Medical Education, 1981, **15**, 315–322

Clinical problem-solving by medical students: a cross-sectional and longitudinal analysis

V. R. NEUFELD, G. R. NORMAN, J. W. FEIGHTNER AND H. S. BARROWS

The Programme for Educational Development and The Departments of Medicine and Family Medicine, Faculty of Health Sciences, McMaster University, Hamilton, Ontario, Canada

Summary

The evolution of clinical reasoning in medical students was studied. A cross-sectional sample consisted of randomly-selected medical students from three classes. Additionally, twenty-two students were observed at yearly intervals from the preclerkship period to the first post-graduate year. Subjects were observed in a clinical examination of a simulated patient, and their thought processes were abstracted from a 'stimulated recall' of the videotaped encounter.

The data were transcribed and coded for computer analysis, yielding several variables characterizing the clinical reasoning process, and four measures of outcome of the encounter. Analysis of variance of differences between students at various educational levels and a doctor criterion group indicated that the majority of the process variables were unrelated to educational level. By contrast, diagnostic and management outcomes were positively related to education. The single process variable which was related to both educational level and outcome was an 'hypothesis aggregate score', a measure of the content of the student's diagnostic hypotheses.

The results of the study indicate that the problem-solving or clinical reasoning process remains relatively constant from medical school entry to practice. This observation has important implications for clinical teaching and evaluation.

Correspondence: Dr V. R. Neufeld, Faculty of Health Sciences, McMaster University, 1200 Main Street West, Hamilton, Ontario, Canada L8S 4J9.

0308-0110/81/0900-0315 $02.00

Key words: *PROBLEM SOLVING; STUDENTS, MEDICAL/*psychol; EDUCATION, MEDICAL, GRADUATE; EDUCATION, MEDICAL, UNDERGRADUATE; DIAGNOSIS; CROSS SECTIONAL STUDIES; LONGITUDINAL STUDIES

Introduction

In the past two decades there has been a trend towards teaching and learning based on the notion of the clinician as problem-solver. Curricula have been developed to teach problem-solving (Spaulding, 1969; Ways *et al.*, 1973), and major efforts have been expended by medical schools (McGuire & Babbott, 1967; Rimoldi, 1961; Berner *et al.*, 1973) and certifying bodies (Senior, 1976; Lamont, 1972) towards the development of instruments which test problem-solving ability. These efforts are founded on certain implicit, though as yet unsupported assumptions about the nature of the clinical problem-solving process. Among these is the assumption that problem-solving can be taught, as evidenced by the evolution of curricula to develop problem-solving ability. The development of measures of problem-solving assumes that this is a general skill, and can be isolated from relevant knowledge. Yet there is no reason to suppose that this ability is any more amenable to instructional approaches than, for example, intelligence, and there is some evidence that problem-solving ability may be closely related to content.

Some studies have examined the problem-solving approach of clinicians. Studies by Barrows & Bennett (1972), Elstein *et al.* (1978), and Kassirer & Gorry (1978) have demonstrated that doctors generated hypotheses early in the clinical encounter,

that these hypotheses were limited in number, and that they exerted a powerful influence on the sequence and nature of clinical data acquired during the encounter.

However, there is a paucity of similar research into the evolution of the clinical reasoning process in medical students. Several small sub-studies by the workers at Michigan State University (Elstein *et al.*, 1978) suggested that medical students also generated early hypotheses although no detailed analysis of this observation was done. Several studies using 'low fidelity' problem simulations such as cards (Rimoldi, 1961), or patient management problems (McGuire & Babbatt, 1967) have been reported. Although these studies have demonstrated a relationship between performance and educational level, the nature of the simulation and the scoring methods used suggest only that students solve problems more correctly as a result of the education process. The studies have not shown whether this improvement is due to increased medical knowledge, increased exposure to similar problems, or an artifact of the simulation method (Goran *et al.*, 1973), and record little about the specific features of the process. The report by Barrows & Bennett (1972) is the only published study of the clinical reasoning process by trainees, using a high fidelity simulation (simulated patients in an actual clinical setting). This study, limited to first-year residents in neurology, demonstrated that some residents used hypotheses less predominantly than doctors, and were more 'rote' in their data gathering.

Methods

Study population

There were two components to the study (Fig. 1). In a cross-sectional component, a cohort of students was randomly selected from each of three medical classes in the three year McMaster curriculum. Thirteen students were selected from Year 1 (0–12

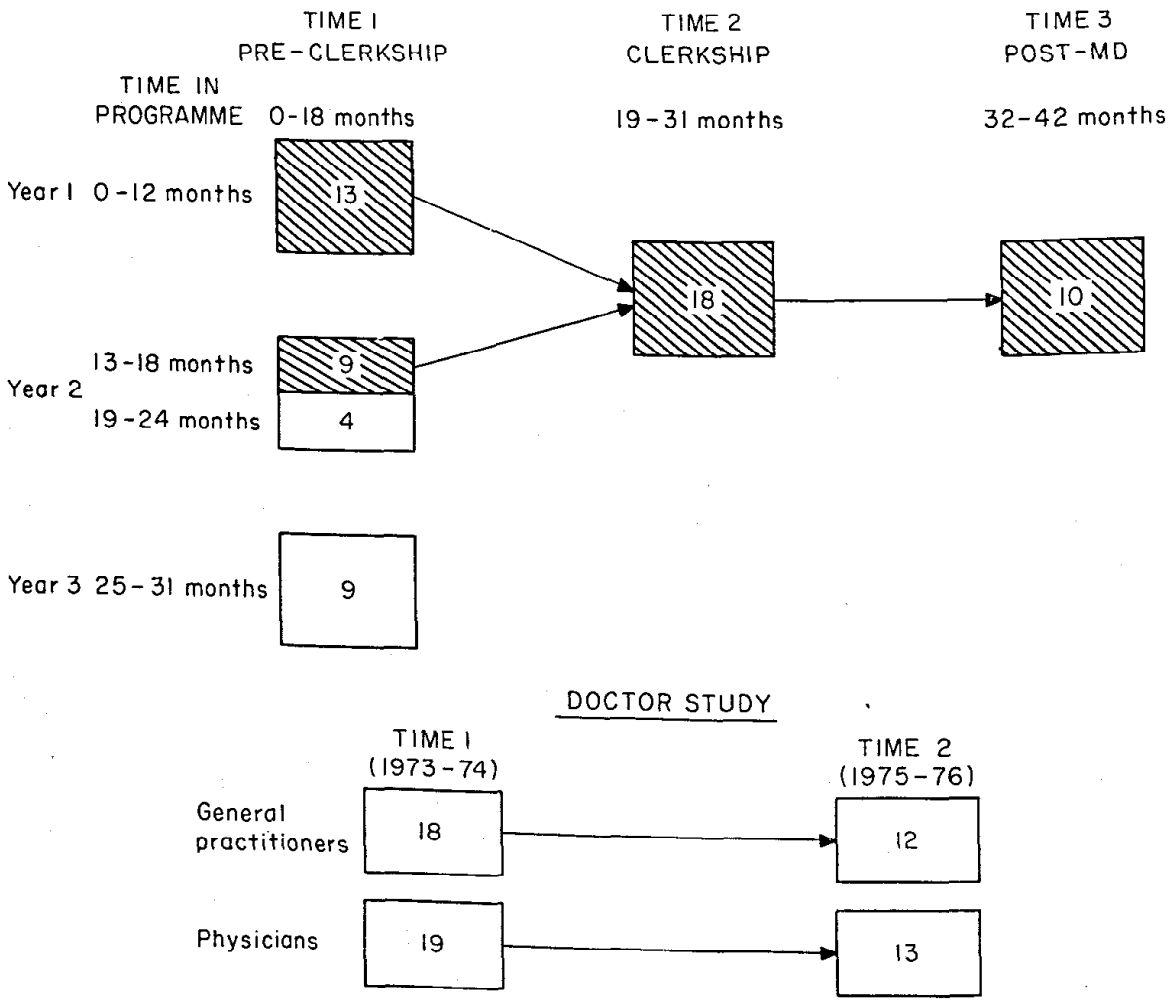

FIG. 1. □, Cross-sectional study; ▨, longitudinal study.

months), thirteen from Year 2 (13–24 months), and nine from Year 3 (25–36 months). These students were studied concurrently within the first year of the study.

In the longitudinal component of the study, three time periods were designated: Time 1 (0–18 months), corresponding to the pre-clerkship segment of the McMaster MD programme, Time 2 (19–31 months), the clerkship, and Time 3 (32–42 months), the first post-graduate year. Twenty-two students from Years 1 and 2 of the cross-sectional study, who had completed less than 18 months of their training (Time 1) were designated the study cohort for the longitudinal component, and were studied at annual intervals. Eighteen of the original twenty-two students were available for study at Time 2, and ten at Time 3. Compliance in the originally selected sample was 100%; dropouts at Time 2 and Time 3 were primarily due to geographic re-location.

The data from the students studies were contrasted with those from a similar study, reported elsewhere (Feightner *et al.*, 1977) of the clinical reasoning of general practitioners and general physicians randomly selected from the Hamilton area. Eighteen general practitioners and nineteen physicians were involved in this study. Each doctor completed one patient encounter in 1973–74. Twenty-five of the original group completed a second encounter in 1975–76, resulting in a total of sixty-two doctor encounters for the comparison.

Experimental procedure

Each student examined one simulated patient and documented the findings in a medical record, then reviewed a videotape of the encounter. Four cases were used in the study; in each case the simulated patient portrayed an accurate and consistent representation of the history, physical findings, and emotional state of the original patient upon which the simulation was based.

The cases represented a range of diagnostic and management challenges. They are briefly described as follows: (1) A 67-year-old man complaining of recent aching chest pain. He is in acute distress, short of breath and complaining of central chest pain with coughing or deep breathing. He is anxious about the condition of his heart, as he had a similar episode for which he was in hospital 2 months previously. Physical examination reveals a slightly elevated blood pressure. (2) A 22-year-old physical education student with progressive difficulty in walking, stocking-and-glove sensory loss, and some weakness in the distal leg muscles. She minimized the severity of her illness. (3) A 30-year-old woman appears depressed and complains of long-standing aching pain in the epigastrium, and non-specific tiredness. (4) An anxious 22-year-old woman with sudden paralysis in both legs. She had previous episodes of blindness in one eye and numbness on one side of the body. She has several positive findings on neurological examination in strength, sensation, reflexes and cranial nerve function.

At the beginning of an experimental session, each student was given a brief orientation to the simulated setting, his role, and the sequence of events. The student then examined the simulated patient in an examining room equipped with one-way glass. This encounter was videotaped and observed by two members of the research group. The first, a clinician, attempted to determine on the basis of the questions asked, the reasoning of the student. The second monitor rated the interaction between the student and patient using a +2 and −2 scale, with +2 representing an empathic, helpful intervention by the student, and −2 an unhelpful intervention.

Immediately afterwards, the student completed a medical record. He then reviewed the videotape with the clinician observer recalling his thinking processes during the encounter using the videotape to stimulate his recall of many points during the encounter. The clinician facilitated this process with non-directive, open-ended questions. This 'stimulated recall' was, in turn, monitored by two non-clinician researchers. The first dictated a capsule summary of the student's thoughts as recalled by the subject, keyed to time in the encounter; the second dictated the actual history taking and physical examination events.

Within no more than 48 hours of the encounter, the dictated information on history, physical examination, significant findings elicited from the patient, student-patient interaction, and stimulated recall were incorporated into a typed transcription of the encounter and marked by the subject. The student was asked to indicate which questions and physical examination manoeuvres were directly testing diagnostic hypotheses. He was also asked to weight each elicited finding against each stated hypothesis on a +2 to −2 scale, with a +2 representing a finding which highly confirmed a specific hypothesis, and −2 a totally disconfirming finding.

Following return of these documents, the encounter was coded for computer analysis. Each encounter generated 100–400 cards of numerical information. Initial analysis was conducted on each encounter to develop a number of summary statistics (e.g. time of first diagnostic hypothesis, proportion of available findings elicited). These variables were then used in examining the differences between the approaches of the students and of doctors observed in the parallel study which used the same clinical problems. These data were analysed using analysis of variance methods, with a post-hoc comparison between each student cohort and the physician criterion group.

Results

Since similar trends are seen in both the cross-sectional and longitudinal components of the study, we will display mainly the longitudinal study.

Data gathering activity in relation to educational level is shown in Table 1. No significant differences in the mean duration of the encounter were detected; however, pre-clerkship and post-graduate students asked significantly fewer questions than the doctor group. These differences are also seen in the number of history questions and physical examination manoeuvres, both of which are somewhat lower in the pre-clerkship and post-graduate group. All groups reported that approximately half the questions asked were directly testing hypotheses.

A 'significant finding' was defined as information which was judged to be necessary for diagnosis and management of the patient by the investigators and was mentioned as important by doctors in the parallel study. This definition resulted in a list of 24–37 significant findings per case. Variables describing the amount and type of significant findings obtained are shown in Table 2.

The 'thoroughness' measure is the proportion of available significant findings which were elicited. Pre-clerkship students elicited significantly fewer findings than the clinicians. However, no differences were found in the remaining groups. The 'efficiency' measure is the ratio of thoroughness to the total time of the encounter. Here, the pre-clerkship students did not perform as well as the remaining groups.

The 'significant finding integration' score expressed the average number of diagnostic hypotheses against which each finding was weighted. This measure reflects the fact that several hypotheses were being considered at the same time, a phenomenon known as parallel processing. The results show that each group was using parallel processing

TABLE 1. Data-gathering activities

	Time 1	Time 2	Time 3	Doctors
Duration of encounters	1754 sec. (355)	1660 (410)	1453 (499)	1729 (606)
Number of history questions	122* (35)	140 (37)	111* (32)	160 (59)
Number of physical examination manoeuvres	16·8* (13)	37·5 (26)	15·7 (9)	34·2 (29)
Proportion (%) hypothesis-testing questions	0·45 (0·16)	0·47 (0·16)	0·49 (0·12)	0·49 (0·19)

* $P<0·05$. Standard deviations in parentheses.

TABLE 2. Measures of data elicited

	Time 1	Time 2	Time 3	Doctors
Thoroughness	0·57* (0·14)	0·62 (0·10)	0·59 (0·09)	0·68 (0·10)
Efficiency	0·35† (0·09)	0·41 (0·13)	0·49 (0·22)	0·45 (0·15)
Significant finding integration	2·87 (1·19)	3·20 (1·45)	3·00 (1·27)	2·87 (1·05)
Average weight of significant findings	0·35 (0·27)	0·36 (0·27)	0·46 (0·17)	0·39 (0·22)

* $P<0·005$, † $P<0·05$. Standard deviations in parentheses.

to an equal extent. Finally, the average weight assigned by subjects to elicited findings was about 0·4 on the previously described +2 to −2 scale, suggesting that all groups tended toward seeking confirmatory data. Thus, although there were small differences in the amount of information elicited, the manner in which the information was used in making judgements among competing diagnostic hypotheses appeared similar in students and doctors.

The results summarized in Table 3 show that the generation of multiple hypotheses early in the encounter is a fundamental feature of the problem-solving process, not only of doctors (Barrows & Bennett, 1972; Elstein *et al.*, 1978), but also of medical students at all educational levels. The first hypothesis is generated by all student samples on the average within 30 seconds of eliciting the chief complaint. The correct hypothesis, when considered, is defined approximately 7 to 11 minutes into the encounter, again with no significant differences across groups. The total number of hypotheses generated over the course of the encounter is about six, again consistently across educational level. A small difference is present in the average specificity of hypotheses, with the pre-clerkship group using slightly more general hypotheses, based on a 4-point scale developed by the investigators for each proposed hypothesis.

The 'hypothesis aggregate score' (Norman *et al.*, 1977), is a measure of the content of the hypotheses. Each hypothesis mentioned by the doctors is assigned a weight based on the proportion of the doctors who mentioned it. These weights are then aggregated over all the hypotheses mentioned by an individual student or clinician, with an appropriate denominator to obtain a score ranging from zero to one. This score reflects the similarity between the hypotheses of an individual subject and a pool of 'ideal' hypotheses derived from the criterion group. This aggregate score was found to be related to educational level, in contrast to the other characteristics of the process of hypotheses generation described above.

Thus, only small differences were observed in the number of questions asked and the amount of clinical information obtained. However, the manner in which the data were used, and the measures of hypothesis generation were similar for all groups. The only measure of the clinical reasoning process which was consistently related to education level was the hypothesis aggregate score, a measure of the content of the diagnostic hypothesis.

In examining the outcomes of the encounter, four measures were defined: accuracy of diagnosis; management (including all investigations and consultations); therapy (statements directly related to patient management); and doctor-patient relationship (the average interaction level obtained from the comments of the second monitor). The first three outcomes were scored using the aggregate score technique described above (Norman *et al.*, 1977). The outcomes for the four groups are shown in Table 4. Data from both the longitudinal and cross-sectional study components are displayed.

There were consistent significant differences in all measures across groups for both the longitudinal and cross-sectional analyses. Of some interest is the finding that the graduate group scored significantly higher than the clinicians in management and therapy, using scores derived from the clinicians themselves. Thus in general, a marked and consistent improvement in the outcome scores was evident with increasing education.

TABLE 3. Characteristics of early hypothesis generation

	Time 1	Time 2	Time 3	Doctors
Time of first hypothesis	34 sec. (41)	27 (37)	29 (47)	28 (44)
Time of correct hypothesis	662 (731)	655 (787)	427 (515)	383 (445)
Number of hypotheses	6·04 (1·67)	6·39 (1·85)	5·60 (1·57)	5·53 (1·82)
Average specificity of hypotheses (1=specific diagnostic entity; 4=symptom)	2·13* (0·45)	2·06 (0·36)	2·00 (0·50)	1·91 (0·40)
Hypothesis aggregate score	0·58† (0·23)	0·60† (0·17)	0·78 (0·20)	0·81 (0·12)

* $P<0·05$; † $P<0·005$. Standard deviations are in parentheses.

TABLE 4. The outcomes of the encounter

	Longitudinal component Time 1	Time 2	Time 3	Doctors
Diagnosis aggregate score	0·44*	0·72	0·81	0·81
	(0·35)	(0·26)	(0·24)	(0·21)
Management aggregate score	0·38*	0·50‡	0·77‡	0·71
	(0·29)	(0·26)	(0·18)	(0·17)
Therapy aggregate score	0·23*	0·34†	0·69‡	0·63
	(0·29)	(0·28)	(0·22)	(0·21)
Interaction level	0·18†	0·10‡	0·30	0·43
	(0·38)	(0·55)	(0·32)	(0·47)
	Cross-sectional component Year 1	Year 2	Year 3	Doctors
Diagnosis aggregate score	0·33*	0·63	0·74	0·81
	(0·35)	(0·29)	(0·19)	(0·21)
Management aggregate score	0·35*	0·49†	0·68	0·71
	(0·28)	(0·28)	(0·29)	(0·17)
Therapy aggregate score	0·23*	0·33†	0·48	0·63
	(0·28)	(0·36)	(0·35)	(0·21)
Interaction level	0·10†	0·50	0·60	0·43
	(0·47)	(0·48)	(0·50)	(0·47)

* $P<0·001$; † $P<0·01$; ‡ $P<0·05$. Standard deviations in parentheses.

To determine whether certain aspects of the process of clinical problem-solving could be identified as important predictors of the outcome measures, a multivariate multiple regression analysis was conducted using the outcome measures as dependent variables. All the process measures, and months of training were used as independent variables. Overall, the process measures accounted for 56% of the variance in diagnostic score, and about 40% of the variance in the other outcome measures. However, only two variables made a significant contribution to the regression: education level, accounting for about 20% of the variance in all outcomes; and the hypothesis aggregate score, which accounted for an additional 12% of the variance in diagnosis, and 8% in the average interaction level score.

Similar results were obtained from a separate regression analysis of the doctors: the hypothesis aggregate score accounted for 23% of the variance in the diagnostic measure.

Thus, the content of the diagnostic hypotheses advanced during the course of the encounter, is a central feature of the problem-solving process. The one variable which was related (except for a few minor differences) to educational level was also the content of the diagnostic hypotheses. This variable, in turn, was the major determinant of diagnostic outcome once the effect of the educational level was removed.

Discussion

We have found that the clinical reasoning process in medical students is remarkably similar in its basic structure to that of doctors. Senior students and doctors obtained slightly more significant data than junior students, but the central elements were similar among students at several educational levels and the doctor criterion group. These include: (a) the number and timing of diagnostic hypotheses, (b) the weight of findings against hypotheses, (c) the extent of parallel processing of findings against, multiple hypotheses, and (d) the extent to which the questions on history and physical were used to test hypotheses.

However, one process measure did change consistently with education—the content and specificity of the diagnostic hypotheses. Furthermore, this measure was the best predictor of the diagnostic outcome. These outcomes were, in turn, strongly correlated with increasing education.

There are a number of potential limitations to the generalizability of these findings. The problem-based McMaster curriculum is unusual, and the admissions process, designed to select students who will thrive in this program, may result in an atypical pool of accepted applicants (Neufeld & Barrows, 1974). Thus it is not known the extent to which these students are representative of the general population

of medical students. In general, however, our findings are consistent with studies of the growth of clinical expertise which show the effect of experience on competence (Mazzuca *et al.*, 1980).

The sample of patient problems was small, and it is possible that with a broader range of problems, differences would emerge which were not apparent in the present study. It is also conceivable that the variables selected were not sufficiently sensitive to detect changes in clinical reasoning with education. Again, however, our findings are consistent with those of Elstein and colleagues (1978).

To the extent that these findings are valid, they have revealed certain features of the problem-solving process as a function of educational level. The diagnostic hypotheses advanced early in the encounter are a central feature of clinical problem-solving of both medical students and doctors. The dynamics of this process—the number of hypotheses, their time of generation, the number of questions related to each hypotheses—are remarkably consistent in all groups. However, the content and accuracy of the diagnostic hypotheses does change significantly with increasing education.

This description of the problem-solving process makes it clear that problem-solving cannot be viewed as a single general skill to be taught and evaluated. Rather, it is an integration of several competencies including (1) the application of knowledge and clinical experience retrievable from memory for use in an encounter with a patient; we think this factor is a direct antecedent of the diagnostic hypotheses, (2) the elements of the clinical reasoning process—the weighting of elicited data against these hypotheses, and the search for additional data, (3) the interviewing and physical examination techniques required to obtain information from the patient, and (4) the interpersonal attributes needed to establish rapport and maintain communication. The interaction between these professional competencies, the characteristics of the patient, and the features of a particular illness, lead to measurable outcomes for a given encounter.

We now have evidence that to a remarkable degree, many of the components of the clinical problem solving process are present in medical students at the outset of their training and are quite similar to those of doctors. Clinical teachers should therefore recognize and make more use of these characteristics in their teaching.

We suggest that instruction should be directed, not so much at 'the problem-solving process' *per se*, but at those behaviours which are amenable to change. These include the acquiring of knowledge and experience, and the learning of clinical data-gathering techniques. These components can be isolated in an educational programme. Important personal attributes are likely in place long before a student reaches medical school, hence the importance of a selection system whch looks for these qualities and of a learning environment which allows them to be expressed. The results of the present study indicate that the basic elements of the clinical reasoning process may be present from very early in undergraduate education. Whether these features can be identified at admission, and whether stable individual differences in these skills exist, remains to be demonstrated. The clinical reasoning process can be incorporated into an instructional program by the use of clinical problem situations. Clinical teachers can expect students to be able to generate hypotheses, though these may be quite general at first, and students can be encouraged to ask for clinical information which confirms or disconfirms their hypotheses.

References

BARROWS, H.S. & BENNETT, K. (1972) The diagnostic (problem-solving) skill of the neurologist. *Archives of Neurology*, **26**, 273–7.

BERNER, E.S. *et al.* (1973) A new approach to evaluating problem-solving in medical students. *Journal of Medical Education*, **49**, 666–72.

DUNCKER, K. (1945) On problem-solving. *Psychological Monographs*, **58**, 270.

ELSTEIN, A.S. *et al.* (1978) *Medical Problem-Solving: An Analysis of Clinical Reasoning*, Harvard University Press, Cambridge, Massachusetts.

FEIGHTNER, J.W. *et al.* (1977) Solving problems: how does the family physician do it. *Canadian Family Physicians*, **23**, 67–71.

GORAN, M.J. *et al.* (1973) The validity of patient management problems. *Journal of Medical Education*, **48**, 171–2.

KASSIRER, J.P. & GORRY, G.A. (1978) Clinical problem-solving: a behavioural analysis. *Annals of Internal Medicine*, **89**, 245–55.

LAMONT, C.T. (1972) The use of simulated patients in a certification examination in family medicine. *Journal of Medical Education*, **47**, 789.

MAZZUCA, S.A. *et al.* (1980) *Assessing the Evaluation of Patterns of Clinical Knowledge Across Levels of Medical Training*. Abstract, American Educational Research Association, 1980 Annual Meeting, p. 120.

MCGUIRE, C.H. & BABBOTT, D. (1967) Simulation technique in the measurement of problem solving skills. *Journal of Educational Measurement*, **4**, 1–10.

NEUFELD, V.R. & BARROWS, H.S. (1974) The McMaster

philosophy: an approach to medical education, *Journal of Medical Education*, **49**, 1040–50.

NORMAN, G.R. *et al.* (1977) Measuring the outcome of clinical problem-solving. *Annual Conference on Research in Medical Education*, **16**, 311–16.

RIMOLDI, H.J.A. (1961) The test of diagnostic skills. *Journal of Medical Education*, **35**, 73–9.

SENIOR, J.R. (1976) *Toward the Measurement of Competence in Medicine*. Report of CBX Project, Philadelphia.

SPAULDING, W.B. (1969) The undergraduate medical curriculum (1969 Model): McMaster University. *Canadian Medical Association Journal*, **100**, 659–64.

WAYS, P.O. *et al.* (1973) Focal problem teaching in medical education. *Journal of Medical Education*, **48**, 565–71.

Received 17 *December* 1980; *accepted for publication* 23 *February* 1981