A method of assessment during vocational training: report of a pilot study

I.M. STANLEY, MRCP, FRCGP

A. BELTON, FRCGP

P. FREEMAN, FRCGP

R.L. KING, FRCGP

A. REED, FRCGP

T. SILVER, FRCGP

J.H. WALKER, FFCM, FRCGP

J. WEBSTER, FSS, MBCS

H.J. WRIGHT, FRCS, FRCGP

SUMMARY. Problems arising from the present rate of failure of vocational trainees in the MRCGP examination are outlined; and the role of formative assessment during training in reducing this rate is considered. A study is described in which trainees in a number of centres were assessed by a method designed to measure specified cognitive abilities and areas of knowledge. The method, based on written papers, provides each candidate with a profile of performance and generates comparative standards. Reliability of marking, the distribution of candidate-scores within and between areas of assessment and techniques for monitoring the effectiveness of questions are reported. Use of the method by College as an educational service to trainees is considered, along with its potential as a Part I MRCGP examination.

Introduction

AT present about one-half of vocational trainees in general practice sit the MRCGP examination at or about the end of their training period.[1] The declared aim of the examination is to 'assess the ability of a candidate to carry unsupervised responsibility for the care of patients in general practice'.[2] Thus failure in the examination carries with it the implication that such individuals have not reached a basic standard; in 1983 some 30 per cent of trainee candidates were placed in this category. For trainees, failure occurs at a critical time of transition from training to practice as a principal, and represents a substantial blow to individual self-esteem.[3] For College, failure by trainees on this scale represents an unwelcome burden on Membership Division resources (including the Panel of Examiners) and invites criticism that inappropriate standards are being applied.[4]

It is clear that there exists a number of ways of reducing the failure rate among trainees while preserving standards held to be appropriate by the Panel of Examiners. One approach, currently under consideration by College, is to provide a largely formative assessment at or about the mid-point of training. This would aim to alert individuals, and those involved in their training, to deficiencies likely to lead to failure in the MRCGP examination. Such an assessment might also have a summative function and be used by College to exclude from the Membership examination individuals with a very high probability of failure. However, to achieve these aims would require an assessment method with certain 'diagnostic' and predictive characteristics. In addition the method would need to be feasible, reliable and valid in other respects.

Background to the present study

In 1979–80 the Division of General Practice at Leeds University devised and tested in undergraduates a method of assessment (IACS) designed to provide each candidate with a profile of achievement in respect of certain areas of knowledge and intellectual skills (or cognitive

I.M. Stanley, Lecturer in General Practice, University of Leeds; A. Belton, Hon. Examination Secretary, Royal College of General Practitioners; P. Freeman, Associate Adviser in General Practice, West Midlands Region; R.L. King, Associate Adviser in General Practice, Wessex Region; A. Reed, General Practitioner, Penrith, Cumbria; T. Silver, Postgraduate Adviser in General Practice, South West Thames Region; J.H. Walker, Professor of Family and Community Medicine, University of Newcastle-upon-Tyne; J. Webster, Principal Lecturer, School of Mathematics and Computing, Leeds Polytechnic; H.J. Wright, Head of the Division of General Practice, University of Leeds.

abilities).[5] The abilities chosen for assessment were directly related to clinical practice and included: observation; hypothesis formulation and testing; planning of follow-up care and of preventive care; and critical reading. The knowledge assessed related to epidemiology, statistics and mechanisms of drug action and drug interaction.

Assessment was by written paper and in use the method proved feasible, reliable and valid. In particular, high levels of reliability of marking were demonstrated and candidate achievement in one ability or area of knowledge did not predict performance in any other.

In 1981 Council of College approved a proposal to investigate the method as a possible Part I MRCGP; Membership Division subsequently set up a small working group of experienced examiners. The group has adapted the content of IACS to the assessment of trainees at or about the mid-point of vocational training. In this paper we report the findings of a pilot study of the method undertaken during 1983.

Format of the assessment

Candidates are required to complete four written papers in approximately three hours. The papers incorporate 11 areas of assessment, each relating to a particular category of knowledge or cognitive ability. Thus:

Paper One (for which 60 minutes are allowed) tests hypothesis formulation; selectivity of history taking; selectivity in physical examination; selectivity in the choice of investigations; and knowledge of epidemiological and statistical definitions.

Paper Two (60 minutes) tests understanding of: psychosocial factors; determinants of follow-up; principles of preventive care; and knowledge of a range of common clinical conditions seen in hospital and general practice.

Paper Three (40 minutes) tests the ability to comprehend and interpret the findings of a published article from a medical journal.

Paper Four (20 minutes) tests the ability to observe abnormalities when shown clinical slides.

The majority of questions seek a specified number of one-line responses to clinical situations. Multiple choice questions are used to assess clinical knowledge. Short answer questions and true/false options are used to test critical reading. Free response is sought to the clinical slides.

Marking schedules are detailed and explicit; one mark is awarded for each correct response and negative marking is employed in multiple choice items. The schedules have been derived from the responses of the Panel of Examiners of College. Not less than 25 marks are available in each of the 11 assessment areas.

Method

Vocational trainees in six centres in England were identified as willing to take the assessment on an experimental basis. In the event, 46 trainees completed it under examination conditions, standardized as far as possible, and supervised by members of the group.

All the completed papers were first marked by one of the group (I.M.S.) whose reliability of marking has been measured with similar papers from undergraduates. Thereafter the papers were marked independently by eight second markers, comprising one experienced College examiner (P.F.) familiar with the method, and seven newly recruited examiners with no previous experience of it. Each second marker received a batch of papers containing equal numbers of the four papers but with not more than one paper from each candidate; in all approximately 24 papers. In addition to marking schedules, detailed marking instructions were provided.

Findings

Data analysis was undertaken on the Leeds University Amdahl computer and was principally concerned with the reliability of marking and the achievement of candidates within and between areas of assessment.

Candidate characteristics

Data on seniority was provided by 40 of the 46 candidates: six were in the first year, 11 in the second year and 23 in the third year of vocational training. Experience of general practice at the time of assessment varied from 0–12 months (mean 4.48 months, standard deviation 3.33, $n=46$).

Inter-marker reliability

The scores achieved by candidates in each of the 11 areas at first marking were correlated with the scores gained at second marking. The results are shown for each of the eight second markers, with the number of points of comparison and the level of statistical significance (Table 1). In interpreting these figures for inter-marker correlation we are in agreement with Ebel: 'Most test constructers are reasonably well satisfied if their tests yield reliability coefficients in the vicinity of 0.90. The reliability coefficients obtained for teacher-made tests tend to fall considerably short of this goal.'[6]

Table 1. Correlation between area scores awarded by first marker and by eight second markers. (Numbers of comparisons in parentheses.)

Second marker	Correlation coefficient		Significance
1	0.87	(58)	$P<0.001$
2	0.87	(50)	$P<0.001$
3	0.89	(60)	$P<0.001$
4	0.90	(59)	$P<0.001$
5	0.90	(54)	$P<0.001$
6	0.90	(60)	$P<0.001$
7	0.93	(60)	$P<0.001$
8 (P.F.)	0.95	(59)	$P<0.001$

The chosen study design maximizes the opportunities, with a limited number of candidates, to make comparisons between first and second markers: to measure the reliability of markers independently of areas and of candidates. However this design limits the capacity to measure such reliability *within* areas; the number of available comparisons (six) is too small to make claims about any one area. Nevertheless, inspection of our data on inter-marker reliability within areas shows it to be consistent with high correlation between marker one and marker two.

Given inter-marker correlations of this order, scores from both markings are substantially the same. In the analyses which follow data from the first marking alone are used.

Achievement within and between areas

Candidate achievement within areas of assessment is shown in Figure 1 as the distribution of candidates about the mean score in each area. Three categories above and below the mean show the number of candidates placed in relation to one and two standard deviations. Two factors should be borne in mind in interpreting these distributions: first, our candidates had no prior knowledge of the assessment, its content or methods; second, volunteers are likely to be of average or above average ability. From the expected numbers in each category it can be seen that certain assessment areas (for example T1) were more 'difficult' than average while others were less so (T4). Similarly, certain areas (T2) discriminate between candidates better than others (T9). Data on the effectiveness of individual questions will be referred to later; such data make possible the 'fine-tuning' of difficulty and discrimination within each area.

Assessment area	Below			Mean		Above
	−2SD	−1SD			+1SD	+2SD
T1	3	1	22	12	8	0
T2	2	4	16	19	4	1
T3	0	6	20	14	5	1
T4	2	6	11	23	3	1
T5	0	6	17	15	7	1
T6	0	8	16	11	11	0
T7	1	7	14	18	6	0
T8	1	7	13	19	5	1
T9	0	11	11	14	10	0
T10	3	3	13	22	5	0
T11	1	7	13	21	4	0

31[a] (68%)

44[a] (95.5%)

T1 = hypothesis formulation
T2 = selective history taking
T3 = selective physical examination
T4 = selective investigations
T5 = definitions
T6 = psychosocial factors
T7 = follow-up care
T8 = preventive care
T9 = clinical medicine
T10 = critical reading
T11 = observation
[a] Numbers expected with normal distribution.

Figure 1. *The distribution of candidates about the mean in 11 areas of assessment (*n = *46 candidates)*

The extent to which the 11 areas operated independently of one another was examined by correlating scores achieved in one area with those achieved in the remaining areas. The findings are shown in Figure 2 as a correlation matrix. It can be seen that correlations are low, exceeding 0.4 in only two instances out of 55 (T5 with T7 and T9); even at this level of correlation not more than 19 per cent of variation in one score is predictable from another. Thus, in general, the scores achieved in any one area do not predict the scores in any other; the 11 areas are in large part acting as independent assessments.

Candidate achievement and experience in general practice

Analysis of candidate performance in this assessment failed to show any statistically significant relationship between achievement and seniority (that is, first, second or third year of training) or between achievement and the number of months of experience in general practice. This suggests that the timing of the assessment within a three-year vocational training programme is largely a matter of practical convenience.

Our sample size was adequate to detect a reasonable level of association between experience and performance in the test; and we are confident that if it exists at all in our group of trainees then it must be small. It remains possible that our failure to show such an association will not be confirmed by future studies and is the result of candidates being self-selected.

The effectiveness of individual questions

Analysis of data was also concerned with the discriminating power and level of difficulty of individual questions within areas. For each question the mean score, standard deviation and range of scores achieved were scrutinized to determine the degree of difficulty. The contribution of each question score to performance within an area was examined by correlating, for each candidate, the proportion of available marks achieved for that question with the proportion achieved on the area as a whole.

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11
T1	1.00										
T2	0.09	1.00									
T3	0.23	0.12	1.00								
T4	0.04	0.12	−0.10	1.00							
T5	0.05	0.15	0.26	0.19	1.00						
T6	0.34	0.05	0.08	0.18	0.08	1.00					
T7	0.23	0.08	0.13	0.34	0.42	0.17	1.00				
T8	0.30	−0.12	0.30	0.01	0.38	0.16	0.23	1.00			
T9	0.17	0.12	0.20	0.16	0.43	0.24	0.25	0.12	1.00		
T10	0.14	−0.01	0.09	−0.20	0.30	0.08	0.22	0.17	0.23	1.00	
T11	0.20	0.27	0.16	0.02	0.30	0.33	0.11	0.30	0.14	0.30	1.00

Figure 2. *Correlation between total scores achieved in different assessment areas* (n = 46 *candidates). See Figure 1 for definitions of T1–T11*

In this way a coefficient was derived for each question, representing its discriminating power relative to other questions within the same area.

Time required for marking

The time taken to mark papers by the seven examiners unfamiliar with the method was recorded by them, and the average time for each of the four types of paper is given in Table 2. In this study each examiner marked not more than six papers of each type, but from up to 24 different candidates; this tends to exaggerate the time needed for marking. Given examiners with experience of the method and larger numbers of papers of each type, it is estimated that marking will require, on average, 40 minutes per candidate.

Table 2. Average time taken in marking by the seven examiners unfamiliar with the assessment method.

	Mean time (min)	SD	Number of papers marked
Paper one	21.5	8.8	6
Paper two	17.7	5.8	6
Paper three	8.3	2.8	6
Paper four	5.6	1.1	6
Total	53.1		

Comments by candidates

Immediately after the assessment the written comments of candidates were requested by invigilators; 43 candidates (93 per cent) responded, and provided 15 distinct comments. The 10 comments made most frequently are listed in order in Table 3.

These results represent strong evidence that the time constraints were felt to be too severe. In the light of the pilot study the working group intends to reduce this pressure by removing non-discriminating questions. Ambiguity is a besetting sin of written assessments but, in our experience, is exaggerated in the recollection of candidates. Ambiguous questions do not discriminate well and this provides one method of reducing their numbers and their weighting in the measurement of achievement.[7] Among these comments there remain important criticisms of detail. While almost inevitable in a pilot study of this kind, the correction of technical shortcomings will form an important part of the remaining work of the group.

The representation of individual achievement

Given the distribution of scores and the substantial independence of assessment areas this method lends itself to the presentation of candidate achievement as a profile of performance. Such a profile would show, for each individual, a score within each of the 11 areas set against certain statistical criteria for that area. The mean score of all candidates and the scores representing one and two standard deviations above and below the mean would provide useful criteria of comparison for individuals.

Table 3. Candidates' comments about the assessment

	Number of comments	Percentage of all candidates (*n* =46)
Not enough time/guidance needed on time	23	*(50)*
Certain questions 'woolly' or ambiguous	13	*(28)*
'Critical reading' irrelevant to general practice	13	*(28)*
Wide-ranging, interesting and relevant	13	*(28)*
'Statistics' too difficult or irrelevant	10	*(21)*
Slides not always clear	6	*(13)*
A syllabus and example questions/answers are needed	4	*(8)*
Not enough space to write	4	*(8)*
The glossary (of research terms) is too long/badly arranged	4	*(8)*
Are definitions important?	2	*(4)*

Standards of performance

As the method involves 11 largely independent assessments of each individual, it provides considerable flexibility in the definition of minimum standards of performance. For example, using only statistical criteria, the probability that a candidate's scores will be less than the mean minus two standard deviations in two or more areas, purely by chance, is low (less than three in 1,000). Even lower levels of probability govern the chance occurrence of a candidate's scores lying between one and two standard deviations below the mean in six or more areas. In the present study two candidates were placed below the first of these standards and three (including one of the former) below the second standard. Thus, on the evidence of the pilot study, employing minimum standards of this kind would fail between 4 and 6 per cent of candidates. The appropriateness of these, or other, minimum standards are in part a matter of judgement; in our view decisions should await the outcome of studies of the predictive capacity of the method.

Predictive validity

All the trainees who participated in this study have declared their intention to sit the MRCGP examination during the next two years. We are investigating, and plan to report, the capacity of this method of assessment to predict future performance in the College examination. For this reason, and with the consent of our study participants, we have not provided individuals with detailed feedback on their performance.

Discussion

This is not the place, nor is it our intention in publishing these findings, to argue the case for a Part I MRCGP. Our aim is to provide evidence that a method exists which might reliably be used to assess valid attributes in vocational trainees. The method appears feasible when applied to groups of candidates and is likely to be of low cost; it provides detailed information to candidates on their achievement and useful standards of comparison. Our preliminary conclusion is that, with minor refinement, the method could be used to provide a formative assessment during the course of vocational training.

However, for the method to operate in this way there are two prerequisites: a sufficient number of trainees taking the assessment to provide meaningful standards; and adequate resources to produce, apply and monitor the assessment on a regular basis. Thus in terms of standards and resources the method is better suited to national rather than local initiative and, in our view, is the natural province of College with its tradition of expertise in assessment.

It will be argued that the method could be offered by College as an educational service to trainees wishing to make use of it. Given sufficient numbers (and we have been heartened by the level of interest shown by trainees) we see no fundamental objection to this application. With smaller numbers of volunteers the capacity to monitor the relationships between performance in this assessment (with feedback given to candidates) and performance in the MRCGP examination will be restricted; such candidates are unlikely to be a representative sample of the whole.

Should there be confirmation of the present evidence that the method is predictive of performance in the MRCGP examination,[3] then a prima facie case will exist for it as a mandatory preliminary hurdle. Used in this way the method could provide early warning of the educational needs of vulnerable trainees; moreover it could, directly and indirectly, lessen the pressure on the Membeship examination through a reduction in the proportion of poor or marginal candidates.

The question has been raised, if one examination predicts performance in another, do we need both?[8] To demonstrate significant correlation between performance in different assessments separated in time is not to prove that one is redundant; merely that there is a consistent relationship between the achievement of candidates on both occasions. Moreover, for the potential achievement of candidates on the second occasion to be realized may well require the educational stimulus of a second assessment.

References

1. Walker J H. Quantity, quality and controversy. *J R Coll Gen Pract* 1983; **33:** 545-556.
2. Royal College of General Practitioners. Obtaining and maintaining Membership: a Council discussion paper. *J R Coll Gen Pract* 1981; **31:** 521-524.
3. Adshead D, *et al.* Helping those who fail the MRCGP examination. *Med Teacher* 1984; **6:** 101-105.
4. Norell J S. What every doctor knows. *J R Coll Gen Pract* 1984; **34:** 417-424.
5. Wright H J, Stanley I M, Webster J. The assessment of cognitive abilities in clinical medicine. *Med Educ* 1983; **17:** 31-38.
6. Ebel R L. *Essentials of educational measurement.* Third Edition. New Jersey: Prentice-Hall, 1979.
7. Walker J H, Stanley I M, Venables T L, *et al.* The MRCGP examination and its methods. II: MCQ paper. *J R Coll Gen Pract* 1983; **33:** 732-4.
8. Marinker M. The MRCGP revisited. *J R Coll Gen Pract* 1984; **34:** 529-531.

Acknowledgements

We wish to express our thanks to all those trainees who volunteered to take this assessment, and their course organizers for making this possible; Carole Webster for her meticulous preparation of the papers and collection of data; and to Pauline Dallmeyer, formerly Administrator, Membership Division RCGP, for support and advice.

Address for correspondence

Professor I.M. Stanley, Department of General Practice, University of Liverpool, PO Box 147, Liverpool L69 3BX.

Conceptions outside marriage 1971–81

Pregnancies conceived outside marriage in 1971 led, in roughly equal proportions, to illegitimate births (34 per cent), legitimate births following the mother's marriage after conception (36 per cent) or termination by abortion under the 1967 Act (30 per cent). By 1981, however, the proportions leading to illegitimate births and to terminations by abortion had increased to 41 per cent and 40 per cent respectively, whereas the proportion leading to legitimate births after the mother's marriage had declined to 19 per cent.

The proportion of all the pregnancies of women aged under 20 years which were conceived outside marriage increased from 66 per cent in 1971 to 76 per cent in 1981. Of such pregnancies, the proportions which led to a legitimate birth following the mother's marriage after conception fell steeply from 45 per cent in 1971 to 20 per cent in 1981. The proportions which led to illegitimate births or which were terminated by abortion correspondingly increased over the period from 29 per cent to 39 per cent and from 26 per cent to 41 per cent respectively. Of the pregnancies conceived in 1981 by women aged under 16, 38 per cent led to an illegitimate birth and 57 per cent were terminated by abortion. The proportion of pregnancies conceived outside marriage by women aged 20 to 24 increased from 19 per cent in 1971 to 29 per cent in 1981. The trends in the proportions of such pregnancies leading to different outcomes were similar to those for teenage girls.

Source: Office of Population Censuses and Surveys. Conceptions inside and outside marriage, 1969 to 1981. *OPCS Monitor* 1984; FMI 84/6.