Medical Teacher, Vol. 22, No. 5, 2000

Investing in descriptive evaluation: a vision for the future of assessment

LOUIS N. PANGARO
Uniformed Services University of the Health Sciences, Bethesda, MD, USA

Introduction

Over the next decades, assessment will emphasize authentic methods that focus on the way in which trainees and physicians conduct themselves in the real, not simulated, care of patients. This evolution is desirable, but not inevitable, and it will require resources and interest comparable to that given in the last decades to standardized patient methods. There will be increasing trust of assessment methods which recognize the primacy of evaluations by teachers and supervisors, an increasing reliance on a descriptive vocabulary for clinical competence, and, it is hoped, an escape from the tyranny of numbers over words. While the exciting work done in quantified performance assessment—such as that using standardized patients to evaluate interviewing, counseling or physical examination skills, and that using interactive computer programs to evaluate decision-making skills—will continue, these methods will be recognized as impractical for daily formative evaluation and feedback, and they will be seen as supplements to summative faculty judgments, necessary but certainly not sufficient to pronounce a student or postgraduate trainee as competent for the next level of responsibility. I see a search for more authentic assessment of competence of trainees in real time, in which we can meet our responsibility to all three of our constituencies: the public, the student and the teacher. So, in this paper I will describe and justify desirable goals in the field of assessment for the next decades, barriers to achieving them, and an overall strategy for giving descriptive evaluation the rigor that has been achieved elsewhere. To repeat, this is the direction in which we *can* go in the next decade, where we *should* go, and not necessarily where we will go.

I shall use the term 'descriptive' evaluation to apply to that using, primarily, *words* to summarize a student's level of competence, in other words to describe it. This is contrasted with assessment techniques whose summary encapsulation of achievement yields a score, typically a number; these we can call 'quantitative'.

Desirable assessment goals

Over the coming years it will be essential that we see high standards of reliability and validity applied to descriptive evaluation of students by teachers on a day-to-day basis. On rotations in the clinical setting, what we call 'clerkships', these *in vivo* assessments are often given a substantial weight in determining a student's grade. A survey of internal medicine clerkship directors established that the majority of a student's grade in the medicine clerkship (on average, 62%) is based on teacher's evaluations, but the same study revealed that clerkship directors typically felt that they had less confidence in this component of grading than in the end-of-clerkship examinations, which we might call *in vitro*. More educational research into descriptive assessments from teachers will achieve a more credible evaluation of competence and therefore help meet our responsibility to society. But, additionally, it will also enhance ongoing formative evaluation so that students will be given feedback to help them improve continuously. Certainly, this will help meet our obligations to students. Finally, this will depend upon improving the evaluation skills of teachers. We will have to give teachers better skills and tools to describe, assess and communicate what we expect of students, and this will certainly depend on a better vocabulary to describe the progress expected of trainees as they advance in clinical responsibility.

Generally, educators have used an analytic approach to defining expectations, that is, we separate competence into separate parts called skills, knowledge and attitudes. Recently, the Association of American Medical Colleges (AAMC) has initiated the Medical Schools Assessment Project (MSOP; AAMC, 1999) which divides competence into four components: duty, altruism, skills and knowledge. The MSOP illustrates the power of this approach to guide those who prepare and manage curriculum for trainees. It specifies both goals within each domain, and objectives that can be used to implement the goals. Under each of these domains; there may be five to ten components. An analytic model (from the Greek, ana-lysis, to loosen up) breaks up competence into components. My own clerkship evaluation form has 15 areas of competence spread out over five domains; but is it practical to ask teachers and students to carry in their heads dozens of objectives to be met during training? The division of competence into these components is quite useful for those of us who plan and manage an educational curriculum. But I have become convinced that just having a detailed evaluation form, with detailed prescription of objectives and goals, in no way assures that students and teachers can use it effectively. In fact, this detailed listing of expectations is merely a way of stating goals, and is not an evaluation tool at all. Consistent evaluation of students in their natural habitat, clinical rotations, will require an approach that is more 'portable', and yet still encompasses the complexity of developing competence.

What we need is a credible terminology that can be used readily by students and teachers so that, across different disciplines (medicine, surgery, pediatrics, etc.), in different

Correspondence: Louis N. Pangaro, MD, Professor of Medicine, Uniformed Services University of the Health Sciences, 4301 Jones Bridge Road, Bethesda, MD 20814-4799, USA. Tel: 202 782 4924; fax: 202 782 7363; email: loupang@aol.com

ISSN 0142-159X (print)/ISSN 1466-187X (online)/00/050478-04

hospitals and training sites, and from year to year, there can be 'reliable' evaluation, that is, one in which there would be consistency among evaluators. This would not only achieve credible summative evaluation at the end of a clerkship (or at the end of a year), but also credible formative evaluation, that is, feedback which anticipates their final grading criteria. In other words, we need a vocabulary that reflects growing expertise. In medical education, we still do not have a universally agreed-upon vocabulary to describe the final goal of our curricula: competence. For purposes of this paper, I will use competence to mean a *synthesis* of all the attributes necessary to do the task for which one is being trained. Competence, then, is the ability to give to a specific situation all that properly belongs to the situation and no more. It is both synthetic and selective, encompassing and discriminating. It takes place in real time, and on a day-to-day basis.

The Department of Medicine at the Uniformed Services University has developed a vocabulary that is 'synthetic' (etymologically, 'putting things together'). This synthetic vocabulary encompasses the growing responsibilities of a trainee developing competence, from early clinical experiences in medical school to the more demanding tasks of a resident (or registrar) in postgraduate training; we refer to the progress from 'reporter' to 'interpreter' to 'manager-educator' (Pangaro, 1999) and refer to it as the 'RIME scheme' for short. The underlying construct is the same as for most, perhaps all, human productivity—observation, reflection, action—but couched in terms that have been quite successful with students and teachers with varying degrees of experience. Each step represents a synthesis of abilities:

- '*Reporter*': the student can accurately gather and clearly communicate the clinical facts on his/her own patients. Mastery in this step requires the basic skill to do a history and physical examination and the basic knowledge to know what to look for. It emphasizes day-to-day reliability, for instance, being on time, or follow-up of a patient's test results. Implicit in the step is the ability to recognize normal from abnormal and the confidence to identify and label a new problem. This step certainly requires a sense of responsibility, and achieving consistency in 'bedside' skills in dealing directly with patients and other professionals. These skills are often introduced to students in their pre-clinical years, but now they must be mastered as a 'passing' criterion.
- '*Interpreter*': Making a transition from 'reporter' to 'interpreter' is an essential step in the growth of a third-year student, and often the most difficult. At a basic level, the student must prioritize among problems identified during his/her time with a patient. The next step is to offer a differential diagnosis. Because a public forum can be intimidating to beginners, we define success for students as offering at least three reasonable diagnostic possibilities for new problems. Follow-up of tests provides another opportunity to 'interpret' the data (especially in the clinic setting). Postgraduate trainees, on the other hand, should show increasing accuracy as they have seen more instances of specific problems. This 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. The student has to make the transition, emotionally, from 'bystander' to see him/herself as an active participant in patient care.
- '*Manager*': This step takes even more knowledge, more confidence and more judgment in deciding when action needs to be taken, and to propose and select among options for patients. Once again we do not require students to be 'right' as often as our residents, so we ask them to include at least three options in their diagnostic and therapeutic plan. A key element is to tailor the plan to the particular patient's circumstances and preferences.
- '*Educator*': Success in each prior step depends on self-directed learning, and on a mastery of basics. To be an 'educator' in our framework means to go beyond the required basics, to read deeply, and to share new learning with others. Defining important questions to research in more depth takes insight. Having the drive to look for hard evidence on which clinical practice can be based, and having the skill to know whether the evidence will stand up to scrutiny are qualities of an advanced trainee; to share leadership in educating the team (and even the faculty) takes maturity and confidence.

I would use the word "performance" to mean what the trainee does when under specific test conditions and when he or she knows that he is 'onstage'. It may be useful to distinguish 'performance' from 'performing'. 'Performance', as I said, is what one does under specific conditions, for instance, during a test or while being watched. But 'performing' is ongoing and continuous; the word itself indicates activity rather than a finished product. I make the distinction because the term 'performing' is much closer to what we need to assess. To know that a student is competent we need to observe the student performing *in vivo*, not an isolated performance under *in vitro* test conditions. Another way of saying this is that the assessment of competence requires a whole series of performances: that in each moment of interaction with a patient, the competent physician must bring many qualities to bear, and what they will require varies from patient to patient. What we want to assess is not simply whether the student can do a specific, prompted task—for instance, asking questions to determine whether a patient is depressed—but whether the student will do such a task, and have the sensitivity to do it effectively even when not prompted or observed, or rather, not when observed by a teacher, but by a patient. This leads to a deeper critique of simple reliance on *in vitro* assessment of trainees under test conditions.

At bottom, medicine is a performing art. Medicine only exists when a doctor is taking care of a particular patient. A textbook of medicine is no more than a script that must be enacted for a particular person. A famous textbook of medicine such as "Harrison's", bears the same relationship to medicine, as the script of the play *Hamlet* bears to a performance of *Hamlet* in front of a specific audience. In other words, medicine is a performing art, and the critical issue is that authentic performance requires an authentic audience. Medicine exists only when we are taking care of a particular patient and, in a deeper sense, a physician only exists as a physician when taking care of a patient. People will suffer whether or not physicians exist. Students can learn without teachers. The physician's existence is contingent upon the presence of specific patients, and a

teacher's existence is contingent upon the presence of specific students.

More authentic assessments require observations of the student in real settings in an ongoing way. Just about everyone would accept that regular observation for prolonged periods of a trainee working with patients would have more 'face validity' than most tests. The problem is that we need to achieve reliability and precision in these observations as a step to achieving valid assessment.

Recent developments in performance assessment achieve a level of authenticity and reliability that is impressive. Computerization of multiple-choice examinations, especially those with sequential and adaptive testing as implemented in the United States by the National Board of Medical Examiners (and reported at meetings of AMEE in prior years), is an impressive feat. In traditional, analytic terms multiple-choice tests are measuring one aspect of competence, specifically, knowledge. It may be helpful to see these tests in a synthetic way, or as a final common pathway of many skills, including the cognitive ability to know what information is worth remembering, the personal skill to manage one's time successfully enough to study, and also, certainly, the commitment to ongoing, self-directed learning. In the language of the analytic model, skills and attitudes are assessed in this test almost as much as knowledge. And recent developments in what I would call quantified tests, standardized patient examinations and computer case simulations, have been especially impressive. The psychometric qualities of these tests have shown that performance-based assessment can take a step closer to true authenticity without sacrificing statistical power. The problem is that such performance-based assessments consume resources; they are expensive and depend on a high level of technology. They are not readily applied in developing countries. Even more universally, they are not applied by teachers at all, but by 'academic managers' like many of us—course, clerkship and program directors. And typically, they are applied only once or twice per year because of their expense and logistical problems. So, my own vision is that we will place more resources in developing a terminology of progress toward competence, and will invest in the faculty development necessary to make such a vocabulary more universal, and finally we will give our course or clerkship directors the resources to do this in real time on an ongoing basis.

At USU we have had success in joining our RIME vocabulary scheme with formal evaluation sessions in which we sit down with teachers every few weeks to discuss students' progress while they are still on our medicine rotations. These evaluation sessions were introduced in our setting 20 years ago (Noel, 1987), and our recent work has demonstrated enhanced sensitivity for detecting problems in both the student's fund of knowledge (Hemmer & Pangaro, 1997) and in professionalism (Hemmer *et al.*, 2000), and for predicative validity of poor performance using the first postgraduate year as an outcome measure (Lavin & Pangaro, 1998).

Barriers to achieving desirable goals

Yet there remain barriers to accepting the validity of descriptive evaluation of competence. First and foremost is a belief that words are 'subjective' and that numbers are 'objective'. In the United States at least, the widespread yet inappropriate use of these terms, 'subjective' and 'objective', has become a barrier to how we evaluate students at all levels of their training. Despite the cautions that the search for 'objectivity' has significant trade-offs (Norman *et al.*, 1991), its pursuit continues at the expense of descriptive methods of assessment. When we say that an assessment method is 'objective', what we mean is that it is disinterested (not biased as to rating) and highly quantified, or 'objectified' (Van der Vleuten *et al.*, 1991). But continued use of the term 'objective' for an assessment tool that yields a number (or percentage, or score above or below a mean) gives it a status in our somewhat scientific community that is often denied to observations by teachers. This trend will be corrected in the next decade.

Using the descriptive RIME vocabulary and based on our formal evaluation sessions, I have personally given feedback to 3000 medical students in the last 15 years. Several patterns demonstrate the problems with our continued use of the terms subjective and objective, rather than descriptive and quantified. When a student is told something that he or she does not like, the first response is often "Yes but that's subjective. My teachers simply didn't like me." Another example is that teachers often have their own confidence undermined by this belief that what they express using words is subjective. The teacher certainly does not want to be inaccurate in conveying his/her impressions of a student or, worse, does not want to harm a student's career. But the problem is not that his/her observations are incorrect, but that the teacher may not have sustained contact with the student, or enough contact to have sampled the student's abilities with a wide variety of patients. In other words, the teacher is instinctively aware that the number of observations he/she has of the student may not have achieved sufficient reliability, so is not comfortable in telling the clerkship director or even the student, much less in giving a grade. The RIME vocabulary is one step to inter-teacher consistency and willingness to document their assessment, and thus achieve a reliable sample of observations.

How can we achieve credible, authentic assessment?

What, then, will be the means to overcome these barriers and achieve credible, authentic assessment of students in an ongoing way that will, first, be valid enough for summative assessment and, second, be portable enough for formative assessment—feedback—by teachers? First, we need to have a vocabulary describing the student's progress that is widely acceptable to teachers and students, and which reflects the goals of the community. The traditional analytic construct—skills, knowledge and attitudes—indicates domains, but does not provide concise expression of goals. The RIME scheme is one possible step in creating a synthetic, portable vocabulary that can be used. As linked with the use of formal evaluation sessions, it has the asset of having being been studied with more rigor than most descriptive systems.

But a vocabulary is only an initial step. Once a descriptive terminology of developing competence becomes accepted, it can be used in systematic studies of assessment between teachers and clerkships within an institution, and even between institutions. In this way we will be able to achieve 'best evidence medical education' (BEME) (Harden *et al.*, 1999). Certainly, we can begin specific studies on the

applicability of the Uniformed Services University (USU) system of formal evaluation sessions, linked with the RIME scheme, to other settings. The project for Reliable and Valid Assessment of the Group on Educational Affairs (GEA) of the AAMC has begun an inter-school study to see whether the enhanced validity of the USU system can be reproduced elsewhere. As a first step in achieving convincing credibility for this system of descriptive evaluation our controlled studies are useful, but much more work is necessary to establish it as one standard of care. We have had the confidence to use our descriptive evaluation method as one outcome for a prospective, randomized trial of a significant change in our core internal medicine curriculum (Pangaro, 1995). Alternatively, there may be other descriptive evaluation systems that can achieve comparable or superior reliability and validity. Each school will have to follow the 'best evidence' that is applicable to its own setting and goals.

Conclusion

Our own system, sitting down with teachers on a regular basis to describe a student's progress from reporter to interpreter and manager-educator has helped us meet the obligation for honest evaluation in our own patient-care community. It has also helped meet our obligation to students by providing them with consistent, ongoing feedback in real time. Finally, we have invested time in developing our own teachers as evaluators of clinical performance, and given them confidence in their own ability to assess, describe and document what a student is and should be doing. I offer this model as one way to enhance the credibility of descriptive clinical assessment over the next decade.

Notes on contributor

Louis N. Pangaro is Professor and Vice-chairman for Educational Programs, Department of Medicine, Uniformed Services University F. Edward Hebert School of Medicine, Bethesda, MD, USA.

Note

The opinions expressed herein are those of the author and do not necessarily reflect the US Department of Defense or any federal agency.

References

Association of American Medical Colleges (AAMC) (1999) Learning objectives for medical student education—guidelines for medical schools: report I of the Medical School Objectives Project, *Academic Medicine*, 74, pp. 13–18.

Harden, R.M., Grant, J., Buckley, G. & Hart, I.R. (1999) BEME Guide no.1: Best Evidence Medical Education, *Medical Teacher*, 21, pp. 553–562.

Hemmer, P.A. & Pangaro, L.N. (1997) the effectiveness of formal evaluation sessions during clinical clerkships in better identifying students with marginal funds of knowledge, *Academic Medicine*, 72, pp. 641–643.

Hemmer, P.A., Hawkins, R., Jackson, J.L. & Pangaro, L.N. (2000) Assessing how well three evaluation methods detect deficiencies in medical students' professionalism in two settings of an internal medicine clerkship, *Academic Medicine*, 75, pp. 167–173.

Lavin, B. & Pangaro, L.N. (1998) Internship ratings as a validity measure for an evaluation system to identify inadequate clerkship performance, *Academic Medicine*, 73, pp. 998–1002.

Noel, G. (1987) A system for evaluating and counseling marginal students, *Journal of Medical Education*, 62, pp. 353–355.

Norman, G.R., Van Der Vleuten, C.P. & De Graaf, E. (1991) Pitfalls in the pursuit of objectivity: issues of validity, efficiency and acceptability, *Medical Education*, 25, pp. 546–547.

Pangaro, L.N. (1999) A new vocabulary and other innovations for improving descriptive in-training evaluations, *Academic Medicine*, 74, pp. 1203–1207.

Pangaro, L.N., Gibson, K., Russell, W., Lucas, C. & Marple, R. (1005) A prospective, randomized trial of a six-week ambulatory internal medicine rotation, *Academic Medicine*, 70, pp. 537–541.

Van Der Vleuten, C.P., Norman, G.R. & De Graaf, E. (1991) Pitfalls in the pursuit of objectivity: issues of reliability. *Medical Education*, 25, pp. 110–118.