

Moving Up the Pyramid: Assessing Performance in the Clinic



More than 18 years ago, George Miller introduced a framework for the assessment of medical students and residents, “Miller’s Pyramid”¹ (Figure 1). In the accompanying address to the Association of American Medical Colleges, he advocated the evaluation of learners for their skills and abilities in the 2 top cells of the pyramid, in the domains of action, or performance, reflecting clinical reality. Miller argued that the demonstration of competence in these higher domains strongly implies that a student has already acquired the prerequisite knowledge, or Knows, and the ability to apply that knowledge, or Knows How, that make up the base of the pyramid. Basic clinical skills (Shows How) are those that can be measured in an examination situation such as an objective structured clinical examination (OSCE). However, the professionalism and motivation required to continuously apply these in the real setting (Does) must be observed during actual patient care.

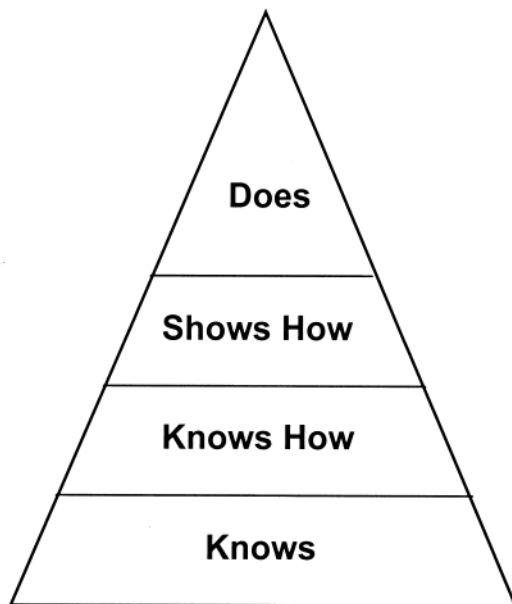


Figure 1. Miller’s pyramid¹. From *Academic Medicine* 1990;65:S63-7; with permission from Wolters Kluwer Health.

The component that Miller argued is the most vital aspect of measurement, what the learner does in clinical practice, has been the most difficult to capture. Almost 2 decades later, we are still struggling with the need to develop reliable and valid methods of assessing learners in the clinical setting.

In the meantime, there have been many advances in the lower echelons of the pyramid. In the domains of Know and Know How, the Medical Council of Canada² and the National Board of Medical Examiners³ have made great strides in the art of the multiple choice examination, the Key Feature examination, and computer-based, adaptive examination. In a pair of landmark publications, Tamblyn, *et al* have provided good evidence of the predictive validity of these assessments to outcomes in clinical practice^{4,5}. The OSCE examination has become so ubiquitous that it has been claimed to define the expectations of practice itself⁶. In the name of reliability, the standardized patient has overtaken the real patient for the purposes of certifying examinations⁷.

However, as outlined in the article by Susan Humphrey-Murto, *et al* in this issue of *The Journal*, the quality of assessment in the clinical setting lags far behind⁸. They state that the most frequently used instrument, the Intraining Evaluation Report (ITER), is completed by the resident or clerkship director, who may have had little personal experience of the learner, and at a time removed from many of those observations. This leads to a migration towards the center of the ubiquitous Likert scale, as the director is reluctant to label the student as being either exceptional or substandard in any specific item. Given that these forms are usually completed at the end of the rotation, and lack information clearly anchored to the performance of the learner, they have little value as a formative, or feedback, instrument that might contribute to the student’s education.

An initiative that shows promise in providing the opportunity for more timely and accurate assessment of clinical learners is the use of logbooks or assessment cards that must

See Resident evaluations: use of daily evaluation forms in rheumatology ambulatory care, page 1298

be completed by the clinical supervisor after each teaching session or patient encounter⁹. In most cases, these cards have scales or categories that include the domains of medical expertise (history taking, physical examination, problem formulation, and diagnosis), communication, professionalism, and written comments. The cards or logbooks may be kept by the resident, to be turned in at the end of the rotation. They represent a clear opportunity for timely feedback to be given to the student or resident.

For a resident engaged in a clinical rotation on an inpatient service, there are usually a limited number of clinical supervisors in attendance. Each resident may have only one or 2 per rotation. Interrater reliability for the encounter forms is therefore not a problem in terms of the consistency of the feedback provided.

Almost all rheumatology teaching takes place in the ambulatory setting. In most large academic institutions, the resident will rotate between the clinics of a number of different rheumatologists. The number of teachers filling out rating forms for each resident is therefore greater. In order to provide a fair assessment, and to give clear feedback to the learner, there must be reasonable agreement between teachers on the subscales of these forms. If reliability can be proven, then the assessment becomes valuable not only for providing feedback, but also becomes more fair as a summative measure of the resident.

Humphrey-Murto and colleagues have tested the hypothesis that a regular assessment instrument can be constructed for use in the outpatient rheumatology setting that can be reasonably reliable with a sensible number of observers and a reasonable number of observations. They have also surveyed the 2 classes of users, the teachers and the learners, to determine the acceptability of this means of assessment.

The authors created a 12-item evaluation form to be used by rheumatology preceptors at the end of each clinic for internal medicine residents. They also asked the residents and preceptors about their perception of the assessment process and of the quality of feedback both before and after this initiative began.

In their article, Humphrey-Murto, *et al* have shown that their evaluation forms have achieved acceptable interrater reliability, with an average of only 8.73 forms per resident, while using a wide range of scores with healthy standard deviations on all categories (Table 1, Reference 7). At the same time, both learners and teachers reacted positively to the experience of taking part in this exercise.

Teachers were more likely to say that they gave regular feedback to the residents on their histories and examinations after the institution of the new forms. A similar change was not perceived on a before-and-after basis by the residents, although those taking part did appreciate the feedback they were given. Perhaps the requirement to complete these forms acted as a gentle reminder to the teachers that time spent in the clinic represents an educational as well as a clinical

experience. This realization would represent an important result for this initiative.

As argued above, an assessment needs to be both reliable and valid. There can be no validity without reliability, so this characteristic of their measure has to be established first. Paradoxically, an instrument whose individual scales have little variation across multiple learners, while reliable, is not likely to be valid. Intuitively we understand that not all our students are the same on all characteristics. Similarly, without a useful range of points on each scale (e.g., poor to superior on history taking), there is no opportunity for a student to receive useful feedback, or to improve. This, I would argue, is an important aspect of validity. So it is with interest that we see that there was a wide range of subscale values used, with a healthy, broad range in standard deviation for each.

Further assessment of validity was unsuccessfully attempted by the authors. They predicted that there would be positive correlation between the scores received by the residents on the evaluation forms and on an OSCE examination completed after the rotation. An insufficient number of the residents completed the examination. Although the correlation was positive, the results were not statistically significant. This should be repeated. The establishment of a reliable instrument does not prove validity. However, it would have been impossible even to attempt to perform this comparison if the instrument were not reliable. Therefore, such comparisons, and any claims of validity, would not be possible with most current ITER.

A reason for the current popularity of OSCE examinations has been the poor performance of the ITER as a measure of clinical ability. If there were another form of reliable and valid performance assessment, it would likely become popular for both feedback and assessment. Other advances in this area include the application of the CANMEDS structure to broaden the scope of assessment¹⁰, the 360-degree assessment to identify the influence of the resident on patients or other healthcare professionals¹¹, and the use of the mini-Clinical Evaluation Exercise to improve feedback¹². The use of these, in parallel with the continuous performance assessment discussed in this issue⁸, may provide further reliability and validity to this domain.

In summary, the structured real-time assessment of residents in the clinical setting is likely to represent a similar change in culture for the assessment of "Does" as the creation of the OSCE examination did for "Shows How" in the last 2 decades. The authors of this article should also be congratulated for putting rheumatology, a paradigm of ambulatory clinical practice, on the leading edge of this initiative in medical education.

STEPHEN AARON, MD, FRCPC,
Professor, Department of Medicine,
University of Alberta,
Edmonton, Canada

REFERENCES

1. Miller GE. The assessment of clinical skills/competence/performance. *Acad Med* 1990;65:S63-7.
2. Page G, Bordage G. The Medical Council of Canada's Key Features Project: A more valid written exam of clinical decision making skills. *Acad Med* 1995;70:104-10.
3. Case SM, Swanson DB, Ripkey DR. Comparison of items in five-option and extended matching formats for Assessment of Diagnostic Skills. *Acad Med* 1994 69S:1-3.
4. Tamblyn R, Abrahamowicz M, Dauphinee WD, et al. Association between licensure examination scores and practice in primary care. *JAMA* 2002;288:3019-26.
5. Tamblyn R, Abrahamowicz M, Dauphinee D, et al. Physician scores on a national clinical skills examination as predictors of complaints to medical regulatory authorities. *JAMA* 2007;298:993-1001.
6. Hodges B. Christmas stocking OSCE! Variations on a theme by Harden. *Med Educ* 2003;37:1134-40.
7. Verghese A. Culture shock — patient as icon, icon as patient. *N Engl J Med* 2008;359:2748-51.
8. Humphrey-Murto S, Khalidi N, Smith C, et al. Resident evaluations: the use of daily evaluation forms in rheumatology ambulatory care. *J Rheumatol* 2009;36:1298-303.
9. Al-Jarallah KF, Moussa MAA, Shehab D, Abdellah N. Use of interaction cards to evaluate clinical performance. *Med Teach* 2005;27:369-74.
10. Royal College of Physicians and Surgeons of Canada. CANMEDS Physician Competency Framework 2005. [Internet. Accessed February 26, 2009.] Available from: <http://rcpsc.medical.org/canmeds/CanMEDS2005/index.php>
11. Lockyer J. Multisource feedback in the assessment of physician competencies [review]. *J Contin Educ Health Prof* 2003;23:4-12.
12. Hatala R, Ainslie M, Kassen BO, Mackie I, Roberts JM. Assessing the mini-Clinical Evaluation Exercise in comparison to a national specialty examination. *Med Educ* 2006;40:950-6.

J Rheumatol 2009;36:1101-3; doi:10.3899/jrheum.090085