The role of evaluation and examinations in Colleges of General Practice*

PROFESSOR CHRISTINE MCGUIRE†

SUMMARY. The process of evaluation described represents a more rational approach to the development of examinations that will yield more reliable and valid information on many aspects of the competence of individuals and on the effectiveness of programmes. It will also clarify all levels of professional education in developing and maintaining competence.

This approach has been greatly helped by the application of a new technology of simulation which promises to become an effective tool in educational decision-making for purposes of programme evaluation and of individual self-assessment, certification, and licensing.

Introduction

As a professional organisation a College of General Practitioners, like any other specialty group, is charged by society with responsibility for establishing and enforcing commonly agreed upon standards of professional competence and conduct. In order to discharge that responsibility it is necessary, on the one hand, to measure accurately the relevant attributes of each individual who aspires to membership in the association, and on the other, to assess the adequacy of educational programmes designed to help physicians achieve and maintain the knowledge and skills necessary to meet professional standards for medical practice.

Current status of evaluation

In most professional organisations the responsibility for certifying the competence of individuals is delegated to one committee, that for programme accreditation to a second, and for programme development to yet another. In many countries these committees operate as autonomous fiefdoms, all too rarely burdened by any great accumulation of valid and reliable evidence relevant to rational decision-making. Furthermore, the shocking fact is that, for purposes of the judgments to be made, the data that are collected by these organisations are almost always totally inadequate, usually irrelevant, and frequently downright misleading.

Programme approval, for example, is generally based on information about what the teacher teaches and the setting in which he does so, not about what the learner learns and the settings in which he is capable of applying that learning. Decisions about programme development suffer from equally serious inadequacies.

The supporting data, if any, are usually limited to the enthusiasm of planners and participants about some existing similar programme or to information about what participants want—not what they need and the most efficient and effective ways of

*Adapted from a paper read at the fifth World Conference on General Practice, Melbourne, October, 1972, and published in full in the *Australian Family Physician*, Vol. 2, April, 1973, pp 183–188; excerpted and reprinted by permission of that journal.

†Professor of Medical Education, Associate Director and Chief, Research and Evaluation, Center for Educational Development, University of Illinois, College of Medicine, Chicago, Illinois, USA.

Journal of the Royal College of General Practitioners, 1974, **24**, 766—777

meeting those needs. In short, the most relevant data for programme evaluation, namely systematic evidence about the quality of the individual product each turns out, and the cost of producing that product, are rarely available to inform the judgments that will and must be made.

Thus, it is obvious that data about individual performance must be collected as much for purposes of programme development and assessment for individual guidance, licensing, or certification. However, the data that are now available for the accreditation of both individuals and programmes are seriously deficient. Their most obvious defect arises from the limited nature of professional competence sampled in most examinations. Every systematic analysis of conventional certification examinations and self-assessment programmes has led inexorably to the conclusion that, at best,* they measure predominantly the ability of the physician to recall, and possibly to express in coherent prose, fragments of information solicited by the examiner. This limitation has been found to be equally characteristic of oral, essay, and multiple choice examinations as commonly used.

Such measurement would be defensible only if it could be shown that scores on a test sampling the ability to recall detailed information, rapidly and under stress, were highly correlated with ability to analyse clinical problems or skill and judgment in managing patients. Unfortunately, such is not the case. Repeated studies of both students and practitioners indicate that the amount of information that can be recalled under conventional examination conditions bears little relation to the more complex intellectual tasks of interpreting data and solving problems, and even less relation to skills of interviewing, examining, and instructing patients. For this reason some medical schools and a few specialty boards have recently undertaken a radically different approach to both individual and programme evaluation.

New approach to the evaluation and examination process

The first step in this newer approach requires that the *purpose of the evaluation be clearly specified*: licensing and certification to determine whether an applicant meets required standards; self-assessment to identify individual strengths and weaknesses as a guide to further learning; or programme evaluation for planning, development, and accreditation.

Once the purposes of an evaluation have been set out it is essential to *decide precisely which components of physician competence are to be measured* in the proposed examination.

Two techniques of making this determination are now being employed: the first relies heavily on the opinion of experts in identifying the body of knowledge and the technical skills needed to provide adequate medical care. The second represents an empirical approach based on detailed job analyses or on the collection of many hundreds of specific anecdotes describing critical incidents illustrative of various types of effective and ineffective performance by physicians. From these data, the critical performance requirements of the specialty are derived; these define the aspects of physician competence to be measured.

The next step in a scientific approach to individual assessment and programme evaluation entails *defining each of the essential components of competence in operational or behavioural terms*. For example, one component of competence is that of gathering

*Given the usual examining procedures it seems clear that the information actually sampled in any given examination is likely to be arbitrary, and is frequently highly variable from candidate to candidate. More often than not, the specific questions asked appear to be a reflection not of the structure of modern medicine but rather of the composition of the examining body. Finally, repeated studies monotonously reveal that the grades or scores yielded by this process are not reproducible from one examiner to another, and that they often reveal more about the biases of the examiner than they do about the competence of the examinee.

adequate data on which to base a diagnosis. This in turn requires that the physician is able to take a competent history. In operational terms, an effective history can be defined as one in which, for example, the physician elicits a full statement of the presenting complaint, explores important leads, is able to communicate fully with the patient without offending his sensibilities.

Similarly, ' knowledge ' can be operationally defined as the ability to *recall* facts and concepts, to *interpret* data and to *solve* relevant problems. This type of definition of the behaviour that constitutes effective performance provides a common set of criteria which each observer can then apply uniformly in judging the adequacy of any individual's performance.

Once the essential components of competence have been identified and each has been defined in precise behavioural terms, the fourth step in the recommended approach involves the development of *appropriate instruments for collecting the data* that force the individual to show whether he can perform in the desired manner. Such test situations may range from new types of paper and pencil exercises through new types of oral and practical examinations, up to and including objective audits of hospital and office records and careful, systematic observations of actual performance in an office or clinic setting.

As the tests, audit forms, observational checklists, and other instruments are being developed, it is important to consider the next step—*data collection.* Who is to be observed, tested, or audited, and under what conditions? In reaching these decisions it is absolutely essential that provision be made for recording test response and audit or observational data in a descriptive, completely objective manner so that predetermined standards can be uniformly applied in making judgments about individuals or programmes.

However, in order to make such judgments it is necessary to *set appropriate standards of acceptable performance.* Typically, this step is taken *after* the collection of data and standards are set *ex post facto* by means of a system that inevitably dooms to failure a more or less fixed percentage of individuals in every group, irrespective of the quality of their performance—a procedure that should lead any prudent man to try to associate himself with a stupid group where his own achievements will be made to look *relatively* good. But this is only one of many ludicrous implications of a system in which the percentage of failures is fixed and the level of performance required for success is elastic.

A more rational approach requires that standards of acceptable performance be made to vary not with the quality of the group being tested but rather in accord with the nature of the tasks included in a test or observation. For some tasks—for example, prescription writing—100 per cent accuracy in avoiding lethal doses is not too much to ask; for others—for example, heart transplants—much lower levels of success may represent superb performance. In short, what is indispensable is a system of setting *a priori* standards which are based on the imperatives of caring for patients in the real world.

Once the examinations have been administered or the observations made, and the standards have been set in the manner described above, the next step in the recommended approach requires that provision be made for *summarising and reporting* the performance of the group and of each individual in a manner that yields adequate feedback to all interested parties: the individuals who are being assessed, the agencies who are certifying candidates or accrediting programmes and the persons who plan and implement those programmes.

For the purposes of certification and licensing it will be enough simply to indicate whether each individual's performance is satisfactory or unsatisfactory and the number of individuals who fall in each category. However, for purposes of diagnosis and self-

assessment both the individual being tested and those who design educational programmes will require a detailed profile of each individual's strengths and weaknesses; and programme administrators may need a cost-benefit analysis which indicates the expenditure of time and money incurred in bringing about any reported improvement in knowledge, skills, professional habits, or attitudes.

Finally, it is necessary to *make a judgment* on the basis of the data provided from the evaluation process. The nature of this judgment will depend on the purpose for which the whole process was initially undertaken. For purposes of certification the judgment will be a simple one: either the individual is certified or he is not. For purposes of guiding further learning, the judgment may entail the development of a detailed educational prescription for each individual. For the purposes of programme planning and assessment, the judgment may take the form of a decision to modify a programme to eliminate it altogether, or to replace it with an entirely different approach to the subject.

Simulation technique in testing competence

In order to implement this newer approach to the assessment process it has been necessary to devise new techniques of assessing the competence of physicians. The most interesting of these are based on the same principles of simulation that underlie the executive games; war games, space and aircraft simulators now generally familiar. However, these principles are only now being introduced in professional education and assessment. Perhaps such limited exploitation is in part due to the vision of astronomical costs conjured up by the word ' simulator.'

Reduced to its essence, however, simulation simply consists in placing an individual in a realistic setting where he is confronted by a problematic situation that requires his active participation in initiating and carrying through a sequence of inquiries, decisions, and actions. Each of these activities triggers appropriate feedback which can be used for subsequent decisions about pending action. These decisions will in turn modify the problem in different ways depending on the unique configuration of reactions and interventions each person makes. In this way a problem can be evolved through many stages until it is terminated when the individual reaches an acceptable resolution or is faced by disruptive alternatives brought about by the unique configuration of his own choices and actions.

Recently this idea has been captured in several methods of simulation that are economically and technologically feasible to use for measuring three critical components of physician competence: skill in interpreting clinical and laboratory data, judgment in patient management, and skills and attitudes in relating effectively to patients and colleagues.

Simulations for the assessment of interpretive skills

In the assessment of interpretive skills it is clear that provision must be made to test the physician's ability to interpret data that come to him through many sensory modalities: visual data, auditory data, or data from palpation and percussion to name but a few. Consequently, to assess these important skills simulated clinical and laboratory data are presented to the physician in realistic form and he is simply asked to interpret them. These simulations may be based on *photographic reproductions* of x-rays, gross and microscopic specimens, lesions, ECG or EEG tracings—all presented in a form that imitates life as closely as possible. *Sound simulators* developed for playback of high fidelity recordings of heart, breath and abdominal sounds through individual stethophones are used to present auditory data. Data that involve a combination of sound, colour and movement are presented in high quality *films or videotapes* showing the the physician-patient interview or certain aspects of the physical examination.

More recently several *three-dimensional models* have been developed; varied pathology can be inserted in these models to require the physician to demonstrate both his skill in eliciting the findings and his ability to interpret them. Perhaps the most elegant of the three-dimensional models now in existence is that developed by the University of Southern California for the training of anaesthetists—SIM I, a *computer-managed robot*, which can be programmed to present a variety of combinations of findings that can be modified in an almost infinite number of ways.

Using any or all of these modalities, the physician being tested can be required to demonstrate that he can describe precisely the findings presented in auditory, visual, or other forms, that he can interpret the significance of these findings, that he can anticipate other findings that might be related to those demonstrated, that he can predict what might exacerbate or relieve the symptoms, and that he can make some initial judgments on their significance and interrelations.

Simulations for the assessment of problem-solving skills

In the assessment of problem-solving skills it is possible to incorporate clinical and laboratory data of the type described above in various formats that closely imitate the situation confronting the physician managing a patient.

For example in the *written format*—widely known as the " rubout examination " or the " PMP " i.e., the patient management problem—the clinical situation is introduced with a brief statement of a complaint such as a physician might elicit from the patient himself, from a friend or relative who accompanied him or from a colleague who referred him. Information about the patient's condition is couched in realistic terms appropriate to its source. It is not presented in classic textbook or conference style. It may consist of only a few lines or a brief film clip in which the patient indicates his reasons for ' seeing the doctor.'

The statement of the problem is followed by a list of specific interventions or of general strategies; from these the physician selects his initial approach, e.g., further history, physical examination, laboratory evaluation or immediate therapy. He records his decision by erasing the opaque overlay or developing the latent image on a specially treated answer sheet to reveal instructions directing him to the section of his choice.

In each new section possible interventions that may yield further information about the patient are listed. The physician selects as many or as few as seem appropriate in light of the specific circumstances obtaining at that stage of the problem. Once again, as he records each decision, the results of that intervention are revealed in realistic verbal or visual form.

Both the stages in the work-up and the responses to each intervention are designed to simulate an actual clinical situation in the office or hospital setting. For example, an order for a specific test reveals a laboratory report; an order for an x-ray, electroencephalogram, electrocardiogram or audiogram, directs the physician to a photographic reproduction of the x-ray or tracing; a colour plate is provided in response to an order for a smear, culture or biopsy; the patient's reaction is reported in response to medication. Unless a consultation is requested, no interpretation of these data is offered and none is explicitly demanded; as in a conventional clinical setting, data are ordinarily provided in response to any request and these data are available for consideration in making subsequent decisions about the management of the patient.

In this way a problem is carried through many stages, in each of which the physician must make further choices based on the *specific* responses of the patient which were evoked by *his own* earlier decisions. The complications which must be managed will depend, as they do in the office or clinic, on the unique combination of specific procedures chosen at earlier stages. If an approach entails harmful interventions at any stage or

fails to encompass measures essential to the recovery of the patient, feedback is given regarding the complication that has developed and opportunity is provided to take appropriate measures to rectify the patient's condition.

If these remedial measures are inadequate, the problem may be terminated and the reader advised that the patient has suffered a relapse and has been sent to another hospital, or has been referred to a consultant or has died. Alternatively, if the corrective measures are adequate, the physician is directed to a new section of the problem representing the next stage in the work-up and management of that patient. Each simulation is thus composed of many sections which correspond to the various stages in the diagnostic work-up and treatment of the patient described, and at each stage alternate sections may be included to provide for differing approaches that physicians may wish to follow.

This format has been found suitable for problems ranging from emergency to long-term chronic disease situations; for problems that focus primarily on diagnosis as well as those that are primarily concerned with management and may involve special considerations in monitoring the patient's progress and revising therapy accordingly; and for problems that require the use of paramedical and community resources for resolution as well as those in which the immediate organic disease is the primary consideration.

The same technique has been used in the development of *computer-assisted simulations* of the clinical encounter—known as—CASE. In these computer simulations as developed at the University of Illinois College of Medicine the physician being tested seats himself at a computer terminal and is presented with a brief initial description of the patient's presenting complaint. It is then his task to determine what data he needs in the patient's work-up. He does so by interrogating the computer directly in unconstrained natural language, just as he would question a live patient, and the computer responds to his inquiries in the same manner. Similarly, he can indicate each aspect of the physical and laboratory examination that he would undertake with this patient. Again, the computer responds in appropriate natural language with a description of the physical finding elicited or with a laboratory report. On completion of his work-up, the physician records his working diagnosis and his plan of management and receives feedback from the computer regarding the congruence between his approach and that of expert consultants.

Third, the *computer-managed robot*, SIM I, described above, can be programmed not only to display various combinations of findings, but also to represent a number of problems and to respond appropriately to different interventions of the anaesthetist. Indeed, SIM I, will do everything including vomit on the physician who gags him when inserting an endotracheal tube. The robot may even " die " if the physician manages him inappropriately or responds inadequately to changes in his condition. However, the lovely thing about this patient is that he never gets tired of being examined or treated and he can be revived as often as required in the training and assessment of anaesthetists.

Finally, realistic clinical problems can be posed in the form of *live simulations*. In these simulations, an actor, a housewife or the examiner, himself, can be " programmed " to simulate a patient in an interview setting. In a simulated diagnostic interview it is the physician's task to interview the programmed patient and elicit essential data from him; in a management or therapeutic interview it is his task to cope with the total problem the patient may present. Skills in data gathering, in crisis management, triage, and office and patient management have been explored by a variety of such live simulations. In the telephone simulation, for example, the physician may be seated in an office-like setting when the telephone rings and a frantic voice on the other end of the line reports some emergency—an accident, a child suspected of ingesting poison, or a suicide threat

or the like. The physician is required to respond as he would to a real-life emergency and the adequacy of his response can be reliably judged.

In summary, written simulations, computer simulations, automated robots and live simulators can all be used to present clinical problems in a realistic form that requires the physician to respond as he would in the office or clinic setting, to receive life-like feedback from the simulator and to cope with the patient's responses to a sequence of inquiries and interventions, much as he must do in his actual practice.

Simulation for the assessment of interpersonal skills and attitudes

Finally, the physician's interpersonal skills and attitudes can be assessed by simulated interviews and conferences in which an individual—again, not necessarily an actor—is programmed to take the part of a patient, a colleague or other member of the health team. These simulations differ from the live simulations of clinical problems only in their focus: in the former, the emphasis is on solving a complex medical problem; in the latter, the situations are chosen to assess the physician's communication skills, his sensitivity to patient and colleague needs, and his attitudes toward these important people.

For example, in what purports to be a straightforward diagnostic interview, the simulated patient may, if appropriately handled by the physician, provide cues that the real problem is a deep-seated personal, family, or emotional one. Analogous simulated interviews with colleagues have been developed to deal with referral and consultation requests and even those ticklish situations in which colleagues differ sharply about the management of a particular patient.

Others have been designed to require the physician to communicate with one or more members of the health team: to give instructions to a simulated nurse, or to request assistance for his patient from a simulated dietitian or social worker, or to make a presentation and respond to colleagues in a simulated staff conference. Clearly, in such settings the physician's sensitivity to cues, his skill in gaining the understanding and co-operation of others and his professional attitudes are soon apparent.

Simulation—a developing field

Only a few of the many types of simulations now in use have been noted above. Others under development include simulations of the hospital encounter where time and cost as well as organic or psychosocial problems must be managed. Simulations are possible of community and preventive health problems, simulations of office and practice management and many others employing everything from the most simple resources available in any community to the most advanced computer and technical facilities available only in the most sophisticated settings. But where it has been introduced simulation, irrespective of the format, gives promise of being a powerful tool in both the instruction and evaluation of physicians.

REFERENCES

Abrahamson, S., Denson, J. S., & Wolf, R. M. (1969). *Journal of Medical Education*, **44**, 515–519.

Barrows, H. (1972). *Simulated Patients.* Springfield, Illinois: Charles C. Thomas.

Charvat, J., McGuire, C., & Parsons, V. (1968). *A Review of the Nature and Uses of Examinations in Medical Education.* Public Health Paper No. 36. Geneva: World Health Organisation.

Friedman, R. B. (1973). *Journal of Medical Education*, **48**, 92–97.

Gordon, M. (1971). *Computerized Automated Mannequin, A Self-Testing-Training Method.* Progress Report, American Heart Association 44th Annual Scientific Session, Anaheim, California, November 11–14. (mimeo available from Section of Cardiology, Department of Medicine, University of Miami, School of Medicine, Miami, Florida).

Harless, W. G., Drennon, G. G., Marxer, J. J., Root, J. A. & Miller, G. E. (1971). *Journal of Medical Education*, **46**, 443–448.

Levine, H. G. & McGuire, C. H. (1970). *Journal of Educational Measurement*, **7**, 63–74.

Levine, H. G. & McGuire, C. (1970). *Journal of Medical Education*, **45**, 700–705.
McGuire, C. & Solomon, L., eds., (1971). *Clinical Simulations: Selected Problems in Patient Management*. New York: Appleton-Century-Crofts & Co.
McGuire, C., Solomon, L. & Bashook, P. (In Press). *Written Simulations: Their Construction and Analysis*.
McGuire, C. & Wezeman, F. H. (In Press). *Simulation in Instruction and Evaluation in Medicine* WHO/EDUC Series. Geneva: World Health Organization.
Penta, F. B. & Kofman, S. (1973). *Journal of Medical Education*, **48**, 442–445.
Williamson, J. Personal Communication. Baltimore, Maryland: Johns Hopkins University, School of Public Health.