



EVALUATION

Examining and Recording Clinical Performance: A Critique and Some Recommendations

KEN COX

*School of Medical Education, University of New South Wales,
Australia*

ABSTRACT *Clinical performance is too complex and interactive for measurement. Judgment is always necessary for its assessment. Experienced clinicians judge trainee performance on many small details. This clinical judgment turns on the trainee's handling of important details in the patient and the malady.*

But the recording of performance retreats to categories and checklists that contain nothing of those critical details or the trainee's judgment. Checklists are incapable of identifying what actually happened, and 'could do' categories have no predictive accuracy in asserting what cases a trainee can actually manage. Clinical examinations have even been subverted by the naïve, pseudorational error that competence is defined by obedience to doing exactly what someone else expects you to do in every case, as in an OSCE examination.

Cases are the unit of clinical practice. The clinical curriculum should be comprised of the critical core cases the trainee must be able to handle in each discipline. Case management, procedural skills and professional behavior can be assessed accurately only in the context of daily clinical work. Formal examinations lack the range of cases and open-ended time that allow examiners to explore a trainee's case knowledge and judgment. Habitual behavior can be assessed only by observing habitual behavior in everyday practice. Assessment and recording should take place only in real world settings, focused on performance on the core cases trainees must be competent to manage.

The assessment of complex performance in all fields is difficult, not just in clinical work. In this paper I first describe what actually happens in current formal clinical examinations. Second, I examine current reductionist attempts at measurement of the candidate's performance. Third, I propose that the assess-

Ken Cox, OAM, MD (*Hon causa*), MA, MB, MS, FRCS, FRACS, FACS, Emeritus Professor of Surgery, Former Head, School of Medical Education, University of New South Wales, Sydney 2052, Australia. Tel: (612) 9817–2902. Fax: (612) 9385–1526. E-mail: Ken.Cox@unsw.edu.au

ment of performance shift from formal examinations on a few selected patients to the critical core cases doctors must manage in daily clinical work. Fourth, I assert that clinical performance should be assessed within the context of daily work.

Clinical Judgment

Start with what actually happens during formal, final clinical examinations. As an examiner, what do you notice when you watch a trainee's clinical performance? Do you have a plan for what to look at, and what to look for? Or do you, in reality, notice that some tasks are being done well and others are not? And do you record your assessment later when you think back about what you saw?

My observations as a clinical examiner from undergraduate to specialist college levels are that examiners behave like other humans, making their assessments on the spot from lots of small details within the performance. Only when the candidate leaves do they turn to their recording forms, translating their specific observations on details into judgments that are recorded as scores. But we neglect to test the accuracy of that very serious judgment on final year students by correlating it with their subsequent performance as an intern during the next year.

Stay with the joint examiners in clinical examinations for a little. What is going through the mind of your colleague as he or she watches? You can glean something of that from four outputs from the assessment:

- comments made after the candidate has left;
- the score each assigns to the performance;
- the content of the discussion around details of the performance by which each justifies their score (which will probably be adjusted if disagreement arises); and
- what they note on their examiners' sheets to justify their agreed score later at an examiners' meeting.

The first identifies what caught their attention. The second places a number on their assessment of the algebraic sum of strong and weak aspects of the performance. The third comprises an interaction between two (occasionally three) clinicians, often of different clinical and examination experience and of different rank in the hierarchy. Each identifies what they see as critical points in the performance, with the agreed score inevitably influenced by their professional relationship. The fourth provides an *aide memoire* of the cases seen, recording those critical points on which the assessment turned, in order to justify their scoring.

All four outputs turn on clinical judgment. None of the four uses categories or checklists (whether these are provided at the examinations or not). None of

the four sums the scaled scores on each category of performance to total overall scores.

Examiners' scores are expressed as one of an arbitrary set of numbers or scale scores, or as a percentage. Scales or number sets are not counting anything – they are merely agreed markers about passing or failing within which examiners' assessments are placed. Even percentages are only partly 'numerical,' since they sit within arbitrary boundaries designed for subsequent administrative actions, such as pass/fail decisions, or awarding of honors or distinctions.

Unfortunately, that administrative pass/fail decision has no follow through. The gaps in performance are not documented, or fed back to the candidate, because that is not the examiners' task or responsibility. Candidates scraping through on 50% may never have the missing 50% addressed in their intern year. Who knows if they continue as a 50% doctor for the rest of their career? Unfortunately also, the gaps in performance across all candidates are rarely recorded systematically for feedback to the clinical schools. Review of those gaps could guide both more thorough teaching of common weak areas, and some revision of the contents of the examination. Undergraduates' performance of procedural skills is not formally examined; their teaching and assessment consequently become a serious and inescapable responsibility during the intern year.

Checklists

Is what I assess the same as what you assess? Do we agree on 'what's important,' and how much 'weight' to give each aspect of clinical performance? Examiners and supervisors of undergraduate or in-service training cannot know what each other is thinking while observing a performance. Consequently, they frequently convene to discuss exactly what each believes is important, and to agree upon what should be assessed, in order to achieve some uniformity in their judgments.

The recording of that assessment poses difficulties, however. Clinical performance is patient- and context-specific, interactive as doctor and patient respond to each other, and complex in the number of significant variables being considered. Consequently, every performance is different. Unfortunately, instead of focusing on how to record performance on specific capabilities, these discussions on assessment usually step back from the qualitative details within the performance. What are substituted are generalizations about the capabilities assumed necessary for clinical work. The product is usually a checklist (or some other recording format) which selects the important categories to be watched for, often with a rating scale for the performance within each category. The categories recorded for in-service assessment of graduate trainees use general terms like 'knowledge base of basic and clinical sciences,' 'history taking,' 'preventive care skills' and 'ethical/legal principles.'

Unfortunately, these pseudorational categories are too general to predict the candidate's future performance, which is the whole point of the examination!

First, some of the categories assess *inputs* (knowledge, use of the clinical 'tool kit'), when the examination of performance is about *processes* and *outputs*.

Second, some checklist categories deal with 'principles' and very general skills ('emergency,' 'team work,' 'time management,' 'counseling'). But these skills become 'real' only during the 'hands-on' management of a particular patient, the details of which are not required in ticking that category. I know of no research that validates the assumed relationship of these general categories with actual performance. Boxes and scales contain nothing of the details of observed performance in managing a case that clinical examiners notice, discuss and record. In the extreme example of the 'objective structured clinical examination' (OSCE), examiners are denied the opportunity to make judgments on what information the candidate seeks, and on the criteria for scoring.

Third, clinical examinations cannot generalize from one case to predict performance on another case because clinical knowledge and performance are case-specific. Correlations of scores within candidates across OSCE stations may be less than 0.1 (though correlations across blind scores of clinical examiners using clinical judgment are usually around 0.7) (Newble & Swanson, 1988). That is, the statistical connection between checklist scores and case performance is weak. This point is serious if the examination or in-service assessment cannot predict whether trainees can manage critical core cases. The statistical response to low correlations is to increase the size of the sample of cases on which the candidate is tested, but that is logistically impossible in formal examinations. That larger sample size and range of cases are available within the intern year, however.

Fourth, the frequent arguments and disagreements during development of an MCQ or OSCE examination are glossed over as examiners choose the 'true answer' the candidate must select, even though it wasn't true for many of the examiners before they debated!

Fifth, and most serious, the ticks on the forms provide no useful guidance on how to correct any flaws detected in clinical performance.

Cases as the Clinical Curriculum

From this point, I focus on interns to avoid any assumption that a program at one clinical level can be applied automatically at other levels. Nevertheless, many of the principles can be applied from undergraduates through to include continuing education. The intern year epitomizes the transition from book knowledge to working knowledge (Cox, 1992a,b), from student as observer to intern as responsible for management choices, from novice watching ward procedures to manually skilled resident, and the shift from preregistration status to formal registration to unsupervised practice.¹ This shift from 'knowing' to 'doing'

requires a wide range of performance skills that can be grouped as case management with its associated procedural skills, person management, and self-management (Cox, 1999).

Cases are the unit of clinical practice, of consultation between doctors, of presentations in clinico-pathological conferences, and increasingly of continuing education, especially in general (family) practice. The writing of guidelines expresses a current attempt to devise agreed case management plans. I believe strongly that both the undergraduate and graduate clinical curricula should and will shift from disease management to case management. More significantly, the case is the unit of medical memory, whether stored mentally as a 'case prototype' (Bordage & Zacks, 1984), an 'illness script' (Boshuizen & Schmidt, 1992; Custers *et al.*, 1996), or an episode or recurrent 'story' (Bordage & Lemieux, 1991; Norman *et al.*, 1992).

Since performance is case-specific, we must decide which of the 300-odd cases doctors may face are critical cases that interns must learn to manage by the end of their preregistration year. Divide that into each clinical attachment. In each specialty, the team can choose their set of 'core cases' – road traffic accident, acute abdominal pain, post-operative fluid balance, and so on – which interns have to handle. The full set of cases for the whole year can be negotiated among teams. This list of cases becomes the agenda for performance assessment during the year, whether within hospital or in ambulatory care.

Case knowledge contains many components. Within each case, clinicians have implicit or explicit case management plans that may begin with what is urgent or serious or dangerous, say, in a child with asthma. Experience teaches what is important to look at and what to look for. Experts cull optimal subsets of the most powerful evidence on which to make a quick provisional diagnosis. Clinical teams can define the specific clinical features they include in their case prototype, and in the variations within that prototype. Everyday clinical work with the intern provides opportunities for passing on the 'practical wisdom' of tips and anecdotes that illustrate the traps for young players. The rich detail in the trade-offs within clinical judgment can be argued through in each investigational and treatment decision.

Supervision during graduate training offers the opportunity for a thorough assessment of a trainee's strengths and weaknesses over a long period in real world clinical work (Cox, 1997a). During that time performance is observed over a range of cases. Each allows extended discussion on how to manage such a patient. The assessments of trainees at various stages along their progress (as undergraduates, interns, or as graduate trainees in a college program) may be based on close personal observation, if the supervisor is directly responsible for ensuring that the trainee's clinical activities are sound. But who is usually the observer – supervisor or senior resident? Current supervision is not always adequate.

Case competence can be assessed semi-formally on current core cases within an agreed recording format by a pair of supervisors (or supervisor and senior

resident), for example, posing a sequence of ‘*What would you do if ...?*’ questions about each case. Those who pass this ‘case management test’ are ‘signed up’ on that case. Those who don’t pass get immediate and exact feedback on the areas they need to work on before they are reassessed (Cox, 1997b).

Each intern thus develops a progressive recorded profile of core cases in which performance has been assessed as competent. If the intern finishes that attachment without a positive assessment on one or more core cases, the responsibility rests with him or her to seek such cases, and to improve performance until assessed as competent and signed up.

Procedural Skills

While lists of procedural skills to be mastered during medical courses have been set out frequently enough, some components are missing in many medical schools. The steps to be followed within each procedure need to be specified in some detail, a task requiring agreement among clinicians and nursing staff (who have usually already spelled them out, but which clinicians may fail to follow). With that agreed sequence on paper as guide, trainees can observe an expert demonstration (perhaps videotaped), and then practice under supervision. During the intern year a team of clinician (perhaps a senior resident) and nurse can sign up interns judged competent to perform the procedure safely unsupervised.

Logbooks are not always popular. If interns are to be judged as performing all skills effectively and safely, however, logbooks are an inescapable responsibility of both interns and clinical supervisors. I wonder whether Pap smears are not done, or not done well, because they were never taught properly, if at all. Intra-hospital iatrogenic mishaps are so frequent that their costs to patients cannot be shrugged off.

Personal Development

The clinical task requires management of the ‘patient as person,’ the patient as a ‘case of something,’ and self-management (of interpersonal sensitivity and intrapersonal awareness). Medical school curricula stress personal development of students’ professional behavior, but conspicuously fail to ensure that it happens, or that it is assessed. Habitual behavior can be assessed only by observing habitual behavior.

Working side by side clinically over a period of time allows clinical supervisors considerable insight into a trainee’s conscientiousness and commitment. If those working with interns are senior residents and not supervisors, then senior residents must also contribute their observations on the intern’s behavior. How good is the intern’s judgment that they’re out of their depth and need help? Do

they get out of bed at night when they're called? Technical errors may be forgiven, but not those of dishonesty or avoidance of responsibility (Bosk, 1979).

Currently, behavioral flaws are underreported and underdocumented. Yet, these are central to safe practice (honesty, admitting errors, accepting responsibility, thoroughness in handover to others, and sensitivity to the needs of patients, relatives and colleagues). 'Best practice' in maintaining conscientious and ethical practice can be assessed and certified only while still under supervision. Incidents of defective behavior cannot be 'measured,' but should be thoroughly documented as part of the supervisor's report.

Current recording of professional behavior is rarely adequate. Some supervisors add comments about the trainee on their in-service assessment form. Some refer to specific instances of performance (usually about flaws, rarely about excellence). These qualitative data are not compulsory and are not systematic, though they do add color to the 'profile' of the trainee. Without any detail or substantiation from particular instances, however, tick-a-box scales and occasional comments cannot sustain any anecdotal case (e.g. in a court of law) for denying interns registration, or excluding them from progressing in vocational training programs.

The critical issue to me, however, is not just that behavioral flaws are reported, but what is done about them. The qualitative data reflecting a clinician's concerns about an intern are usually insufficient to block their passage to registration, allowing any with a flawed personality to practice unsupervised for the rest of their career. While the clinician's first step is constructive feedback to the trainee on unacceptable behavior, such admonitions may be insufficient to 'correct' the defects. Clinicians need skills in mentoring, and listening thoughtfully to their intern as a trusted friend, before a path for their improvement can be mutually agreed upon. Alas, clinicians usually complain about interns, but do little or nothing to guide their personal development. These educational capabilities within clinical supervisors seem some distance away.

Conclusions

Performance is complex. So much is going on that we are forced to rely on our judgment to identify what was done well, and what was not. That judgment should record qualitative detail, as far as is practical.

Arbitrary 'could do' categories and scores are too distant from the messy 'can do' details of everyday practice to specify what the intern can actually do. Recording forms using general categories comprise sets of 'good ideas' but with little utility, since they are inexact, cannot predict future performance, or change anything for the better.

Clinical work requires case management (including procedural skills), person management and self-management. The clinical curriculum should define a set of core cases within each practice discipline to ensure learning of essential case

knowledge and case management plans. Assessment of performance in case management and procedural skills can be conducted simply by paired examiners in the local work context.

Only those working closely with trainees can know how they respond interpersonally and intrapersonally in managing everyday clinical tasks. That intimate work relationship provides the opportunity for transferring not only practical knowledge and skills on the important cases trainees must handle, but also the expected professional behavior. Flaws in performance should be dealt with on the spot. If supervisors do not or cannot supervise or assess intern performance thoroughly, that responsibility must be assigned to others (such as senior residents) who themselves must be thoroughly trained for that set of tasks.

Notes

1. I recognize that this shift begins earlier than the intern year in some programs. The principles I am discussing can apply to whenever the shift actually occurs.

References

- BORDAGE, G. & LEMIEUX, M. (1991). Semantic structures and diagnostic thinking of experts and novices. *Academic Medicine*, 66, S70–S72.
- BORDAGE, G. & ZACKS, R. (1984). The structure of medical knowledge and the memories of medical students and general practitioners. *Medical Education*, 18, 406–416.
- BOSHUIZEN, H.P.A. & SCHMIDT H.G. (1992). On the role of biomedical knowledge in clinical reasoning by experts, intermediates and novices. *Cognitive Science*, 16, 153–184.
- BOSK, C. (1979). *Forgive and remember managing medical failure*. Chicago: Chicago University Press.
- COX, K. (1992a). What doctors need to know: a note on professional performance. *Medical Journal of Australia*, 157, 764–768.
- COX, K. (1992b). What surgeons know: a note on clinical working knowledge. *Australian & New Zealand Journal of Surgery*, 62, 836–840.
- COX, K. (1997a). Looking in the wrong direction. *Australian & New Zealand Journal of Surgery*, 67, 829–833.
- COX, K. (1997b). Work-based learning. *British Journal of Hospital Medicine*, 57, 265–269.
- COX, K. (1999). *Doctor & patient—exploring clinical thinking*. Sydney: UNSW Press.
- CUSTERS, E.J.F.M., REGEHR, G. & NORMAN, G.R. (1996). Mental representations of medical diagnostic knowledge: a review. *Academic Medicine*, 71, S55–S61.
- NEWBLE, D.I. & SWANSON, D.B. (1988). Psychometric characteristics of the objective structured clinical examination. *Medical Education*, 22, 325–334.
- NORMAN, G.R. *et al.* (1992). Expertise in visual diagnosis: a review of the literature. *Academic Medicine*, 66, S78–S83.