In-training assessment developments in postgraduate education in Europe

Cees van der Vleuten*† and Bas Verhoeven‡

*Department of Educational Development and Research, Maastricht University, Maastricht, The Netherlands
†King Saud University, Riyahd, Kingdom of Saudi Arabia and
‡Department of Surgery, Maastricht University Medical Centre, Maastricht, The Netherlands

Key words
assessment, clinical performance, direct observation, portfolio, postgraduate training.

Correspondence
Professor Cees van der Vleuten, Department of Educational Development and Research, Faculty of Health, Medicine and Life Sciences, Maastricht University, P.O. Box 616, 6200 MD, Maastricht, The Netherlands. Email: c.vandervleuten@educ.unimaas.nl

C. van der Vleuten PhD; **B. Verhoeven** MD, PhD.

Accepted for publication 13 January 2013.

doi: 10.1111/ans.12190

Abstract

Aim: This paper reviews changes that are underway in postgraduate medical education in various European countries. Training in the workplace is a very effective way of learning, but it has many imperfections. Changes in in-training assessment are proposed to remedy some of these.

Assessment tools: The focus is on a set of assessment tools for performance in authentic work-based contexts. These tools include direct performance measures of single clinical events (mini-Clinical Evaluation Exercise, Direct Observation of Practical Skills, Objective Structured Assessment of Technical Skills, Case-based Discussion, Mini-Peer Assessment Tool) and performance measures over a period of time (Multi-Source Feedback), based on judgement by one or more knowledgeable assessors (supervisor, other healthcare professional, patient, trainee himself/herself). Quantitative and qualitative information from single assessments is first and foremost used to promote learning, but also aggregated across a large sample of contexts and assessors in order to obtain an overall picture of a trainee's progress. Aggregating instruments, such as the portfolio, can be used to collect, support and assess outcomes in terms of competencies achieved. We will describe this set of instruments and provide theoretical background as well as our own practical experiences.

Discussion: A central message is that the utility of assessment methods lies very much in the (understanding of) the users. Therefore, our concern is more with the actual implementation of change than with the assessment technology per se. If we fail in our efforts to implement real change, postgraduate education may be at risk for bureaucratization and trivialization. We nevertheless are excited to see change happening in the right direction, but remain patient, not expecting very quick wins.

Introduction

In-training assessment (ITA) in postgraduate medical education in Europe is on the eve of major change. Speaking of Europe as a uniform geographic entity is tempting but not always accurate. More particularly from an educational perspective, European countries are not progressing at the same pace. Change is most marked in the north-western parts of Europe, with the United Kingdom clearly leading the way. While the whole of Europe is in the midst of implementing a uniform two-cycle system of higher education, including medical education,[1] there is no similar initiative for postgraduate education. Currently, the requirements of postgraduate medical education with regard to programme content and structure, accreditation and licensing vary both within and between countries. Postgraduate specialty training may be preceded by 1 or 2 years of general clinical training, as in the United Kingdom, or follow immediately after completion of undergraduate medical training, as in the Netherlands, although it is customary for Dutch trainees to take up a temporary post as a resident-not-in-training. Most surgical training programmes in the Netherlands last 5–6 years. The last year of the current programme in general surgery consists of 'differentiation', with trainees spending 80% of their time working in upper/lower gastrointestinal surgery (GI), trauma surgery, vascular surgery or paediatric surgery. In the first 5 years, there are 3-month rotations in the intensive care unit (ICU) and the emergency department, while the remaining time is divided between the operating room, wards, outpatient clinic, on-call duty and special courses on subjects like Advanced Trauma Life Support (ATLS) and laparoscopy. Over the course of the 6-year programme, trainees are supervised by surgeons on a rotation basis, and the intensity of supervision diminishes over time. After the 6-year programme, many surgeons undertake an additional training programme of 2 consecutive years in subspecialties like surgical oncology, upper/lower GI, trauma surgery, vascular surgery or paediatric surgery. Postgraduate training in thoracic surgery, orthopaedic surgery as well as reconstructive surgery is preceded by the first 2 years of the general surgical programme. Surgical specialties like ear, nose and throat; gynaecology; neurosurgery and ophthalmology have separate programmes. This bird's-eye view of the variegated landscape of postgraduate training in one specialty area in one country illustrates why it is complicated to write about the European perspective. Setting aside these reservations, we have no doubt that change is on the horizon. In this paper, we attempt to outline the main trends. We hope the readers will forgive us for using illustrations from our own context. It should be Dutch.

Why change occurred is occurring

The simplest way of explaining why change is inevitable is to point out that mere work experience in clinical practice does not suffice to train competent medical specialists for today's and tomorrow's health care. The apprenticeship model was the traditional approach, but without disparaging the power of this model, it is obvious that the complexity of modern health care in tertiary settings is a far remove from the world of the dyadic interaction between apprentice and master. From an educational perspective also, there is good evidence that working in clinical practice is simply not enough. Let us give an example. A British study found that in the first year of postgraduate training, 8.4% of trainees' prescriptions contained errors, 1.89% of which were lethal.[2,3] Remarkably, if not disturbingly, these percentages increased for second year trainees (10.3%; 1.57% lethal). Although the authors found no simple causal explanation, they advocated for better integration of theory and practice and just-in-time (rather than just-in-case) training programmes. Whatever the cause of the unexpected increase in prescribing errors, imperfections in postgraduate training programmes are widely documented. They include sparse feedback,[4] too much or too little responsibility and numerous near accidents,[5] limited direct observation of trainee patient interactions,[6] a clash between production and learning often leading to flat learning curves[4] and unfavourable learning climates,[7] and if these shortcomings do not offer ample incentive for change, there are the increasing calls for specialists to possess competencies beyond the traditional domains of medical knowledge and skills. Providing optimal care also depends on general competencies in areas like communication, team functioning, information finding, professionalism, ethics, scholarship, etc. Whenever things go wrong in clinical practice, more often than not these general skills are involved.[8]

We have now outlined the reasons behind the sense of urgency that has motivated regulators responsible for postgraduate training to initiate proposals for change. Next, we will consider what sorts of changes are required.

Proposed interventions

We see three fundamental strategies for change. The first one is improved *structuring of the learning process*. This typically involves translating competency frameworks and standards into a formal curriculum structure. An example is the Canadian Medical Education Directives (CanMEDS) framework,[9] which was adopted fairly recently to guide changes in *all* postgraduate specialist training programmes in the Netherlands (and other parts of Europe). All disciplines in the Netherlands were required to write a curriculum plan. For general surgery, this resulted in 44 themes, clustered around general surgery, upper/lower GI, vascular surgery, trauma surgery, lung surgery and paediatric surgery. Several themes specifically address patient safety, handover, education and research. Further, key procedures have been identified for each cluster, and it is explicitly stipulated that the provision of structured written feedback is obligatory in assessments of procedural skills. Trainees have to keep track of all their activities, exams, assessments and feedback in their (electronic) portfolio. Additionally, specific courses must be offered, such as ATLS, Basic Surgical Exam, X-ray health physics, ICU exam Fundamental Critical Care Support, and supervisors have to regularly provide constructive written feedback, schedule meetings with trainees and offer trainees opportunities to individualize their learning paths.

The second strategy is to enhance *feedback and reflection*. This is where ITA comes in. Assessment has to be continuous and (in)formative, based on the assessor's first-hand experience and direct observation and give an active role to the trainee. Formal supervision sessions, progress reviews and progress decision moments are introduced as components of a structured assessment programme. We will elaborate on the ITA in more detail later on.

First, we turn to the final strategy for change: *organizational* interventions. These are prompted by the realization that plans on paper do not in and of themselves change training practices. More focused implementation strategies are required. Staff development programmes form an integral part of such strategies. Furthermore, in a wider organizational perspective, systems for quality assurance of the curriculum need to be put in place. Quality assurance means awareness of the input resources of programmes, evaluation of the quality of programmes (for which evaluation instruments must be developed) and a governance structure to monitor that necessary programme changes are actually pursued. This is the organizational counterpart of ITA: in-training evaluation of training programmes in order to ensure that continuous plan-do-check-act cycles are implemented and ultimately (but not surprisingly) result in successful accreditation.

These strategies can be observed to be taking place in a number of European countries, albeit at a different pace, as we noted before. But most importantly, there is a general sense of urgency and changes are taking place that will impact on the training system as a whole.[10]

Before returning to ITA and its crucial role, we will first discuss some developments in assessment theory that underpin the current developments in assessment of clinical performance in the workplace.

Assessment theory

Substantial progress has been made in assessment technology and theory in recent decades.[11,12,13] We will now attempt to describe the main lessons in a nutshell. We have learned that not standardization or objectification but sampling enables reliable assessment. Clinical competence in whatever form or subdomain is inherently context specific. We therefore need to sample across many contexts to be able to make generalizable inferences that can help predict future performance. On a similar note, sampling across many different assessors in these contexts can overcome the subjectivity of assessments. We have learned that assessment drives learning, but that it does not necessarily lead it in the desired direction. Classical summative (end-of-course) assessments, although effective in achieving competency gains,[14] provide little information to guide learning. If assessment is to have a positive effect on learning, it must provide information that is meaningful to learners. This strongly depends on both the quality of the feedback and how it is used. Ongoing assessment, or ITA, should foster feedback that offers incentives to improve performance and informs trainee action. We have also learned that holistic judgements of complex competencies, for example professionalism, can have validity and be even superior to reductionist checklists, which only too easily turn learning into ritualistic clinical performance. We have learned that any form of assessment entails compromise. We cannot have individual assessment moments that are reliable, valid, acceptable, feasible and meaningful, all at the same time. We want a collection of different assessments, preferably arranged purposefully in an assessment programme, obtained with many different tools and sampling across a wide range of contexts, patients and assessors. Different purposes call for different compromises, but compromises are inescapable. Individual assessments are viewed as representing single data points. Single data points in an assessment programme should promote learning and the one thing on which compromise is unacceptable is learning value. Feedback should be served maximally in all individual assessments. By aggregating across data points, we gain an overall picture containing a wealth of information. Promotion decisions (reliable and valid) must always be underpinned by aggregated information, never by a single data point. We have learned that aggregating information from individual data points is preferably based on a meaningful classification of content. Only too often are we tempted to aggregate across apples and oranges, simply because they were assessed using the method (e.g. we average across a resuscitation and a communication station in an Objectively Structured Clinical Evaluation (OSCE) assuming they have something in common). However, it would be much wiser to aggregate across methods in a way that makes sense. This is where competency frameworks come in. Traditional assessment approaches tend to be governed by isomorphic skill by method combinations: each skill has its preferred assessment method (and no other one). In modern assessment programmes, however, a variety of methods is considered suitable to assess a single competency and vice versa: a variety of competencies can be soundly judged using one single method.

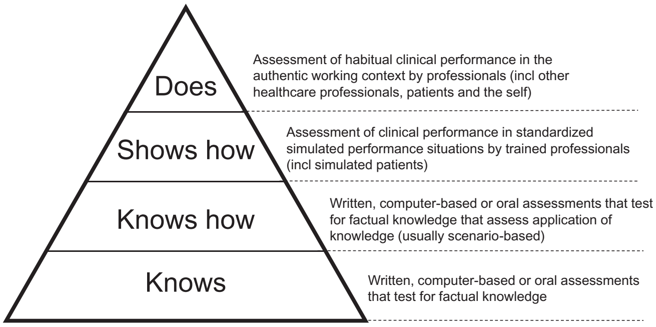
Does
Assessment of habitual clinical performance in the authentic working context by professionals (incl other healthcare professionals, patients and the self)
Shows how
Assessment of clinical performance in standardized simulated performance situations by trained professionals (incl simulated patients)
Knows how
Written, computer-based or oral assessments that test for factual knowledge that assess application of knowledge (usually scenario-based)
Knows
Written, computer-based or oral assessments that test for factual knowledge

Fig. 1. Miller's pyramid and types of assessment used for assessing the layers.

These developments have given rise to modern assessment technologies, specifically tailored to the clinical workplace. Figure 1 shows a familiar assessment framework representing a simple model showing competencies in ascending order of complexity with the most suitable assessment methods. The first three layers (knows, knows how and shows how) are typically assessed by standardized tests (such as Multiple Choice Questions (MCQs) or OSCEs). The top of the pyramid calls for non-standardized assessment in the workplace. The ultimate goal is assessment of habitual performance by many different (subjective) assessors across a range of different contexts. Each individual assessment is primarily intended as a source of feedback to trainees and should therefore be learner oriented, remedial and developmental, whereas aggregation across assessments reveals a rich picture of competency development that can be used for (intermediate and final) decision making about progression and promotion. Aggregation can be done across the layers of the pyramid, all depending on the competency structure used. Although a good assessment programme uses instruments from all layers of Miller's pyramid, our discussion is confined to the top of the pyramid because that is where most developments in postgraduate medical education in Europe are taking place. We do however wish to clearly state that limiting the assessment programme to work-based ITA alone is insufficient. An integral programme needs to be complimented with more standardized forms of assessment sampling the other lower layers of the pyramid. These other methods of assessment may either be internal to the training program or external as a mandatory requirement from a professional organization as is the case in many countries.

Developments in in-training work-based assessment

The new approaches to assessment based on the developments in assessment theory described in the previous section can be categorized as direct performance measures and aggregation methods. We will discuss these categories separately and link them to our own experiences and reflections.

Direct performance measures

Within this category, two classes of instruments can be distinguished based on the duration of the assessment period: *single event measures* and *global performance measures*.

Single event measures are based on observation of clinical performance in one single event (patient encounter, procedure, handover). During and/or after observing the event, the clinical teacher completes a generic rating scale (i.e. one that is not specific to the case/patient in question) and additionally provides brief written narrative feedback. In most cases, the common procedure is for teacher and trainee to sit down after the assessment and talk about the feedback from the rating form and the narrative comments. The best-known instrument in this category is the mini-Clinical Evaluation Exercise.[15] Ratings are given on history-taking skills, physical examination skills, humanistic qualities/professionalism, clinical judgement, counselling skills, organization/efficiency and overall clinical competence. Several other but similar instruments address specific competencies, qualities or skills. Direct Observation of Practical Skills[16] assesses procedural skills; Objective Structured Assessment of Technical Skills[17] is used for surgical techniques in the operating theatre; the Professionalism Mini-Evaluation Exercise is tailored to aspects of professional behaviour in clinical performance.[18] A method that is extensively used in the United Kingdom is the Case-based Discussion (CBD), in which a brief presentation of a case by a trainee is followed by a structured interview with an assessor. Although CBD involves no direct observation (one could challenge its position at the top of Miller's pyramid for this), it has its own place in the toolkit for in-training work-based assessment. A web search on the above acronyms will yield a host of different concrete examples of rating scales that are used all over the world, including in Europe.

Solid evidence for the usefulness of these instruments is slowly emerging,[19,20] although more research is needed. Estimating reliability requires large scale datasets,[16] but so far all research seems to point to the magic number of eight encounters under the assumption of a different assessor per encounter. We will now discuss some of the lessons we have learned from our involvement in the practical implementation of work-based assessment instruments.

Giving feedback is a skill that has to be learned and maintained. Feedback is well researched,[21] and it is known that the way it is delivered can make a dramatic difference to its effect.[22] In our experience, the feedback delivery skills of many clinicians fall short with regard to promoting acceptance of and learning from feedback. Feedback training programmes are therefore imperative. Such programmes typically simulate or videotape an assessment and the related feedback. Interestingly, the communication skills underlying good feedback quite closely resemble the communication skills for doctor–patient encounters: stimulating reflection, open-ended questioning, organizing information, formulating and checking on an action, etc. Another similarity with the patient encounter is the importance of a safe environment, which the assessor can help to establish. The dialogue between assessor and trainee is probably the single most determining factor for the success of this type of assessment.

As narrative, qualitative information is considered more useful than numerical ratings; we invite assessors to verbalize their impression of the trainee from the assessment, in addition to completing the numerical rating form. We advise them to focus on one or only a few major points, always explicitly describe how the feedback connects to the trainee's actions (and follow-up) and document their feedback on paper. This takes time, we agree. But, we are strongly in favour of few assessments that are done well and offer usable feedback as opposed to numerous superficial evaluations, which offer feedback of very poor quality if any.

It is also important to incorporate assessment as an established component of the routines of clinical practice. We again stress the importance of brevity (the 'mini' epithet deserves maximum attention). Nevertheless, observation and feedback use up time, one of the scarcest commodities in hectic clinical workplaces. It is therefore probably unrealistic to expect departments to sustain a meaningful workplace assessment programme unless steps have been taken to ensure that the programme is firmly embedded within the routine clinical flow. That is why we are ardent supporters of fixed assessment moments within outpatient itineraries. Also, we champion rotating appointments of assessors who are responsible for assessment within a given time window. Having talked at some length about single event measures, we now turn our attention to global performance measures.

Global performance measures assess clinical performance over longer periods of time (e.g. 6 months or a whole year). The classic supervisor rating would fit into this category, if only it sampled across assessors. Multisource Feedback (MSF) or 360 degree feedback seems today's most popular exponent of this category.[23] Again, a web search will return a plethora of MSF questionnaires. Once an appropriate questionnaire has been decided on, they are sent to a variety of assessors: usually the trainee him/herself and others who have relevant experience of the trainee's performance, such as the clinical supervisor, other healthcare professionals, peers, support staff and patients. Assessors are usually selected by trainee and supervisor together. Although variable numbers of assessors have been cited as prerequisite for a reliable outcome,[16,23,24] an approximate number of eight assessors seems generally adequate. All aspects we discussed in relation to single event measures apply in equal measure to MSF, but there are some additional remarks to be made.

Feedback stands a better chance of being accepted if the assessors have credibility in the eyes of the trainee. Involving the trainee in selecting assessors can increase the credibility of the resulting information and thus the likelihood of feedback acceptance. Confidentiality and anonymity of assessor specific feedback all contribute to making MSF work.

Feedback can trigger strong emotions. A moderator can make it easier for a trainee to deal with these. Stimulating critical reflection in a dialogue and giving the trainee time to respond to feedback are important, as is follow-up.

The 'less is more' principle is perhaps even more apt in relation to MSF. MSF is an intense and laborious process for all parties involved. Doing it too frequently is likely to jeopardize the usefulness of the whole process.

If done well, trainees place high value on this type of evaluation. It is a good idea to reduce the administrative burden by using electronic tools. Such tools can also safeguard anonymity, aggregate information and present it in attractive web-based graphical feedback reports.

Aggregation methods

Aggregation methods sample performance across a longer period of time or even continuously. Two much-used instruments are the logbook and the portfolio. Portfolios have become particularly popular and are gaining ground in postgraduate training programmes.[25] In the Netherlands and in the United Kingdom, for example, all programmes require their trainees to keep a portfolio that is used to monitor progress, coach the trainee and assess outcomes. The portfolio has its origins in architecture and the arts. Artefacts are assembled and presented to potential clients or supporters of the artist's work. This concept has been transferred to education, where the trainee presents evidence of work done and competence gained. This can take the form of assessment results, video materials, lists of procedures performed, academic work (e.g. critical appraisal of a topic, a presentation), etc. Periodically, the trainee reflects on the evidence, shares these reflections with their supervisor or coach, draws conclusions and plans new learning activities accordingly. Ultimately, the portfolio is assessed (e.g. by a committee) in order to determine if the desired outcomes have been attained. So, the portfolio is both learning and assessment instrument. It reverses the burden of evidence. The learner is responsible for proof of competence, not the teacher/supervisor (as is the case in all other forms of assessment).

We said earlier that portfolios are becoming increasingly popular. More research around portfolios is emerging, and some good reviews have been published.[26,27,28] A few essential lessons have been learned, and we now add some lessons from our own practice.

Bureaucracy is to be avoided at all cost, as it is a sure way to dampen enthusiasm. Physical size is no criterion for a good portfolio. Huge shoeboxes full of evidence which nobody looks at are nothing but a waste of space. A good portfolio is *lean and mean*. Evidence should be limited, but more importantly, it must be used for reflection. Writing reflections is a tedious process, which only very few people find enjoyable. Keeping things brief and feasible is a wise strategy for effective portfolio use.

Just as with the other methods we discussed, the interactions around the instrument define the portfolio's success. The portfolio should be shared with others in some sort of social interaction. This will usually involve the supervisor or coach, but a peer group is equally acceptable. Essential is the discussion around portfolio content and the ensuing reflections. Without periodical discussion, reflection and monitoring, the portfolio is reduced to a paper tiger with no appreciable learning value. But, the crucial condition for success is assessment. Only a portfolio that is assessed stands a chance of being taken seriously by all involved. So, portfolios work best if the functions of monitoring, coaching and assessment are ensured and combined.

Computer technology can ease the administrative burden of portfolio and even add value. For example, together with the University of Manchester, we have developed a web-based electronic portfolio that facilitates all related assessment activities (using all above described instruments), allows documentation of materials (e.g. of surgical procedures), reflections, mentoring reports, etc. and, on top of that, generates overviews and feedback profiles that chart trainee progress. Obviously, synthesizing information across instruments is only feasible if all instruments address the same competencies. The electronic version of the portfolio is highly appreciated by all users.

Discussion

We introduced this paper by explaining why change in postgraduate medical education is inevitable, and we argued that the workplace is a powerful learning environment but with many imperfections. We have outlined assessment strategies that can address some of these imperfections. Central to our views is the notion that assessment should first and foremost deliver feedback to promote and support learning and that the trainee must be given an active and reflective role.

Apart from feedback, ITA programmes enable high-stakes decisions. In fact, successful and full implementation of a portfolio approach would obviate the need for a further high-stakes summative exam. Such an exam would be just another point measurement in time and thus unable to provide as much or as rich information as the data collected longitudinally and continuously in the portfolio. This is not to say that standardized tests are superfluous. They are important, but they should be incorporated into an overall programme of assessment and preferably integrated in ITA as a programme. If external assessment is maintained, internal assessment procedures should be predictive for these external exams. The portfolio is the umbrella that unifies formative and summative assessment, offers optimal encouragement for learning, while at the same time delivering high-quality information about a trainee's development as a basis for robust decisions. There are various strategies for maximizing the robustness of these decision,[29] all related to procedures to make decisions credible and trustworthy.[11]

Assessment and instruction are largely overlapping in this approach. But, herein also lies the most dangerous pitfall. A general lesson from our experience is that the value of instruments depends more on the users than on the instruments themselves. The seriousness and the proficiency with which the assessment tasks are carried out determine their learning value to the trainee. If assessment is not conducted seriously, if filling out forms becomes a goal in itself (because of an external requirement), learning value suffers. Like so many assessment exercises, the process then becomes ritualized, with people jumping through hoops and ticking boxes without meaningful learning taking place. The baby will be thrown out with the bathwater.

The problem therefore lies not so much with the 'technology' or the 'most appropriate' assessment instrument, the problem lies with implementation. Speaking from our experiences in the Netherlands, it has been a great pleasure to witness the changes in postgraduate specialist training programmes. All disciplines, including surgery, are implementing their new plans. It is immediately apparent that this will take time. If success depends on the users, and if the users are asked to change their behaviours, it is imperative that the users should gain a clear understanding of their role in the learning and assessment process. This is a process that is not to be taken lightly and one that cannot be rushed. It takes time, effort and perseverance. We should not forget that clinical supervisors learn like any other group of learners; they need to understand why change is so important, they need to practise new methods and they need feedback and time and help to digest it. Like any other learner, simply telling them what and how they are expected to change is unlikely to bring about any real and durable change. Several initiatives have been taken in many different institutions to ensure that teachers receive feedback on their work. For example, the MSF approach has been used to evaluate clinical supervisors' teaching and supervisory qualities.[30] All in all, the risk of failure due to lack of understanding, commitment and skill in those who actually have to run and deliver the programme highlights the relevance of the organizational interventions we described earlier.

In conclusion, postgraduate medical education is indeed being transformed in some European countries. We are fascinated to watch as these developments unfold, and we are optimistic. Coming from a medical school that was among the first adopters of problem-based learning (in 1974), we have witnessed dramatic changes in undergraduate 'lecture hall and classroom' based teaching. School-based learning has changed in medical schools across the globe, and teaching has professionalized to an astonishing degree. We think that today we are on the brink of a similar change in work-based learning, particularly in the postgraduate arena. Like any change, do not expect it to happen overnight. But also, do not be surprised when it actually comes about.

References

1. Patricio M, den Engelsen C, Tseng D *et al*. Implementation of the Bologna two-cycle system in medical education: where do we stand in 2007? – results of an AMEE-MEDINE survey. *Med. Teach.* 2008; **30**: 597–605.
2. Dornan T, Ashcroft D, Heathfield H *et al*. *An in Depth Investigation into Causes of Prescribing Errors by Foundation Trainees in Relation to Their Medical Education. EQUIP Study*. Manchester: Hope Hospital (University of Manchester Medical School Teaching Hospital), 2009.
3. Tully MP, Ashcroft DM, Dornan T *et al*. The causes of and factors associated with prescribing errors in hospital inpatients: a systematic review. *Drug Saf.* 2009; **32**: 819–36.
4. Sargeant J, Armson H, Chesluk B *et al*. The processes and dimensions of informed self-assessment: a conceptual model. *Acad. Med.* 2010; **85**: 1212–20.
5. Teunissen PW, Boor K, Scherpbier AJ *et al*. Attending doctors' perspectives on how residents learn. *Med. Educ.* 2007; **41**: 1050–8.
6. Holmboe ES. Faculty and the observation of trainees' clinical skills: problems and opportunities. *Acad. Med.* 2004; **79**: 16–22.
7. Boor K, Scheele F, van der Vleuten CP *et al*. How undergraduate clinical learning climates differ: a multi-method case study. *Med. Educ.* 2008; **42**: 1029–36.
8. Papadakis MA, Teherani A, Banach MA *et al*. Disciplinary action by medical boards and prior behavior in medical school. *N. Engl. J. Med.* 2005; **353**: 2673–82.
9. Frank JR, Danoff D. The CanMEDS initiative: implementing an outcomes-based framework of physician competencies. *Med. Teach.* 2007; **29**: 642–7.
10. ten Cate O, Scheele F. Competency-based postgraduate training: can we bridge the gap between theory and clinical practice? *Acad. Med.* 2007; **82**: 542–7.
11. Van der Vleuten CP, Schuwirth LW, Scheele F *et al*. The assessment of professional competence: building blocks for theory development. *Best Pract. Res. Clin. Obstet. Gynaecol.* 2010; **24**: 703–19
12. Van der Vleuten CPM, Schuwirth LWT. Assessment of professional competence: from methods to programmes. *Med. Educ.* 2005; **39**: 309–17.
13. van der Vleuten CP, Schuwirth LW, Driessen EW *et al*. A model for programmatic assessment fit for purpose. *Med. Teach.* 2012; **34**: 205–14.
14. Larsen DP, Butler AC, Roediger HL 3rd. Test-enhanced learning in medical education. *Med. Educ.* 2008; **42**: 959–66.
15. Norcini JJ, Blank LL, Arnold GK *et al*. The mini-CEX (Clinical Evaluation Exercise): a preliminary investigation. *Ann. Intern. Med.* 1995; **123**: 795–9.
16. Wilkinson JR, Crossley JG, Wragg A *et al*. Implementing workplace-based assessment across the medical specialties in the United Kingdom. *Med. Educ.* 2008; **42**: 364–73.
17. Reznick R, Regehr G, MacRae H *et al*. Testing technical skill via an innovative 'bench station' examination. *Am. J. Surg.* 1997; **173**: 226–30.
18. Cruess R, McIlroy JH, Cruess S *et al*. The Professionalism Mini-evaluation Exercise: a preliminary investigation. *Acad. Med.* 2006; **81** (10 Suppl): S74–8.
19. Hawkins RE, Margolis MJ, Durning SJ *et al*. Constructing a validity argument for the mini-Clinical Evaluation Exercise: a review of the research. *Acad. Med.* 2010; **85**: 1453–61.
20. Pelgrim EA, Kramer AW, Mokkink HG *et al*. In-training assessment using direct observation of single-patient encounters: a literature review. *Adv. Health Sci. Educ. Theory Pract.* 2011; **16**: 131–42.
21. Archer JC. State of the science in health professional education: effective feedback. *Med. Educ.* 2010; **44**: 101–8.
22. Shute VJ. Focus on formative feedback. *Rev. Educ. Res.* 2008; **78**: 153–89.
23. Lockyer J. Multisource feedback in the assessment of physician competencies. *J. Contin. Educ. Health Prof.* 2003; **23**: 2–10.
24. Williams RG, Klamen DA, McGaghie WC. Cognitive, social and environmental sources of bias in clinical performance ratings. *Teach. Learn. Med.* 2003; **15**: 270–92.
25. Van Tartwijk J, Driessen EW. Portfolios for assessment and learning: AMEE Guide no. 45. *Med. Teach.* 2009; **31**: 790–801.
26. Butler P. *A Review of the Literature on Portfolios and Electronic Portfolios*. Palmerston North, New Zealand: Massey University College of Education, 2006.
27. Driessen E, van Tartwijk J, van der Vleuten C *et al*. Portfolios in medical education: why do they meet with mixed success? A systematic review. *Med. Educ.* 2007; **41**: 1224–33.
28. Tochel C, Haig A, Hesketh A *et al*. The effectiveness of portfolios for post-graduate assessment and education: BEME Guide No 12. *Med. Teach.* 2009; **31**: 299–318.
29. Driessen EW, Van der Vleuten CPM, Schuwirth LWT *et al*. The use of qualitative research criteria for portfolio assessment as an alternative to reliability evaluation: a case study. *Med. Educ.* 2005; **39**: 214–20.
30. Lombarts MJ, Arah OA, Busch OR *et al*. [Using the SETQ system to evaluate and improve teaching qualities of clinical teachers]. *Ned. Tijdschr. Geneeskd* 2010; **154**: A1222.