Can undergraduate assessment be both realistic and reliable?

R. C. GODFREY*, I. C. McMANUS**

Introduction – what is wrong?

'Examinations are now-a-days the instruments which shape training in medicine, as elsewhere. When they are compulsory, as in medicine, it would need the genius of originality to escape from their influence...'[1]

(Anonymous, 1884)

Examinations, as was indeed the case over a century ago, still dominate medical education. However, awkward questions are now, at last, being asked about undergraduate examinations. Are they reliable? Are they testing the right skills? Could they be passing unsuitable people? Might they be lacking in objectivity? What is their very purpose?

It is a truth generally acknowledged that examinations are the most powerful driving force for undergraduates;[2] and as such they have attracted the maximum critical attention of reformers. Extraordinary as it may seem, there are still medical schools which take the greatest pains to be up-to-date in their clinical practice, but are still examining by those most unreliable of indicators: the clinical case and the unstructured viva.

By way of introduction, here is the fate of three undergraduate students in recent examinations at Southampton:

Case 1. The student, an extremely bright, diligent and articulate young woman, was given a long case of deep venous thrombosis in a post operative patient. After hearing an excellent diagnosis and in-depth discussion of this problem, the examiner asked why clinical examination of the central nervous system had not been mentioned. The student replied that this seemed a clinically inappropriate task in this particular case, and she was accordingly graded B instead of the expected A.

Case 2. This student, a shy, but careful and thoughtful young man, saw a case of stroke, with a past history of transient ischaemic attacks and a myocardial infarction. As the past history seemed clearly relevant, he chose to present it first. After a few moments the examiner interrupted, saying that the presenting complaint should be described first. When the student tried to say why he was giving the history this way, he was told not to argue. His further performance was poor, and he failed.

Case 3. This student had completed one of the best intercalated BSc degrees in recent years, and had already made important research steps in the field of molecular immunology, in which he intends to make his career. In his finals 'distinction viva' the examiners were a psychiatrist and a public health physician, both of whom questioned solely within their specialised fields. The student became exasperated, convincingly lost his potential distinction, and nearly walked out in disgust.

In all three cases, there might be dispute about the justice of the outcomes, but the very fact that there is more than one tenable view reflects poorly on the objectivity of these examination encounters. In each case part of the problem arises because neither examiners nor examinees have any clearly stated aims, objectives or criteria against which performance is being assessed. In other words, it is a game without rules. Worse still, the social conditions almost guarantee disaster for some of the candidates; presentation to senior figures of the medical establishment under exam conditions is surely one of the most stressful moments of life, apt to bring out the best in some undergraduates, but equally likely to produce cessation of logical thought and near total temporary memory loss in others.

The manner, appearance and personalities of the principal players in a clinical exam inevitably produce subjectivity. But perhaps even more depends on choice of content.[3] Honest practitioners know that there are both strong and weak areas in their clinical brains; it can hardly be fair to judge an undergraduate on a mere handful of cases, especially if they happen to be the arcane curiosities often brought in for clinical exams.[4]

Written exam papers are just as bad. The search for greater objectivity is of course laudable, but the multiple choice question (MCQ) paper – albeit that its use is probably mainly driven by its greater ease for examiners – now excessively dominates

***Department of Medicine, Southampton General Hospital, Southampton SO16 6YD**
****Academic Department of Psychiatry, St. Mary's Imperial College School of Medicine at Norfolk Place, London W2 1PG**

*DR. R.C. GODFREY
**PROFESSOR I.C. McMANUS

(Correspondence to Dr. Godfrey)

the scene. Banks of questions, refined over the years, are used to test the memories of students at their outer limits and beyond, whilst ignoring the central core of basic medicine. Questions are selected statistically as good 'discriminators' – the unpredictable catch, the clever double negative.[3] In addition, all examiners have their idiosyncratic pet questions which miraculously are supposed to separate supposed sheep from supposed goats, whilst forgetting that a well-organised course should result in a uniformly high pass rate. Adequate formative assessment *en route* (in effect continual feedback upon performance), rather than 'do-or-die' assessment at the end, can happily result in a situation in which most, or even all, students pass and yet high standards are maintained.

Obsession with ranking and competition leads to another fallacy: 'the more demanding our exams, the harder our students will work, the more distinctions they will win, the higher will be our status in the league table of medical schools, and the better all our students will perform'. No advanced intellectual powers are needed to spot the *non sequiturs* in each part of this educational equivalent of the trickle-down theory. Yet this same argument causes examiners to spend disproportionate effort discriminating amongst the top 20% of high flying students. Meanwhile the difficult questions for the high achievers frighten and humiliate the 80% who want to be competent doctors rather than academics or researchers.[6]

To use educational jargon, there are serious doubts about the reliability and the validity of medical examinations. One thing can be sure, the average exam would fail abysmally the strict criteria for reliability and validity laid down for most laboratory tests. Here then are a few ideas which should be considered be any examiner.

Making the examination more realistic

A straw poll of junior doctors on the attributes needed to be effective in their daily clinical work would almost certainly reveal something along these lines (items listed roughly in order of priority):

Communication skills (with patients and staff)
When and where to seek help
Treatment of emergencies
Team work, organisation and adaptability
Good clinical technique
Differential diagnosis and selection of appropriate tests.

The last two are probably measured moderately well by current clinical examinations, but what about the rest of the skills listed above? Most undergraduate exams concentrate on a limited part of practical doctoring almost to the total exclusion of others which may seem (and probably are) more relevant and useful to the young practitioner. Of course medical education is not only about training non-consultant hospital doctors (NCHDs), it is also about a general education in medical science. Nevertheless, most medical students will actually have to be NCHDs. Table 1 proposes some practical suggestions which may increase examination realism.

Table 1: Introducing realism to examinations

Skill to be tested	*Method*
Communication skills	Standardised patients Video of consultation, with feedback
Seeking help	Use of resources during exam (including books) Allowing interval to consult resources (e.g., the McMaster 'Triple Jump')
Emergency management	Simulations (mannequins, actor) Computer programmes
Team work, etc.	Project work in small groups Group production of reports
Clinical Technique	Standardised patients Observed history taking and clinical examination
Differential diagnosis	Tests of clinical reasoning. 'Key features' approach.

The methods of testing suggested in Table 1 do not all fit readily into the format of a single examination, and they are probably better spread about different parts of the course as part of continuous assessment. As a nineteenth-century editorialist put it, 'The tactics which have so often succeeded on the turf are the best here also – a good steady pace all through the race, and a rush at the end'.[7] If, for practical and administrative reasons, some form of final examination is desirable then many different types of skill can be assessed by the objective structured clinical examination (OSCE).[8,9]

Perhaps the most important innovation in medical examinations in recent years has been the use of standardised patients – typically professional actors, or sometimes lay members of the hospital staff or community volunteers, who have been trained to play patients with particular conditions. They were originally introduced because of their strength in teaching and assessing communication skills,[10] particularly in situations calling for sensitivity and tact such as the breaking of bad news to a cancer patient and spouse, or advising a couple recently discovered to be infertile. However, they can also simulate classical physical symptoms (and even sometimes signs) of conditions such as ischaemic heart disease and peptic ulcer, making them ideal for use in examinations. Their advantages are that one can be fairly sure that all students are being assessed on the same information; they can often be made extremely rich in medical detail – a social history, family history and a complex personality, on top of a standard clinical condition;[11] they do not try to 'help' weaker students for whom they feel sorry; and special arrangements do not have to be made to have them in the ward at the right time.

The use of books and other reference sources during examinations may seem a bizarre suggestion, but what doctor has to face practice without any access to literature and colleagues? The real skill comes in setting realistic questions and tasks, to which the examinee must respond by using available resources efficiently to produce a reasoned and integrated answer – of course not just a precis of passages taken verbatim from a textbook or journal. In a so-called 'triple jump', students have a 24-hour period in which to assimilate and integrate information regarding a problem presented to them. This could be a multi-faceted clinical problem, or an invitation to study and criticise an original paper.[12]

Simulations, especially using increasingly sophisticated mannequins, are invaluable for the testing of resuscitation skills, together with many emergency procedures such as intubation, suturing and catheter placement. How many new house officers feel fully prepared for these procedures at qualification? Similarly, computers are inevitably going to find a larger place in examinations, offering the ultimate in objectivity. An imaginatively designed interactive programme on the electrolyte changes in diabetic ketoacidosis, or blood gases in different types of respiratory failure could make excellent test material, with its own built-in marking system.

Project work (including in-depth case studies) now rightly occurs more often in medical curricula; it stimulates a self-directed approach to learning and can also help to develop written and verbal presentation skills. In addition there is a suggestion that self-directed learning acquired at medical school continues well into clinical practice.[13] Why not also encourage students to work together in small groups on their projects, write their reports together, and be examined together? Much can be learnt about the social skills and interactions of the participants.

The 'key features' approach[14] seems to be a good principle on which to develop clinical skills early in a medical career. Applied to a clinical assessment, it means that examiners should be checking that students can focus on the common constellations of key symptoms and signs which occur as features of conditions such as heart failure, stroke and pneumonia. For example a student might be asked *'which combination of two physical signs distinguishes a pneumothorax from other causes of acute breathlessness?'*

Can clinical reasoning be tested?[3] It is interesting that attempts to stimulate the clinical decision-making process in so-called patient management problems (PMPs) proved difficult to set and to mark. Experienced clinicians could not agree on the best management plan, and inexperienced students often seemed to do as well or better than senior experienced doctors in parallel testing. Also the performance in tests thought mainly to measure critical reasoning correlated positively with tests of factual knowledge, spoiling the notion that ace clinicians have some divine gift. Probably the best way to bring clinical reasoning into a test, whatever its format, is to make the content as realistic as possible. Nearly every real-life patient calls for a weighing of probabilities and evidence when making clinical decisions.

Improving the reliability of examinations

Several pitfalls surrounding existing examinations have been mentioned already; the worst are lack of content specificity, use of small unrepresentative samples, subjectivity, and norm-referenced (rather than criterion-referenced) marking. Table 2 puts forward some straightforward guidelines aimed to reduce these problems.

Table 2: How to improve reliability in examinations

Problem	*Solution*
Lack of content specificity	Specify content to be assessed in clinical cases and orals. Ensure content reflects course objectives.
Small samples	Increase spread of samples, especially clinical cases. Use key features approach to enable wide cover.
Subjectivity	Construct careful marking schemes for essays, modified essays, and short answers. Check variability by parallel marking. Replace clinical case exams with OSCEs. Check inter-rater variance in OSCEs and orals. Increase use of computers.
Norm-referenced marking	Marking of all components to be strictly criterion-referenced. Separate exams for distinction candidates.

Many of the recommendations of Table 2 seem self-evident or even naive, yet how many medical schools take pains to reduce subjectivity by these simple expedients? A point in the table which is rather less generally appreciated concerns the need for a wide sampling of cases with different conditions if clinical tests are to be acceptably reliable. At least 20 cases are recommended in important exams, occupying several hours of testing time. Clearly an OSCE with 20 or more stations for a class of 150 students raises difficult, though not insuperable, logistic problems.

So far this paper has dealt exclusively with the design and execution of examinations. We hope that the present grading of students by constant testing will give way to a more enlightened system in which examinations have a less dominant place. The ultimate, no exams at all, may not be achievable, and indeed may not even be desirable; there is some evidence that examinations play an important educational role by helping

students to integrate otherwise disparate information that has been acquired at widely separated parts of a course. If that is to happen, however, then the examinations must themselves encourage integration.

A plea for simple tests of vocational ability – and some free time for education

Doctors are sometimes accused of being rather dull, usually only happy in each other's company, talking shop. Could this be possibly the result of their over-pressurised training at a time of life when character and interests are being moulded. It is difficult for medical schools to resist passing on the exponential growth in medical knowledge uncritically to their undergraduates, rather than equipping these intelligent people to think for themselves.

Moran Campbell once made the comment in reply to a question about producing better doctors: 'we're not here to train doctors; we're educators'. In other words we should stop making our programmes resemble those of exam-oriented schools and turn them into universities. How? By reducing the exam burden to what is necessary to protect the general public (e.g., a national licensing examination like that of Canada) and leaving much more space for special interests to flower according to the local interests of staff and students.

A sceptic might argue that all our demands for reliability and validity are utterly unrealistic. The answer to that should be obvious by now. If an exam is not reliable and valid then of course it is utterly unrealistic; and it can only have validity if its content is realistic. Any other form of examination is a ritual carried out for the benefit of the examiners rather than the appropriate assessment of students.[15]

References

[1] Anonymous. Medical training and liberal education (Editorial). *British Medical Journal* 1884; **ii:** 546-547.

[2] Newble, D.I. and Jaeger, K. The effects of assessments and examinations on the learning of medical students. *Med Educ* 1983; **17:** 165-171.

[3] van der Vleuten, C.P.M., and Newble, D.I. How can we test clinical reasoning? *Lancet* 1995; **345:** 1032-1034.

[4] Saunders, K.B. The MRCP (UK) exam: an examiner's view. *J Roy Coll Physicians* 1994; **28:** 369.

[5] Fleming, P.R. The profitability of 'guessing' in multiple choice question papers. *Med Educ* 1988; **22:** 509-513.

[6] McManus, I.C. Examining the educated and the trained. *Lancet* 1995; **345:** 1151-1153.

[7] Anonymous. A word to students (Editorial), *Medical Times and Gazette* 1870: 312-313.

[8] Harden, R.M., Stevenson, M., Downie, W., and Wilson, G.M. Assessment of clinical competence using objective structured examinations. *British Medical Journal* 1975; **1:** 447-451.

[9] Lowry, S. (1993) Assessment of students. *British Medical Journal* 1993; **306:** 51-54.

[10] McManus, I.C., Vincent, C.A., Thom, S. and Kidd, J. (1993) Teaching communication skills to clinical students. *British Medical Journal;* **306:** 1322-1327.

[11] Pololi, L.H. Standardised patients: as we evaluate, so shall we reap. *Lancet* 1995; **345:** 966-968.

[12] Blake, J.M., Norman, G.R., and Smith, E.K.M. Report card from McMaster: student evaluation at a problem-based medical school. *Lancet* 1995; **345:** 899-902.

[13] Bennet, K.J., Sackett, D.L., Haynes, R.B., and Neufeld, V.R. A controlled trial of teaching critical appraisal of clinical literature to medical students. *JAMA* 1987; **257:** 2451-2454.

[14] Bordage, G. and Page, G. An alternative approach to PMPs: the 'key features' concept. In Hart, I.R. and Harden, R.M., eds. Further developments in assessing clinical competence. Montreal. Heal Publications, 1987: 59-75.

[15] Godfrey, R.C. Undergraduate examinations – a continuing tyranny. *Lancet* 1995; **345:** 765-767.