Evaluation of a computerized assessment package for general practitioner trainees

JOHN B DONALD

DONALD THOMSON

SUMMARY. A computerized assessment package for general practitioner trainees has been developed in order to measure the impact of teaching during the trainee year and to help identify areas of possible weakness in the knowledge base of trainees. The programme consists of an electronic questionnaire asking for background educational details, confidence rating scales, 60 multiple choice questions divided into 11 educational areas and a patient management problem to add variety and interest and to test decision making skills. A printout is produced at the end of the test summarizing the trainee's results and comparing these with the previous cohort of trainees. The programme was designed to be interesting and stimulating to the trainee by using colour and graphics, ensuring it was easy to use and providing instant feedback of results in comparison with their peer group. It was also designed to require the minimum intervention by trainers. Three sets of 70 trainees undertook the test in consecutive years with each trainee completing the assessment twice, once at the beginning of the trainee year and again towards the end of the year. In addition, a group of trainees completed a manual questionnaire asking them to rate certain aspects of the test. The results of the assessments showed a significant increase in knowledge in nearly all topic areas between the first and second tests. In general, the trainees' response to the test was positive with 63% stating it was useful in pinpointing areas of possible knowledge deficiency.

The computer package described has been shown to be an effective and acceptable method for some aspects of trainee assessment.

Keywords: vocational training assessment; assessment techniques; computer applications.

Introduction

A PHASED evaluation project (PEP) was devised by a group of doctors from the south east of Scotland faculty of the Royal College of General Practitioners in 1986. This group constructed a curriculum for trainees by creating check lists for each of the successive phases of the trainee year and these have been published.[1]

As part of the phased evaluation project it was decided to devise a computer based package of assessment for trainees in their general practice year. The two main aims of this assessment were that it should be used as near the start of the trainee year as possible to act as a diagnostic assessment to establish a baseline for core knowledge; it should then be repeated between seven and nine months later. It was hoped that having identified areas of knowledge deficiency in the first phase of the trainee year, it might be possible to show that some of these deficiencies had been corrected in the final phase. In addition, it was hoped to impart some computer literacy to trainees as although the general applications of computer aided learning are in their infancy with few examples in the literature,[2] a high proportion of practices have a computer (Ryan MP, personal communication, 1992) and most trainees will probably be exposed to a computer at some stage during their trainee year.

The decision to adopt this approach was based on the ability of computers to perform immediate analysis and provide feedback of performance to the trainee in an unbiased and as far as possible, objective way. The approach also allowed trainers to remain distanced from the tests, thus avoiding personal bias and extra work.

The reliability and validity of different assessment techniques in postgraduate education is difficult to establish with confidence and many different methods have been used. The overall consensus suggests that a number of different methods should be used in parallel.[3,4] Although a recent report has suggested that the Manchester rating scales may have some validity,[5] and they are recommended for national use,[6] there is, at present, no consistent approach by different regions of the United Kingdom.

In the development of the phased evaluation project computer programme, all existing methods of assessment[3,4,7,8] were examined and three chosen: confidence scales, multiple choice questions and patient management problems. These were the three methods which could be adapted most easily to computer presentation and marking. It was always considered that these methods should be taken in conjunction with other methods such as the Manchester rating scales, an objective structured clinical examination (OSCE) and video analysis of consultations, in order to provide as complete and objective a picture as possible of the capabilities of the trainee.

Method

Assessment design

Once the three methods of assessment had been chosen, evaluation of various programme generating software packages was carried out.[9,10] However, these packages were insufficiently flexible to cope with the layout, analysis and reporting which was required and recommended.[11] The high level language QuickBasic® (Microsoft) was therefore chosen as this gave complete flexibility of design and once the programme was compiled it was secure and could not be broken into.

The design of an electronic questionnaire and confidence scales was relatively easy, but the multiple choice questions and a patient management problem required considerable input from many sources.[12-17] A review of the literature indicated that an independent true/false format for multiple choice questions was generally adopted with minus one for wrong answers, a zero for 'do not know' and plus one for correct answers in order to penalize guesswork.[12,13,15] Each trainer in the south east Scotland area was asked to produce two multiple choice questions on specific topics and these were collated, validated and tested together with many other multiple choice questions from various sources by the steering group of the phased evaluation project. Questions were chosen on the basis of their relevance and importance to general practice. In addition, multiple choice

J B Donald, BA, FFARCS, MRCGP, general practitioner, Livingston, West Lothian, medical adviser, Forth Valley Health Board and RCGP Stuart faculty fellow 1987–92. D Thomson, FRCP, FRCGP, senior lecturer, Department of General Practice, University of Edinburgh.
Submitted: 7 January 1992; accepted: 20 May 1992.

questions relating to specific specialties were sent to the appropriate specialists for their comments and feedback. Eleven key areas relevant to general practice were identified and a final group of 60 multiple choice questions was chosen and divided into these 11 areas.

A similar exercise was undertaken to construct, test and validate a general practice oriented patient management problem.[16,17] Use was made of branching techniques, and different types of responses were required for different questions: a yes/no response, multiple correct responses graded by marking the 'degree of correctness' and single correct responses. Marking for each question was carried out using a sliding scale from plus two for an ideal course of action to minus two for an answer that was obviously wrong or even harmful. Use was also made in the patient management problem of medical test results such as biochemical values, audiometry graphs and electrocardiograph recordings which add interest to the test and help to break up a purely text based programme.

Use of the programme

Candidates switch on the computer and type 'PEP'. There are a few introductory screens and then they are presented with the electronic questionnaire asking for basic details. Each field is validated to eliminate spurious responses and candidates can correct their responses at the end of each screen. All the inputs are summarized on screen before moving to the next part of the test. The questionnaire is followed by a single screen of confidence ratings, one for each of the 11 key areas determined for the multiple choice questions. The candidates rate their confidence in these areas using a scale from one to nine and typing in an integer.

The 60 multiple choice questions follow, and as the candidates go through the test, their incorrect responses or those to which they reply 'do not know' are highlighted, so that there is an element of formative assessment in the programme. When the multiple choice questions are complete the individual marks for each topic and the overall mark are shown, as well as a bar chart which shows the trainee's mark in relation to the previous 70 trainees who have completed the test. The candidates then proceed to the patient management problem where instructions tell them which keys to press to access the various branches of the problem and the screens of graphic information. At the end of this part of the test a bar graph of comparative results is again shown.

Overall results of the multiple choice questions and patient management problem are shown on a further bar graph and two copies of a printout are produced. This shows the results of the electronic questionnaire, the confidence ratings, the results of the multiple choice questions by topic and overall, with symbols to show where the candidate lies in relation to the previous 70 trainees in terms of plus or minus one or two standard deviations, the results of the patient management problem and the overall mark (multiple choice questions and patient management problem).

The programme was designed so that it was impossible to exit from the assessment without rebooting the computer, to ensure that each candidate finished the test. The time taken to complete the test was approximately 90 minutes.

Each trainee completed the test twice (excluding the electronic questionnaire the second time), once at the beginning of the trainee year and again towards the end. The multiple choice questions and the patient management problem questions were identical in the two tests but answers to the patient management problem were given only during the second test, as it was felt that giving the answers on the first occasion would invalidate the use of the patient management problem in the subsequent test. Using the same multiple choice questions was not considered to be a problem as a candidate's memory for individual responses was unlikely to be reliable over an eight month period.

After the first year of testing and refining the questions and building up a bank of approximately 70 trainees, two further cohorts of trainees went through the test and the results were evaluated over a two year period. The results were not saved to disc but were entered into a combined database/spreadsheet programme (*Ability Plus*®, Migent) for subsequent analysis.

Candidate's access to the programme

One of the problems of computer assisted learning and assessment is access to suitable hardware to run the programmes. Many options were considered as the programme is easily transportable on a floppy disc. However, it was decided to make the programme available on only one computer and to ask each trainee to come to one general practice surgery to carry out the test. This avoided problems of computer compatibility.

Therefore at the beginning of the trainee year all the candidates were given a specific time to do the test, time off from their practices was allowed and travelling expenses refunded. Up to four trainees can complete the test in one day and all of the trainees in the region can complete the test in four to six weeks. This has proved completely successful with 100% completion rate by all trainees in the region and has the advantage of one person overseeing the candidates and computer, helping with any problems and collecting the results.

Computer equipment required

The programme can be run on any IBM® PC compatible computer running MS-DOS® (Microsoft), version 3.1 or higher, with or without a hard disc. Two versions of the programme are available on the same disc, one being mainly text based for use on a computer with a monochrome screen without suitable graphic adaptors and the other for computers with suitable colour screens and graphic adaptors.

Manual questionnaire

A manual questionnaire was circulated to one of the groups of trainees who had completed their first test, to assess the ease of use and acceptability of the test. Their responses were graded on a Likert scale of one to five — strongly negative to strongly positive response.

Statistical analysis

Statistical testing was performed using the Student's *t* test owing to the relatively small size of the samples involved.

Results

Three annual cohorts of trainees completed the test twice (1989–91). On the first occasion this was done within two months of starting their training and on the second occasion about eight or nine months into the trainee year. Results from the first cohort were discarded as the programme changed too much in response to feedback from these trainees. This left a total of 121 trainees over two years who completed the assessment programme twice (eight of the 121 were newly appointed principals, that is very recent ex-trainees).

The results were analysed in many ways, but most of the analysis concentrated on comparing the results of the multiple choice questions and patient management problem, the overall results and the self assessed confidence ratings at both the first and second tests, and various factors derived from the electronic questionnaire such as experience, postgraduate qualifications, vocational training and sex. Table 1 shows that the only signifi-

cant relationship between performance in the test and various trainee characteristics was whether the candidate was a UK graduate or not. In addition, trainees scored significantly higher on the second test than on the first. An attempt to relate the performance in the individual topics of the multiple choice questions with either experience or postgraduate qualification in that topic showed that only the relationships between medical experience of 12 months or more and a postgraduate qualification in medicine, and the score on the medical questions were significant (Table 2).

There was a significant increase in the trainees' scores between the two tests for all topics of the multiple choice questions except geriatrics, and for the patient management problem (Table 3). These improvements went hand in hand with increasing confidence as shown by the confidence ratings on Table 3. The confidence ratings that increased by the greatest amount were in the areas of psychiatry, dermatology, ophthalmology, ear, nose and throat medicine, administrative/social/legal aspects of general practice and drug therapy. These were also the topics, apart from drug therapy, where the majority of candidates' confidence ratings at the first test were 40% or less (Table 4). Table 4 shows that trainees' assessment of confidence did not correspond very closely to their actual results, even when they rated themselves to be at the lower and upper ends of the confidence rating scale. Table 4 is more useful in showing the areas where trainees felt the least confident, although this was not reflected in their actual scores for that topic.

The mean age of the 121 candidates sitting the first test was 28 years (range 24–49 years) and the mean time between qualification and sitting the first test was four years (range one to 15 years). The mean time taken for the 121 first tests was one hour 20 minutes with a minimum of 43 minutes and a maximum of three hours five minutes.

Considering the multiple choice questions, the greatest spread of results was shown for the questions in the area of administration/social/legal aspects of general practice (−25 to +100). Other wide ranging marks were for the questions on dermatology (−4 to +96), ear, nose and throat medicine (−20 to +80) and drug therapy (−16 to +95). The most consistent results were shown in the area of medicine (+20 to +80).

Table 1. Mean score for the first and second tests on the multiple choice questions and the patient management problem, by characteristics of the trainees.

Trainee characteristic	Mean score (%)[a]
Sex	
Male ($n=52$)	50
Female ($n=69$)	50
Vocational training	
None ($n=8$)	50.5
Formal ($n=33$)	50.5
Self-planned ($n=80$)	49.5
UK graduate	
Yes ($n=89$)	51
No ($n=32$)	46.5***
Total postgraduate experience (months)	
⩾41 ($n=49$)	51
<41 ($n=72$)	49
Year of training	
1990 ($n=62$)	51
1991 ($n=59$)	49
Time of test ($n=121$)	
1st test (October)	44
2nd test (May)	55***

n = number of trainees. [a] For multiple choice questions plus patient management problems. Student's t test: *** $P<0.001$.

Table 2. Mean score on the multiple choice questions on a topic, by trainees' experience or postgraduate qualification.

Experience/qualification in topic[a]	Mean topic score (%)[b]
Experience in medicine	
⩾12 months ($n=53$)	62
<12 months ($n=68$)	55***
Postgraduate qualification in medicine	
Yes ($n=13$)	63
No ($n=108$)	55**
Postgraduate qualification in paediatrics	
Yes ($n=43$)	54
No ($n=79$)	50
Postgraduate qualification in obstetrics	
Yes ($n=55$)	76
No ($n=66$)	72
Experience of psychiatry	
⩾3 months ($n=43$)	46
<3 months ($n=78$)	40

[a] In all other topics, there were insufficient numbers of trainees who had either experience or a postgraduate qualification to make statistical comparisons valid. [b] Mean of first and second tests. Student's t test: **$P<0.01$; ***$P<0.001$.

Table 3. Variation in scores between the first and second test for the multiple choice question topics and the patient management problem, and mean confidence ratings for the 11 topics.

	Mean score (%) ($n=121$)		Mean confidence rating (%) ($n=121$)	
	Test 1	Test 2	Test 1	Test 2
Multiple choice questions				
Medicine	47	56***	60	59
Surgery	49	55**	57	59
Paediatrics	41	55***	54	57
Obstetrics	68	79***	60	64
Psychiatry	34	50***	45	51
Dermatology	36	52***	33	47
Ophthalmology	41	56***	32	44
ENT	26	45***	40	50
Geriatrics	26	30	59	56
Ad/soc/legal	22	53***	45	62
Drug therapy	42	53***	42	53
Patient management problem	49	55**		

n = number of trainees. ENT = ear, nose and throat medicine. Ad/soc/leg = administration/social/legal aspects of general practice. Student's t test: **$P<0.01$; ***$P<0.001$.

The manual questionnaire was circulated to all 62 trainees who had completed their first test in October 1989. The results of the questionnaire (Table 5) show that, in general, the trainee response was strongly positive although only 32% felt it influenced their training plan. In the same questionnaire twice that percentage considered that it pinpointed possible deficiencies in their knowledge. Rather surprisingly, 42% of trainees said they actually enjoyed the experience.

Only one of the 300 doctors who have now done the test could not complete the test owing to difficulties in operating the computer, and had to repeat it at a later date when help with the computer was available.

Table 4. Variation in scores on the multiple choice questions for trainees with a high confidence rating (70% or more) and a low rating (40% or less) at the first test.

Topic	Mean score on multiple choice questions at first test (%) for trainees with confidence ratings of ⩾70%	⩽40%
Medicine	*61* (*n* = 19)	*51* (*n* = 19)**
Surgery	*55* (*n* = 12)	*47* (*n* = 20)
Paediatrics	*58* (*n* = 24)	*47* (*n* = 34)**
Obstetrics	*88* (*n* = 69)	*72* (*n* = 21)**
Psychiatry	*50* (*n* = 12)	*49* (*n* = 71)
Dermatology	*66* (*n* = 5)	*48* (*n* = 81)**
Ophthalmology	*59* (*n* = 15)	*54* (*n* = 91)
ENT	*47* (*n* = 24)	*44* (*n* = 63)
Geriatrics	*37* (*n* = 14)	*29* (*n* = 32)
Ad/soc/legal	*53* (*n* = 14)	*45* (*n* = 81)
Drug therapy	*54* (*n* = 21)	*46* (*n* = 42)

n = number of trainees. ENT = ear, nose and throat medicine. Ad/soc/legal = administration/social/legal aspects of general practice. **$P<0.01$.

Table 5. Trainees' responses to the computerized assessment.

Question	% of trainees making positive response[a] (*n* = 62)
Did you find the test useful?	*65*
Did it pinpoint possible knowledge deficiencies?	*63*
Was it relevant to general practice?	*66*
Did you enjoy it?	*42*
Has it influenced your training plan for the trainee year?	*32*
Would you do the test again?	*74*
Did you have any difficulty in operating the computer?	*3*
Did you find the patient management problem useful?	*52*

n = number of trainees. [a]Score of four or five on Likert scale.

Discussion

The results of this study should be interpreted with caution owing to the inherent and well recognized problems with the reliability and validity of the test procedures used in the assessment programme and to the possibility of 'cueing' either from the same candidate doing the test twice or from other candidates who have done the test at a different time. As the intention of the assessment is more formative than summative it is stressed in the introduction to the test that the main intention is educational and that trying to gain higher marks by cueing is counterproductive.

The lack of demonstrable relationships between trainees' experience or postgraduate qualification in a topic and their score in that topic could be a function of the inadequacy of the test procedures themselves. It is certainly debatable whether a small number of multiple choice questions and a patient management problem are sufficient to pinpoint accurately deficiencies in knowledge in specific areas. It must be acknowledged that at best, the results of the assessment can only act as a rough guide, but to increase the number of multiple choice questions for each topic to a level where the result would be more meaningful, would result in an unacceptable increase in the time needed to take the test. It must also be acknowledged that the significant increase in performance between the two tests could be a result of factors other than better training. For example, all candidates perform better in multiple choice question tests with practice and there may have been an element of cueing and recall in the second test.

In spite of the numeric results presented here this computerized assessment should be seen first and foremost as an educational, formative tool to be used in conjunction with other test procedures. The computer programme has stood the test of time in this region and enthusiasm for it has not lessened. One of the main reasons for this is that it virtually runs itself with only a small amount of secretarial input and virtually no doctor or trainer input which is the main problem with many other test procedures, such as the objective structured clinical examination. It has also been modified recently for use by non-training doctors and for use throughout the UK.

References

1. Scottish Council for Postgraduate Medical Education. *Learning and teaching general practice.* Edinburgh: SCPME, 1986.
2. Stanley I, Stephens C. Teaching problem handling in general practice: a computer assisted learning software package for medical students. *Br J Gen Pract* 1991; **41:** 155-158.
3. Freeman J, Byrne PS. *The assessment of postgraduate training in general practice.* 2nd edition. Guildford: Society for Research into Higher Education, 1976.
4. Association of Course Organisers. *The assessment of trainees in vocational training for general practice.* Ripon: ACO, 1988.
5. Difford F, Hughes R. Experience of using rating scales for the assessment of vocational trainees in general practice. *Br J Gen Pract* 1991; **41:** 360-364.
6. Joint Committee on Postgraduate Training for General Practice. *Assessment and vocational training for general practice. Final report of a JCPTGP working party.* London: JCPTGP, 1987.
7. Fabb WE, Marshall JR. *The assessment of clinical competence in general family practice.* Lancaster: MTP Press, 1982.
8. Wakeford R (ed). *Directions in clinical assessment.* Cambridge University Press, 1985.
9. Beech G (ed). *Interactive learning on the IBM-PC.* Wilmslow: Sigma Press, 1986.
10. Desch LW. Use of commercial 'authoring systems' for medical education. *Med Educ* 1986; **20:** 417-423.
11. Bartram D, Beaumont JG, Cornford T, *et al.* Recommendations for the design of software for computer based assessment — summary statement. *Bull Br Psychol Soc* 1987; **40:** 86-87.
12. Harden R. *Constructing multiple choice questions of the multiple true/false type. Association for the Study of Medical of Education booklet no. 10.* Dundee: Centre of Medical Education, University of Dundee, 1979.
13. Fleming PR. *The administration of a multiple choice question bank. Association for the Study Medical of Education booklet no. 19.* Dundee: Centre of Medical Education, University of Dundee, 1984.
14. Harden R. *Preparation and presentation of patient management problems (PMPs). Association for the Study of Medical of Education booklet no. 17.* Dundee: Centre of Medical Education, University of Dundee, 1983.
15. Anderson J. *The multiple choice question in medicine.* London: Pitman Press, 1982.
16. McGuire CH, Soloman M, Bashook PG. *Clinical simulations: selected problems in patient management.* 2nd edition. New York: Appleton Century Croft, 1977.
17. Grace M, Hanson S, Fincham SM,*et al.* A scoring technique for computerised patient management problems. *Med Educ* 1977; **11:** 335-340.

Acknowledgements

I thank Alastair Donald, Graham Buckley, the trainers in the south east of Scotland faculty and Stuart Pharmaceuticals for valuable help throughout the phased evaluation project. Special thanks are due to Anne Simpson for overseeing the programme and for data entry.

Address for correspondence

Dr J B Donald, Howden Health Centre, Howden, Livingston, West Lothian EH54 5TP. Enquiries about the programme to: PEP Office, Leith Mount Surgery, 46 Ferry Road, Edinburgh EH6 4AE or call the RCGP trainee hotline on 071-823 8645.