**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; "Competencies'
- 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 school 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:

<table>
<thead>
<tr>
<th>Educational process</th>
<th>Aristotle</th>
<th>Clinical Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assessment = making observations</td>
<td>Observation</td>
<td>History and Physical diagnosis</td>
</tr>
<tr>
<td>about learners</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluation = determining learner's</td>
<td>Reflection</td>
<td>Diagnosis</td>
</tr>
<tr>
<td>success for expected level</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grading/Feedback = taking an action</td>
<td>Action</td>
<td>Therapy</td>
</tr>
<tr>
<td>- administrative/societal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- educational intervention</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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. (Battistone)

**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
**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 (Gage, 1994). 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 I scores (Durning, 2003).

**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 students performance into a step-wise, such as, second-year level, third-year level, fourth-year level, internship level; or, reporter, interpreter, manager/educator in parenthesis, Pangaro, 1979; Battistone, 2002).

**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 criterion-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 wound closure, the number of sutures used, and the time taken to do the suturing. The examiner then decides whether the standard has been met, and how well (cries 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. A survey by the Clerkship Directors in Internal Medicine revealed that on average, 62% of the clerkship grade was derived from narrative clinical evaluations and the remainder was from final examination scores. (Magarian and Mazur, 1990)

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" may be a better term for MCQs or OSCEs (Norman, 1991) 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 (Case and Swanson, 1996).

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 (Roop, Pangaro, 2001; Lavin, Pangaro, 1999). 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 (Pangaro, 1999) 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. 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 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.
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 (Pangaro, 2000). 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) and rely on summary observations of teachers to see whether they can put things together.

"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. 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. 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 Chapter I3: 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
Reliability and Validity; Feasibility

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, (Walsh and Betz, 1990).

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 (Haynes, 2001). 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. Construct 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 on these subjects. (Downing, 2003, 2004, 2004)

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 the priorities. Nonetheless, it is preferable to develop reliable and valid tools; then try to make them work.

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.
References


Gage, NL Hard Gains in the Soft Sciences (Bloomington, Ind.: CEDR, Phi Delta Kappa, 1985).


Pangaro L, A New Vocabulary and Other Innovations for Improving Descriptive In-training Evaluations, *Academic Medicine* 74: 1203-1207, 1999