The objective structured clinical examination for general practice: design, validity and reliability

DONALD M. THOMSON, MRCP, MRCGP
Lecturer in General Practice, University of Edinburgh

SUMMARY. This paper describes and analyses an experimental clinical examination for general practice. Differences in the results obtained by fourth year students, fifth year students and trainee general practitioners highlight some of the dilemmas of medical education. It is concluded that examinations which include clinical tests can increase the credibility of general practice examinations, can be reliably applied to small groups of candidates and therefore have considerable potential for formative assessment. The constraints of assuring inter-examiner reliability in a multicentre design, together with its labour intensive nature, would however make this type of examination difficult to adapt to the needs of national end-of-training examinations.

Introduction

MANY doctors are concerned about examination methods because they perceive a credibility gap between the methods and the reality of clinical practice. The widely stated view that a 'clinical' examination could increase the validity of examinations in general practice is often expressed without an appreciation of the very low levels of reliability that are achieved in traditional clinical examinations, where students are assessed on different patients with different problems by examiners using ill-defined marking criteria. In this respect clinical examinations contrast with, for example, multiple choice question examinations, which are a very reliable method of assessment — but only of factual knowledge — and which are therefore valid only for a limited part of the teaching objectives. Thus there is an apparent conflict between examinations which have good technical specifications (for example, reliability) and examinations which 'feel right' (one aspect of validity).

Although knowledge must form the basis of clinical practice and its acquisition the basis of medical teaching, the increasing volume of this knowledge has produced an imbalance in medical education.[1,2] This emphasis on the acquisition of factual/technical knowledge is often accentuated by the design of assessment methods which in the pursuit of reliability are limited predominantly to the assessment of knowledge. In addition, this imbalance in assessment is of importance since tests of knowledge, by themselves, are poor predictors of clinical performance.[3]

In the debate about the curing and caring aspects of medicine[4-6] there is, on the one hand, a legitimate anxiety that many students, both undergraduate and postgraduate, fail to achieve a 'safe' standard of knowledge; on the other hand society would appear to condemn the medical profession less for a lack of skill in combating disease but more for ignorance and insensitivity in coping with the patient's ideas, concerns and expectations.[7,8] Thus assessment methods must address not only the issue of safety and competence (the scientific aspect) but also that of interpersonal skills and attitudes (the humanitarian aspect).

Medical teachers require to develop examination methods which seek to balance the important conflicts of medical examinations — between reliability and validity, between the scientific and the humanitarian aspects of medicine and between what is examinable and what is important. This paper describes an exploration of these ideas.

Conceptual basis of the curriculum and examination

The most obvious characteristic of general practice is its breadth of clinical content. Much teaching and learning in general practice is necessarily opportunistic. The constraints of curricular time and the needs of inexperienced students, however, demand that teachers of general practice at undergraduate level evolve a rational basis for a selected curriculum.

Over the last three years the Department of General Practice at Edinburgh University has created a curriculum based on content and on process. 'Content' is a familiar concept and is normally expressed as 'teaching topics'. Educational 'process' is a less familiar concept and is normally expressed as 'knowledge, skills and attitudes', though more accurately as the 'cognitive, psychomotor and affective domains' and their component attributes. Five attributes appropriate to teaching in the context of general practice were selected from the traditional medical educational taxonomy[9-11] — 'knowledge', 'interpretation' and 'problem-solving' from the cognitive domain, 'communication' from the psychomotor domain and 'attitudes' from the affective domain. These five attributes were placed in a hierarchy — knowledge, interpretation, communication, problem-solving, attitudes — thus employing the principle, normally associated with the cognitive domain, that each higher level subsumes all lower levels. This hierarchy may be attractive to many teachers who find traditional educational concepts obscure and impractical. It implies that medical education is a progression which, although based upon knowledge, ascends above this level. It breaks down the barriers between the cognitive, psychomotor and affective domains and therefore emphasizes the interdependence of the attributes. Most importantly the ascending levels correlate well with the progressive difficulty and importance of teaching and assessment in these areas.

The requirement of this curriculum was therefore for an examination format which could test the five attributes, each of which requires a different test type; a format which could therefore accommodate both written and clinical tests.

The objective structured clinical examination was originally designed to increase the reliability of the traditional hospital clinical examination.[12] It is a circuit of 'stations' around which candidates move at five-minute intervals. Each test is designed to examine a component of clinical performance, each station utilizing a test of high objectivity. Its purpose is to retain the validity of a clinical examination but to minimize the variability attributable to the examiners by standardizing the problems, the marking criteria, the grading and, where appropriate, the patients.

Method

The core of the examination was a circuit of some 10 tests, with up to 16 stations, the additional stations being information stations (where the candidates were primed for the next station), rest stations and experimental stations (where new tests could be tried out). Thus each examination could accommodate up

to 16 candidates. An example of an examination is shown in Table 1. (Full details of questions and information briefs can be obtained from the author.)

Cognitive domain

The six tests in the cognitive domain (two each of knowledge, interpretation and problem-solving) utilized test types of known levels of reliability. The test types ranged through multiple choice questions, tests of factual recall, data interpretation tests (including clinical photographs and laboratory results) and short patient management problems. Each type used true/false, single word or short sentence answers and derived their objectivity both from the test designs and from 'ideal' answer schedules prepared and tested prior to the examination.

Psychomotor domain

There were two tests of communication skills: at one the candidates were observed taking a history, at the other giving advice. Each station was preceded by an information station where a clinical record or a hospital discharge letter was provided. A patient 'profile' with physical, psychological and social components was devised for each and the patient roles were created by an actor or actress. Each examiner was provided with a patient profile, marking criteria and a marking schedule.

Affective domain

The two tests of attitudes were preceded by a 'provocation' — either a clinical situation or a controversial text. One test was designed as a structured oral with standard lead questions, the other as a written test. Both were designed to test whether the candidate displayed an appropriate blend of a humanitarian attitude (a caring approach to the needs of patients) and a scientific attitude (a critical approach to the management of a clinical problem).

Thus each examination contained three 'live' stations (where examiners were present) and seven 'written' stations (where the answers were scored after the examination). A series of stations can be linked and thus examine different aspects of one clinical problem. At the end of the examination the candidates were provided with the 'ideal' answers and given feedback from both examiners and actors.

Results

Over one academic year the department ran 14 examinations (nine undergraduate and five postgraduate) of this design. A total of 154 candidates took part: 51 fourth year students, 54 fifth year students and 49 trainee general practitioners. The trainee general practitioners had a mean postgraduate experience of 3.2 years.

Face validity

A high level of acceptability to the candidates was achieved despite the threatening nature of what was, to all of them, a novel experience. When asked if the examination was fair, 61% of 148 candidates found it 'fair', 30% were 'neutral' and 9% thought it 'unfair' (derived from a linear analogue scale).

Difficulty and discriminating capacity

The results are displayed as frequency distributions of performance (mean of two tests) in the five attributes (Table 2) and demonstrate the different levels of difficulty and discriminating capacity of the tests in each attribute.

Test and examiner reliability

The reliability of the least objective tests, those of communication and attitudes, was assessed. At 181 of the communication tests and at 90 of the attitude tests two examiners were present. There were moderate correlations of high statistical significance between examiners in tests of these attributes (Pearson's r = 0.60 for communication and 0.71 for attitudes, $P<0.001$). Candidates undertook two tests of each attribute. There was no statistically significant relationship between the two tests for either attribute (r = 0.12 for communication and 0.03 for attitudes).

Table 1. Example of a plan of an objective structured clinical examination: 16 candidates, five minutes per station, running time 90 minutes.

	Station no.	Test	Subject	Mark	Simulated patient	Examiner
Attribute measured						
Knowledge	1	Multiple choice questionnaire	Anxiety/diabetes/otitis media	/10		
	8	List	Advantages of illness	/10		
Interpretation	2	History	Epilepsy/compliance			
	9	Photograph	Infectious mononucleosis	/10		
Problem-solving	5	Modified essay question	Dyspnoea/conflict	/10		
	12	Modified essay question	Upper respiratory tract infection/prescribing	/10		
Communication	11	History	Anxiety/pregnancy	/10	IR	JS
	4	Advice	Myocardial infarction	/10	PO	JH/MP
Attitudes	6	Written	Social prescribing	/10		
	14	Oral	Malingering	/10		JH/DM
Additional stations						
Information	3		Preparation for 4			
	10		Preparation for 11			
	13		Preparation for 14			
Rest	7 and 16					
Experimental	15					

Table 2. Frequency distribution of the marks of 154 students for the five attributes of the objective structured clinical examination.

	No. of students						
	Objective tests			Criterion referenced tests			
Mark	Knowledge	Interpre-tation	Problem-solving	Communi-cation	Attitudes	(Grade)	Overall
10	0	3	0	0	0		0
9	6	8	3	1	1		0
8	16	27	7	4	3	(Excellent)	0
7	29	32	24	24	32	(Very good)	20
6	31	39	44	63	60	(Good)	71
5	35	27	37	50	49	(Adequate)	53
4	22	14	26	12	8	(Inadequate)	10
3	13	3	13	0	1	(Awful)	0
2	2	1	0	0	0		0
1	0	0	0	0	0		0
0	0	0	0	0	0		0

Table 3. Correlation matrix (Pearson's *r*) for the five attributes of the objective structured clinical examination.

	Knowledge	Interpreta-tion	Communica-tion	Problem-solving	Attitudes	Total
Knowledge	1	0.09	0.08	0.09	0.02	0.55
Interpretation		1	0.25**	0.32***	0.05	0.66
Communication			1	0.21**	0.17*	0.52
Problem solving				1	0.17*	0.64
Attitudes					1	0.38
Total						1

* $P<0.05$, ** $P<0.01$, *** $P<0.001$.

Construct validity

Construct validity (whether a test is examining the correct attribute) is difficult to demonstrate by direct measures but indirect measures can provide some indication of this parameter.

A correlation matrix (Table 3) provides a measure of construct validity. The low correlations suggest that the attributes under test are separate attributes. The statistically significant but low correlations between interpretation, communication and problem-solving provide some evidence of a relationship between these attributes.

Mean performance in each of the attributes for each of these cohorts is displayed graphically against the candidates' length of experience (Figure 1) and demonstrates a linear acquisition of knowledge, interpretation and problem-solving; but not of communication skills in which the fifth year students did less well than their fourth year colleagues. There was no apparent difference in 'attitudes' between the three cohorts.

Predictive validity

In the two years following this experimental exercise 25 of the 49 trainee general practitioners sat the membership examination of the Royal College of General Practitioners (although over a range of time intervals between the two examinations). The correlations demonstrated a low positive but statistically significant relationship between the MRCGP results and the cognitive elements (knowledge, interpretation and problem-solving) of the examination (Pearson's $r = 0.39$, $P<0.05$). No significant relationship was found between the MRCGP results and the psychomotor (communication) or affective (attitudes) elements of the examination ($r = 0.20$ and -0.08). The small number of candidates, the variation in the time before sitting the MRCGP, and the moderate reliability of the clinical examination data do not support a more detailed analysis of this data.

Discussion

Examinations are one of the most important influences upon what a student learns, not least because they present visible evidence of the values of the discipline. In the absence of a well-defined curriculum examinations form a major stimulus to learning and passing them inevitably becomes the goal for students and teachers alike. Valid and reliable examinations have therefore as much power to support medical education as do examinations of poor reliability or low validity to distort and weaken it.

The objective structured clinical examination is less an original method than an original package of established examination methods. Other clinical examinations in general practice have been developed — for example by the Royal Australian College of General Practitioners — and circuit examinations are familiar to most medical undergraduates. The objective structured clinical examination has now been applied to several disciplines and one such examination has been described in general practice.[13] Its advantages and disadvantages have been summarized by Fabb and Marshall.[14]

Reliability

The tests of communication and attitudes are open to criticism on the basis of their relatively subjective nature. The inter-examiner correlations in communication and attitudes ($r = 0.60$ and 0.71) were in excess of those normally achieved in traditional clinical examinations (which would not be difficult) but fell below the minimum standard of 0.9 recommended by Ebel.[15] It may be argued that in the context of the validity versus reliability debate these figures are acceptable. Since the examiners were inexperienced in this examination format these figures give ground for cautious optimism that training and experience will lead to higher levels of reliability.

There is, however, an important distinction between setting

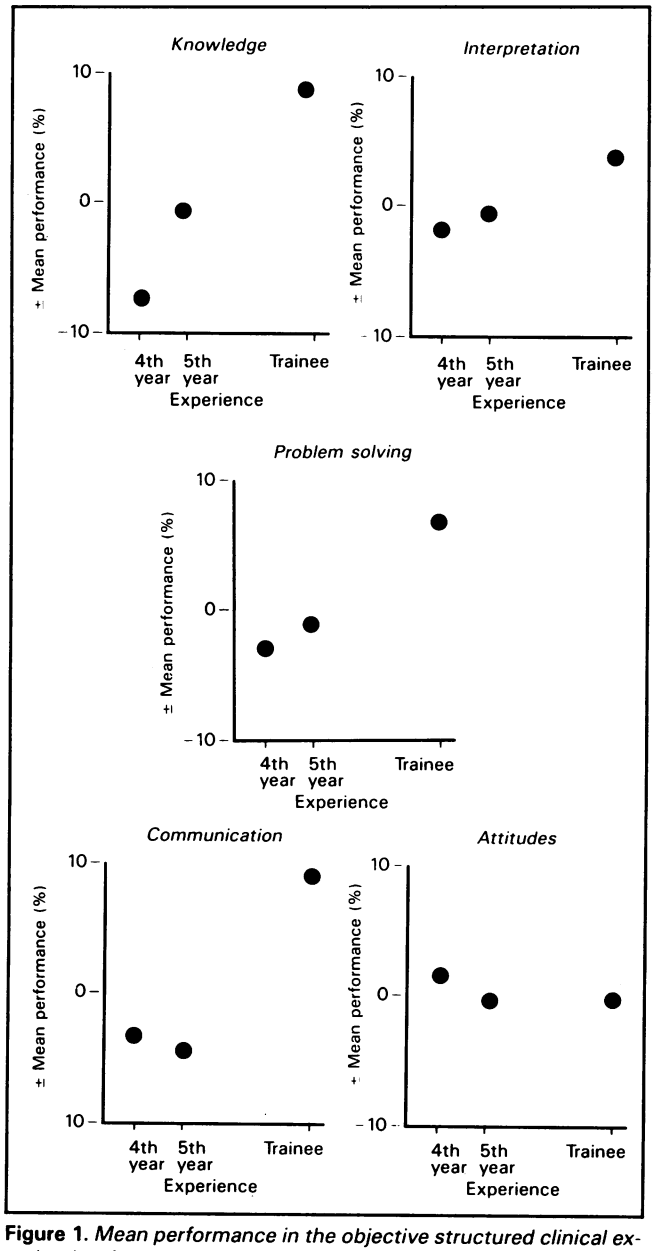

Figure 1. *Mean performance in the objective structured clinical examination for the five attributes against experience for 51 fourth year students, 54 fifth year students and 49 trainees.*

objective tests and making objective decisions about acceptable levels of competence. Objective tests still require the imposition of a subjective pass–fail mark. Anderson,[16] in commenting on multiple choice questionnaire examinations, has stated: 'examiners who think that objective assessment methods will give them a statistically valid and objective means of setting a pass mark are deluding themselves — the methods are being used to cloak and give an air of verisimilitude to what is essentially still an arbitrary decision.'

Our ability to make a judgement about the attributes of communication or of attitudes was apparently limited more by the small number of tests of these attributes than by the test design or by examiner expertise. This demonstrates one aspect of the problem of 'generalizability',[17] a limiting factor in many assessment techniques, and one of the 'unpleasant psychometric surprises'[18] that limit the conclusions that may be drawn from such tests.

Our interpretation of this finding is that the ability to demonstrate communication skills is dependent not only on the possession of that skill (process dependent) but also on the candidate's knowledge of the illness concerned (content dependent). This finding agrees with other research in medical education where there is poor evidence that any attribute (for example 'problem-solving') is a discrete entity and where it has proved difficult to demonstrate that a candidate's performance in any one test is better than 'case-specific' (a further aspect of 'generalizability').

We believe that although skill in 'history-taking' is relatively independent of content, certain aspects of communication (for example 'advice-giving' and 'reassurance') are substantially content dependent. There is therefore an important qualification to be made to the statement that 'communication skills can be taught early in the medical curriculum'.

Validity

The validity of an examination reflects not only the design of its component tests and the relationship between the examination and the educational objectives (the responsibilities of the examiners) but also the relevance and specificity of the educational objectives (the responsibility of the discipline as a whole). It is often the absence of specified educational objectives which results in an inconclusive debate on the validity of any examination in general practice.

The narrow range of total marks compared with the wider range in each attribute confirms that each student learns differently from similar educational experiences (the 'leopard effect'). This underlines the need for examinations to assess across the breadth of a curriculum not only on content but also in process.

The results of performance in each of the five attributes for the candidates at different levels of experience is compatible with evidence from published sources that undergraduate education is predominantly concerned with factual knowledge and its application,[1,2] that certain aspects of communication skills deteriorate during undergraduate education[19] (although apparently substantially improve through postgraduate experience) and that there is no evidence to show that medical education improves attitudes[20] (indeed in certain respects it may be associated with a deterioration, for example the development of cynicism).

Predictive validity (whether the results can predict future performance) is an idealized concept. Although one might wish to compare performance in an examination against a final and absolute performance indicator the only available realistic comparison is one against performance in a subsequent examination. The concept of predictive validity is flawed in other respects. It assumes that students start their education at the same level of experience and learn at a similar rate. Neither is true in postgraduate education for general practice. However, since major national examinations such as the MRCGP predominantly assess in the cognitive domain the results of this project provide evidence of appropriate construct validity of the cognitive elements of our examination.

Curriculum

One of the best arguments for experimenting with a new examination is that the insights from that experiment lead to a review of teaching methods and objectives. The most important consequence of this experiment has been to clarify some of the problems of undergraduate teaching of general practice and to lead us to seek more effective methods of teaching communication skills and attitudes.

Practicality

All clinical examinations require considerable resources in preparation and execution. The objective structured clinical examination is no exception. This 16 station/10 test examination for 16 candidates requires six members of staff (one coordinator, two actors and three examiners) and a considerable amount of space. The difficulties of maintaining consistency and concentration for actors and examiners deserves emphasis as does the requirement for training of examiners. The examination thus will be viewed differently by those who are already committed to a clinical examination and who are therefore concerned with improving reliability in contrast to those who have avoided using clinical examinations and who view the logistic problems with concern.

Conclusion

Examinations which include clinical tests can increase the credibility of general practice examinations but must do so within the constraints of inter-examiner reliability and practicality. It is suggested therefore that this assessment technique would be best employed not as a multi-centre end-of-course pass/fail ('summative') examination but as in-course ('formative') assessment on small groups of candidates, where the aim is to define, for each candidate, those areas where further study would be appropriate.

Notwithstanding the considerable potential merits of the objective structured clinical examination for formative assessment its effect upon students and teachers in stimulating debate, self-examination and peer comparison has led us to conclude that one of its most valuable applications is as a teaching technique. The development of 'competitive learning' a teaching module based on this examination will be the subject of a subsequent paper.

Only one interpretation of the objective structured clinical examination has been described and analysed here. The method is of course adaptable to the needs of widely differing educational objectives and is to be recommended to those concerned with the development of medical education as a means of exploring and measuring the conflict between what is examinable and what is important.

References

1. *Report of Royal Commission on Medical Education.* London: HMSO, 1968.
2. Robson JR. In: *Curriculum changes in United Kingdom medical schools.* Dundee: ASME, 1973.
3. Freeman J, Byrne PS. *The assessment of vocational training for general practice. Reports from general practice 17.* London: Royal College of General Practitioners, 1976.
4. Wyn Pugh E, Lloyd GJ, McIntyre N. Relevance of educational objectives for medical education. *Br Med J* 1975; **3:** 688-690.
5. Gordon D. Curing medicine and caring medicine. *Aust Fam Phys* 1978; **7:** 777-784.
6. Horder J, Ellis J, Hirsch S, *et al.* An important opportunity. *Br Med J* 1984; **288:** 1501-1511.
7. Pendleton D, Schofield T, Tate P, Havelock P. *The consultation; an approach to learning and teaching.* Oxford University Press, 1984.
8. Ackroyd E. In: Wells N (ed). *The second pharmacological revolution. Responsibilities to the consumer.* London: Office of Health Economics, 1982.
9. Bloom BS. *Taxonomy of educational objectives: the classification of educational goals. Handbook 1. Cognitive domain.* London: Longmans, 1956.
10. Krathwohl DR, Bloom BS, Masia BB. *Taxonomy of educational goals. Handbook 2. Affective domain.* London: Longmans, 1964.
11. Charvat J, McGuire C, Parsons V. *A review of the nature and uses of examinations in medical education. Public health papers no. 36.* Geneva: WHO, 1968.
12. Harden RM, Gleeson FA. ASME medical education booklet no. 8. Assessment of clinical competence using an objective structured clinical examination (OSCE). *Med Educ* 1979; **13:** 39.
13. Hall-Turner WJA. An experimental assessment carried out in an undergraduate general practice teaching course (OSCE examination). *Med Educ* 1983; **17:** 112-119.
14. Fabb WE, Marshall JR. *The assessment of clinical competence in general family practice.* Lancaster: MTP Press, 1983.
15. Ebel R. *Essentials of educational measurement.* New Jersey: Prentice-Hall, 1979.
16. Anderson J. *The multiple choice question in medicine* (2nd edn). London: Pitman, 1982.
17. Wakeford R (ed). *Directions in clinical assessment.* Cambridge University Press, 1985.
18. Newble DI. The assessment of clinical competence — a perspective from 'down under'. In: Hart IR, Harden RM, Walton HJ (eds). *Newer developments in assessing clinical competence. Conference proceedings, World Federation of Medical Education.* Montreal: Heal, 1986.
19. Sanson-Fisher R, Maguire P. Should skills in communicating with patients be taught in medical schools? *Lancet* 1980; **2:** 523-526.
20. Rezler A, Flaherty J. *The interpersonal dimension in medical education.* New York: Springer, 1985.

Address for correspondence

Dr D.M. Thomson, Department of General Practice, University of Edinburgh, Levinson House, 20 West Richmond Street, Edinburgh EH8 9DX.