can be designed to test any particular facet of students' clinical skills including ability to take a directed history, to perform various aspects of the physical exam, to interpret laboratory and x-ray studies, and to make clinical decisions about specific clinical problems.[272] In its simplest format, an OSCE can present clinical case scenarios including laboratory data, chest x-rays, EKGs, and photographs of clinical findings, but not include standardized patients.

OSCEs have the disadvantage of being very labor intensive and costly because standardized patients must be recruited, trained, and compensated. Standardized patients can be individuals who are taught to simulate a particular disease or who have stable physical findings. Although many Internal Medicine Clerkships have added an OSCE as an end-of-clerkship examination, it may be less expensive and time-consuming to administer a cross-clerkship OSCE at the end of the third year. In addition, a centralized, multidisciplinary examination is likely to be of higher quality than department-based end-of-clerkship examinations.

Lessons from Research on Clerkship Exams

Locally generated examinations may be based on different premises and may test different characteristics than NBME Subject Tests. For example, a locally generated diagnostic pattern recognition examination, administered at the end of the clerkship, features brief patient vignettes that describe classic presentations of common diseases.[273] Students identify the correct diagnosis from an extended matching list of between 16 to 26 diagnoses. The overall correlation with the NBME Medicine Subject Test is quite high, although about 10% of students performed about a standard deviation higher and 10% a standard deviation lower than on the Subject Test. This suggests that the students' ability to recognize common diagnostic problems may be independent of their ability to perform well on a knowledge-based examination such as the NBME Medical Subject Test.

The similarities between the Subject Exam and USMLE Step 2 may allow for some inferences as significantly more research has been done on the USMLE Step 2 than has been done on the Subject Exam. The validity of the USMLE Step 2 Clinical Knowledge Exam has recently been examined by addressing the degree to which experts view exam content as clinically relevant and appropriate to Step 2.[274] The underlying principle of USMLE Step 2 is to assess whether an individual can apply medical knowledge, skills and understanding clinical science essential for the provision of safe and effective patient care under supervision (i.e., a new intern on the first day of internship). Cuddy and co-investigators asked 27 experts to individually rate the clinical relevance and appropriateness of 150 questions. They demonstrated that 92% of expert judgments indicated that the item content was clinically relevant, 90% indicated that the content was appropriate for Step 2 and 85% indicated the content was used in clinical practice. The regression analysis indicated that the more difficult items and the more frequently used items were considered more appropriate for Step 2. The results indicated the majority of item content is clinically relevant and appropriate, providing important validation support for the USMLE Step 2 Exam.

NBME Subject Tests and end-of-clerkship examinations reflect cumulative knowledge, including that acquired from basic science courses and prior clinical experiences. Therefore, they are not pure assessments of the knowledge acquired during a particular clerkship. To determine what knowledge was acquired during a clerkship itself, assessment of prior knowledge or control for prior experience must be documented. In practice, this is usually done only for medical education research purposes. Clerkship directors recognize that students in the latter half of the

year are likely to have significantly increased knowledge, skills and attitudes compared to students early in the year. One reasonable approach, therefore, is to evaluate student Subject Test performance based on the timing of the clerkship, and use a "normative" method to compare one student with others taking the clerkship at the same time (see Chapter 6, Section 2). Literature on this subject has produced variable results. In one study, different versions of the NBME Subject Test were given on the first and last day of a medicine clerkship to all students over 2 consecutive years.[275] Students' mean scores were equally low at the beginning of the clerkship regardless of when they took the clerkship or which other clerkships they had previously taken. However, students in the second half of the year had greater improvement in performance at the end of the clerkship than their colleagues at the beginning of the year. The experience of many clerkship directors nationally has shown that students' performance on the NBME Subject Tests and other examinations tends to be higher as experience increases during the third year.[248, 262-264] Since this reflects higher student achievement later in the year, a "criterion-based" system of evaluation would accept a rise in clerkship grades later in the year as an appropriate reflection of this. Furthermore, relating subject examination performance to patient experiences and, by inference, curriculum, has proven elusive: students having an ambulatory experience in third year did not perform better than those who did not on questions categorized by clerkship directors as "ambulatory" in content.[276]

A recent study has indicated that the NBME Subject Test might be one of the more discriminating indicators when calculating student grades.[277] This study involved an Obstetrics and Gynecology Clerkship, which has many similarities to other clerkships. Student final performance was calculated by weighing clinical performance 60%, formal presentation 10%, oral exam 10% and the NBME Subject Test score 20%. Of these four indicators, only the NBME Subject Test score was normally distributed. The clinical performance score and the formal presentation scores were highly skewed, the oral examination scores was slightly skewed. The NBME Subject Test was the most highly correlated (r=.86) with the overall clerkship performance; much higher than the Clinical Performance, formal presentation score, or oral examination. The NBME Subject Test explained 74% of the variance in the overall clerkship performance. Although this study was done in an Obstetrics and Gynecology clerkship, there are lessons for all clerkships as clinical evaluations also tend to be highly skewed. Grade inflation has also been reported in the Internal Medicine Clerkship.[100] The clinical assessment is only a marginal discriminator of final performance because of the high degree of clustering at the high end of the performance spectrum. This author also found a very weak correlation between the NBME Subject Test scores and all the other major categories of assessment tools that they use. This would indicate that different categories of performance indicators are functionally independent measures of clinical achievement.

In-clerkship exams have been assessed to see if they can identify students with insufficient knowledge during the medicine clerkship.[98, 278-279] This concept is similar to the in-training evaluation that is a common feature in internal medicine residency programs. In-clerkship tests have identified students who are at risk of failing an end-of-clerkship examination. Unfortunately, counseling did not improve final examination pass rates.[280]

There is much interest in the relationship between student performance on the new USMLE Clinical Skills Exam and the USMLE Step 2 Clinical Knowledge Exam. Prior to the implementation of the Clinical Skills Exam, the National Board of Medical Examiners performed a study examining this relationship in 858 fourth year medical students participating in the 2002 Clinical Skills Examination Field Test and 6,372 international medical graduates who took the exam for the first time.[281] The results show only modest correlations, ranging from .16 to .38, between scores from a standardized patient exam of clinical skills and those from the multiple-

choice Clinical Knowledge Exam. One should exercise some caution in interpreting this result because the stakes for the US students taking this exam were relatively low (it was not required for licensure as it is now). This study provides evidence that the Clinical Skills Examination provides information about examinees that is not available in the Clinical Knowledge Exam.

Section 11 Writing Multiple-choice Questions

Ruth-Marie E. Fincher, M.D.

The purpose of this section is to assist faculty in constructing high-quality, multiple-choice questions (items) that evaluate students' knowledge, and their ability to apply it to clinical situations. Improvements and innovations in multiple-choice item formats, especially extended matching, allow item writers to simulate real clinical cases more closely than previously.[282-285] Learning to write multiple-choice items using the formats of the National Board of Medical Examiners (NBME) produces higher-quality items. Therefore, all faculty who write multiple-choice items for examinations should master the principles of item writing.[286]

The NBME uses items that are "one best answer" (type A, or matching); therefore, I will discuss only these types. I recommend avoiding other types of multiple-choice items, such as K-type (1, 2, and 3 only, 1 and 3 only, etc), multiple true/false, or A-B-Both-Neither. You should also provide the reference range of laboratory values unless you are testing the students' recall of the values. Generally, the goal is to assess the students' understanding of any divergence from normal, not whether they know the normal values. Normal values are given on USMLE Step examinations and NBME Subject Tests. In addition to the following discussion on writing multiple-choice items*, Improving Student Assessment: Evaluation in the Basic and Clinical Sciences* by Case and Swanson is an excellent reference[287] and it is available online (www.nbme.org/about/publications.asp)

Multiple-choice, One-best-answer Items

Multiple-choice, one-best-answer items require the student to select the single best response.

One-best-answer Item Types:

> *Single item stem* (Type A) consists of a single stem, usually with four or five response options. You may use three or more than five options.
>
> *Extended Matching* (Type R) consists of a set of 2 to 10 items, with 5 to 26 matching option responses.

Both item types can test recall or application of clinical knowledge. Strive to test application of knowledge rather than merely recall of information.

Recall vs. Application of Knowledge Examples:

<u>Example #1</u>: Recall

Which of the following is the most common physical finding in patients with pulmonary embolus?

A. Jugular venous distension
B. Right ventricular heave

C. S3 gallop
D. Tachypnea*
E. Unilateral leg swelling

This question tests recall of an isolated fact. In contrast, the following item requires the student to apply knowledge to a clinical situation, rather than simply recall a fact. (The correct answer is indicated by the asterisk - *)

Example #2: Application of Knowledge

A 66-year-old woman had the abrupt onset of shortness of breath and left-sided pleuritic chest pain 1 hour ago. She had been recovering well since a colectomy for colon cancer 2 days ago. Blood pressure is 160/90 mm Hg, pulse is 120/min, and respirations are 32/min. She is diaphoretic. Breath sounds are audible bilaterally, but inspiratory effort is decreased, S1 and S2 are normal, and jugular veins are not distended. Which of the following is the most likely cause of her acute condition?

A. Acute myocardial infarction
B. Dissecting aortic aneurysm
C. Pneumonia
D. Pneumothorax
E. Pulmonary embolus*

Example #3: Application of Knowledge (Higher order)

The following example tests higher order application of knowledge than the previous question.

A 66-year-old woman comes to the emergency department because of dyspnea and pleuritic chest pain for 1 hour. Her blood pressure is 160/90 mm Hg, pulse is 120/min, and respirations are 30/min. Her lungs are clear to auscultation. S1 and S2 are normal, and no murmur or gallop is heard. Electrocardiogram shows sinus tachycardia and nonspecific ST and T wave changes. Portable chest x-ray shows poor inspiratory effort and clear lung fields. Laboratory studies show:

WBC	$12000/mm^3$
Hematocrit	44%
Arterial blood gas (room air)	PO_2 62 mm Hg
	PCO_2 30 mm Hg
	pH 7.52

Which of the following is the most appropriate next diagnostic step?

A. Cardiac catheterization
B. CT scan of the chest
C. Echocardiogram
D. Pulmonary arteriography
E. Ventilation-perfusion lung scan*

The student must suspect the most probable diagnosis (pulmonary embolus) and determine the next appropriate diagnostic procedure.

Patients usually present with signs and symptoms, not a diagnosis. Therefore, write examination questions that replicate the process of clinical problem solving. Questions such as "Which of the following is true about polymyalgia rheumatica?" or worse yet, "Which of the following is not true about polymyalgia rheumatica?" do not elicit clinical thinking. They are a series of true-false statements.

Constructing One-best-answer Questions (Type A)
The stem and question or lead-in statement

The stem should be a clinical vignette, whenever possible, and consist of all or some of the following, in this order:

- Patient's age and gender (include race only if it is important to the question)
- Presenting symptom(s)
- Pertinent history (be sure time sequences are clear)
- Pertinent physical
- Pertinent laboratory findings

The question or lead-in statement at the end of the stem should be clear and answerable without having read the options for the answer.

Examples of good lead-in questions/statements are:

Which of the following is the most likely diagnosis? or The most likely diagnosis is:
Which of the following is the most appropriate next step in treatment?
Which of the following is the most likely explanation for the patient's findings?
Treatment with which of the following could have prevented the patient's condition?
Which of the following put the patient at risk for this condition?

Examples of poor lead-in statements or questions are:

Each of the following statements about_____ is correct except:
Which of the following statements about_____ is correct?

Questions or lead-in statements such as these are imprecise and nearly always contain heterogeneous options (e.g., mixture of diseases, laboratory data, mechanisms of disease, treatments, complications).

The Responses (Options)

The stem should be longer any options, as demonstrated:

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

A. XXXXXXXXXX
B. XXXXXXXXXX
C. XXXXXXXXXX
D. XXXXXXXXXX
E. XXXXXXXXXX

Good responses (options) should:

- Be homogeneous, i.e., responses should be all diagnoses, all tests, or all mechanisms of disease, etc. Do not mix categories of responses ! This is a cardinal sin of item writing!
- Be approximately the same length.
- Have a grammatically correct ending to the lead-in statement or answer to the question.
- Not include "None of the above" and "all of the above."
- Not be tricky or picky questions. The goal is to assess knowledge and its application, not test-taking ability.
- Be alphabetized. Item writers have a tendency for the correct answer to be in one position more frequently than the others (e.g., “B”).

Example #4: Flawed one-best-answer item

A 7-year-old girl is brought to a physician’s office by her mother complaining of chronic abdominal pain, irritability and crankiness. Her mother also hints there may be family problems. Which of the following would be most helpful to aid understanding of this patient’s problem?

A. Elicit further information about the family problems and other potential stressors.
B. Perform a physical examination.
C. Reassure the mother that it is a normal phase her daughter is going through.
D. Refer the girl to a gastroenterologist.
E. Refer parents for marital counseling.

The flaws in this item include:

- The clinical findings in the stem are inadequate
- The stem does not pose a clear question
- One cannot arrive at the correct answer without looking at the options
- The options are heterogeneous
- The distracters (wrong answers) are not of similar length or complexity
- The wording in the stem is unclear. Who is complaining of pain, patient or mother?

Example #5: Better written one-best-answer item

A 20-year-old woman, accompanied by her mother, comes to the emergency department because she has had chest pain for 2 hours. The pain began while she was sitting at home, and was accompanied by palpitations, light-headedness, and difficulty breathing. Four months ago, while at a mall, she experienced the sudden onset of similar symptoms. She has had three similar attacks while shopping, each of which spontaneously resolved after 10 minutes. For the past month she has been afraid to leave the house because she feared recurrence of symptoms. Physical examination, blood glucose, and EKG are normal. Which of the following is the most likely diagnosis?

A. Generalized Anxiety Disorder
B. Hypochondriasis
C. Panic disorder*
D. Simple phobia
E. Social phobia

Constructing Matching Items (Type R)

Matching items more closely resemble actual clinical situations than other multiple-choice-item formats. Consequently they evaluate students' diagnostic and management skills more accurately. Extended matching questions offer the opportunity to write items that cross disciplines because of the larger number of possible options (Examples 7 and 8). For example, an item about causes of altered mental status could have options that include cardiac, pulmonary, metabolic, psychiatric, and neurologic possibilities. This technique counters the tendency of students and faculty to compartmentalize knowledge by specialty.

Extended matching items require a computer-readable answer sheet that has more than the usual five choices (A-E). Some answer sheets allow up to 10 responses (A-J); others allow up to 26 options (A-Z).

Elements of a well-constructed extended matching set

- Theme
- Lead-in statement
- Option list
- At least two item stems.

Theme: Each matching set needs a theme, for example, chest pain, depressive symptoms, or abdominal pain. Identify the theme before writing a matching set.

Lead-in-statement: This statement tells the student the theme of the set and what to do. For example, "For each patient with chest pain, select the most likely diagnosis," or "For each patient with fever, select the most appropriate next diagnostic test."

List of options: Make a list of possible responses for your theme. There must be at least five responses, but the number is limited only by the number of options on the computer scan sheet. Forcing students to choose from longer lists of possible options more closely simulates real clinical situations. The option list should include only one option type (e.g., diagnoses, drugs). Options that cross disciplines make the set more closely resemble a real-life situation.

Item Stems: Clinical vignettes used in matching questions should be similar to those used in the stems of Type A questions. Each vignette in a set should contain the same amount and type of information. Most stems in R-type items are no more than five lines long. A knowledgeable student should be able to determine the correct response by reading the stem, without looking at the list of options.

Example #6: Extended matching item set

Theme: Fatigue

Options:

A. Cushing's disease
B. Acute intermittent porphyria
C. Congestive heart failure
D. Major depression
E. Epstein-Barr virus infection
F. Folate deficiency
G. Dysthymic disorder
H. Hyperthyroidism
I. Hypothyroidism
J. Mitral valve prolapse
K. Lyme disease
L. Bipolar disorder
M. Substance abuse
N. Vitamin B12 deficiency

Lead-in*:* For each patient with fatigue, select the most likely diagnosis.

Stems*:*

1. A 19-year-old woman has had fatigue, fever, and sore throat for 1 week. Her temperature is 38.3^{o}C (101^{o}F). Examination reveals cervical lymphadenopathy and splenomegaly. Leukocyte count is 5000/mm3 (80% lymphocytes, many of which look atypical). Serum aspartate amino-transferase (AST, SGOT) is 200 IU/L. Serum bilirubin concentration and serum alkaline phosphatase are within the reference range. (Answer: E)

2. For the past 2 months, a 50-year-old woman has been fatigued and “lacked energy.” She has gained 15 pounds in the same time interval. Physical examination reveals delayed deep tendon reflexes. (Answer: I)

Example #7: Extended matching item set

Theme*:* Chest Pain

Options*:*

A. Angina pectoris
B. Aortic stenosis
C. Costochondritis
D. Dissecting aortic aneurysm
E. Gastro-esophageal reflux
F. Herpes zoster
G. Mitral valve prolapse
H. Myocardial infarction
I. Panic disorder
J. Pericarditis
K. Pulmonary embolus

Lead-in*:* For each patient with chest pain, select the most likely cause.

Stems*:*

1. For 1 hour, a 72-year-old man has had worsening chest pain that feels like "someone tearing my chest." The pain radiates to his back. Blood pressure is 160/90 mm Hg in the right arm and 105/70 mm Hg in the left arm. A murmur of aortic regurgitation, not previously present, is heard. (Answer: D)

2. For 12 hours, a 28-year-old woman has had anterior chest pain made worse by deep breathing or lying supine. She has had systemic lupus erythematosus for 4 years. There is a friction rub over her left sternal border. (Answer: J)

3. A 48-year-old man has recurrent episodes of burning chest pain located at the level of the lower sternum. The episodes last 15 to 45 minutes and are frequently relieved by antacids. The pain is often precipitated by lying supine or eating a large meal. (Answer: E)

4. A 60-year-old woman has severe burning, left-sided chest pain that radiates from the mid-sternum around the left side of the chest to the back. Touching the skin over the involved area lightly with a Q-tip causes an unpleasant, burning sensation. Lung and cardiovascular examinations are normal. The skin over the affected area appears normal. Electrocardiogram is normal. (Answer: F)

5. A 22-year-old man has had persistent chest pain over, and to the left of, the upper sternum for 3 days. He describes it as a "nuisance ache," aggravated by lifting weights. Blood pressure, and cardiac and lung examinations are normal. The pain is reproduced by pressure at the junction of the upper left thoracic ribs and the sternum. (Answer: C)

Table 6.11.1. Item-writing Recommendations	
Guideline	**Practical Suggestions**
Address important concepts	Follow an examination blueprint
Write clinical vignettes whenever possible	Present information in order: age and gender, history, physical examination, laboratory data
Write a focused question or lead-in statement	Should be able to answer without reading options Ask: "Can question be answered without looking at the options?" e.g., "What is the most likely diagnosis?" or "The most likely cause is:"
Write homogeneous options	All diagnoses, all laboratory tests, all outcomes It is easy to write homogeneous options if the lead-in is focused.
Alphabetize options	This addresses the tendency for a certain option to be correct more often than others
Avoid ambiguous phrases	e.g., may, usually, frequently, rarely How often is usually? Frequently? Rarely?
Avoid absolute phrases	E.g., always, never These are almost always wrong answers
Avoid implausible or inconsistent options	Options should be: Plausible and not deceptive Same length Same perspective (e.g., all positive or all negative)
Avoid overlapping numeric responses	e.g., A. 10-25 B. 20-40
Do not use "all of the above" or "none of the above"	Include more than one point in the same response if necessary e.g., A. Murmur, fever, fatigue rather than A. Murmur B. Fever C. Fatigue D. All of the above
Avoid unnecessarily long or tricky options	Test knowledge, not ability to interpret what item means
All options should be the same relative length	Correct answer is usually the longest
Be sure options are grammatically correct with lead-in	Grammatically incorrect responses are almost always incorrect
Do not repeat a word from the lead-in in the options	Cues the correct answer

Constructing the Examination

As the clerkship director, you are responsible for the quality of the examination even if you do not write most of the items. Ensure that the individual items are high quality and that the overall examination assesses knowledge of the important concepts of the clerkship. Topics that are easy to write examination questions for may not be the most important topics to assess on the examination. It is helpful to follow these steps when constructing an examination:

- Develop an examination blueprint. The blueprint should list the topics to be covered on the examination (e.g., chest pain, fatigue, prevalence, specificity, ectopic pregnancy) and the domain to be assessed (e.g., definition, diagnosis, management, interpretation of data).

- Teach faculty how to write multiple-choice items consistent with the format used on the examination. Even 1-hour item-writing workshops are helpful.

- Ask faculty to submit items that follow the examination blueprint. Many of the items probably will be poorly written and will require considerable editing.

- Select the items for the examination during a test construction meeting where the contributing authors read their items aloud, followed by discussion, questions, improvements, and an accept/reject decision. Faculty are likely to resist the process because they must submit items far enough ahead of the administration date to allow time for review and editing, and because item writing becomes a "public," rather than "private," process. Remind the group the goal is to produce the highest quality items, not to defend one's submissions. The process is worth the effort and produces higher quality examinations.[286, 288]

If this approach is not feasible, edit submitted items and return them to the author for further input or approval. After you are satisfied with the quality of the examination, ask a colleague to review the examination critically prior to its administration. Ask the colleague to read the examination as if through the eyes of a student. Reviewing the examination without the answer key helps to highlight ambiguities that might otherwise be missed.

Summary

Written examinations are an important modality for assessing clinical competence. While many medical schools use externally developed examinations, such as the National Board Subject Tests, some departments develop their own clerkship examination. Examination items should cover important concepts and assess students' ability to apply knowledge to clinical situations or solve a clinical problem. They should not primarily test recall of factual information, and they should minimize the likelihood that test-wise students will be able to answer items correctly without clinical knowledge or problem-solving skills.

Section 12. Setting Standards in Clerkship Examinations

Julia Corcoran, MD

Clerkship Tests

Tests add a great deal to a clerkship. They serve several purposes – a motivator for student study, an assessment of student knowledge, an objective contribution to grade assignment and a tool for evaluating clerkship teaching. Nearly all clinical clerkships rely heavily on faculty appraisals of student performance for grade assignments; most also add a testing component. This section complements Chapter 6, Section 9 which describes available testing methods, and focuses on how *individual* exams are graded. Section 13 extends this discussion to how grades of specific examinations are *combined* with teachers' grades into a summary clerkship grade.

Testing vs. Performance Appraisal

Testing can be more "objective" than performance appraisals in a clinical clerkship. Tests lack the "halo effect" that an energetic, youthful student can generate. They can also sample a larger portion of the knowledge domain, allowing the student to demonstrate a breadth as well as depth.

Performance appraisal by medical faculty, notoriously, have a limited range of responses, tend to be skewed toward the upper end of the scale and often lack specific descriptors or examples of the student's clinical performance. Even when the appraisal form is well constructed (Likert-type scale of seven to nine, behaviorally anchored evaluation points), more than 7 faculty evaluations are necessary to balance out doves, hawks and indifferent appraisers.[33] Furthermore, it is difficult to obtain a critique that can be used for feedback. "The best student I have ever worked with" doesn't help the student constructively any more than the lack of negative critique for the struggling student. Clerkship tests can be used to pick out the good student from the poor student and to offer feedback rarely forthcoming from faculty appraisals of student performance. (for contrasting discussion of teachers' evaluations, see Chapter 6, Section 3)

Reliability and Validity

For tests to serve the purposes noted in the first paragraph they must have reliability and validity. Reliability is a quality of a specific test. Reliable tests perform consistently (well or poorly) within and between groups of students. They correctly identify students who are achieving at, above and below an accepted level. Statistical models can help assess the reliability of tests (alpha), test items (p) and evaluators (inter-rater reliability). These statistical models help to determine how much of the variance in scores is related to student knowledge of the subject vs. test flaws and other unexplained sources. (for definitions of terms, see also Chapter 6, Section 2)

Validity is not a quality of the test score, but of the meaning and inferences drawn from that score. Older concepts of validity concentrated on validity of content (representation of the subject area), construct (is the test proper for the skill i.e. an essay to judge writing) and criterion (does the information correlate with other findings). The current concept of validity is that of a unified hypothesis. Validating a test is the process of collecting empirical data and logical arguments to support that our inferences (grades assigned) are correct.[289] Ultimately, validity is achieved when grades correctly indicate a certain level of skill or knowledge in the clerkship domain. Validating requires correlations and corroborations rather than straightforward statistical analysis. Kerfoot et al. present a flowchart for the iterative process of validating a test for students on a Urology rotation.[290] An example of corroborative evidence for validity of a Pediatrics final examination is provided by a correlation study by Hijazi et al.[291]

Testing Options: What, When, If and How

Tests should be appropriate to the situation, and several examples follow. In a 2-week or 4-week elective clerkship on a subspecialty service with a single student involved, it would be difficult to develop and validate a long multiple choice examination (MCQ). In that setting, a collection of short essays or an oral examination (constructed responses - CR) may be more useful at the end of the rotation (summative). In an 8-12 week rotation, however, it might be quite possible to develop and validate local MCQ exams, administer a National Board of

Medical Examiners Subject Test (NBME Shelf Test), create an objective structured clinical examination (OSCE) with standardized patients or an oral examination with faculty. (see also Chapter 6, Section 10) During longer clerkships, even first-day testing (formative) might also be possible in order to identify students at risk of failing the clerkship final examination. Hemmer et al.[292] and Denton et al.[210] have explored this approach in the Medicine clerkship. Early feedback increases detection, but not necessarily improvement in student performance. Nonetheless, identifying these students and trying to make sure they have the necessary preparation also adds validity to the process.

Table 6.12.1 summarizes the pros and cons of the commonly used test formats.

Planning the Test: Blue Printing

It seems rather obvious that tests should be planned to reflect the course. Documentation of that plan is referred to as the test blueprint. This documentation serves as evidence toward the validity of the test.

To develop a blue print, the faculty must establish what knowledge content is important, which skills are important and which domains are to be examined (e.g. book knowledge, clinical skills, technical skills, communication skill, professionalism), the level of cognition demanded (recall, application, problem solving, etc) and the relative importance of each element. The content of the examination should be linked tightly to the curriculum of the clerkship. For example, asking random Pediatric questions on an Obstetrics and Gynecology examination detracts from the validity and reliability of the Ob-Gyne examination by adding unnecessary variance. Student scores will depend on factors other than knowledge about Ob-Gyn. Blue prints should be developed for all tests, whether MCQ, CR or OSCE.

An example of a blue print for our Third-Year Surgery Clerkship midterm is found below. The percentage of content for each subspecialty was determined by the amount of time and objectives devoted to each within the clerkship. In addition, we surveyed our objectives in each of the subspecialty domains to determine where it fell on the Bloom's taxonomy of mastery of knowledge: recall (remembrance or recitation of facts), application (able to apply knowledge such as interpretation of tables, etc) or problem solving (using knowledge in a novel situation to solve the problem, also called synthesis).[61]

Table 6.12.1. Pros and Cons of Different Testing Formats		
Test Type	Pros	Cons
Locally developed MCQ	-Well accepted format -Can be adapted to the local content of the clerkship -No "direct" purchase cost -Easily keyed and graded -Straight forward to administer	-VERY difficult to produce sufficient number of MCQ's measuring higher order function without professional test writing, editing and vetting -Difficult to keep the question pool from getting out to the students
NBME Subject Test	-Well accepted format -Good psychometric information available -Graded for you (2 week lag) -Constantly changing question pool -Helps to prepare student for NBME Step 2 written exam -Straight forward to administer	-Cost (around $32 per student) -May not reflect the local curriculum
Locally developed Constructed Response – written	-Accepted format -Easier to develop items to measure higher order function -Well suited to subspecialty elective rotations	-Time consuming to grade (recommend typed into to computer to alleviate handwriting issues…) -Difficult to key incorrect aspects of an answer
Locally developed Constructed Response - oral	-Accepted format -Easier to develop items to measure higher order function -Prepares student for ABMS-style oral examinations -Well suited to subspecialty elective rotations	-Hard to corral sufficient faculty to administer for large clerkship -Hard to uniformly grade (inter-rater problems)
Objective Structured Clinical Exercise (OSCE)	-Evolving into an accepted format on the clerkship level -Prepares student for NBME Step 2 Clinical Exam -Can be used in formative and summative testing situations -Can test higher order function	-Requires appropriate facility and expertise to develop questions, train patients, etc -Developing check lists (keys) to favor higher order function can be hard -Expensive to produce – more so for summative than formative testing -Difficult to get the reliability of a MCQ test

Once the blue print is developed, it should be disseminated to the departmental curriculum committee to confirm that it reflects the intentions of the curriculum design. The faculty should review the blueprint for instructional design and item development. MOST importantly, the students should see the blueprint, so they understand the expectations of the clerkship examinations.

Table 6.12.2: Sample Test Blue Print: 60 question Multiple-Choice Midterm Examination in Third-Year Surgery Clerkship Northwestern University Feinberg School of Medicine				
Content	Recall (58%)	App (38%)	Prob Solv (4%)	TOTALS number (%)
	number of questions	number of questions	number of questions	
G.Surgery	9	6	1	16 (27)
ENT	6	4	0	10 (17)
NS	2	1	0	3 (5)
Ophthal	1	2	0	5
Ortho	6	4	0	17
Plastic	1	1	0	3
Urol	5	4	1	17
CT	1	1	0	3
Vascular	2	1	1	6
TOTALS	33	24	3	60 (100)

Giving the blue print to the students is not equivalent to handing them the content of the examination. It does eliminate a good portion of the “what am I thinking” mind game and allows them to concentrate on learning/studying pertinent information. This, in turn, decreases the amount of variance in student scores due to factors surrounding content of the examination and increases the validity of the using the test scores in grading.

Constructing the Test

Once the blue print has been approved, developing items and assembling them into a test is the next task. This administrative work includes setting a time line and assigning topics to the subject matter experts – your faculty and collecting the items. Busy clinicians can be hard to pin down, but persistence pays off. The specifics of how to write well multiple-choice items and develop standardized patient cases are covered elsewhere in this text. This section concentrates on how to make the test happen. Because the NBME subject exam series is well known to CDs, this section strongly reflects my personal experience putting together a locally developed formative multiple-choice test and locally developed summative OSCE.

Concise instructions help the faculty members develop items that fit into the blue print. I have found a letter of intent useful. In it, I outline the subjects to be covered, the format to be used and send a copy of the pertinent course objectives and a similar item in another domain which can be used as a template if necessary. Lead-time is essential. Each year I start almost 6

months in advance and always seem to be working until the last week before test administration. Routine follow-up with electronic correspondence and contact with office assistants is helpful. Visiting the faculty in their office, operating room or clinic can also be important.

Once the items have been submitted, editing begins. The first round of editing includes checking that items match the blue print and course objectives. The verbiage must be clear, without double negatives, double entendre and red herrings. Item flaws such as colloquial language, misspellings, and grammatical errors must be eliminated.[293] Medical transcriptionists make good copy editors in this situation – they have heard the vocabulary but cannot read into the meaning.

The test items should then be put in order of administration and an answer key developed. I - administered the test to the core faculty both to check the items and check the key. I, intuitively, chose not to use residents for maintenance of confidentiality of test contents, but they certainly could help.

These steps for collecting, editing, assembling and keying a test are the same regardless of the test format. Documentation of each of the events also feeds into the validity evidence.

Administration and Security

In order to decrease the variance in scores because of cheating, it must be sought for vigilantly and prevented. Cheating takes many forms – advance copies of examinations, direct copying of answers from an unwitting student or even cooperative cheating among a group of students. The best solution to this problem is prevention.

Once assembled, the paper tests should be kept under lock and key. I prefer to do all work on any given test on the computer “off the network,” so that hacking is less likely. Electronic copies of the test are kept on compact disc under lock and key. Constructed response and standardized patient situations are less likely to be affected than MCQ tests by advance leaks of test content.

When giving pencil and paper examinations, students should be seated in a well-lighted space with ample space between students. Tests should be numbered and counted at the end of the test to assure that no copies leave the room. Students should be informed that “passing questions on” does not help them or future students and that this type of behavior in NBME and ABMS examinations is considered copyright fraud.

Test scores when available should be reviewed promptly. A quick look at the descriptive statistics including the low score, high score, distribution of scores and mean will tell the test administrator a lot about irregularities. Scores should reflect the range of student abilities. Clustering of scores at a certain number, an unusually high number of students achieving high or low grades or a sudden change in the range of scores can indicate irregularities. When such numbers show up, the test should be reviewed for any errors in the answer key, the test itself or its assembly. If all is well with test and key, then the possibility of cheating must be entertained. Students who have performed well earlier in the clerkship should do well toward the end and vice versa.

Misadministration can cause irregularities, which should be documented by the exam proctor. Examples include pairing a wrong key with a test, loss of power during testing with projected or audio stems, loss of a standardized patient or a students' response sheet, loss of data entered into a computer system.

The next step in testing is to record the scores and submit them for inclusion in the grading process. Scores should be reported to the students and to the faculty. Grades, and sometimes scores, should be reported to the Dean's Office. Spreadsheets work quite well in place of an old-fashioned grade book. These sheets should be considered privileged data and kept under lock and key as well. Much like patient information, student information should be considered protected personal data.

Evaluating the Test

Different tests require different evaluations. As mentioned above, a quick glance at the descriptive statistics surrounding any test will give a rough guide as to its performance as an instrument to discriminate between students' achievement levels. The broader distribution curves will accentuate differences between students, as long as the source of that variance is the test itself and not irregularities. Statistical reliability analysis can shed more light on the examination's performance.[294]

Cronbach's alpha or the KR-20 are reliability analyses suitable for multiple choice examinations and larger OSCE examinations. Alpha is a measurement of how well the test assesses a single body of information. The test is tested against itself, internally – essential broken in half and the halves compared with as many combinations of questions as possible. The math is complex but can be done by computer programs such as SPSS, which can be place on personal computers. These stats are often provided by the school's computer center when they score scan-able "bubble sheets".

Acceptable alpha for classroom MCQ tests is in the range of .7 to .8. Alpha for high-stakes examinations such as the USMLE runs around .9. One way to increase the reliability of the test and raise alpha is to increase the number of items on the examination, rather than rewriting the questions themselves. For this reason, alpha for OSCE's tend to be lower, in the range of .6, for a 12-station format.

Another way of considering evaluation of an OSCE is factor analysis. While alpha is used to assess whether a test covers a cohesive unit of knowledge, factor analysis can help determine whether multiple scales (read skills in an OSCE) are being sampled. Through a data-reduction statistical method, factor analysis determines which items act similarly, that is test a single skill (load on a single axis in the statistical parlance). Chesser et al. reported analyzing a high-stakes OSCE with factor analysis to help the faculty determine the pass-fail cut point.[295]

As an example, in our first two years of administering a summative OSCE examination in a multidisciplinary surgery clerkship, it took us a while to build up the check lists and clinical stations. The alpha we calculated was consistantly in the range of .48 to .53. We were concerned about what we were really testing. Factor analysis revealed to us that we had four axes. Looking at how the items loaded onto the 4 axes, we noted that we seemed to have items which grouped around skills in 1. physical examination, 2. clinical problem management, 3. communication skills and 4. motor skills. Knowing these qualities of the items, we felt we could

comment on the students mastery of the skills. Written testing was used to evaluate of knowledge content.

Another test evaluation method is item analysis. The performance of the item is measured grossly by its difficulty “p”, which is the percentage of students who correctly answer the examination. Items with p<.3 are items which students are guessing on randomly. Items with p>.8 are easy and most students know them. Items between these extremes are the items that help discriminate between students. The point biserial (p-bis) is a discrimination index reflecting which students correctly answer the questions, those in the upper, middle or lower third. A p-bis of >.20 indicates that a question discriminates well between high, mid and low performers.

Items that are too easy, too hard or non-discriminating add noise into the variance. If a test is purely to mark achievement, it can be composed of pure mastery questions with a p>.8. But if we are to use the test for assigning ranks (read grades) than discrimination should be sought.

Determining Pass/Fail Cut Off

Several methods to determine pass/fail points are available. Historically norm-referenced pass/fail schema were acceptable – that is, grading “on the curve.” A problem with norm referencing is that no one has made a decision about what should be known and must be known. Criterion-referenced schemas develop an absolute numerical pass-fail point. Different schema produce different cut off points.

When developing long, professional, high-stakes tests these methods are worth the investment of time, energy and money. In the classroom, some simplified combination methods are available – the Hofstee method and the Direct Borderline method (an Angoff variant).

The Hofstee method requires the faculty to choose a maximum percentage correct, that even if everyone achieved this score it would be acceptable for all to pass. The minimum percentage correct, that even if everyone achieved this score, all should fail. The lowest number of failures acceptable and the highest number of failures acceptable are plotted. The cumulative distribution curve is plotted against these points to determine the cut off point. This schema has elements of relative and absolute design and seems well understood by faculty and students. Cusimano and Rothman applied this technique and compared to other schema on a fourth-year summative OSCE.[296] They found that it gave consistent and acceptable statistical and real data.

Downing compares the Direct Borderline method to the Nedelsky, Hofstee and Ebel methods in establishing pass/fail cut offs for two basic science MCQ examinations and finds that it performs favorably compared to the more established methods.[297] In the Direct Borderline method the clerkship director/classroom instructor ask the question for each item whether the borderline failing/passing student would get the question correct – yes or no – add the ayes for cut off score. Unlike the Angoff method, probabilities and panels are not used making the Direct Borderline method achievable for the course director. Because the direct borderline method relies on a single or small number of faculty, it is particularly useful for a clerkship examination. At the same time that the faculty reviews the test and checks the answer key, the question can be rated for the pass-fail point of a borderline student.

Item Banking

Item banking refers to keeping track of test items in an orderly format. After going to the trouble of developing high quality test items, keeping track of them only makes sense. In addition to storing the item itself, statistical information about the items performance (p and p bis), when the item was used (clerkship, quarter, year) and qualities of the question (subject matter, level of knowledge mastery, type of question: MCQ, CR, OSCE). Item banking can be as simple as an Excel spreadsheet or complicated database set up by the school's department of education and computer services. (I have the former…)

There are several benefits to this kind of organization of data about test items. By examining the performance of an item over time, outdated items can be removed and poorly performing items can be culled as better items are developed, increasing the overall quality of your test items in the bank.

One can create similar examinations that are not identical examinations. Using a test blue print and an item bank, one can pull different questions of similar difficulty and subject matter but which haven't been used in the past 2 clerkships. This type of rotation of items increases the security of the test as no two tests are identical each clerkship but allows the faculty to give tests which can be documented equivalent.

If multiple clerkship directors, regionally or nationally, keep data about items in a similar manner, items can be traded between institutions so that your faculty need not come up with fresh items every year.

Summary

Objective testing is an important component in assigning grades to clinical clerks. Attention to detail of the test content, test format and test administration can add validity evidence. Evaluation of the test with the goal of continual improvement can decrease random variance, increasing reliability and adding validity evidence as well. The inferences we can draw in order to grade our students can only be as good as the tests we base them upon.

Additional resources:

- Gronlund, NE. Assessment of Student Achievement. Seventh edition. Allyn and Bacon, Boston, 2003
- Haladyna, TM. Developing and Validating Multiple-Choice Test Items. Second edition. Lawrence Erlbaum Associates, Publishers, Mahwah, New Jersey, 1999

Section 13. Converting Evaluations into Grades

Michael Battistone, MD

Introduction

The clinical performance of students on clerkship rotations involves a complex constellation of attributes and behaviors. These attributes and behaviors are not physical properties (e.g., height, weight) that can be measured directly. Rather, clinical performance can be considered a psychological construct, or theoretical concept, that provides a framework for interpreting and discussing the meaning of observed performance within a certain context. Another example of a psychological construct is "emotional intelligence." This is difficult to measure, yet is accepted in popular culture and recognized in academic settings. Clinical performance is a construct in which knowledge, skill, attitudes, and values can be discussed; these latter terms denote specific domains - frames of reference that serve to focus observations and guide reflection.

In considering the issues involved in deriving grades from evaluations, it is important to reiterate the distinction between evaluation and grading (Chapter 6, Section 2). This section discusses issues involved in converting assessments and evaluations into final grades, and offers examples for of how to do the conversion. Evaluation of the performance of an individual (student, teacher) or program (course, curriculum) involves critical reflection on observations made during the period of evaluation — reflection that considers these observations within the framework of the expectations of the evaluators (preferably that of the program). Evaluations of student performance are strengthened when specific observations are reported in detail and interpreted in the context of course requirements.

Grading, on the other hand, involves the assignment of rank to a set of evaluations; often the performances are classified across ordinal categories (Honors, High Pass, Pass, Low Pass, Fail or A-B-C-D-F), or expressed quantitatively as points or percentages.

There are many reasonable methods by which evaluations may be converted into grades, and each institution and clerkship may have needs that require unique solutions. To assist those responsible for developing solutions in the form of policy, some common issues are identified below and discussed dialectally ("point and counterpoint"). Specifically, the topics to be addressed are 1) whether those who evaluate students should also grade them, 2) whether evaluations of performances later in a clerkship or later in the clinical year should count more toward a student's final grade, and 3), whether normative or criterion-based (fixed standard) approaches should be used in determining honors grades. Examples and descriptions of how these issues are being addressed at one institution (University of Utah School of Medicine) are included. Finally, there is a review of several approaches used in computing the final grade, and a brief discussion of some of the issues involved in grades that may be considered "Low Pass" or "Fail."

Links Between Evaluators and Graders—Should Teachers Grade?

Question for clerkship directors: should teachers grade or simply evaluate?

Point: Those who evaluate should be those who grade.
At many institutions, evaluators (faculty, and often housestaff) are expected to submit both evaluations and grades, which are often documented in a single instrument. (see also Chapter 6, Section 3) There is "face validity" to this method (i.e., it seems valid), and though students

may dispute a grade, their complaints often appeal to issues of communication or observation (e.g., unclear expectations or infrequent or inadequate sampling of performance), rather than to an argument that this approach is fundamentally unfair. Similarly, faculty and residents have traditionally accepted the responsibility to grade students who have been assigned to work with them on clinical rotations. If this model is chosen, the clerkship director and departmental education committee will determine the percentage of each student's grade to be allocated to each level of teacher (resident, attending), and the final grade will primarily be a calculation of input from the graders.

Counterpoint: Those who evaluate should not be expected to grade.
The rationale for this approach is based on the view that there is some aspect in the relationship between the trainee and the evaluator that introduces bias, and thereby poses a substantial threat to validity.[67] A recent study found that 60% of the variance in student performance ratings was attributable to the interaction between rater and student, and the dependability of individual ratings was very low (4%).[298]

In deciding how to incorporate the recommendations of individual teachers into a final summative grade, the clerkship director must weigh the experience of the teacher, the kind and amount of contact with the student, and how much confidence there is in the individual assessments themselves. This last factor involves dealing with variance in teacher behaviors, and can be difficult to quantify. Some faculty may feel pressure to "inflate" grades, particularly in settings where they are giving feedback to students directly ("face-to-face").[299] Although not all evaluators succumb to the pressures to inflate grades, there is variance across the range of raters. On the other hand, some instructors may defend their practice of rarely, if ever, awarding a perfect score with the comment, "no medical student is perfect"; others may typically give high marks because they are afraid of discouraging the student if they do otherwise. Under these conditions, grades tell us less about the performance of the individual student than they do about the operational constructs and philosophy of the rater. This is a very serious problem, and must be addressed if the validity of evaluations is to be preserved.

We recommend two approaches to consider in a system in which the evaluators also grade: 1) to include a systematic and ongoing method of training raters (ideally in "real time") to use your evaluation and grading techniques, and 2) to develop mechanisms to detect, identify, and modulate the extent of rater leniency or severity. With the growth of electronic and online evaluation, some computer programs may assist in this task, and it may be possible to adapt currently available software for use in your existing evaluation method.

Some clerkships have addressed these threats to validity by separating the processes of evaluation from grading. In these systems, evaluators are instructed to focus their efforts on producing detailed written documentation of their observations of student performance. All student evaluations are then reviewed by a grading committee that may include the clerkship director. This method seeks to protect the evaluators from the pressures of grade inflation, and to preserve the ability of the clerkship director to serve as student advocate.

Example of one method and discussion:

Regardless of whether evaluators are charged with the responsibility of grading, the links between evaluation and grading should be strong, and the process of converting evaluations to grades should be systematized under the auspices of the department; in some schools the dean's office provides guidelines to insure inter-departmental equivalence. Resources should be directed to ensure adequate periods of observation of students by faculty and residents, and

to the timely collection of valid evaluations. The clerkship director may want to identify a minimum time of exposure (e.g., two weeks) that is required before a rater can be expected to submit an evaluation. However, the length of the assigned tour of duty may not be the most significant factor in determining whether a given evaluation is credible. Valid evaluations of clinical performance are predicated on valid observations of clinical performance. There may be substantial variance across raters in regards to their efficiency and effectiveness in observing their students—it may be that one rater can submit a credible evaluation after a period of only one week, while another rater may not be able to do this despite having been assigned 3 weeks with the team. If there have been several changes in evaluators over a given period of observation (for example, a change in supervising residents), it may be reasonable for the raters to submit a single evaluation that represents their shared opinion. Since students may eventually contest how grades are calculated, we recommend that the clerkship Handbook should only indicate the consideration in grade calculation, and that the department will allocate grading input based on these (see Chapter 6, Section 14).

When teachers do "grade" it is especially important to guide their observations. As a method for achieving consistent input from teachers, the Internal Medicine clerkship at the University of Utah uses formal evaluation sessions, which are held every 3 weeks during inpatient rotations (see also Chapter 6, Section 3). These meetings are moderated by the clerkship director (or an assistant clerkship director) and are attended by ward residents (postgraduate year-2 and PGY-3) and the faculty on ward service at the time. At Utah interns are not required to attend the sessions or to submit evaluations; their impressions are communicated through their supervising resident. Each evaluator is asked to verbally describe their students' performance using the Reporter-Interpreter-Manager-Educator (R-I-M-E) framework,[20] and to provide specific observations to substantiate their rating. Evaluators are directed to conclude their critiques by identifying a specific "next step" to which students' future efforts should be directed. The vocabulary terms identified and discussed at the evaluation session are converted to numerical ratings; this is discussed in more detail in the section below.

Immediately after the evaluation session, the students meet individually with the clerkship director to review the written evaluations, as well as the verbal comments. Each student appointment is scheduled for fifteen minutes, and strategies for achieving the prescribed "next step" are discussed.

Considering Student Progress Within the Course—Should Later Evaluations Count More?

Question for clerkship directors: should later evaluations count more?

Point: The final grade of clinical performance should be one in which each individual evaluation (grade) is weighted equally.
Each observer has something to offer in the evaluation of a student. Not counting each evaluation toward the final grade risks diminishing the importance of each individual period of observation, in the eyes of both the students and the instructors. In addition, the dependability of the final evaluation increases with the number of ratings—the more evaluations included, the more dependable the grade will be.

Counterpoint: The final grade should emphasize student performance at the end of the clerkship.

In many cases, student performance is seen to improve over the course of a clinical rotation. It may be that the trajectory of student growth (the learning curve) may be more predictive of future performance than any individual point on the line. Students who start strong, but do not progress, may be at greater risk for future marginal performance than students who struggle initially, then "get it" and demonstrate rapid improvement. Certainly students who demonstrate declining performance raise concern. If each individual evaluation is graded and given equal weight, valuable information about student development may not be captured and reflected in the final grade.

Examples and discussion

This topic actually includes two separate issues; the slope of the learning curve, and the final level of performance. In addressing the first, it is important to know whether the specific evaluation system being used is responsive to student growth. In comparing a global numeric rating system with the R-I-M-E descriptive vocabulary,[133] we found that the descriptive method was superior to the numeric system in demonstrating student progress through the nine-week inpatient portion of the clerkship. If any given evaluation method is relatively unresponsive to student growth, it would lack construct validity for the premise that students do in fact improve, this issue would be moot, and it would be important to include as many ratings as possible (assuming valid methods of observation and evaluation) so as to ensure adequate dependability.

In addressing the final level of performance that the student achieves, the clerkship director must also consider whether the same grading standard should be applied uniformly across the year. Data suggest that, particularly for the NBME examinations, students who are assessed later in the clinical year perform better than students who are tested earlier.[265, 300] However, this effect may not hold for all forms of assessment; a 2000 report describing an OSCE with case content linked to learning objectives in an ambulatory care clerkship did not demonstrate an effect linked to student maturation.[301]

Considering Student Progress Within the Academic Year — Should Later Grades be Adjusted for Time of Year?

Question for clerkship directors: should the same grading standard for clinical performance evaluations apply in the first half of third year as in the second half?

Dynamic grading criteria (a system in which adjustments are made over the course of the year) require a different approach than static criteria. For example, at the University of Utah, in converting the R-I-M-E descriptors to numeric grades adjustments are made to account for the level of experience the student has had in the clerkship year to that point (Table 6.13.1). By the end of our clerkship rotation, a student's performance must consistently be rated at the level of Reporter in order to receive a grade of "Pass." Since we do not currently award grades of "Low Pass" or "Marginal Pass," there is no numerical equivalent for sub-Reporter work (identified as the "Observer" rank in our system) for the final period of evaluation. If performance in the final three weeks (which is always taken in the second semester) is judged to be at the Reporter level, a grade of 2.5 (out of a possible 4.0) is given. If the same performance is observed in the first semester, the score is slightly higher (2.75). As shown in the Table, similar adjustments are made for each of the descriptors. For a grade of "Honors," a rating of Manager is required. This converts to 3.5 (the numeric criterion for Honors) for the second semester ("meets expectations for Honors"), and 3.75 in the first semester ("exceeds expectations for Honors").

Table 6.13.1				
Descriptor	**NUMERIC EQUIVALENT (4.0 POSSIBLE POINTS)**			
	First Semester		Second Semester	
Observer	0	“PASS”	0	“FAIL”
Observer/Reporter	2.0		0	
Reporter	2.75		2.5	“PASS”
Reporter/Interpreter	3.0	“HIGH PASS”	2.75	
Interpreter	3.25		3.0	“HIGH PASS”
Interpreter/Manager	3.5	“HONORS”	3.25	
Manager	3.75		3.5	“HONORS”
Manager/Educator	4.0		3.75	
Educator	4.0		4.0	

Table 6.13.1 illustrates how a criterion-based (fixed standard) framework can still be used to equilibrate grades based on the time of the academic year; in other words, the standard is fixed within each half of the year. The table lists the numeric equivalents for a given descriptive rating for the first and second semesters of the Internal Medicine Clerkship. Intermediate steps between descriptors (e.g., Reporter/Interpreter) are used when student performance is deemed to be in transition between vocabulary terms.

Normative vs. Criterion-based (fixed standard) Approaches

Question: should clerkship directors compare students to each other or to fixed standards?

Point: Honors grades should reflect the top x % of medical student performances, for any given year, irrespective of how strong their performances are (normative)
This approach is often favored by certain “consumers” of the evaluation and grading process — the deans who will write summary letters and the residency program directors who review them. Normative grading indicates where an individual student ranks in relation to their peers, although the confidence one places in this information should not exceed their confidence in the validity of the methods used to derive the ranking. (See Chapter 6, Section 2, for definitions.)

This method may also be favored by students if the evaluation and grading schemes are strict. If the top 20% of the class will receive Honors regardless of the individual score, this may include more students than under a criterion-based system.

Counterpoint: Honors grades should indicate the number of students who meet the stated requirements for Honors (criterion-based, fixed standard).
Students are more likely to endorse this practice when the criteria for Honors are believed to be attainable. It promotes an atmosphere of collaboration instead of competition, and may be more motivating. Criterion-based systems focus attention on course goals, and facilitate teachers and course directors in working as student advocates. If the criteria and standards are too “easy” and do not challenge the students sufficiently, or if student performance relative to the criteria is not rigorously assessed, there is a risk that so many Honors ratings may be distributed that the distinction of this grade is lost.

Many schools use nationally-normed subject examinations from the National Board of Medical Examiners (“shelf” exams). Although these tests are scored and reported normatively, their

contribution to the individual student's grade often is criterion-referenced. Here are some examples of policies based on this approach:

> "In order to be eligible for an "Honors" rating in the clerkship, the student's score on the NBME pediatrics subject exam must be in the top quartile."
>
> "Students whose score on the NBME psychiatry subject exam is within the bottom decile will be considered to have failed the examination, and will receive a grade of "Incomplete" for this clerkship. Their grade will remain Incomplete until they have repeated the exam and scored above the bottom decile."

Example and Discussion:

Grading systems may combine normative and criterion-based approaches, seeking to capture the strengths of each. In the Internal Medicine clerkship at the University of Utah, we prefer criterion-referenced methods in grading individual components of the overall grade (e.g., clinical performance evaluations, the NBME medicine subject examination, and formal case presentation). This provides a context for evaluators to assess and describe student performance relative to department goals, facilitates meaningful feedback to students, and primes the process of strategic planning with students as they consider how they will work toward achieving the prescribed "Next Step".

At the University of Utah, a student's final grade is computed as follows:

Clinical evaluations:	70%
OSCE :	10%
NBME medicine subject exam:	20%
Formal Case Presentation	Pass/Fail
TOTAL	100%

The clinical performance evaluations are provided by faculty and residents who have worked with the student during the evaluation period, as described above. R-I-M-E descriptors are generated in formal evaluation sessions and transformed to the numerical scale shown in Table 6.13.1. We use a normative approach in retrospective (end-of-year) evaluation of the clerkship itself. We consider the percentage of students receiving Honors (for each component of the course, as well as for the final clerkship grade) to be a useful and meaningful measure of the robustness of our combination of evaluation methods. We have confidence in awarding Honors for an overall grade if 20-30% of students each year achieve this, as defined by their performance across the array of criterion-referenced assessment tools. If less than 20% receive Honors, we retroactively adjust to a normative approach to identify additional students so that no less than 20% of the class will receive Honors (though this will also trigger a comprehensive course evaluation process). If more than 30% receive Honors, methods of observation, evaluation, and grading are reviewed, and new assessment tools and techniques are considered for the following year.

ACE Groups' Use of Methods and Examinations

Clerkship directors may find some guidance from colleagues in other clerkship groups in the Alliance for Clinical Education (ACE) for how to calculate grades from data provided by clerkship directors' groups.

In April 1996, the Clerkship Directors in Internal Medicine (CDIM) Evaluation Task Force conducted a survey of internal medicine clerkship directors as to the methods of evaluation that were being used in their schools. These findings (Table 6.13.2) were presented later that year at CDIM's 7th Annual National Meeting.

Table 6.13.2. Internal Medicine Clerkship Assessment Methods

Evaluation Method	% of Schools	Mean % of Grade (SD)	Range (%)
Teachers' Evaluations	99	63 (20.5)	10-100
NBME Subject Exam	84	22 (11.8)	0-50
Analytic Examination (Free response)	34	20 (11.8)	5-50
OSCE	21	19 (78.5)	0-33

In a 1999 follow-up survey, the CDIM Evaluation and Research Committee queried clerkship directors as to the types of quantifiable examinations they used, their satisfaction with each approach, and the weight they gave each method when calculating the final grade (Table 6.13.3). The response rate for this survey was 89%, and show that most clerkship directors used the NBME medicine subject exam and stipulated a minimum score. Satisfaction was measured using a 5-point modified Likert-scale (1 = very satisfied, 2 = moderately satisfied, 3 satisfied, 4 = moderately dissatisfied, 5 = very dissatisfied). These data were presented at CDIM's 10th Annual Meeting.

Table 6.13.3. Internal Medicine Clerkship Final Examination Methods

	NBME	Faculty Written Exam	OSCE
% of Clerkships using this method (n)	83 (90)	27 (29)	28 (30)
% of Clerkships requiring minimum score to pass	80 (72)	66 (19)	63 (19)
Minimum score required to pass (mean +/- SD)	59 (2.21)	66 (15.3)	69
% of the Final Grade (mean (+/- SD, range))	24 (10.3, 0-50)	22 (10.2, 7.5-50)	15 (10.5, 0-33)
Clerkship Director Satisfaction with Method	2.1	2.0	1.9

In addition to CDIM, national organizations of directors of other clerkships have studied this issue. Table 6.13.4 presents the results of a 2003 survey of the Association of Directors of Medical Student Education in Psychiatry (ADMSEP) which sought to identify the primary methods of student assessment in psychiatry clerkship programs at 141 accredited U.S. and Canadian allopathic medical schools.[302]

Table 6.13.4. Psychiatry Clerkship Grading Methods		
Evaluation Method	% OF SCHOOLS	MEAN % OF GRADE (RANGE %)
Teachers' Evaluations	99	54 (10-100)
NBME Subject Exam	75	31 (0-100)
Departmental Exam	37	22 (0-50)

The duration of a clerkship may influence the selection of evaluation methods. A recent survey of 150 U.S. and Canadian clerkship directors in psychiatry (Table 6.13.5) showed that shorter clerkships were more likely to use the NBME subject test, and less likely to incorporate OSCEs or oral examinations.[303]

Table 6.13.5. Psychiatry Clerkship Grading Methods in Relation to Clerkship Length			
Grading Method (% using)	**Clerkship Length**		
	Four-weeks	Six-weeks	Eight-weeks
NBME	77	74	83
OSCEs	14	20	14
Direct observation	14	29	10
Oral examination	5	24	42
Logbooks	14	18	29

Finally, the use of the NBME subject examination has also been studied in the context of the surgical clerkship. The results of a survey of surgery clerkship directors were presented at the 2004 Annual Meeting of the Association for Surgical Education,[57] they are included in Table 6.13.6, with comparative data from the medicine and psychiatry surveys.

Table 6.13.6. Use of NBME Subject Exam Across Clerkships			
	Surgery	**Medicine**	**Psychiatry**
Prevalance of use of NBME test (%)	91	83	75
Prevalance of minimum score requirement (%)	88	80	N/A
Minimum passing score (mean, raw)	58	59	59
Mean Contribution to Final Grade (%)	34	24	31

Failing Grades

The criteria for failing a component of the clerkship or the clerkship in its entirety should be clearly stated (see Chapter 6, Section 14). This is unambiguous when the criteria can be quantified (e.g., a score on the NBME subject examination less than or equal to 58, a cumulative score from clinical performance evaluations less than 50% of the total possible points, etc.). In addition, careful descriptive evaluations (e.g., "This student was consistently unprepared, unreliable, and interacted poorly with patients and with the other members of the team.") can serve to trigger administrative action that can lead to a summary judgment of failure

by an appropriate supervisory body (e.g., a departmental committee), even without requiring formal calculation of a grade.

Conclusion

Key issues in converting evaluations to grades include deciding who should evaluate and who should grade, how to account for student progress within the course, and what role normative or criterion-referenced approaches should play.

The assessment, evaluation, and grading of performance has been described by Pangaro in the context of the classical sequence of “observation-reflection-action” (see Chapter 6, Section 2). The process of converting evaluations into grades is the process of transitioning from reflection into action. Each institution and clerkship will develop individualized methods that best serve their specific needs, though assuring that good observations of students are made by faculty, residents, or any other evaluator, is essential. Valid grades can only be derived from valid evaluations; valid evaluations must be based on valid observations. Ensuring that valid observations are made is perhaps the most important factor, and also the greatest challenge.

Section 14. Legal Aspects of Failing Grades

Thomas Jamieson, MD, JD, Paul Hemmer MD, MPH , and Louis Pangaro, MD

The challenge of formally evaluating learners is among the most important and solemn responsibilities of both medical school faculty and institutional departments. Medical schools and their faculty are accountable to society and the most important grading decisions for clerkship directors are at the pass-fail threshold. Certainly, distinctions between higher grades may have consequences for students and those program directors who select interns. Nevertheless, society at large is chiefly interested in those departmental grading decisions that lead to a medical school's promotion (or progress) committee's consideration and review of student performance, possibly leading to dismissal (or “attrition”) from medical school. Medical school applications are declining in number in federal, state, and private institutions placing perhaps an even greater burden on clerkship directors and clinical faculty of identifying those who are not ready for the level of independence expected of a new graduate.[304]

The reluctance of evaluators to judge a student as deficient "on the record" inevitably creates a tension with an institution's duty to maintain acceptable standards in its graduates and, ultimately, to protect the public.[100] While students may argue and threaten legal action if they are failed in clerkship rotations, to date there is little legal precedent to support them. While adequate notice and hearing requirements must be met, and the process must be fair and reasonable, the courts, based on United States Supreme Court holdings, have generally avoided imposing judicial trappings on medical schools or other institutions of higher learning.[103]

The purpose of this section is to review principles for giving failing grades, with an emphasis on legal guidelines. The general principles of substantive and procedural due process may also apply to distinctions between higher grades, but generally, these have not been tested in the courts.

Background

Legal Concepts

A well worn aphorism of the American legal system is that "ignorance of the law is no defense". Whether legally informed or not, medical educators owe particular legal obligations, mainly constitutional protections, to their students and house officers in training. The Fourteenth Amendment expressly states, "… (no state) shall deprive any person of life, liberty, or property without due process of law; nor deny to any person within its jurisdiction the equal protection of the law…"[305] The United States Supreme Court has spoken clearly on the extent of the basic constitutional rights of learners enrolled in public institutions and has specifically articulated a standard for medical learners of procedural and substantive due process that we will now address.

Procedural and substantive due process

“Due process of law” has two distinct forms. "Procedural due process" is a constitutional guarantee of procedural fairness, which, at a minimum, entitles a party whose rights are to be affected to be notified, and, in some circumstances, to be heard. "Substantive due process" is a constitutional guarantee of protection from arbitrary and unreasonable action. Perhaps it is convenient to think of due process generally in terms of "why" measures are being taken against a student (substantive due process) and "how" those measures are being imposed (procedural due process). Clerkship directors need to be familiar with two landmark Supreme Court cases that go to the core of due process and should shape a department's approach to giving failing grades.

In Regents of the University of Michigan v. Ewing (1985),[306] the United States Supreme Court accepted a medical school litigant's invitation to "assume the existence of a constitutionally protected property right in a (medical student's) continued enrollment".[307] The case was brought by a former student who was dismissed after failing Part One of the National Board of Medical Examiners (now known as the United States Medical Licensing Examination), and was who not permitted a retake of the examination. The medical school's own promotional pamphlet stated:

> "Everything possible is done to keep qualified medical students in the medical school. This even extends to taking and passing National Board Exams. Should a student fail either part of the National Boards, an opportunity is provided to make up the failure in a second exam."[306]

In Michigan v. Ewing, the Supreme Court held that while denial of an opportunity to retake the exam (a historic precedent for an institution that had permitted 39 prior examination failures a retake) "may constitute evidence of arbitrariness" it was not probative (i.e., “diagnostic”) in itself but only a single fact, among other facts, for the jury to consider. In supporting the school in their final ruling, the Supreme Court noted that the medical school considered the student's *entire* record in reaching the decision to dismiss, not merely the results of a single examination.[306]

One implication of the Michigan v. Ewing ruling is that clerkship directors and their departments are entitled to consider the entirety of a student’s clerkship record in making a grading decision, and as importantly, should *state explicitly* that they have done so. Furthermore, the Supreme Court held that reasoned decision-making in academic matters is, per se, not arbitrary and capricious. Moreover, the burden of proof is on the student to show that a decision was not factually based and was made irrationally. Unless there is evidence that a decision was arbitrary or capricious, courts will not overturn faculty decisions. We strongly recommend that

departments have a committee to review all potentially failing grading decisions, and that all available sources of information about a student - the entirety of the student's clerkship record - be considered.

The ruling in the Ewing case reaffirmed the Supreme Court's prior conclusions in University of Missouri v. Horowitz (1978) that enrollment in medical school is a basis for invoking constitutional protections of due process, and they also reinforced the principle of judicial non-interference in academically-based decisions.[308] In Missouri v. Horowitz, the student was dismissed from medical school in her final year, after receiving negative evaluations and unsatisfactory clerkship grades in multiple clerkships during her third and fourth clerkship rotations. The medical school's Council on Evaluation (progress or promotions committee) had reviewed the student's performance throughout the final two years of medical school and had taken action to place the student on academic probation and notice prior to the decision to dismiss the student. The Court held that the medical school had certainly met, and even exceeded, the Constitutional requirements of the 14th amendment for procedural due process.

Importantly, Missouri v. Horowitz held that procedural due process requirements owed students for *academic* dismissals were less than those for *disciplinary* dismissals. In other words, for matters that are considered by an institution to be academic, which includes professional behavior, the burden of proof is on the student to show that due process has been violated. Whereas, if the institution considers that a student's problem is disciplinary rather than academic, then the burden of proof is on the institution. We recommend, therefore, that performance problems such as multiple, unexcused absences be considered academic, even if the student can fulfill his/her duties on those days when present. Saying that the student is "competent" but has a disciplinary problem, might cause confusion in the presumption of judicial non-interference in the medical school's judgment about what constitutes acceptable academic performance.

Indeed, student -perceived inequities alleging that testing and grading policies are unfair, or an allegation that dismissal is premised on an instructor's incompetence, face a daunting legal standard. While an egregious case of institutional ineptitude is conceivable, courts repeatedly offer a presumption of legitimacy to professional schools' decisions as to students' academic fitness. Further, courts appear to regard the grading of in-house examinations as a matter of academic discretion and typically do not apply or impose a formalistic standard. Strong policy considerations militate against the intervention of courts in controversies relating to an educational institution's judgment of a student's academic performance.[306] However, if a student can demonstrate that a usual grading process was truncated, or ignored, a court may hold that an institution acted arbitrarily, in other words, that due process was denied.

Process: how much is due?

Courts have not attached extensive procedural requirements to failing and/or dismissing a medical student. Students may request, or even demand, such procedural accommodations as an open hearing, the presence of legal counsel, recorded proceedings, or at least a written record of Promotions (Progress) committee deliberations. However, such procedural amenities have not been held due in court decisions. In fact, courts have consistently been leery of the "undue judicialization" of an administrative hearing in the academic environment viewing this as largesse, a burden on institutions, and an improper allocation of resources.[309] There may be exceptions, of course, if an institution has codified its own policy of procedural measures, or in the uncommon circumstance where a student is also facing criminal charges stemming from the incident in question.[310] Perhaps perspective and summative clarity to the issue may be found in

the words of U.S. Supreme Court Justice Byron White when he noted that in an academic dismissal "the Due Process Clause requires, 'not an elaborate hearing before a neutral party,' but simply 'an informal give-and-take between student and disciplinarian' which gives the student 'an opportunity to explain his version of the facts'".[310]

Breach of contract

Although lacking the formality of an expressed written contract, the "promise" between an institution and a student may be enforced under the theory of implied contract. The provisions expressed or implied between the institution and the student derive mainly from the official documents promulgated by the school. Courts may carefully examine student and faculty handbooks, policy statements, rules and regulations. Please recognize that courts may bind you to what you write down in your handbook. Also, importantly, an institution may be bound by the procedures afforded students in previous cases, even if the procedures are unwritten and even if the facts of the cases are somewhat dissimilar.[307]

For their part, students agree to follow an institution's rules and risk penalty - including suspension or dismissal - if they do not. Clerkships should be explicit that they expect students to read and follow policies and procedures that are written in a student's guidebook (whether provided in paper copy, or on a departmental web site). The institution, in exchange for the student's tuition payment, implicitly agrees to provide the academic programs and support services reasonably necessary for students to perform successfully. Although the traditional judicial deference to academic decisions persists, courts appear increasingly willing to hear cases involving aggrieved students under “contract”, a legal theory which may be applied at both public and at private institutions. The emergence of contract as a cause-of-action serves to underscore the need for departments and clerkship directors to periodically audit written materials which outline policies, rules, and regulations and to actually do what they say they will do. In short, every administrator of an academic program must "read the document" of promulgated policies. The occasional failed student who prevails against an institution may do so by exposing the school for breeching its own procedural guarantees.[311]

The Future? Shifting legal strategies by medical students

In recent years litigation by some aggrieved students has broadened the basis for bringing a legal action against an educational institution. The time-honored traditional cause of action, a denial of basic constitutional due process rights, remains oft litigated but may now be only one of several legal theories advanced by a failed student as plaintiff.[312] Defamation, educational malpractice, denial of statutorily conferred rights (such as the Americans with Disabilities Act, and Title VII, IX violations) and breach of contract are some of the legal theories broadening the list of allegations which may be directed at medical schools. (For further discussion please see Chapter 10: Working with Students with Difficulties: Academic and Nonacademic)

Whether actions based on these, or other legal theories will soften the judicial deference afforded medical schools is unclear but this does appear to be a developing area of law. Certainly clerkship directors need to be aware of the grading policies articulated in their own course books, handbooks, or other course guidelines.

Assigning a grade

A failing student may insist that a single evaluator unduly influenced a grade outcome. In settings where a faculty member (or houseofficer) is acting as the university's proxy, the faculty member fulfills one of the functions involved in an academe's "four essential freedoms." Universities have the "freedom" to decide "who may teach, what may be taught, how it shall be taught, and who may be admitted to study."[313] Because grading is pedagogic, the assignment of the grade is subsumed under the university's freedom to determine how a course is to be taught.[314] Therefore, a student should understand that individual faculty members may recommend a grade, but the responsibility for final grade determination rightfully (and legally) rests with the university (and, by proxy, the clinical department). Furthermore, faculty may need to be advised that they do not have a First Amendment right to expression via a school's grade assignment procedures.[315] At the Uniformed Services University of the Health Sciences we inform and summarize the issue in our third year internal medicine clerkship handbook by noting,

> "We guide the faculty and house staff in assessing how well you have met clerkship goals. Their role is evaluation; final responsibility for grading rests with the Department."[316]

Faculty are often wary of recommending low grades, and their reasons include a fear of litigation.[100] They can be encouraged to be honest in their evaluations - though not actually protected against litigation being instigated - by reminding them that departmental process is responsible for assigning grades, and that their own recommendation is simply that, a recommendation. Moreover, if is the "entirety" of the record[306] which will be used by the department to determine the final grade, then the observation of individual faculty member is very unlikely to be the basis of the final decision, unless the comments clearly documented an obviously egregious breach.

Recent years have seen an expanding role for electronic communications in the evaluative process of students. Faculty must be aware that e-mail entries are not confidential conversations. In fact, just as every other entry in a student's file, e-mail is viewed as a "written document" and, therefore, e-mails are "discoverable" in a lawsuit. "Discovery" is a pre-trial device whereby one party can obtain facts and information about a case from the other party to assist in trial preparation. Importantly, institutions should not foster a culture of editorialized informal discussion in evaluative e-mails but rather one of passing of honest and properly worded responses.

Recommendations

The preceding discussion has outlined that while courts have generally deferred to institutions of higher learning in their reasoned judgments about the fitness of students to continue in their training, it is also clear that students are owed both substantive and procedural due process. As clerkship directors, this impacts our grading policies and procedures, which should be evident to students, teachers, and the institution. We provide some general guidance below:

Institutional policies:
General policies and procedures at the level of the institution are often implicit in the curriculum in the first two years. For instance, students are instructed about the ethical behavior and professionalism expected in relationships with patients in their very first year of school. Policies on cheating and plagiarism alike are, likewise, introduced early in the curriculum. These need

not be repeated in detail in orienting third-year students, but we do recommend a general statement during each orientation, and in the student's handbook, that ethical and professional behavior are expected.

Consistency in Orientation:
The clerkships' handbook becomes an important tool whereby students are informed about the processes to be followed in their grading. Both the content of these handbooks and the orientations that students attend at the start of each rotation (or, if relevant, at each clerkship site) constitute part of the due process to which students are entitled. (Please see Chapter 10: Working with Students with Difficulties: Academic and Nonacademic and Chapter 16: The Clerkship Orientation). We recommend that all orientation materials, including handbooks, be reviewed and endorsed by a departmental committee, that includes the Department Chair, and also agreement by the on-site clerkship directors who may be responsible for the dissemination.

Clerkship-Specific Expectations:
Clerkship directors need to be explicit about expectations specific to their own rotations. In the first two years, for instance, showing up for lecture is not mandatory, but in third year it typically is—this should be explicit. Other examples include, but are not limited to, expectations for taking night call, the number and promptness for submitting written histories and physicals, and the rapidity with which basic textbook material should be mastered.

Working with Teachers:
As part of the faculty development process (see Chapter 8: Faculty Development) teachers should learn how to apply departmental expectations to the evaluation of individual students. This will avoid the concern of a student that the evaluation was arbitrary. As much as possible, summative evaluation should be based on multiple teachers (or, in the case of private practice rotations with a single physician, on multiple, documented observations. Faculty should learn how their evaluations contribute to, rather than determine, a final summative grade by the department, so that they will not withhold from the department any observations about the student which are concerning.

Reviewing the Entirety of the Record:
Clerkships benefit from having a regular education committee to review the clerkship records of performance for students who are in jeopardy of receiving a failing grade.[21, 105, 279, 317] We recommend that a group decision, reviewing the entirety of the individual student's clerkship record, is an established way of guaranteeing procedural due process. Clerkships are allowed to decide what materials are to be the bases for these decisions, and this may include evaluations (including "recommended grades") from teachers, examination scores, and 360° evaluations from nurses and others. Consistency is strongly recommended, since once a department has established an expectation for what materials are to be reviewed, this may well constitute the precedent for judgment about other students in the future.

Section 15. Feedback

Andrew Albritton, MD and Lisa E. Leggio, MD

Introduction

According to the Liaison Committee on Medical Education standard ED-30,[318] "The directors of all courses and clerkships must design and implement a system of formative and summative evaluation of student achievement in each course and clerkship." Providing feedback to students is one of the most important responsibilities of an educator. Without feedback, students can only gauge performance by trial and error. Frequently the students learn about problems at the end of an educational experience when it is too late to make corrections. Students can better meet expectations and develop the desired knowledge, skills, attitudes, and behaviors when given timely, constructive feedback. Feedback should follow directly from formative evaluation. The purpose of feedback is to reinforce appropriate behaviors and to correct mistakes and misconceptions. Conveying specific observations and insights about students' performance guides them in enhancing future performance toward meeting or exceeding expectations. This section will outline recommendations for the practice of feedback based on established principles.[319-322]

Basic Elements of Effective Feedback

First, create a climate of mutual trust and respect. The teacher and students should have common goals. Effective feedback takes place in an appropriate location that offers privacy. Second, make sure students clearly understand the expectations for the clerkship. Review the goals and objectives and the criteria for evaluation. During the educational experience, students usually want to know about their performance as it relates to a projected grade and what steps are necessary to reach the next level. Third, effective feedback is timely and occurs on a regular basis. After a specific incident, provide feedback as soon as possible after the event occurs. Sometimes students may not realize that feedback is taking place. Labeling feedback as "feedback" will avoid this situation. Be supportive when giving feedback. Consider using the "feedback sandwich" by giving constructive feedback between positive behaviors at the beginning and end of the session. Keep in mind when this technique is overused students begin to hear a compliment and think "Uh oh, what's coming next!"

Feedback is presented as information; it is formative, and uses verbs and nouns. In contrast, evaluation is presented as judgment, tends to be summative, and uses adverbs and adjectives. An example of feedback is "Your differential diagnosis did not include leukemia which is important to consider in a patient with easy bruising." An example of evaluation is, "Your differential diagnoses are inadequate."

Guidelines for Structuring Feedback

The following steps help ensure that feedback is effective in helping students improve their performance:

- Solicit feedback from students. Ask students to self-assess their performance in the various areas that are to be evaluated, as well as what skills they should further develop. Feedback can be solicited from students by asking questions like, "How did things go?" or "What went well?"

- Share your actual observations with students regarding the skills, attitudes, or behaviors that they are performing well.
- Help students identify specific areas for improvement and suggest next steps to improve performance. For example, "If you had an opportunity to do it again, what would you do differently?"
- Ask students if they understand or have any questions about the feedback. Using interactive feedback has two helpful benefits: it allows students to verbalize the problem, thus "saving face"; secondly, it creates an opportunity to see if students have any insight into how things are going – this is very useful information especially if students lack insight.

Characteristics of Effective Feedback

Effective Feedback should be

- Timely and take place close to when the event occurred. Consider postponing giving feedback if students are under a lot of stress such as post-call, just before an exam, or when ill. One should also postpone giving feedback if he or she is angry or lacks adequate time to provide thoughtful feedback.
- Given in an appropriate location – not in front of others unless giving general feedback to a group.
- Descriptive and nonjudgmental. Talk about "what" the students did, rather than "who" they are. Focus on the behavior and not the person.
- Based on direct observations of specific skills, attitudes, or behaviors rather then generalizations, interpretations, or assumed intentions.
- About decisions and actions, not assumed intentions or interpretations.
- Reinforcing what is done right.
- Focused on areas that students can control or change.
- Ensuring the students improve while maintaining their self-respect.
- Limited to what students can use. Avoid "feedback overload" caused by providing too many suggestions during a session. Instead, make two or three important points and schedule additional feedback sessions to address other areas for improvement.
- Given in such a way that students understand it and know how to take the next step.

Examples of Feedback

A student presents a patient during rounds or in clinic. The presentation includes the pertinent information, but was lengthy and disorganized. Comments on the structure of process will be presented within brackets. In the appropriate setting, ask the student how he or she feels about his or her presenting skills [*soliciting feedback from the student*]. Tell the student that you are giving feedback about the presentation on that particular patient [*labeling feedback*]. "The presentation included all the pertinent information [*positive feedback*]. To improve the organization of the presentation, omit the extraneous information [be specific about the information that was not necessary to include] to make the presentations more concise and focused [*suggestions for improvement*]." Ask the student if the feedback was helpful and if he or she has any questions [*making sure the student understands the feedback*].

After observing a student's physical exam on a child with congenital heart disease, first ask the student how they felt the encounter went. If the student expresses concern about not hearing the heart murmur, say "Talking to the child about her new puppy seemed to make her more comfortable in the beginning [*something done well*]. After you examined her ears, she was crying and you were not able to hear her heart murmur [*something that could have been done*

better]. Try listening to the heart first while the patient is calm and quiet before looking in the ears [*action plan*]."

Challenges in Giving Feedback

Students who do not respond to feedback

First determine why the students did not incorporate the feedback. Did the students not recognize they were being given feedback or are the students resistant to feedback? By always labeling feedback, the problem of students not recognizing the feedback can be avoided. Sometimes students do not fully understand how to incorporate the feedback. Asking students if they understand the feedback usually addresses this issue. Students who are resistant to feedback can be challenging. One strategy is giving "feedback about the feedback." The process is the same as giving feedback with an emphasis on the importance of incorporating feedback. (see also Chapter 10: Working with Students with Difficulties: Academic and Nonacademic)

Issues regarding professionalism

Have a clear understanding of the issues before meeting with students. Request that the individual(s) who raised the concern about students' professional behavior document the issues in writing. Meet with the students and ask them to describe what happened. For example, after a hectic night on call, a student makes the comment, "I have good news, Mr. Smith died." The resident makes the attending physician aware of the student's unprofessional comment, and the attending physician contacts the clerkship director to address the student's lack of professionalism. First, ask the student to tell you what happened from his or her perspective. The student may have meant that Mr. Smith does not have to suffer any longer. The issue is one of communication and not professionalism. Giving feedback about behavioral problems provides an opportunity to truly help students gain insight into how they are perceived by others. Always document in writing what took place during meetings regarding professional concerns. With serious professional or behavioral problems, strongly consider having the Associate Dean for Student Affairs attend the meeting.

The angry student

Before attempting to give feedback to angry students, the cause of the anger must be addressed. For example, "you seem angry; would you like to talk about it?" Sometimes students are so angry that their behavior is escalating to the point of being inappropriate. When this happens, taking control of the situation is critical. There are several strategies to consider. Tell students that they first need to calm down and get control of their anger before the situation or concern can be discussed. In some situations, a more aggressive approach may be necessary. Even though the facts of the students' concerns may not be clear, address their inappropriate behavior. For example, "Since I have not worked with you in the clinical setting, I cannot comment about your clinical skills. However, based on this meeting, I can make several comments about your professionalism." After making the statement, pause for the students' reaction. Be prepared to answer the question "What do you mean?" This is the opportunity to give the students feedback about how to handle these types of situations in a more professional manner.

Feedback About the Clerkship

In addition to having a feedback session at the end of the clerkship, periodically obtaining feedback from the students during the clerkship experience is also important. By soliciting

feedback, students' concerns are frequently identified before major problems develop. Be open and receptive to their problems and even criticisms. Consider scheduling a time to meet with the students or spend a few minutes before a conference asking for feedback about how things are going or if there are any concerns or problems. Students are more likely to be open or honest in a group than individually. Group feedback gives you an opportunity to determine whether the issues are related to one or a few students, or if they are of a more general concern.

Changing the Culture for Feedback

One of the most common complaints clerkship directors receive from students is "I never got feedback." Changing the culture about feedback may be one of the most challenging endeavors for educators. With the increased clinical demands on faculty and the eighty-hour work week for residents, finding time for feedback is becoming more difficult. Consider encouraging faculty to have "feedback rounds" once a week instead of teaching rounds. It provides the team an opportunity to talk about what went well in the care of the patients for that week, and in the case of an unexpected death or transfer to the intensive care unit, the team could explore what, if anything could have been done differently in the care of the patient. Having feedback rounds provides an opportunity to address any issues or concerns the team may be encountering and for the attending physician to obtain feedback about his or her role. With weekly feedback rounds, the attending can also meet individually with the students and residents to give feedback.

References

1. Wayne DB, Butter J, Siddall VJ, Fudala MJ, Lindquist LA, Feinglass J, Wade LD, McGaghie WC. Simulation-based training of internal medicine residents in advanced cardiac life support protocols: a randomized trial. Teach Learn Med. 2005;17(3): 210-216.
2. Wayne DB, Butter J, Siddall VJ, Fudala MJ, Wade LD, Feinglass J, McGaghie WC. Mastery learning of advanced cardiac life support skills by internal medicine residents using simulation technology and deliberate practice. J Gen Intern Med. 2006;21 (In press).
3. McGaghie WC. Evaluation of learners. In: McGaghie WC, Frey JJ (eds.). Handbook for the Academic Physician. New York: Springer-Verlag, 1986:125-146.
4. McGaghie WC, Downing SM, Kubilius R. What is the impact of commercial test preparation courses on medical examination performance? Teach Learn Med. 2004;16:202-211.
5. Dawes RM. House of Cards: Psychology and Psychotherapy Built on Myth. New York: Free Press, 1994.
6. Hubbard JP. Measuring Medical Education: The Tests and the Experience of the National Board of Medical Examiners, 2nd ed. Philadelphia: Lea & Febiger, 1978.
7. Issenberg SB, McGaghie WC, Gordon DL, Symes S, Petrusa ER, Hart IR, Harden RM. Effectiveness of a cardiology review course for internal medicine residents using simulation technology and deliberate practice. Teach Learn Med. 2002;14:223-228.
8. Pugh CM, Srivastava S, Shavelson R, Walker D, Cotner T, Scarloss B, Kuo M, Rawn C, Dev P, Krummel TH, Heinrichs LH. The effect of simulator use on learning and self-assessment: the case of Stanford University's e-pelvis simulator. In: Westwood JD, Hoffman HM, Mogel GT, Stredney D (eds). Medicine Meets Virtual Reality 2001. Amsterdam: IOS Press, 2001:396-400.
9. Pugh CM, Youngblood P. Development and validation of assessment measures for a newly developed physical examination simulator. J Am Med Inform Assoc. 2002;9:448-460.
10. Issenberg SB, McGaghie WC, Petrusa ER, Gordon DL, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. Med Teach. 2005;17(1):10-28.
11. The Medical School Objectives Writing Group. Learning objectives for medical student education—guidelines for medical schools: Report I of the Medical School Objectives Project. Acad Med. 1999;74:13-18.
12. ACGME Outcome Project. Toolbox of Assessment Methods. September 2000. http://www.acgme.org accessed 02/23/05.
13. Arnold L. Assessing professionalism behavior: yesterday, today, and tomorrow. Acad Med. 2002;77:502-515.
14. Papadakis MA, Hodgson CS, Teherani A, Kohatsu ND. Unprofessional behavior in medical school is associated with subsequent disciplinary action by a state medical board. Acad Med 2004;79:244-249.
15. Bogner MS (ed). Human Error in Medicine. Hillsdale, NJ: Lawrence Erlbaum Associates, 1994.
16. Kohn LT, Corrigan JM, Donaldson MS (eds). To Err is Human: Building a Safer Health System. Washington, DC: National Academy Press, 2000.
17. Brannick MT, Salas E, Prince C (eds). Team Performance Assessment and Measurement: Theory, Methods, and Applications. Mahwah, NJ: Lawrence Erlbaum Associates, 1997.
18. Salas E, Fiori SM (eds). Team Cognition: Understanding the Factors that Drive Process and Performance. Washington, DC: American Psychological Association, 2004.

19. Carnahan D, Hemmer PA. Descriptive evaluation. Sec. 3 of Ch. 6 Evaluation and grading of students. LD Pangaro and WC McGaghie (eds.) In: Guidebook for Clerkship Directors, 3rd ed.
20. Pangaro LN. A new vocabulary and other innovations for improving descriptive in-training evaluations. Acad Med. 1999;74:1203-1207.
21. Lavin B, Pangaro LN. Internship ratings as a validity outcome measure for an evaluation to identify inadequate clerkship performance. Acad Med. 1998;73:998-1002.
22. Denton GD, DeMott C, Pangaro LN, Hemmer PA. Narrative review: use of student-generated logbooks in undergraduate medical education. Teach Learn Med. 2005;17 (In press).
23. Turnbull J, MacFadyen J, van Barneveld C, Norman G. Clinical work sampling: a new approach to the problem of in-training evaluation. J Gen Intern Med. 2000;15:556-561.
24. Tugwell P, Dok C. Medical record review. In: Neufeld V, Norman G (eds). Assessing Clinical Competence. New York: Springer Publishing Co., 1985:142-182.
25. Case SM, Swanson DB. Constructing Written Test Questions for the Basic and Clinical Sciences, 3rd ed. Philadelphia: National Board of Medical Examiners, 2002.
26. Downing SM. Assessment of knowledge with written test forms. In: Norman GR, van der Vleuten CPM, Newble DI (eds). International Handbook of Research in Medical Education, Part Two. Dordrecht, The Netherlands: Kluwer Academic Publishers, 2002:647-672.
27. Jozefowicz RF, Koeppen BM, Case S, Galbraith R, Swanson D, Glew H. The quality of in-house medical school examinations. Acad Med. 2002;77:156-161.
28. Tekian A, McGuire CH, McGaghie WC (eds). Innovative Simulations for Professional Competence Evaluation. Chicago: Department of Medical Education, University of Illinois at Chicago, 1999.
29. Gaba D. Human work environment and simulators. In: Miller RD (ed). Anesthesia, 5th ed. Philadelphia: Churchill Livingstone, 2000:2613-2668.
30. Petrusa ER. Clinical performance assessments. In: Norman GR, van der Vleuten CPM, Newble DI (eds). International Handbook of Research in Medical Education, Part Two. Dordrecht, The Netherlands: Kluwer Academic Publishers, 2002:673-709.
31. McGaghie WC, Renner BR, Kowlowitz V, Sauter SVH, Hoole AJ, Schuch CP, Misch MS. Development and evaluation of musculoskeletal performance measures for an objective structured clinical examination. Teach Learn Med. 1994;6(1):59-63.
32. Stufflebeam DL. The checklists development checklist (CDC). Western Michigan University, Evaluation Center. Available at: http://www.wmich.edu/evalctr/checklists/cdc.htm (Accessed 04/25/2005).
33. Williams RG, Klamen DA, McGaghie WC. Cognitive, social and environmental sources of bias in clinical performance ratings. Teach Learn Med. 2003:15(4):270-292.
34. Mancall EL, Bashook PG (eds). Assessing Clinical Reasoning: The Oral Examination and Alternative Methods. Evanston, IL: American Board of Medical Specialties, 1995.
35. Hunt DD, Maclaren CF, Scott CS, Chu J, Leiden LI. Characteristics of dean's letters in 1981 and 1992. Acad Med. 1993;68:905-911.
36. Hunt DD, Maclaren CF, Scott CS, Marshall SG, Braddock CH, Sarfaty S. A follow-up study of the characteristics of dean's letters. Acad Med. 2001;76:727-733.
37. Payne BC. The medical record as a basis for assessing physician competence. Ann Intern Med. 1979;91:623-629.
38. Ramsdell JW, Berry CC. Evaluation of general and traditional internal medicine residencies utilizing a medical records audit based on educational objectives. Med Care 1983;21:1144-1153.
39. van der Vleuten CPM, Swanson DB. Assessment of clinical skills with standardized patients: state of the art. Teach Learn Med. 1990;2:58-76.

40. Lin P, Miller E, Herr, G, Hardy C, Sivarajan M, Willenkin R. Videotape reliability: a method of evaluation of a clinical performance examination. J Med Educ. 1980;55:713-715.
41. Shepherd D, Hammond P. Self-assessment of specific interpersonal skills of medical undergraduates using immediate feedback through closed-circuit television. Med Educ. 1984;18:80-84.
42. Pratt D, Magill, MK. Educational contracts: a basis for effective clinical teaching. J Med Educ. 1983;58:462-467.
43. Hafler JP, Lovejoy FH. Scholarly activities recorded in the portfolios of teacher-clinician faculty. Acad Med. 2000;75:649-652.
44. Simpson D, Hafler J, Brown D, Wilkerson L. Documentation systems for educators seeking academic promotion in U.S. medical schools. Acad Med. 2004;79(8):783-790.
45. O'Sullivan PS, Reckase MD, McClain T, Savidge MA, Clardy JA. Demonstration of portfolios to assess competency of residents. Adv Health Sci Educ. 2004; 9:309-323.
46. Kaplan SH, Ware JE. The patient's role in health care and quality assessment. In: Goldfield N, Nash DB (eds). Providing Quality Care: The Challenge to Clinicians. Philadelphia: American College of Physicians, 1989:25-69.
47. Center for Creative Leadership. 360 by Design. Available at: http://www.ccl.org (Accessed 02/28/2005).
48. Miller GE (ed). Teaching and Learning in Medical School. Cambridge, MA: Harvard University Press, 1961.
49. Newman C. The Evolution of Medical Education in the Nineteenth Century. London: Oxford University Press, 1957.
50. Frederiksen N. The real test bias: influences of testing on teaching and learning. Am Psychol. 1984;39:193-202.
51. Newble DI, Jaeger K. The effect of assessments and examinations on the learning of medical students. Med Educ. 1983;17:165-171.
52. Good M-J D. American Medicine: The Quest for Competence. Berkeley: University of California Press, 1995.
53. Battistone MJ, C Milne, M A Sande, LN Pangaro, PA Hemmer, TS Shomaker, The Feasibility and Acceptability of Implementing Formal Evaluation Sessions and Using Descriptive Vocabulary to Assess Student Performance on a Clinical Clerkship, Teach Learn Med 2002;14(1):5-10.
54. Gage, NL Hard Gains in the Soft Sciences, Bloomington, IN: CEDR, Phi Delta Kappa, 1985.
55. Durning S, Pangaro L, Denton GD, Hemmer P, Wimmer A, Grau T, Gaglione MA, Moores L, Inter-site Consistency as a Standard of Programmatic Evaluation in a Clerkship with multiple, Geographically Separated Sites, Acad Med. 2003;78:S36-S38.
56. Hemmer PA, Szauter K, Allbritton TA, Elnicki DM. Internal medicine clerkship directors' use of and opinions about clerkship examinations. Teach Learn Med. 2002;14(4):229-235.
57. Turner DA., Anderson KD. Surgery Shelf Exam Failure: Remediation Strategies and Policies in Medical Schools, in Abstracts From the Proceedings of the 2004 Annual Meeting of the Association for Surgical Education (ASE), Teach Learn Med. 2005;17(3):297-303.
58. Levine RE, Carlson DL, Rosenthal RH, Clegg KA, Crosby RD. Usage of the National Board of Medical Examiners Subject Test in Psychiatry by U.S. and Canadian clerkships. Acad Psych. 2005;29(1):52-57.
59. Norman GR, Van der Vleuten CP, De Graaff E. Pitfalls in the pursuit of objectivity: issues of validity, efficiency and acceptability. Med Educ. 19911;25(2):119-126.
60. Roop S, Pangaro L, Measuring the Impact of Clinical Teaching on Student Performance during a Third Year Medicine Clerkship, Amer J Med. 2001;110 (3):205-209.

61. Bloom BS, Ed. Taxonomy of educational objectives: The classification of educational goals: Handbook I, cognitive domain. New York ; Toronto: Longmans, Green, 1956.
62. Dreyfus H, Dreyfus S. Mind over Machine. New York: Free Press, Macmillan, 1986;16-51.
63. Pangaro L, Investing in Descriptive Evaluation: a vision for the future of assessment. Med Teach. 2000;22(5):478-481.
64. Walsh WB, Betz NE. Tests and Assessment, Prentice Hall, New Jersey, 1990;49-70.
65. Haynes SN. Clinical Applications of Analog Behavioral Observation: Dimensions of Psychometric Valuation, Psychological Assessment 2001;13:73-85.
66. Downing SM. Validity: on meaningful interpretation of assessment data. Med Educ. 2003;37(9):830-837.
67. Downing SM, Haladyna TM. Validity threats: overcoming interference with proposed interpretations of assessment data. Med Educ. 2004;38(3):327-333.
68. Downing SM. Reliability: on the reproducibility of assessment data. Med Educ. 2004;38(9):1006-1012.
69. Schuwirth L, van der Vleuten CPM . ABC of learning and teaching in medicine Written assessment. Brit Med J. 2003;326:643-645.
70. Williams RG, Dunnington GL, Klamen DL. Forecasting Residents' Performance - Partly Cloudy. Acad Med. 2005;80(5):415-422.
71. Marienfeld RD RJ. Subjective vs. objective evaluation of clinical clerks. NEJM. 1980;302:1036-1037.
72. Awad SS LK, Aoki N, Awad SH, Berger DH. Does the Subjective Evaluation of Medical Student Surgical Knowledge Correlate with Written and Oral Exam Performance? J Surg Res. 2002;104(1):36-39.
73. Tonesk X. The Evaluation of Clerks: Perceptions of Clinical Faculty. A Summary of Issues and Proposed Actions. In; 1983; Washington, DC: The Association of American Medical Colleges; 1983.
74. Simpson MA. Medical student evaluation in the absence of examinations. Med Educ. 1976;10:22-26.
75. Eisner EW. The Educational Imagination. New York: MacMillan Publishing Co., Inc.; 1979.
76. Tonesk X. AAMC program to promote improved evaluation of students during clinical education. J Med Educ. 1986;16(Part 2):83-88.
77. Hunt DD. Functional and dysfunctional characteristics of the prevailing model of clinical evaluation systems in North American medical schools. Acad Med. 1992;67:254-259.
78. Tonesk X. An AAMC pilot study by 10 medical schools of clinical evaluation of students. Journal of Med Educ. 1987;62:707-718.
79. Appel J, Friedman E, Fazio S, Kimmel J, Whelan A. Educational Assessment Guidelines: A Clerkship Directors in Internal Medicine Commentary. Am J Med. 2002;113(2):172-179.
80. Magarian GJ, Mazur DJ. Evaluation of students in medicine clerkships. Acad Med. 1990;65:341-345.
81. McLeod PJ. Undergraduate Clinical Education in Internal Medicine at Canadian medical schools. Acad Med.1994;69:55-57.
82. Herbert WN, Cummings RV, Droegemueller W. Profile of student clerkship administration in obstetrics and gynecology. Obstet Gynec. 1990;76:153-155.
83. Schwiebert LP. How required US family medicine clerkships and preceptorships evaluate medical students. Fam Med. 1996;28:559-564.
84. Zahn CM NS, Armstrong AY, Satin AJ, Haffner WHJ. Variation in medical student grading criteria: A survey of clerkships in obstetrics and gynecology. Am J Obstet Gynecol. 2004;190(5):1388-1393.
85. Elnicki DM Ainsworth M, Magarian GJ, Pangaro LN. Evaluating the internal medicine clerkship: a CDIM commentary. Am J Med. 1994;97:I-VI.

86. Holmboe E. Faculty and the Observation of Trainees' Clinical Skills: Problems and Opportunities. Acad Med. 2004;79(1):16-22.
87. Tonesk X. Editorial: Clinical Judgment of Faculties in the Evaluation of Clerks. J Med Educ. 1983;58:213-214.
88. Grim DR, Miller MD. Criteria for evaluating performance of third-year medical students. Fam Med. 1993;25:388-390.
89. Maxim BR Dielman TE. Dimensionality, Internal Consistency and Interrater Reliability of Clinical Performance Ratings. Med Educ. 1987;21:130-137.
90. Stillman RM. Pitfalls in evaluating the surgical student. Surgery. 1984;96:92-95.
91. Metheny WP. Limitations of physician ratings in the assessment of student clinical performance in an obstetrics and gynecology clerkship. Obstet Gynecol. 1991;78:136-140.
92. Silber CG NT, Paskin DL, Eiger G, Robeson M, Veloski JJ. Do Global Rating Forms Enable Program Directors to Assess the ACGME Competencies? Acad Med. 2004;79(6):549-556.
93. Williams RG, Colliver JA, Dunnington GL. Assuring the reliability of resident performance appraisals: More items or more observations? Surgery. 2005;137(2):141-147.
94. Arnold L. Assessing Professional Behavior: Yesterday, Today, and Tomorrow. Acad Med. 2002;77(6):502-515.
95. Papadakis MA, Loeser H, Healy K. Early Detection and Evaluation of Professionalism Deficiencies in Medical Students: One School's Approach. Acad Med. 2001;76(11):1100-1106.
96. Tonesk X Buchanan RG. Datagram: Faculty perceptions of current clinical evaluation systems. J Med Educ. 1985;60:573-576.
97. Stemmler EJ. Promoting improved evaluation of students during clinical education: a complex management task. J Med Educ. 1986;16:75-81.
98. Hemmer PA, Pangaro LN. The effectiveness of formal evaluation sessions during clinical clerkships in better identifying students with marginal funds of knowledge. Acad Med. 1997;72:641-643.
99. Hemmer PA, Pangaro LN. Using formal evaluation sessions for case-based faculty development during clinical clerkships. Acad Med. 2000;75(12):1216-1221.
100. Speer A Solomon DJ, Fincher RM. Grade inflation in internal medicine clerkships: results of a national survey. Teach Learn Med. 2000;12(3):112-116.
101. Irby DM Milam S. The legal context for evaluating and dismissing medical students and residents. Acad Med. 1989;64:639-643.
102. Gough HG Hall WB, Harris RD. Evaluation of performance in medical training. J Med Educ. 1964;39:679-692.
103. Dubovsky SL. Coping with entitlement in medical education. NEJM. 1986;315:1672-1674.
104. Magarian GJ. Evaluation and Grading. In: Fincher RM, ed. Washington, DC: American Association of Medical Colleges; 1996.
105. Hemmer PA, Hawkins R, Jackson J, Pangaro LN. Assessing how well three evaluation methods detect deficiencies in professionalism during a clerkship. Acad Med. 2000;75(2):167-173.
106. McLeod PJ. Faculty assessments of case reports of medical students. J Med Educ. 1987;62:673-677.
107. Noel GL HJ, Caplow MP, Cooper GS, Pangaro LN, Harvey J. How well do internal medicine faculty members evaluate the clinical skills of residents. Ann Intern Med. 992;117:757-765.
108. Kalet A EJ, Kowlowitz V. How well do faculty evaluate the interviewing skills of medical students? J Gen Intl Med. 1992;7:499-505.
109. Carline JD, Paauw DS, Thiede KW, Ramsey PG. Factors affecting the reliability of ratings of students' clinical skills in a medicine clerkship. J Gen Intern Med. 1992;7:506-510.

110. Kreiter CD, Ferguson K, Lee Won-Chan, Brennan RL, Densen P. A generalizability study of a new standardized rating form used to evaluate students' clinical clerkship performances. Acad Med. 1998;73:1294-1298.
111. Keynan A, Friedman M, Benbassat J. Reliability of global rating scales in the assessment of clinical competence of medical students. Med Educ. 1987;21:477-481.
112. MacRae HM, Vu NV, Graham B, Word-Sims M, Colliver JA, Robbs RS. Comparing checklists and databases with physicians' ratings as measures of students' history and physical examination skills. Acad Med. 1995;70:313-317.
113. Miller JM Jr, Smith IK, Sosnowski JR, Hester LL. Evaluation of student performance in an obstetrics and gynecology clerkship. J Reprod Med. 1982;27:443-446.
114. Dawson-Saunders B, Paiva RE. The validity of clerkship performance evaluations. Med Educ. 1986;20:240-245.
115. Hull AL. Medical student performance: A comparison of house officer and attending staff as evaluators. Eval Health Prof. 1982;5:87-94.
116. Littlefield JH, Harrington JT, Anthracite NE, Garman RE. A description and four-year analysis of a clinical clerkship evaluation system. J Med Educ. 1981;56:334-340.
117. Campos-Outcalt D, Witzke DB, Fulginiti JV. Correlations of family medicine clerkship evaluations with scores on standard measures of academic achievement. Fam Med. 1994;26:85-88.
118. Case SM, Ripkey DR, Swanson DB. The relationship between clinical science performance in 20 medical schools and performance on Step 2 of the USMLE Licensing Examination. Acad Med. 1996;71:S31-S3.
119. Lawrence PF, Nelson EW, Cockayne TW. Assessment of medical student fund of knowledge in surgery. Surgery. 1985;97:745-749.
120. Greenberg LW Getson PR. Assessing student performance on a pediatric clerkship. Arch Ped Adoles Med. 1996;150:1209-1212.
121. Hull AL Hodder S, Berger B, Ginsberg D, Lindheim N, Quan J, Kleinhenz ME. Validity of three clinical performance assessments of internal medicine clerks. Acad Med. 1995;70:517-522.
122. Stenchever MA, O'Toole B, Irby D. Evaluating student performance in an obstetrics and gynecology clerkship. Am J Obstet Gynecol. 1979;134:235-237.
123. Dielman TE, Hull AL, Davis WK. Psychometric properties of clinical performance ratings. Eval Health Prof. 1980;3:103-117.
124. Holmboe ES, Huot S, Chung J, Norcini J, Hawkins RE. Construct validity of the miniclinical evaluation exercise (miniCEX). Acad Med. 2003 Aug;78(8):826-830.
125. Holmboe ES, Hawkins RE, Huot SJ. Effects of training in direct observation of medical residents' clinical competence: a randomized trial. Ann Intern Med. 2004;140(11):874-881.
126. Ten Eyck RP, Maclean TA. Improving the quality of emergency medicine rotation/clerkship evaluations. Am J Emerg Med. 1994;12:113-117.
127. Printen KJ, Chappell W, Whitney D. Clinical performance evaluation of junior medical students. J Med Educ. 1973;48:343-348.
128. Bennett AJ, Arnold LM. Use of a computerized evaluation system in a psychiatry clerkship. Acad Psych. 2004;28(3):197-203.
129. Duque G. Web-based evaluation of medical clerkships: a new approach to immediacy and efficacy of feedback and assessment. Med Teach. 2003;25(5):510-514.
130. Lye PS, Biernat KA, Bragg DS, Simpson DE. A pleasure to work with - an analysis of written comments on student evaluations. Amb Ped. 2001;1(3):128-131.
131. Ainsworth MA, Speer AJ, Solomon DJ. A clinical evaluation form to improve faculty critique of students. Acad Med. 1995;70:445.

132. Littlefield JH. Developing and Maintaining a Resident Ratings System. In: Lloyd JS LD, ed. How to Evaluate Residents. Chicago: American Board of Medical Specialists; 1986.
133. Battistone MJ, Pendleton B, Milne C, Battistone ML, Sande M, Hemmer PA, Shomaker TS. Global descriptive evaluations are more responsive than global numeric ratings in detecting students' progress during the inpatient portion of an internal medicine clerkship. Acad Med. 2001;76:S105-S107.
134. Noel GL. A system for evaluating and counseling marginal students during clinical clerkships. J Med Educ. 1987;62:353-355.
135. Pangaro LN, Hemmer PA, Gibson KF, Holmboe E. Formal Evaluation Sessions Enhance the Evaluation of Professional Demeanor. In: Presented at 8th International Ottawa Conference on Med Educ and Assessment; 1998; Philadelphia, PA; 1998.
136. Roop SA, Pangaro L. Effect of clinical teaching on student performance during a medicine clerkship. Am J Med. 2001;110(3):205-209.
137. Ogburn T, Espey E. The R-I-M-E method for evaluation of medical students on an obstetrics and gynecology clerkship. Am J Obstet Gynecol. 2003;189(3):666-669.
138. Albritton TA, Fincher RM, Work JA. Group evaluation of student performance in a clerkship. Acad Med. 1996;71(5):551-552.
139. Jacobson MJ, Sherman L, Perlman I, Lefferts R, Soroff H. Clerkship site and duration: Do they influence student performance? Surgery. 1986;100(2):306-310.
140. Pangaro LN. Expectations of and for the medicine clerkship director. Am J Med. 1998;105:363-365.
141. Pangaro L, Bachicha J, Brodkey A, Chumley-Jones H, Fincher RM, Gelb D, Morgenstern B, Sachdeva AK. Expectations of and for Clerkship Directors: A Collaborative Statement from the Alliance for Clinical Education. Teach Learn Med. 2003;15(3):217-222.
142. American Association of Medical Colleges. AAMC Educational Outcomes Project. Accessed at www.aamc.org.
143. Institute of Medicine. Health Professions Education: a bridge to quality. National Academy Press. Washington. 2003.
144. Richards BF, Rupp R, Zaccaro DJ, Cariaga-Lo L, Harward D, Petrusa ER, Smith AC, Willis SE. Use of a standardized patient based clinical performance examination as an outcome measure to evaluate medical school curricula. Acad Med. 1996;71:S49-S51.
145. Anderson MB, Stillman PL, Wang Y. Growing use of standardized patients in teaching and evaluation in medical education. Teach Learn Med. 1994;6:15-22.
146. Sloan DA, Donnelly MB, Schwartz RW, Strodel WE. The objective structured clinical examination: the new gold standard for evaluating postgraduate clinical performance. Ann Surg. 1995;222:735-742.
147. Stillman PL, Swanson D, Regan MB, Philbin MM, Nelson V, Ebert T, et al. Assessment of clinical skills of residents utilizing standardized patients. A follow-up study and recommendations for application. Ann Intern Med. 1991;114:393-401.
148. Barrows HS. An overview of the uses of standardized patients for teaching and evaluating clinical skills. Acad Med. 1993;6:443-453.
149. Van der Vleuten CPM, Swanson DB. Assessment of clinical skills with standardized patients: state of the art. Teach Learn Med. 1990;2:58-76.
150. Rethans JJ, Sturmans F, drop R, van der Vleuten C, Hobus P. Does competence of general practitioners predict their performance? Comparison between examination setting and actual practice. BMJ. 1991;303:1377-1380.
151. Hodges B, Regehr G, McNaughton N, Tiberius R, Hanson M. OSCE checklists do not capture increasing levels of expertise. Acad Med. 1999;74(10):1120-1134.
152. Regehr G, MacRae H, Reznick RK, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. Acad Med. 1998;73:993-997.

153. Platt FW, Mcmath JC. Clinical hypocompetence: the interview. Ann Intern Med. 1979;91:898-902.
154. Meuleman JR, Caranasos GJ. Evaluating the interview performance of internal medicine interns. Acad Med. 1989;64:277-279.
155. Beaumier A, Bordage G, Saucier D, Turgeon J. Nature of the clinical difficulties of first year family medicine residents under direct observation. Can Med Assoc J. 1992;146:489-497.
156. Sachdeva AK, Loiacono LA, Amiel GE, Blair PG, Friedman M, Roslyn JJ. Variability in the clinical skills of residents entering training programs in surgery. Surgery. 1995;118:300-309.
157. Pfeiffer C, Madray H, Ardolino A, Willms J. The rise and fall of student's skill in obtaining a medical history. Med Educ. 1998;32:283-288.
158. Stewart MA, McWhinney IR, Buck CW. The doctor-patient relationship and it's effect upon outcome. J R Coll Gen Pract. 1979;29:77-82.
159. Weiner S, Nathanson M. Physical examination. Frequently observed errors. JAMA. 1976;236:852-855.
160. Wray NP, Friedland JA. Detection and correction of house staff error in physical diagnosis. JAMA. 1983;249:1035-1037.
161. Butterworth JS, Reppert EH. Auscultatory acumen in the general medical population. JAMA. 1960;174:32-34.
162. Raferty EB, Holland WW. Examination of the heart, an investigation into variation. Am J Epidemiol. 1967;85:438-444.
163. Mangione S, Nieman LZ. Cardiac auscultatory skills of internal medicine and family practice trainees. A comparison of diagnostic proficiency. JAMA. 1997;278:717-722.
164. Fox RA, Clark CLI, Scotland AD, Dacre JE. A study of pre-registration house officers' clinical skills. Med Educ. 2000;34:1007-1012.
165. Peterson MC, Holbrook JH, Hales DV, Smith NL, Staker LV. Contributions of the history, physical examination, and laboratory investigation in making medical diagnoses. West J Med. 1992;156:163-165.
166. Kirch W, Schaffi C. Misdiagnosis at a university hospital in 4 medical areas. Report on 400 cases. Medicine. 1996; 5:29-40.
167. Hampton JR, Harrison MJG, Mitchell JRA, Prichard JS, Seymour C. Relative contributions of history-taking, physical examination, and laboratory investigation to diagnosis and management of medical outpatients. BMJ. 1975;2:486-489.
168. Bordage G. Why did I miss the diagnosis? Some cognitive explanations and educational implications. Acad Med. 1999;74:S138-S143.
169. Turnbull J, Gray J, MacFacyen J. Improving in-training evaluation programs. J Gen Intern Med. 1998;13:317-323.
170. Duffy DF. Dialogue: the core clinical skill. Ann Intern Med. 1998; 128: 139-141.
171. Johnson BT, Boohan M. Basic clinical skills: don't leave teaching to the teaching hospitals. Med Educ. 2000;34:692-699.
172. Engel GL. The deficiencies of the case presentation as a method of teaching: another approach. N Engl J Med. 1971;284:20-24.
173. Engel GL. Are medical schools neglecting clinical skills? JAMA. 1976; 236: 861-63.
174. Scenes P. The role of faculty observation in assessing students' clinical skills. Contemp Issues Med Educ. 1997;1:1-2.
175. Noel GL, Herbers JE, Callow MP, Cooper GS, Pangaro LN, Harvey J. How well do internal medicine faculty members evaluate the clinical skills of residents? Ann Intern Med. 1992;117:757-765.
176. Herbers JE, Noel GL, Cooper GS, Harvey J, Pangaro LN, Weaver MJ. How accurate are faculty evaluations of clinical competence? J Gen Intern Med. 1989;4:202-208.

177. Kalet A, Ear JA, Kilowatts V. How well do faculty evaluate the interviewing skills of medical students? J Gen Intern Med. 1992;97:179-184.
178. Elliot DL, Hick am DH. Evaluation of physical examination skills. Reliability of faculty observers and patient instructors. JAMA. 1987;258(23):3405-3408.
179. Lindy, Frank J; Farr, James L. Performance rating. Psych Bulletin.1980;87:72-107.
180. Holmboe ES, Fiebach NF, Galaty L, Huot S. The effectiveness of a focused educational intervention on resident evaluations from faculty: a randomized controlled trial. J Gen Intern Med. 2001;16:1-6.
181. Woehr DJ, Huffcutt AI. Rater training for performance appraisal: a quantitative review. J Occupational Org Psych. 1994;67:189-205.
182. Hauenstein NMA. Training raters to increase the accuracy of appraisals and the usefulness of feedback. Pgs. 404-42. In: Performance Appraisal, Smither JW, Ed. Jossey-Bass. San Francisco. 1998.
183. Stamoulis DT, Hauenstein NMA. Rater training and rating accuracy: training for dimensional accuracy versus training for rater differentiation. J Appl Psych. 1993;78:994-1003.
184. Holmboe ES, Hawkins RE, Huot SJ. Direct observation of competence training: a randomized controlled trial. Ann Intern Med. 2004;140:874-881.
185. Kogan JR, Bellini LM, Shea JA. Feasibility, reliability, and validity of the mini-clinical evaluation exercise in a medicine core clerkship. Acad Med. 2003;78:S33-S35.
186. Norcini JJ, Blank LL, Arnold GK, Kimball HR. The mini-CEX (clinical evaluation exercise): a preliminary investigation. Ann Intern Med. 1995;123:795-799.
187. Holmboe ES, Huot S, Hawkins RE. Construct validity of the Mini Clinical Evaluation Exercise. Acad Med. 2003;78:826-830.
188. Holmboe ES, Williams F, Yepes M, Huot S. Feedback and the MiniCEX. J Gen Intern Med. 2004;19(part 2):558-561.
189. MacKoul G. The SEGUE framework for teaching and assessing communication skills. Patient Educ Couns. 2001;45:23-34.
190. Cegala DJ and Broz SL. Physician communication skills training: a review of the theoretical backgrounds, objectives, and skills. Med Educ. 2002;36:1004-1016.
191. Lane JL, Gottlieb RP. Structured clinical observations: a method to teach clinical skills with limited time and financial resources. Pediatrics. 2000;105:973-977.
192. Pangaro LN, Gibson K, Russell W, Lucas C, Marple R. A prospective, randomized trial of a six-week ambulatory internal medicine rotation. Acad Med. 1995;70:537-541.
193. Hemmer PA, Grau T, Pangaro LN. Assessing the effectiveness of combining evaluation methods for the early identification of students with inadequate knowledge during a clerkship. Med Teach. 2001;23:580-584.
194. Constance B, Dawson B, Steward D, Schrage J, Schermerhorn G. Coaching students who fail and identifying students at risk for failing the National Board of Medical Examiners Medicine subject test. Acad Med. 1994;69(10, Suppl.):S69-S71.
195. Salvatori P. Reliability and validity of admissions tools used to select students for the health professions. Adv Health Sci Educ. 2001;6:159-175.
196. McGaghie WC. Assessing readiness for medical education: evolution of the Medical College Admissions Test. JAMA. 2002;288(9):1085-1090.
197. Gilbert G. Predictive validity of the Medical College Admissions Test Writing Sample for the USMLE Examinations Steps 1 and 2. Adv Health Sci Educ. 2002;7:191-200.
198. Hojat M. A validity study of the Writing Sample section of the MCAT. Acad Med. 2000;75(10, Suppl.):S41-S44.
199. Johnson EK, Edwards JC. Current practices in admission interviews at U.S. medical schools. Acad Med. 1991;66:408-412.

200. Stern DT, Frohna AZ, Gruppen LD. The prediction of professional behaviour. Med Educ. 2005;39(1):75-82.
201. Merideth K, Dunlap M, Baker HH. Subjective and objective admission factors as predictors of clinical clerkship performance. J Med Educ. 1982;57(10, Part 1):743-751.
202. Armstrong A, Dahl C, Haffner W. Predictors of performance on the National Board of Medical Examiners Obstetrics and Gynecology subject examination. Obstet Gynecol. 1998;91(6):1021-1022.
203. Gonnella J. An empirical study of the predictive validity of number grades in medical school using 3 decades of longitudinal data: implications for a grading system. Med Educ. 2004;38:425-434.
204. Hemmer PA, Szauter K, Allbritton TA, Elnicki DM. Internal medicine clerkship directors' use of and opinions about examinations. Teach Learn Med. 2002;14(4):229-235.
205. Roop S, Pangaro LN. Measuring the impact of clinical teaching on student performance during a third year medicine clerkship. Am J Med. 2001;110(5):205-209.
206. Myles TD, United States Medical Licensure Examination Step 1 scores and obstetrics-gynecology clerkship final examination. Obstet Gynecol. 1999;94(6):1049-1051.
207. Omori D. Introduction to clinical medicine: a time for consensus and integration. Am J Med. 2005;118(2):189-194.
208. Poremba J. Using second year student performance in clinical skills courses to predict substandard clerkship outcome. 2004; unpublished data.
209. Peitzman S. Comparison of "fact-recall" with "higher-order" questions in multiple choice examinations as predictors of clinical performance in medical students. Acad Med. 1990;65(9, Suppl.):S59-S60.
210. Denton GD, Durning SJ, Wimmer AP, Pangaro LN, Hemmer PA. Is a faculty developed pretest equivalent to pre-third year GPA or USMLE Step 1 as a predictor of third-year internal medicine clerkship outcomes? Teach Learn Med. 2004;16 (4):329-332.
211. American Board of Internal Medicine. Medical Professionalism in the New Millenium: A Physician Charter. ABIM/ACP/European Federation of Internal Medicine, 2004.
212. Datta V, Chang A, Mackay S, Darzi A. The relationship between motion analysis and surgical technical assessments. Am J Surg. 2002;184:70-73.
213. Perry AG, Potter PA. Specimen Collection (Chapter 43). In: Clinical Nursing Skills and Techniques. Mosby. Boston. 1994;1096-1102.
214. Fincher RM, Lewis LA. Learning, experience, and self-assessment of competence of the third-year medical students in performing bedside procedures. Acad Med. 1994;69:291-295.
215. Gronlund NE. Assessment of student achievement. 6th ed. Allyn and Bacon. Boston; 1998.
216. Regehr G, MacRae H, Reznick R, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-type examination. Acad Med. 1998;73:993-997.
217. Martin JA, Regehr G, Reznick R, MacRae H, Murnaghan J, Hutchison C, et. al. Objective structured assessment of technical skill (OSATS) for surgical residents. Br J Surg. 1997;84:273-278.
218. Friedlich M, Wood T, Regehr G, Hurst C, Shamji F. Structured assessment of minor surgical skills (SAMSS) for clinical clerks. Acad Med. 2002;77(10 Suppl):S39-S41.
219. Wanzel KR, Ward M, Reznick RK. Teaching the surgical craft: From selection to certification. Curr Probl Surg. 2002;39:573-659.
220. Browne JP. Validating simulation and assessment devices in surgery: a review. Min Invas Ther Allied Technol. 2000;9:353-356.
221. Downing SM. Validity: on meaningful interpretation of assessment data. Med Educ. 2003;37:830-837.

222. Downing SM. Reliability: on the reproducibility of assessment data. Med Educ. 2004;38:1006-1012.
223. Reznick RK. Teaching and testing technical skills. Am J Surg. 1993;165:358-361.
224. Yudkowsky R, Downing S, Klamen D, Valaski M, Eulenberg B, Popa M. Assessing the head-to-toe physical examination skills of medical students. Med Teach. 2004;26:415-419.
225. Reisner E, Dunnington G, Beard J, Witzke D, Fulginiti J, Rappaport W. A model for the assessment of students' physician-patient interaction skills on the surgical clerkship. Am J Surg. 1991;162:271-273.
226. McCormick DP, Rassin GM, Stroup-Benham CA, Baldwin CD, Levine HG, Persaud DI, et.al. Use of videotaping to evaluate pediatric resident performance of health supervision examinations of infants. Pediatrics. 1993;92:116-120.
227. Dath D, Regehr G, Birch D, Schlachta C, Poulin E, Mamazza J, et.al. Toward reliable operative assessment- the reliability and feasibility of videotaped assessment of laparoscopic technical skills. Surg Endosc. 200418:1800-1804.
228. McGaghie WC. Simulation in professional competence assessment: basic considerations. In: Tekian A, McGuire CH, McGaghie WC (eds.). Innovative Simulations for Assessing Professional Competence. Chicago: Department of Medical Education, University of Illinois at Chicago, 1999.
229. Miller GE. The assessment of clinical skills/competence/performance. Acad Med. 1990;65(9 suppl):S63-S67.
230. Collins JP, Harden RM. AMEE Medical Education Guide No. 13: real patients, simulated patients and simulators in clinical examinations. Med Teach. 1998;20:508-521.
231. Boulet JR, Swanson DB. Psychometric challenges of using simulations for high-stakes testing. In: Dunn WF (ed). Simulators in critical care and beyond. Des Plaines, IL: Society of Critical Care Medicine, 2004.
232. Kalu P, Atkins J, Baker D, Green CJ, Butler PEM. How do we assess microsurgical skill? Microsurgery. 2005;25:25-29.
233. Rifkin WD. Rifkin A. Correlation between housestaff performance on the United States Medical Licensing Examination and standardized patient encounters. Mt Sinai J Med. 2005; 72(1):47-49.
234. Barzansky, B, Etzel S. Educational Programs in US Medical Schools, 2003-2004. JAMA. 2004;292(9):1025-1031.
235. Wagner PJ, Lentz L, Heslop SD. Teaching communication skills: a skills-based approach. Acad Med. 2002; 77(11):1164.
236. Colletti L, Gruppen L, Barclay M, Stern D. Teaching students to break bad news. Am J Surg. 2001;182(1):20-23.
237. Robertson K, Hegarty K, O'Connor V, Gunn J. Women teaching women's health: issues in the establishment of a clinical teaching associate program for the well woman check. Women Health. 2003;37(4):49-65.
238. Adamo G. Simulated and standardized patients in OSCEs: achievements and challenges 1992-2003. Med Teach. 2003;25(3):262-270.
239. Brazeau C, Boyd L, Crosson J. Changing an existing OSCE to a teaching tool: the making of a teaching OSCE. Acad Med. 2002; 77(9):932.
240. Schuwirth LW, van der Vleuten CP. The use of clinical simulations in assessment. Med Educ. 2003;37(Suppl 1):65-67.
241. Karnath B, Thornton W, Frye AW. Teaching and testing physical examination skills without the use of patients. Acad Med. 2002;77(7):753
242. Carpenter JL. Cost analysis of objective structured clinical examinations. Acad Med. 1995;70(9):828-833

243. Brown JA, Abelson J, Woodward CA, Hutchison B, Norman G. Fielding standardized patients in primary care settings: lessons from a study using unannounced standardized patients to assess preventive care practices. Int J Qual Health Care. 1998:10(3):199-206.
244. De Champlain AF, Margolis MJ, King A, Klass DJ. Standardized patients' accuracy in recording examinees' behaviors using checklists. Acad Med. 1997;72(10, Suppl 1):S85-S87.
245. Norcini J, Boulet J. Methodological Issues in the Use of Standardized Patients for Assessment Teach Learn Med. 2003;15(4):293-297.
246. Petrusa ER. Taking standardized patient-based examinations to the next level. Teach Learn Med. 2004;16(1):98-110.
247. Hodges B. Regehr G. McNaughton N. Tiberius R. Hanson M. OSCE checklists do not capture increasing levels of expertise. Acad Med. 1999;74(10):1129-1134.
248. McIlroy JH, Hodges B, McNaughton N, Regehr G. The effect of candidates' perceptions of the evaluation method on reliability of checklist and global rating scores in an objective structured clinical examination. Acad Med. 2002;77(7):725-728.
249. Medical professionalism in the new millennium: a physician charter. Ann Intern Med. 2002;136:243-246.
250. Medical School Objectives Project. Learning objectives for medical student education. Guidelines for medical schools: Report I of the Medical School Objectives Project. Acad Med. 1999;74:13-18.
251. Liaison Committee on Medical Education. Functions and Structure of a Medical School. Standards for Accreditation of Medical Education Programs Leading to the M.D. Degree. Chicago: LCME Secretariat, American Medical Association, 2005.
252. Siraisi NG. Medieval & Early Renaissance Medicine. Chicago: University of Chicago Press, 1990.
253. Puschmann T. A History of Medical Education. New York: Hafner Publishing Co., 1966. (Originally published 1891)
254. Stewart JB. Blind Eye. New York: Simon & Schuster, 1999.
255. Mendoza-Denton R, Ayduk O, Mischel W. Person X situation interactionism in self-encoding (I Am . . . When): implications for affect regulation and social information processing. J Pers Soc Psychol. 2001;80:533-544.
256. Tesser A, Rosen S. The reluctance to transmit bad news. In: Berkowitz L, (ed). Advances in Experimental Social Psychology, Vol. 8. New York: Academic Press, 1975; 194-232.
257. Stern DT, (ed). Measuring Medical Professionalism. New York: Oxford University Press, 2005.
258. Veloski JJ, Fields SK, Boex JR, Blank LL. Measuring Professionalism: a review of studies with instruments reported in the literature between 1982 and 2002. Acad Med. 2005;80: 366-370.
259. Krych AJ, March CN, Bryan RE, Peake BJ, Pawlina W, Carmichael SW. Reciprocal peer teaching: students teaching students in the gross anatomy laboratory. Clin Anat. 2005;18(4):296-301.
260. Ripkey DR, Case SM, Swanson DB. Predicting performance on the NBME Surgery Subject Test and USMLE Step 2: The effects of surgery clerkship timing and length. Acad Med. 1997;72(10, Suppl):S31-S33.
261. Case S, Becker D, Melnick D, Fincher RM. Establishing pass/fail and honors performance standards for the NBME medicine clerkship exam - An application of the Hofstee method. Paper presented at Annual Meeting of the American Educational Research Association. San Francisco, CA, April 1995.

262. Jackson JR, Scott LK, Dismukes WE. The relationship between prior clerkship experience and student performance in medicine clerkships – implications for grading. Med Educ. 1982;16:133-136.
263. Stillman RM. Effect of prior clinical experience on students' knowledge and performance in surgery. Surgery. 1986;100:77-81.
264. Bass EB, Fortin AH 4th, Morrison G, Wills S, Mumford LM, Goroll AH. National Survey of Clerkship Directors in Internal Medicine on the Competencies that should be Addressed in the Medicine Core Clerkship. Am J Med. 1997;102(6):564-568.
265. Baciewicz FA Jr, Arent L, Weaver K, Yeastings R, Thomford NR. Influence of clerkship structure and timing on individual student performance. Am J Surg. 1990;159:265-268.
266. Magarian GJ, Mazur DJ. A national survey of grading systems used in medicine clerkships. Acad Med. 1990;65(10):637-639.
267. Elnicki DM, Lescisin DA, Case S. Improving the National Board of Medical Examiners internal medicine subject exam for use in clerkship evaluation. J Gen Intern Med. 2002;17:435-440.
268. Norcini JJ, Diserens D, Day SC, et al. The scoring and reproducibility of an essay test of clinical judgment. Acad Med. 1990;65(9, Suppl):S41-S42.
269. Day SC, Norcini JJ, Diserens D, et al. The validity of an essay test of clinical judgment. Acad Med. 1990;65(9, Suppl):S39-S40.
270. Elnicki DM. Shockcor WT. Morris DK. Halbritter KA. Creating an objective structured clinical examination for the internal medicine clerkship: pitfalls and benefits. Am Med Sci. 1993;306(2):94-97.
271. Ramsey PG, Shannon NF, Fleming L, Wenrich M, Peckham PD, Dale DC. Use of objective examinations in medicine clerkships. Ten year experience. Am J Med. 1986;81:669-674.
272. Ainsworth MA, Roger LP, Markus JF, Dorsey NY, Blackwell TA, Petrusa ER. Standardized patient encounters: A method for teaching and evaluation. JAMA 1991;67:42-50.
273. Rosebraugh CJ, Speer AJ, Solomon DJ, Szauter KE, Ainsworth MA, Holden MD, Lieberman SA, Clyburn EB. Setting standards and defining quality of performance in the validation of a standardized-patient examination format. Acad Med. 1997;72(11):1012-1014.
274. Grum CM, Case SM, Swanson DB, Woolliscroft JO. Identifying the trees in the forest: Characteristics of students who demonstrate disparity between knowledge and diagnostic recognition skills. Acad Med. 1994;69(10, Suppl):S66-S68.
275. Cuddy MM, Dillon GF, Clauser BE, Holtzman KZ, Margolis MJ, McEllhenney SM, Swanson DB. Assessing the validity of the USMLE Step 2 clinical knowledge examination through an evaluation of its clinical relevance. Acad Med. 2004;79(10, Suppl):S43-S45.
276. Fincher RM, Case SM, Ripkey DR, Swanson DB. Comparison of ambulatory knowledge of third-year students who learned in ambulatory settings with that of students who learned in inpatient settings. Acad Med. 1997;69(10,Suppl):S66-S68.
277. Magarian GJ, Mazur DJ. Does performance on the NBME Part II medicine examination when used as a clerkship examination reflect knowledge acquired during the medicine clerkship? J Gen Intern Med. 1991;6:145-149.
278. Nahum GG. Evaluating medical student obstetrics and gynecology clerkship performance: which assessment tools are most reliable? Am J Obs Gyn. 2004;191(5):1762-1771.
279. Parenti CM. A process for identifying marginal performers among students in a clerkship. Acad Med. 1993;68(7):575-577.
280. Hemmer PA, Markert RJ, and Wood V. Using in-clerkship tests to identify students with insufficient knowledge and assessing the effect of counseling on final examination performance. Acad Med. 1999;74:73-75.

281. Muller ES, Harik P, Margolis MJ ,Clauser BE, McKinley DW, Boulet JR. An examination of the relationship between clinical skills examination performance and performance on USMLE Step 2. Acad Med. 2003;78(10, Suppl):S27-S29.
282. Swanson DB, Case SM. Trends in written assessment: a strangely biased perspective. In R Harden, I Hart, H Mulholland, eds. Approaches to Assessment of Clinical Assessment. Norwich, CT: Page Brothers, 1992.
283. Case SM, Downing SM. Performance of various multiple-choice item types on medical specialty examinations: types A, B, C, K and X. Proceedings of Twenty-Eighth Annual Conference of Research in Medical Education (RIME). 1989;28:167-172.
284. Case SM, Swanson DB. Extended matching items: a practical alternative to free-response questions. Teach Learn Med. 1993;5:107-115.
285. Case SM, Swanson DB, Woolliscroft JO. Assessment of diagnostic pattern recognition skills in medicine clerkships using a written test. In: Hardin R, Hart I, Mulholland H, eds. Approaches to Assessment of Clinical Competence. Norwich, CT: Page Brothers, 1992;459-464.
286. Wallach PM, Crespo LM, Holtzman DZ, Galbraith RM, Swanson DB. (2004) Use of a Cimmttee Review Process to Improve the Quality of Ciurse Examinations, Advances in Health Sciences Education, In press.
287. Case SM, Swanson DB. Improving Student Assessment: Evaluation in the Basic and Clinical Sciences. Philadelphia, PA, National Board of Medical Examiners, 1995.
288. Jozefowicz R, Koeppen BM, Case SM, Galbraith R, Swanson D, Glew RH. The quality of in-house medical school examinations. Acad Med. 2002;77:156-161.
289. Brualdi A. Traditional and modern concepts of validity. ERIC/AE Digest Series EDO-TM-99-10. ERIC Clearinghouse on Assessment and Evaluation. December, 1999.
290. Kerfoot BP, Baker H, Volkan K, Church PA, Federman DD, Masser BA, DeWolf WC. Development of validated instrument to measure medical student learning in clinical urology: A step toward evidence based education. J Urol. 2004;172:282-285.
291. Hijazi Z, Permadasa IG, Moussa AA. Performance of students in the final examination in paediatrics: Importance of the “short cases.” Arch Dis Child. 2002;86:57-58.
292. Hemmer PA, Grau T, Pangaro LN, Assessing the Effectiveness of combining evaluation methods for the early identification of students with inadequate knowledge during a clerkship. Med Teach. 2001;23(6):580-584.
293. Haladyna TM. Developing and Validating Multiple-Choice Test Items. Second edition. Lawrence Erlbaum Associates, Publishers, Mahwah, New Jersey, 1999.
294. Rudner LM. Questions to ask when evaluating tests. ERIC/EA Digest Series EDO-TM-94-06. ERIC Clearinghouse on Assessment and Evaluation, 1994.
295. Chesser AM, Laing MR. Miedzybrodzka ZH. Brittenden J. Heys SD. Factor analysis can be a useful standard setting tool in a high stakes OSCE assessment. Med Educ. 2004;38(8):825-831.
296. Cusimano MD and Rothman AL. The effect of incorporating normative data into a criterion referenced standard setting in medical education. Acad Med. 2003; 78(suppl):S88-S90.
297. Downing SM, Lieska NG, Raible MD. Establishing passing standards for classroom achievement tests in medical education: A comparative study of four methods. Acad Med. 2003;78(Suppl):S85-S87.
298. van Barnveld C. The dependability of medical students’ performance ratings as documented on in-training evaluations. Acad Med. 2005;80:309-312.
299. Coletti LM. Difficulty with negative feedback: face-to-face evaluation of junior medical student clinical performance results in grade inflation. J Surg Res. 2000;90:82-87.
300. Reteguiz JA, Crosson J. Clerkship order and performance on family medicine and internal medicine National Board of Medical Examiners exams. Fam Med. 2002;34(8):604-608.

301. Thomas PA, Shatzer JH. Standardized patient assessment of ambulatory clerks: effect of timing and order of the clerkship. Teach Learn Med. 2000;12(4):183-188.
302. Carlson DL. Clerkship evaluation and remediation. Teach Learn Med. 2004;16(4):392-393.
303. Rosenthal RH, Levine RE, Carlson DL, Clegg KA, Crosby RD. The "shrinking" clerkship: characteristics and length of clerkships in psychiatry undergraduate education. Acad Psych. 2005;29(1):47-51.
304. Association of American Medical Colleges Data Book: Statistical Information Related to Medical Education. AAMC January 2005.
305. United States Constitution, Bill of Rights. Available for review at: http://www.house.gov/Constitution/Amend.html. Last accessed 10 August 2005.
306. Regents of the University of Michigan v. Ewing, 474 U.S. 214 (1985). Available for review at: http://caselaw.lp.findlaw.com/scripts/getcase.pl?court=us&vol=474&invol=214 Last accessed 10 August 2005.
307. Goss v Lopez, 419 U.S. 565 (1975). Available for review at http://caselaw.lp.findlaw.com/cgi-bin/getcase.pl?court=US&vol=419&invol=565 Last accessed 10 Aug 2005.
308. Board of Curators of the University of Missouri v. Horowitz, 435 U.S. 78 (1978). Available for review at: http://caselaw.lp.findlaw.com/scripts/getcase.pl?court=us&vol=435&invol=78 Last accessed 10 August 2005.
309. Wozniak v. Conry, 236 F.3d 888, 891 (7th Cir., 2001).
310. Gorman v. University of Rhode Island, 837 F.2d 7 (1st Cir. 1988).
311. Bergstrom v. Buettner, 697 F.Supp 1098 (D.N.D. 1987).
312. Schaer v. Brandeis University, 735 N.E.2d 373 (Mass. 2000). Available for review at: http://biotech.law.lsu.edu/cases/schools/Schaer_v_Brandeis.htm. Last accessed 10 August 2005.
313. Helms LB, Helms CM. Forty years of litigations involving medical students and their education. Acad Med. 1991;66:1-7.
314. Sweezy v. New Hampshire, 354 U.S. 234 (1957). Available for review at: http://caselaw.lp.findlaw.com/scripts/getcase.pl?court=US&vol=354&invol=234. Last accessed 10 August 2005.
315. Brown v. Armemnti, 247 F.3d 69 (3rd Cir. 2001).
316. Introduction To The Clerkship In Internal Medicine Of The F. Edward Hebert School Of Medicine. Available at: http://www.usuhs.mil/med/clerkshiphandbook2006.pdf Last accessed 10 August 2005.
317. Gaglione MM, Moores L, Pangaro L, Hemmer PA. Does Group Discussion of Student Clerkship Performance at an Education Committee Affect an Individual Committee Member's Decisions? Accepted for publication RIME Proceedings. Acad Med. 2005;80:S55-S58.
318. http://www.lcme.org/functionslist.htm#educational%20program, accesed November 2005.
319. Ende J. Feedback in Clinical Medical Education. JAMA. 1983;250:777-781.
320. Westberg J, Jason H. 1996 Fostering Learning in Small Groups. New York, NY: Springer Publishing Company.
321. Westberg J, Jason H. Fostering Reflection and Providing Feedback; Helping Others Learn from Experience. New York, NY: Springer Publishing Company, 2001.
322. Whitman N, Schwenk T. The Physician as Teacher (2nd Edition). Salt Lake City, UT: Whitman Associates, 1997.

The Royal College of Physicians and Surgeons of Canada
Le Collège royal des médecins et chirurgiens du Canada
Office of Education Bureau de l'éducation

Developing Multiple Choice Questions for the RCPSC Certification Examinations

Editorial revisions:
June 2004

Tim Wood & Gary Cole
Educational Research and Development
Recherche et perfectionnement pédagogiques

Table of Contents

Preface

As a member of an examination board, test committee, or a specialty committee, you may be asked to develop multiple choice questions (MCQs) for an examination. This document is intended to assist you in developing MCQs that are well constructed, reliable, and valid.

The booklet consists of two sections. The first section provides a description of how to write well constructed MCQs for the Royal College and includes information about the characteristics of well constructed questions and common problems that occur when writing them. The second section contains four worksheets. The first worksheet consists of an exercise designed to help you recognize well constructed and poorly constructed questions. The second worksheet consists of item templates that are designed to facilitate the construction of MCQs. The third worksheet consists of a checklist that can also be used to facilitate the construction of questions. The fourth worksheet consists of a form that is used for submitting questions to the Royal College. These worksheets were designed to augment the information found in the first section of this booklet but can also be used as tear away worksheets to help you work on specific tasks.

I. Introduction

MCQ examinations are, arguably, the most reliable, valid, and cost effective method of assessing the clinical competence of candidates, especially for measuring their medical knowledge. From the candidates' perspective, MCQ examinations often consist of questions that are trivial, irrelevant, and ambiguous. Why would this discrepancy in the perceived value of MCQ examinations exist? The main reason is related to how the questions are typically constructed. Many MCQs are constructed to test the simple recall of textbook knowledge, or, in an attempt to make the questions difficult for candidates, test the knowledge of relatively uncommon medical conditions. In addition, questions are often constructed in a language or format that is clear to the author but is ambiguous when read by candidates.

The purpose of this booklet is to describe how to develop MCQs that are constructed in a format that candidates will find clear and relevant.

II. Characteristics of a Well Constructed MCQ

A well constructed MCQ consists of a stem, a lead-in question, and a series of response options. For example consider the following question:

> A 60-year-old man presents with progressive weakness of arms and legs. He reports difficulty climbing stairs or combing his hair. He also has difficulty swallowing, but he has no visual complaints. On physical examination, you note a maculopapular eruption on the eyelids, nose, cheeks, and knuckles. Joint examination is normal. What is the most likely diagnosis?
>
> **a. Dermatomyositis**
> b. Myasthenia gravis
> c. Polymyalgia rheumatica
> d. Rheumatoid arthritis

The first component of the question is called the stem. The stem is actually a clinical case presentation and usually consists of a presenting problem along with relevant signs, symptoms, lab tests, etc. The second component of the question is called the lead-in question. This is the actual question that the candidate is asked to answer. The last component of the question contains the response options. One of the options is chosen to be the correct answer, and the remaining options are called distractors.

This example of a well constructed question has three characteristics that should be noted. First, the question has four response options, one of which is correct. Although many types of multiple choice questions exist, the Royal College recommends the use of the "one best answer" type of question, which has one clearly correct answer and three distractors. Second, note that the question consists of a clinical situation and asks the candidate to use that information to answer the question. This approach emphasizes the application of medical knowledge and makes the question appear to be more clinically relevant and valid to the candidates. The third characteristic deals with the shape of the item. The stem of a well constructed question consists of a clinical case and should contain all of the information that is necessary to answer the question. For these reasons, the stem will tend to be long but the options should be relatively short. The figure below shows the structure of a well constructed MCQ.

Shape of a well constructed question.[1]

Long Stem: consists of a clinical case and all relevant facts.

a.
b.
c. Response Options (short)
d.

III. Steps in Constructing a Well Constructed MCQ

A. Choose a topic for the question

The topic is the theme for a specific question; that is, it is the specific medical knowledge that a question is designed to test. When choosing a topic for a question, focus on one important concept, typically a common clinical problem from your specialty.

In most cases, the topics will be given to you by your examination board chair and will be chosen from the test blueprint. A test blueprint is a guide that is used for creating an examination and consists of a list of the competencies and topics that should be tested on an examination.

B. Choose the appropriate context for the question

Context defines the clinical situation that will test the topic. Context is important because it determines what type of information should be included in the stem and the response options. Consider the following two examples.

Example 1.

Topic: Turner's Syndrome
Context: Physical Examination

An 18-year-old woman presents with primary amenorrhea. Which of the following signs best supports the diagnosis of Turner's syndrome.

a. Hypertension
b. Hirsutism
c. Short Stature
d. Epicanthal folds

[1] Case, S.M. & Swanson, D.R. (1998). Constructing written test questions for the basic and clinical sciences. (pp. 42). National Board of Medical Examiners: Philadelphia

Example 2.

Topic: Turner's Syndrome
Context: Diagnosis

An 18-year-old woman presents with primary amenorrhea. On exam you notice that she is 148 cm tall. In addition, you note that her external genitals are immature and there is no breast development. What is your most likely diagnosis?

a. Turner's Syndrome
b. Mixed Gonadal Dysgenisis
c. Pure Gonadal Dysgenisis
d. Noonan's Syndrome

Notice that both examples are testing the same topic, which is Turner's Syndrome. The context of the questions differs, however, and this difference influences the type of information that is presented in the question. For the first example, the context is a physical examination so the stem and response options contain information likely to be found during a physical exam. For the second question the context is diagnosis, so the stem contains relevant signs and symptoms and the response options consist of potential diagnoses.

Common clinical contexts that could be used for constructing an MCQ include the following: interpreting data, eliciting data (physical exam, history taking), further investigations, diagnosis, initial management, long term care, risk factors, side effects and contraindications, counseling, and ethical issues.

C. Create a stem

1. Use clinical cases

Clinical cases provide a good basis for a stem. The clinical case should begin with presenting a problem and be followed by relevant signs, symptoms, results of diagnostic studies, initial treatment, subsequent findings, etc. In essence, all the information that is necessary for a competent candidate to answer the question should be provided in the stem.

2. Use a clear lead-in question

The lead-in question should give clear directions as to what the candidate should be doing to answer the question. For example, consider the following examples of lead-in questions.

Example 1: Regarding myocardial infarction:
Example 2: What is the most likely diagnosis?

Note that for the Example 1, no task is presented to the candidate. This type of lead-in statement will often lead to an ambiguous or unfocused question. In the second example, the task is clear and will lead to a more focused question. To ensure that the lead-in question is well constructed, the question should be answerable without the candidate having to look at the response options. As a check, cover the response options and try to answer the question.

3. The content should be at an appropriate level of difficulty

Well constructed MCQs should be written at an appropriate level of difficulty to test the knowledge level of the candidates. For Royal College examinations, the questions should be designed to test the knowledge of a resident who is ready to practice their profession competently. In other words, would a specialist on his/her first day of practice know how to answer the question?

Note that testing the appropriate knowledge level of a resident does not mean that a question must be extremely difficult. If the question is testing knowledge that is essential to the practice of the specialty then the question may actually be quite easy.

4. Make the question clinically relevant

Try to focus on problems that would be encountered in clinical practice rather than assessing the candidate's knowledge of trivial facts or on obscure problems that are seldom encountered. The types of problems that you commonly encounter in your own practice can provide good examples for developing questions.

5. Test the application of medical knowledge

Well constructed MCQs should test the application of medical knowledge rather than just the recall of information. Benefits to testing the application of knowledge include the following: the question will be focused on clinically important information rather than trivia, the question will identify those candidates who have memorized factual information but are unable to use that information effectively, and, from the candidates perspective, the validity of the question will be improved. The use of a clinical case as the basis for a question will help ensure that a question tests the application of medical knowledge.

D. Create the correct answer

The correct answer should be clearly correct. If the "best answer" is sought, then this should be stated in the lead-in question.

When creating the correct answer try to avoid clues that would reveal an option as being the correct answer. Some common problems to avoid include the following:

1. the correct answer is longer than the other distractors
2. textbook wording is used for the correct answer but not for the distractors
3. specific determiners (always, never) are used in the correct answer but not in the distractors.

E. Create the distractors

A good distractor should be inferior to the correct answer but should also be plausible to a non-competent candidate. When creating a distractor, it may help to think how an inexperienced resident would respond to the clinical case described in the stem. In addition, try to avoid clues that would reveal a response option as a distractor. Some common problems to avoid include the following:

1. the distractors and the correct answer are not homogenous in content (e.g. the correct answer is a treatment, the distractors are tests).
2. the grammar of the distractors does not match the grammar of the stem.
3. the distractors are not the same length as the correct answer.

IV. Other Guidelines to Consider When Constructing an MCQ Item

A. Avoid the use of "all of the above" as a response

A candidate only has to identify two response options as correct to know that "all of the above" is the correct response. This reduces the value of the question. In addition, "all of the above" implies that there is more than one correct answer. The Royal College recommends that MCQs be constructed so that only one option is correct.

B. Avoid mutually exclusive options

For questions that require a single best answer, options that contradict one another cannot both be correct and therefore mutually exclusive options reduce the number of plausible responses.

C. Avoid overlapping content in the response options

The information in the response options should be independent of one another. For example, imagine that one had a written question about pain management in which the correct answer was to “prescribe an analgesic” and one of the distractors was “prescribe Tylenol”. There is an overlap in the content of these two response options and therefore they are not independent of one another.

D. Avoid imprecise terms like sometimes, frequently, often, etc

The definition of these terms is ambiguous and will cause confusion if used on an examination question.

E. Avoid the use of negative terms in the lead-in question (i.e. all of the following except:)

Negative terms tend to overly complicate a question. In addition, you are primarily interested in whether the candidates know the best response not necessarily the poorest response.

F. Use “none of the above” sparingly

Sometimes it is difficult to create five plausible options and therefore the response "none of the above" can be used. If “none of the above” is used, however, it should be the correct answer on at least 1/5 of the questions used on the examination and it should be clearly correct or clearly incorrect.

Worksheet 1: Examples of MCQ items

The following is an exercise to help you recognize well constructed and poorly constructed MCQs. Some of the questions that follow were well constructed and some were not. Read the questions and if you think one is poorly constructed then list the problems. The last part of this worksheet displays which items we feel were well constructed, which items we feel were poorly constructed as well as a list of problems that have been identified.

1. A 32-year-old unemployed alcoholic who underwent a mastoidectomy as a youth presents with headaches, nausea, vomiting, drowsiness, and confusion. He does not have a fever, but his right eardrum is not visualized and there appears to be some discharge. There is slight neck stiffness as well. What is the most appropriate investigation at this time?

   a. Lumbar puncture
   b. ECG
   c. X-ray the skull
   d. CT scan the head

Problems:

2. When a tendon is cut and repaired, what is the strength of repaired tissue after one year?

   a. almost always less then normal
   b. usually greater than normal
   c. almost the same as normal
   d. more or less than normal, depending on the age of the patient

Problems:

3. According to the guidelines of the American Heart Association, in what way should a patient with a prosthetic heart valve be given prophylactic antibiotic treatment before a surgical procedure?

   a. in routine fashion to everyone
   b. according to the magnitude of the procedure
   c. according to the type of microbial flora most likely to cause endocarditis
   d. only for gastrointestinal procedures

Problems:

4. A 32-year-old woman presents with a 2-week history of diarrhea associated with heat intolerance, sweating and restlessness. Physical examination reveals a blood pressure of 150/60 mm Hg and a pulse of 106/minute. She has a fine tremor of her outstretched arms. Her thyroid is diffusely enlarged, firm and tender. Which one of the following tests will help to establish the etiology of her thyrotoxicosis?

   a. Antithyroid antibodies
   b. Sensitive thyroid-stimulating hormone assay
   c. Free triiodothyronine (T_3)
   d. Radioactive iodine uptake

Problems:

5. In the management of foot ulcers in diabetics, which of the following statements about assessment for arterial revascularization of the lower limbs is TRUE?

   a. It is not beneficial because nonvascular factors, such as neuropathy and infection, minimize the benefits of revascularization
   b. It is not beneficial because artherosclerosis is too widespread for surgical correction to be beneficial
   c. It is not beneficial because arterial revascularization is too limited for surgical correction to be beneficial.
   d. It is advisable because artherosclerosis is sometimes segmental and amenable to surgical correction.

Problems:

6. Pulmonary embolism:

    a. always associated with a fever
    b. is never seen in non smokers
    c. is always confusable with Pneumonia
    d. is treated by administering Heparin

Problems:

7. A 55-year-old man presents with shortness of breath and purulent sputum. There is no history of hemoptysis or chest pain. On several occasions in the past few days he has experienced episodes of feeling hot or cold but there have been no rigors. Chest exam shows hyperinflation and decreased breath sounds without dullness or crackles but with scattered wheezes. Chest radiograph is normal. Spirometry shows the following: FEV1: 1.68 (58% predicted),FVC 2.12 (75% predicted). In managing this patient you would suggest:

    a. intravenous antibiotic
    b. oral theopylline
    c. inhaled bronchodilator therapy
    d. smoking cessation

Problems:

8. Which of the following drugs given in the setting of acute mycardial infarction has not been shown to reduce mortality?

    a. intravenous r-tissue plasminogen activator
    b. intravenous streptokinase
    c. acetylsalicylic acid
    d. nifedipine

Problems:

Potential problems with the questions

Question 1: - this is a well constructed MCQ

Question 2: - uses vague descriptors like “more or less” and “usually”.

Question 3: - the correct answer is longer than the distractors.
- one of the distractors is cued as a wrong answer. The question is asking for methods of administering a treatment and Distractor D is not a method.

Question 4: - this is a well constructed MCQ

Question 5: - the stem is vague and can’t be interpreted without reading the options
- the correct answer is cued because it is the only positive option.
- the shape of the item is incorrect. The response options are almost as long as the stem. A well constructed MCQ has a long stem and short response options.

Question 6: - the lead-in question is unclear
- uses imprecise terms (e.g. frequently)
- options are not homogenous (signs, diagnosis, risk factors, treatments)
- the first distractor is not grammatical when combined with the stem
- the shape of the item is incorrect. The response options are longer than the stem. A well constructed MCQ has a long stem and short response options.

Question 7: - the lead-in question is unclear. Is the question testing the first step in managing the patient, long term care for the patient, or the most effective treatment?

Question 8: - it is not clear what is being measured because of the negative lead-in statement. Is the question testing whether the candidate knows that nifedipine does not reduce mortality or that the other drugs do reduce mortality?

Worksheet 2: Item Templates

Constructing good MCQs can be difficult and some people find an item template to be a useful tool. Item templates are designed to have the structure of a well constructed MCQ but are missing the content of the question. As a question writer, one would choose a particular template and fill in the blanks with the appropriate information. The following item template is an excerpt from a book by Case and Swanson[2] and should prove to be useful for creating questions.

The overall structure of an item can be depicted by an item template. You can typically generate many items using the same template. For example, the following template could be used to generate a series of questions related to gross anatomy:

> A *(patient description)* is unable to *(functional disability*). Which of the following is most likely to have been injured?

This is a question that could be constructed using this template:

> A 65-year-old man has difficulty rising from a seated position and straightening his trunk, but he has no difficulty flexing his leg. Which of the following muscles is most likely to have been injured?
>
> **A. Gluteus maximus**
> B. Gluteus minimus
> C. Hamstrings
> D. Iliopsoas

Many basic science questions can be presented within the context of a patient vignette. The patient vignettes may include some or all of the following components:

> Age, Gender (eg, A 45-year-old man)
> Site of Care (eg, comes to the emergency department)
> Presenting Complaint (eg, because of a headache)
> Duration (eg, that has continued for 2 days).
> Patient History (with Family History ?)
>
> Physical Findings
>
> +/- Results of Diagnostic Studies
> +/- Initial Treatment, Subsequent Findings, etc.

[2] Case, S.M. & Swanson, D.R. (1998). Constructing written test questions for the basic and clinical sciences. (pp. 38 - 39). National Board of Medical Examiners: Philadelphia

Additional Templates

A *(patient description)* has a *(type of injury and location)*. Which of the following structures is most likely to be affected?

A *(patient description)* has *(history findings)* and is taking *(medications)*. Which of the following medications is the most likely cause of his *(one history, PE or lab finding)*?

A *(patient description*) has *(abnormal findings)*. Which [additional] finding would suggest/suggests a diagnosis of *(disease 1)* rather than *(disease 2)*?

A *(patient description)* has *(symptoms and signs)*. These observations suggest that the disease is a result of the *(absence or presence)* of which of the following *(enzymes, mechanisms)*?

A *(patient description)* follows a *(specific dietary regime)*. Which of the following conditions is most likely to occur?

A *(patient description)* has *(symptoms, signs, or specific disease)* and is being treated with *(drug or drug class)*. The drug acts by inhibiting which of the following *(functions, processes)*?

A *(patient description)* has *(abnormal findings)*. Which of the following *(positive laboratory results)* would be expected?

(time period) after a *(event such as trip or meal with certain foods)*, a *(patient or group description)* became ill with *(symptoms and signs)*. Which of the following *(organisms, agents)* is most likely to be found on analysis of *(food)*?

Following *(procedure)*, a *(patient description)* develops *(symptoms and signs)*. Laboratory findings show *(findings)*. Which of the following is the most likely cause?

A *(patient description)* dies of *(disease)*. Which of the following is the most likely finding on autopsy?

A patient has *(symptoms and signs).* Which of the following is the most likely explanation for the *(findings)*?

A *(patient description)* has *(symptoms and signs)*. Exposure to which of the *(toxic agents)* is the most likely cause?

Which of the following is the most likely mechanism of the therapeutic effect of this *(drug class)* in patients with *(disease)*?

A patient has *(abnormal findings)*, but *(normal findings)*. Which of the following is the most likely diagnosis?"

Worksheet 3: Checklists for the Development of MCQs

A Follow these five steps when developing a question.

1. Choose a topic for the question ☐
   Topics are the specific knowledge that the question is designed to test. They are related to the competencies that the examination should be testing.
2. Decide on the context for testing the objective ☐
   Context is the clinical situation (interpreting data, diagnosis, management) that will determine what information should be provided in the question.
3. Write the stem of the question ☐
   The stem should use a clinical case as the basis of the question. It should also contain all relevant information necessary to answer the question and should end with a clear question.
4. Create four response options ☐
   Choose one option to be the correct answer and this option should be clearly correct. The remaining three options are called distractors and should be clearly incorrect but plausible to a weaker candidate.
5. Try the question on a colleague ☐
   Other people often notice problems that the author may have missed or not considered when writing the question.

B Use this checklist when creating a stem.

1.	Is the question related to a topic from the blueprint?	Yes
2.	Is the stem relevant to clinical practice?	Yes
3.	Was a clinical case used as the basis for the question?	Yes
4.	Does the stem consist of all information that a competent candidate will require to answer the question?	Yes
5.	Does the lead-in question clearly indicate how to answer the question?	Yes
6.	Is the stem written at an appropriate level of difficulty?	Yes
7.	Can the question be answered without looking at the options?	Yes
8.	Does the question test the application of medical knowledge rather than recall?	Yes

Worksheet 3 (continued)

C. Use this checklist when creating the response options.

1.	Is there one clearly correct answer?	Yes
2.	Are all the distractors plausible to a weak candidate?	Yes
3.	Are there any obvious clues to the correct answer (all options are homogenous, grammatical, same length, same degree of technical language)?	No
4.	Were terms like "all of the above", and "all of the following except" used?	No
5.	Were terms like “none of the above” used sparingly and if so were the items occasionally correct?	Yes
6.	Were any imprecise terms (frequently, sometimes, often) used?	No

D. Does the item have an appropriate shape?

The stem of a well constructed MCQ usually consists of a clinical case and should contain all of the information necessary to answer the question. For these reasons, the stem will tend to be relatively long but the options should be relatively short. The figure below presents a diagram of how a well constructed MCQ should look.

Shape of a well constructed question.[3]

Long Stem: consists of a clinical case and all relevant facts.

a.
b.
c. Response Options (short)
d.

[3] Case, S.M. & Swanson, D.R. (1998). Constructing written test questions for the basic and clinical sciences. (pp. 42). National Board of Medical Examiners: Philadelphia.

Worksheet 4: Item Submission Form

NAME: _____

SPECIALTY: _____

TOPIC: _____

In the space below please **type** your question**

CORRECT ANSWER:

KEYWORDS:

REFERENCE: Journal - Author(s), volume, page(s) and year
Book - Title, edition, page(s)

CLASSIFICATION:

** ITEMS MUST BE CLASSIFIED AND HAVE A REFERENCE

Multiple Choice (MC) Item Writing Guidelines for Royal College Exams

Content concerns

1. Avoid trick items.
2. Base each item on important content to assess; avoid trivial content.
3. Use relevant material to test higher level learning such as the inclusion of clinical settings. Avoid testing for simple recall.
4. Every item should reflect specific content area as defined by an exam blueprint.
5. Keep the content of each item independent from content of other items on the test.
6. Avoid over specific and over general content when writing MC items.
7. Do not create opinion-based items.
8. Keep vocabulary appropriate for the group being tested.
   - Avoid the use of acronyms.
   - Use nationally accepted terms common to the specialty.
   - Include both SI and traditional Imperial measures when appropriate.
9. The purpose of Royal College examinations is to assess competence. Some questions should be designed to discriminate between competent and non-competent candidates whereas others may be mastery-level questions (questions that test knowledge that all competent candidates should know).

Writing the stem

10. Ensure that the directions in the stem are very clear.
11. Include the central idea in the stem instead of the choices.
12. Avoid window dressing such as excessive verbiage or unnecessary "red herrings".
13. Word the stem positively, avoid negatives such as NOT or EXCEPT. If negative words are used, use the word cautiously and always ensure that the word appears CAPITALIZED and **boldface**.

Writing the choices

14. Develop as many effective choices as you can. Although research suggests three options may be adequate, Royal College examinations require 4 options including only one correct answer.
15. Make all three distractors plausible yet definitively incorrect.
16. Vary the location of the right answer according to the number of choices.
17. Place choices in logical or numerical order.
18. Keep choices independent; choices should not be overlapping.
19. Keep choices homogeneous in content and grammatical structure.
20. Keep the length of choices about equal.
21. *None of the above* should be used carefully.
22. Avoid *All of the above*.
23. Use typical errors of candidates to write your distractors.
24. Avoid giving clues to the right answer, such as
    a. Specific determiners including always, never, completely, and absolutely.
    b. Clang associations, choices identical to or resembling words in the stem.
    c. Grammatical inconsistencies that cue the test taker to the correct choice.
    d. Conspicuous correct choice.
    e. Pairs or triplets of options that clue the test taker to the correct choice.
    f. Blatantly absurd, ridiculous options.
25. Given the high-stakes nature of Royal College examinations, do not use humour.

Formatting concerns

26. Ensure that the Type A format is used (with 4 options and only one correct answer). Do not use the True/False, matching, multiple-correct answer or complex Type K format (that tests logic and reading skills rather than content knowledge).
27. Format the item vertically instead of horizontally.

Style concerns

28. Edit and proof items.
29. Use correct grammar, punctuation, capitalization, and spelling.
30. Minimize the amount of reading in each item.

MCQ Item Writing Guidelines based on:
Haladyna, T., Downing, M., & Rodriguez, M. (2002). A *review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment.* Applied Measurement in Education, 15(3), 309-334.