# A New Vocabulary and Other Innovations for Improving Descriptive In-training Evaluations

*Louis Pangaro, MD*

**Abstract:** Progress in improving the credibility of teachers' descriptive evaluations of students and residents has not kept pace with the progress made in improving the credibility of more quantified methods, such as multiple-choice examinations and standardized patient examinations of clinical skills. This article addresses innovative approaches to making the ongoing in-training evaluation (ITEv) of trainees during their clinical experiences more reliable and valid. The innovations include the development of a standard vocabulary for describing the progress of trainees from "reporter" to "interpreter" to "manager" and "educator" (RIME), the use of formal evaluation sessions, and closer consideration of the unit of clinical evaluation (the case, the rotation, or the year). The author also discusses initial results of studies assessing the reliability and validity of descriptive methods, as well as the use of quantified methods to complement descriptive methods. Applying basic principles—the use of a taxonomy of professional development and statistical principles of reliability and validity—may foster research into more credible descriptive evaluation of clinical skills. *Acad. Med.* 1999;74:1203–1207.

For at least a decade medical educators have been increasing their use of more highly quantified tools to evaluate students and residents in many disciplines.[1–3] This has decreased the teacher's role in certifying professional competence and increased the program director's role through the use of in-training examinations (ITEs, and specifying an extended multiple-choice test) or the NBME subject examination, and objective structured clinical examinations (OSCEs). This increased use of highly quantified assessment tools has yielded important gains, including a sense of "objectivity" that comes with such measurement, a sense of institutional accountability for the competence of an individual learner, and a set of outcomes that might be used to validate the curriculum that produces the graduate. In this article I focus on teachers' "descriptive" evaluations of trainees (i.e., those that use words) and on methods for making descriptive evaluations more reliable, valid, useful, and feasible—in other words, ways to enhance their "credibility."

## Standards for Descriptive Evaluations

This article explores how the basic science of measurement can be applied to improving teachers' descriptive evaluations of learners, as well as how quantified measures can supplement our written pictures of trainees' competence. The goal is to achieve credible evaluation and thereby to meet our obligations to society at large, to students, and to teachers. Emphasis is on those methods that can feasibly be applied in real time by teachers. In particular, I describe what might constitute "in-training evaluation (ITEv)"—ways to improve teachers' descriptions of trainees' performances (using words) and their use of performance grids and (sometimes) global rating forms. I use the term ITEv to distinguish this kind of evaluation from in-training examinations (ITEs) and other quantified tests, such as OSCEs. These are certainly evaluation tools, but they are usually administered by the program director and given only on occasion. The ITEv, as I use the term, provides both ongoing formative evaluation and documentation for subsequent, summative grading. To have credibility, any ITEv system must meet standards of reliability and validity, and it should also be feasible.

**Reliability** is the consistency, stability, or repeatability of results. Technically, reliability is the percentage of the observed variance that is due to true score variance rather than error variance. In other words the "signal" (what we want to measure) should be sufficiently greater than the "noise" (problems with the assessment tool) that we can trust it. Usually, we want at least 80% of the variance to be true score variance (a reliability figure of .8) for high-stakes decisions.

**Validity** is the confidence that we are measuring what we want to measure, and includes, among other issues, both content validity (does our assessment method reflect enough of the domain in question?) and predictive validity (will present results be reflected in subsequent performance?).

**List 1**

**The RIME Framework for Student Progress**

| | |
|---|---|
| **Reporter:** | Consistently good in interpersonal skills; reliably obtains and communicates clinical findings |
| **Interpreter:** | Able to prioritize and analyze patient problems |
| **Manager:** | Consistently proposes reasonable options incorporating patient preferences |
| **Educator:** | Consistent level of knowledge of current medical evidence; can critically apply knowledge to specific patients |

Whether an evaluation method can actually be conducted in our own setting—do we have the time, money and space to do it—is a question of **feasibility**. For descriptive evaluations by teachers, their willingness to actually use the evaluation form or method may be critical. Often this is the limiting step in what we actually choose to do for our trainees, but it is desirable to first look for reliable and valid tools, and then try to make them work.

## A "Synthetic" Vocabulary of Clinical Progress

A basic science of education requires a usable vocabulary for what to evaluate,[4] but, unfortunately, there has been no generally accepted terminology or vocabulary for describing a learner's stage of competency or progress toward independence. Most rating forms used for ITEv usually employ an "analytic" evaluation system, dividing the learner's competence into skills, knowledge, and attitudes. There is evidence from many clinical settings that faculty do not apply descriptive rating forms with consistency or discrimination.[5–8] At worst, this has led to a commonly-held conception that evaluations by teachers are "subjective." Teachers and program directors need to develop or adopt a uniform terminology for ITEvs if these evaluations are to be consistent and useful.

In the Department of Medicine, at the Uniformed Services University of the Health Sciences (USUHS) we describe the progression of trainees using the following terminology: Reporter, Interpreter, Manager, Educator ("RIME") (see List 1). This framework emphasizes a developmental approach, and distinguishes between basic and advanced levels of performance for both ward and clinic rotations. Such a system is "synthetic" rather than "analytic," and each step represents a synthesis of skills, knowledge, and attitudes that have been practiced from the preclinical years of medical school through residency. Since printed evaluation forms are, essentially, just ways of conveying goals to teachers and trainees, a valid method of evaluation relies on the willingness and ability of the community to use it,[5] and a prime virtue of this "RIME" terminology is its portability and ease of use by all teachers, not just by clerkship and program directors. I describe the specific meaning of each term below.

**Reporter.** At the "reporter" level, the trainee can accurately gather and clearly communicate the clinical facts about his or her own patients. Mastery of this step requires the basic skills to obtain a history and do a physical examination and the basic knowledge of what to look for. This descriptor emphasizes day-to-day reliability—for instance, being on time, or following up on a patient's progress. The trainee at this stage has a sense of responsibility and is achieving consistency in bedside skills in interpersonal relationships with patients. These skills are often introduced to students in their preclinical years, but they should be mastered as a "passing" criterion in the third or fourth year of medical school. Certainly, a resident who is *not* a reliable reporter should be given immediate and clear feedback about performance standards required to pass.

**Interpreter.** Making a transition from "reporter" to "interpreter" is an essential and often difficult step in the professional growth of a trainee. At a basic level, the student must be able to prioritize among problems identified in his or her time with the patient. The next step is to offer a differential diagnosis. Follow-up of tests provides another opportunity to "interpret" the data (especially in the clinic setting). This interpreter step requires a higher level of knowledge, and more skill in selecting the clinical findings that support possible diagnoses and in applying test results to specific patients. To move from "reporter" to "interpreter," the learner has to make the transition, emotionally, from being a "bystander" to seeing himself or herself as an active participant in patient care. Most faculty would regard consistency as an interpreter of common medical problems as a passing criterion for interns.

**Manager.** This step takes even more knowledge, more confidence, and more judgment in deciding when action needs to be taken, and proposing and selecting among options for patients. At this stage a trainee must be able to tailor the plan to the particular patient's circumstances and preferences; this requires higher-level interpersonal skills, including the skills needed to educate patients. In procedural or operative specialties, technical and manual skills fit in here, but proficiency in them would not outweigh deficiencies as a reporter or interpreter.

**Educator.** Success in each prior step already depends on self-directed learning and a mastery of basics. To be an "educator" in the RIME framework means a resident must be able to go beyond the required basics, to read deeply, and to share new learning with others. It also means having the insight to define important questions to research in more depth, the drive to look for hard evidence on which clinical practice can be based, and the skill to know whether the evidence will stand up to scrutiny. The advanced trainee,

also has the maturity and confidence to share in educating the team (and even the faculty).

## Moving from Stating Goals to Using Them

Teaching teachers to evaluate students more consistently and according to departmental rather than personal guidelines requires supervision and faculty development initiatives. This can be accomplished in several ways, but the most effective we have found at USUHS has been formal evaluation sessions, where a clerkship director sits down at regular intervals with teachers to evaluate student performances.[9] Combined with the descriptive RIME vocabulary outlined above, we have achieved in the in-training evaluation of medical students on an internal medicine clerkship a reliability of greater than 0.8 (which is sufficient for high-stakes decisions[10]) and a strong predictive validity for ratings by internship directors.[11] Our system of formal evaluation sessions permits trained clerkship coordinators to guide evaluators in applying departmental guidelines consistently. With an additional investment of 15 minutes per student, individual feedback is given by clerkship directors the next day (and response to feedback becomes a subject of future evaluation). The sessions have doubled our sensitivity in detecting students' knowledge deficiencies compared with written evaluations.[12] Even more important, the interaction *develops* housestaff and faculty as observers of student performances, giving *them* feedback on their evaluations. Finally, the system allows for an action plan to be formulated during the session—e.g., adjusting a student's patient load, or coaching another student in case presentation. We believe this evaluation system meets the criteria for credibility (see List 2) and inter-school studies are currently in progress through the Project on Reliable and Valid Assessment of the Group on Educational Affairs of the Association of American Medical Colleges.

**List 2**

**Characteristics of Credible Clinical Evaluation for Trainees**

**Formative evaluation**
- should be based on direct observations of teachers
- should guide teachers in accurately reflecting departmental goals
- should reliably occur at key points in rotation
- should develop a plan for working with teachers to improve skills
- should reflect the framework of the final evaluation

**Summative evaluation**
- should be based on multiple observers or observations
- should be consistent across sites, teachers, rotations
- should accurately reflect institutional goals
- should document mastery or deficiency of core goals

## The Unit of Evaluation: the Rotation, the Case, the Year

### The Rotation

Most clinical experiences for students and residents give the trainee patient-care responsibilities under the supervision of a faculty member for a period of four to eight weeks, and this has become the most common unit of evaluation for ITEv. Most rotations collect written performance evaluations for students and residents, and these, collectively, can achieve reliability provided a sufficient number are obtained. Generally, a minimum of seven written forms must be obtained to have confidence in the reliability of the evaluation.[13] However, this number is often not feasible, the technique does not ensure the validity of the process, and any credibility for formative evaluation is compromised since the results are in only after the learner has finished the rotation. Hence, mid-rotation evaluation is essential at the student level, and desirable at the postgraduate level.

### The Case

In a sense, a global rating represents the sum of a teacher's experience of a learner's case work during a rotation. One feasible alternative to having the teacher summarize and integrate all these observations in a single final evaluation is to document the learner's level of proficiency for *each* case presented to a teacher. This could be done using a brief performance checklist of four to seven items[14] appropriate to the case at hand and the learner's level of training, or asking the teacher to characterize the level achieved by the trainee as "reporter," "interpreter," or "manager–educator" (the two stages are closely linked). A potential advantage of using the case as the unit of evaluation is that a large number of instances or observations may be generated for each trainee, allowing for the achievement of the reliability needed for high-stakes decisions. This approach may also relieve teachers of the pressures that lead to grade inflation, since no single grade yields the final recommended grade. This kind of ongoing documentation of individual professional behaviors would yield reliable evaluations that would hold great promise for validly predicting future performance.

Specific "educational products" of the trainee can be similarly reviewed on a case-by-case basis: an audit of residents' written chart notes can achieve acceptable reliability as well as prompt an improvement in subsequent performance.[15] Individual case write-ups by students, reviewed by preceptors with a standardized checklist, have shown strong agreement

(.85) in ratings with a panel blinded to the student's identity and time of year.[16]

### The Year

Finally, it is possible to consider an entire year of training as the unit of evaluation. This is typically the approach taken for interns and residents; the program director weighs evaluations for all rotations and decides who will advance to the next year. In this model, ITEvs for individual rotations are only formative and provisional. For evaluations of professionalism or discipline-specific areas both sensitivity (enhanced ability to detect trainees who have problems) and specificity (less labeling of those who don't) may be enhanced by the withholding of summative evaluation until a year's worth of observations have been achieved.[17]

## Supplementary ITEv Tools

I believe that the ongoing description of a trainee's performance by teachers should be the core of evaluation and feedback with respect to professional growth and performance. However, two other evaluation tools can be very valuable supplements to ITEv, particularly in the evaluation of knowledge, reasoning ability, and bedside skills in interviewing, physical examination, and communication. Provided they can feasibly be done in a time frame that allows for feedback (and that they do not become a way for teachers to avoid fulfilling their own role in evaluation), these methods, which are briefly described below, can be very useful. Since they test "competence" (what a student can do under test conditions) rather than "performance" (what a student does habitually) they are important, even necessary, but not sufficient for overall evaluation.

**Multiple-choice tests.** Residency programs have used multiple-choice ITEs, typically those prepared by their certifying boards, for approximately decades. The inability of faculty to consistently predict residents' funds of knowledge[3,18] has made these tests a mainstay of trainee evaluation. Their reliability, validity, feasibility, and acceptability are well documented.[19–21] If the unit of evaluation is the academic year, or even the entire residency, these ITEs are truly formative, and could be part of ITEv. Faculty members' abilities to detect students' knowledge deficiencies can be doubled by the use of formal evaluation sessions, although the detection rate remains less than 50%. Overall, though, there is high "specificity" in a teacher's observation that a learner's fund of knowledge is weak, and such an observation is certainly sufficient to justify feedback.[12,18] Most important, we can see performance on a multiple-choice examination as reflecting much more than single item (knowledge) in the analytic attitude–skills–knowledge construct. From a synthetic perspective, such examinations probably test a final common pathway for multiple skills and behaviors, such as the learner's time management, ability to abstract patterns and recognize key features, eagerness to learn, etc. Thus, repeated use of a multiple-choice examination throughout the school year as a progress test[22] or a quarterly profile examination[23] can provide useful, ongoing formative evaluation of multiple competencies.

**OSCEs.** The use of objective structured clinical examinations (OSCEs) of learners' skills as a form of ITEv,[24–26] especially with standardized patients, has been growing in both medical school and residency settings. However, OSCEs, like multiple-choice ITEs, remain under the control of the program or clerkship director, and are rarely used by individual teachers in the real-time formative evaluation of residents and students. Their initial costs can be substantial,[27] but they can feasibly be part of the yearly comprehensive evaluation of competence.

## Conclusion

The very term "in-training" implies an ongoing process in which the observations from ITEvs will be used as formative assessment to generate feedback. Achieving credible evaluation that is reliable, valid, and feasible is an essential task for those who supervise clinical programs for students and residents, and this is true for both formative and summative evaluation. Developing teachers as competent evaluators is essential for generating more frequent and more useful feedback about their progress for students and residents. Other strategies and methods promise to improve the reliability and validity of the descriptive in-training evaluations provided by teachers; these include formal sessions with teachers in ongoing case-based discussions of teaching (with the learner as the "case"), documenting the performances of trainees for each case they participate in, and employing the simple, portable terminology of professional development (RIME) that I described above.

*Dr. Pangaro is professor, Department of Medicine, Uniformed Services University of the Health Sciences, Bethesda, Maryland.*

*Address correspondence and requests for reprints to Dr. Pangaro, Department of Medicine–EDP, USUHS, Bethesda, MD 20814-4799; e-mail: ⟨loupang@aol.com⟩.*

## References

1. Holmboe ES, Hawkins RE. Methods for evaluating the clinical competence of residents in internal medicine: a review. Ann Intern Med. 1998;129:42–8.
2. Hilliard RI, Tallett SE. The use of an objective structured clinical examination with postgraduate residents in pediatrics. Arch Pediatr Adolesc Med. 1998;152:74–8.

3. Wise S, Stagg PL, Szucs R, Gay S, Mauger D. Assessment of resident knowledge: subjective assessment versus performance on the ACR in-training examination. Acad Radiol. 1999;6:66–71.
4. Bloom BS. Taxonomy of Educational Objectives, Handbook 1, Cognitive Domain. New York: Longman, 1956.
5. Metheny WP. Limitations of physician ratings in the assessment of student clinical performance in an obstetrics and gynecology clerkship. Obstet Gynecol. 1991;78:136–41.
6. Resnick RK, Blackmore D, Cohen R, Baumber J, Rothman A, et al. An objective structured clinical examination for the licentiate of the Medical Council of Canada: from research to reality. Acad Med. 1993; 68(10 suppl):S4–6.
7. Thompson WG, Lipkin M Jr, Gilbert DA, Guzzo RA, Roberson L. Evaluating evaluation: assessment of the American Board of Internal Medicine Resident Evaluation Form. J Gen Intern Med. 1990;5:214–7.
8. Noel GL, Herbers J, Caplow M, Cooper G, Pangaro L, Harvey J. How well do internal medicine faculty members evaluate the clinical skills of residents? Ann Intern Med. 1992;117:756–65.
9. Noel GL. A system for evaluating and counseling marginal students. J Med Educ. 1987;62:353–5.
10. Pangaro LN, Jamieson T, Hemmer P, Gibson KF, DeGoes JJ. Descriptive clinical evaluation can achieve reliability comparable to standardized tests. Presented at the Association for Medical Education in Europe Conference 1997, Vienna, Austria.
11. Lavin B, Pangaro L. Internship ratings as a validity outcome measure for an evaluation system to identify inadequate clerkship performance. Acad Med. 1998;73:998–1002.
12. Hemmer P, Pangaro L. The effectiveness of formal evaluation sessions during clinical clerkships in better identifying students with marginal funds of knowledge. Acad Med. 1997;72:641–3.
13. 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;7:506–10.
14. Turnbull J, Gray J, MacFayden J. Improving in-training evaluation programs. J Gen Intern Med. 1998;13:317–23.
15. Holmboe E, Scranton R, Sumption K, Hawkins R. Effect of medical record audit and feedback on residents' compliance with preventive health care guidelines. Acad Med. 1998;73:901–3.
16. Pangaro L, Gibson K, Russell W, Lucas C, Marple R. A prospective, randomized trial of a six-week ambulatory internal medicine rotation. Acad Med. 1995;70:537–41.
17. Papadakis MA, Osborn E, Cooke M, Healy K. A strategy for the detection and evaluation of unprofessional behavior in medical students. Acad Med. 1999;74:980–90.
18. Hawkins RE, Sumption KF, Gaglione MM, Holmboe ES. The in-training examination in internal medicine: resident perceptions and lack of correlation between resident scores and faculty predictions of resident performance. Am J Med. 1999;106:206–10.
19. Sloan DA, Donnelly MB, Schwartz RW, Felts JL, Blue AV, Strodel WE. The use of objective structured clinical examination (OSCE) for evaluation and instruction in graduate medical education. J Surg Res. 1996; 63:225–30.
20. Cox SM, Herbert WN, Grosswald SJ, Carpentieri AM, Visscher HC, Laube DW. Assessment of resident in-training examination in obstetrics and gynecology. Obstet Gynecol. 1994;84:1051–4.
21. Webb LC, Sexson S, Scully J, Reynolds CF, Shore MF. Training directors' opinions about the psychiatry resident in-training examination (PRITE). Am J Psychiatry. 1992;149:521–4.
22. Albano MG, Cavallo F, Hoogenboom R, et al. An international comparison of knowledge level of medical students: the Maastricht Progress Test. Med Educ. 1996;30(40):239–45.
23. Arnold L, Willoughby TL. The "Quarterly Profile Examination." Acad Med. 1990;65:515–6.
24. Dupras DM, Li JT. Use of an objective structured clinical examination to determine clinical competence. Acad Med. 1995;70:1029–34.
25. Schwartz RW, Donnelly MB, Sloan DA, Johnson SB, Strodel WE. The relationship between faculty ward evaluations, OSCE and ABSITE as measures of surgical intern performance. Am J Surg. 1995;169:414–7.
26. Hamadeh G, Lancaster C, Johnson A. Introducing the objective structured clinical examination into a family practice residency program. Fam Med. 1993;25:237–41.
27. Reznick RK, Smee S, Baumber JS, et al. Guidelines for estimating the real cost of an objective structured clinical examination. Acad Med. 1993;68:513–7.