Medical Education, 1985, **19**, 344–356

Knowledge and clinical problem-solving

G. R. NORMAN, P. TUGWELL, J. W. FEIGHTNER, LINDA J. MUZZIN AND L. L. JACOBY

Program for Educational Development, McMaster University, Hamilton, Ontario, Canada

Summary. A consistent finding in the literature on measures of clinical problem-solving scores is that there are very low correlations across different problems. This phenomenon is commonly labelled 'content-specificity', implying that the scores differ because the content knowledge necessary to solve the problems differs. The present study tests this hypothesis by presenting groups of residents and clinical clerks with a series of simulated patient problems in which content was systematically varied. Each subject also completed a multiple choice test with questions linked to each diagnosis presented in the clinical problems. Three of the four problem-solving scores showed low correlations, even to two presentations of the same problem, and no relationship to content differences. None of the scores were related to performance on the multiple choice test. The results suggest that variability in problem-solving scores is related to factors other than content knowledge, and several possibilities are discussed.

Key words: *Problem-solving; *Diagnosis; *Internship; *Clinical clerkship; Canada; Specialties, medical

Introduction

The ability to define and manage clinical problems is viewed as central to clinical competence in medicine (American Board of Internal Medicine, 1979), and is a pervasive theme in educational objectives in most medical schools. This ability is usually viewed as a general skill described by a variety of terms (problem-solving, clinical reasoning, clinical judgement, diagnostic skill, synthesis, etc.) which interacts with, but is distinct from knowledge.

The increasingly important role of clinical problem-solving in undergraduate and postgraduate education and the explicit recognition of the skill by certification and licensing bodies has led to the development of a variety of evaluation methods to assess this skill. They may involve a real patient, in the circumstance where the supervisor directly observes the interaction between physician and patient (Finkel & Norman, 1973), a written problem such as a patient management problem (McGuire & Babbott, 1967), or a computer-based problem, exemplified by the CBX project of the American Board of Internal Medicine and the National Board of Medical Examiners (Senior, 1976).

Considerable research effort has been expended to establish the reliability and validity of these methods. Inter-observer reliability has been demonstrated for direct observation methods (Finkel & Norman, 1973), and a number of studies have explored construct and concurrent validity of patient management problems (Mazzuca *et al.*, 1981; Schumacher, 1983; Case, 1981; Skakun *et al.* 1979). In general, the methods possess adequate internal consistency (McGuire & Babbott, 1967; Helfer & Slater, 1971), and have demonstrated moderate

Correspondence: Geoffrey R. Norman PhD, Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada, L8N 3Z5.

correlation with multiple choice tests (Case, 1981; Skakun *et al.* 1979), which has been interpreted as evidence that they are testing some other important dimension of competence. Construct validity, showing positive change in scores with increasing education has been frequently (Mazzuca *et al.*, 1981; Robinson & Dinham, 1977), but not always (Marshall, 1977) demonstrated. Other studies have used techniques such as factor analysis to provide confirmatory evidence that the methods are assessing one or more general skills (Berner *et al.*, 1977).

However, research conducted over the past few years using a variety of patient formats and scoring approaches has consistently revealed one disquieting finding—there is apparently very little relationship between performance of a student or resident on one problem and his or her performance on a second dissimilar problem. This finding has emerged from studies using live simulated patients (Norman & Tugwell, 1982; Elstein *et al.*, 1978), patient management problems (Berner *et al.*, 1977; Donnely *et al.*, 1982), and computer simulations (Skakun *et al.*, 1979; Norcini *et al.*, 1983), which all consistently report correlations across problems of 0·3 or less.

These low correlations suggest that general skills such as data-gathering, problem-solving, or clinical judgement, if they exist, account for very little of the observed variation in performance. In attempting to explain these findings, some authors have suggested that clinical problem-solving is 'content-specific' (Elstein *et al.*, 1978); that is, the solution of a single clinical problem requires mastery of knowledge specific to that problem, and this knowledge is sufficiently variable that no consistency of performance across problems can be observed. This explanation, although plausible, cannot be confirmed by the observation of a low correlation across problems, since other variables than content knowledge may contribute to this variability. A critical test of the hypothesis of content specificity would require some experimental control over the relevant content knowledge, either by using a concurrent measure of knowledge relevant to the problem or by systematically varying the content of problems from very similar to very dissimilar.

The present study incorporates both these approaches. Samples of subspecialists, residents and clinical clerks were challenged with ten clinical problems portrayed by live simulated patients, and a multiple choice test of relevant content. The clinical problems varied from two presentations of the same problem by different actors at different times, to problems in the same subspecialty with differing chief complaint, different final diagnoses, and problems in a different subspecialty. The objective of the study was to determine to what extent variability in performance across problems could be attributed to different complaints, diagnoses and specialties, and to measures of relevant knowledge, all of which represent variability which could be related to content-specificity.

The study was designed to examine the relationship between content knowledge and problem-solving using two constructs.

(1) The proportion of the total variance in problem-solving scores attributable to: (a) the same problem presented on a second occasion; (b) a second problem with the same complaint and different diagnosis; (c) a second problem with the same diagnosis but a different presenting complaint; (d) a problem in the same subspecialty; and (e) a problem in a different subspecialty.

(2) The relationship between scores on a multiple choice knowledge test and problem-solving scores.

Methods

The study involved a total of thirty subjects; five specialists in rheumatology and five in cardiology or respirology, ten second-year residents in internal medicine and ten second-year medical students. All but one specialist were academic physicians at McMaster University.

Eight simulated patient problems were developed for the study—four in cardiorespiratory and four in rheumatology. The patient problems were selected according to the scheme of Fig. 1, so that each problem shared either a presenting complaint or diagnosis with another problem in the same specialty area.

Protocols were initially developed by specialists associated with the research team; these protocols were then distributed to a small

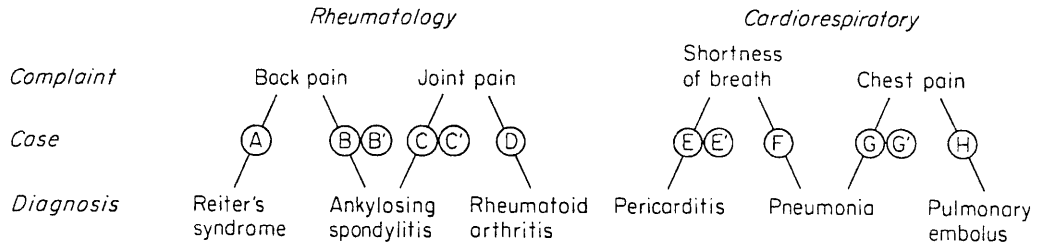

FIG. 1. Experimental design.

group of academic internists who were asked to state a differential diagnosis and indicate whether the problem would present a reasonable challenge to a second-year resident. In this manner, problems went through a series of revisions until they were felt to be satisfactory. Once the protocols were finalized, simulated patients were trained to simulate each problem. For two problems in each specialty area, two individuals were trained to simulate the problem.

In addition to the simulated patient protocols, multiple choice questions were developed relevant to each problem area. Initially, test items were furnished by the National Board of Medical Examiners using the six diagnoses as key words. These test items were then supplemented by items obtained from self-assessment materials and by a few items composed for the study. Each sixty-item questionnaire, one in each specialty area, contained twenty test items per diagnosis.

Each subspecialist saw all four problems in his specialty area in a single afternoon and completed the MCQ in the specialty. Students and residents saw five patients on each of two half-days separated by about 2 weeks. Presentation was balanced so that each half-day the subject saw either three rheumatology and two cardiology cases, or the reverse. Order of presentation was randomized, and the schedule was constructed in such a way that the second patient presenting a problem was seen on the alternate half-day. Subjects also completed the MCQ test at this time, one specialty subtest at each session. Each encounter was videotaped for subsequent review. Subjects also completed a structured medical record containing diagnosis, investigations and management plan.

Data analysis

Definition and validation of variables

The definition of variables to measure aspects of clinical performance is not straightforward. No consensus exists in the literature regarding the appropriateness of various measures of problem-solving, and measures used in past studies range from the total number of questions asked to scores of appropriateness of investigations.

The selection of variables to characterize the clinical encounters was guided by several, sometimes conflicting, principles.

(1) Measures used should be representative of those used in previous studies in order that the results could be generalized to other research.

(2) Measures should focus on cognitive measures of performance. Thus, measures of doctor–patient relationship or interviewing skills would be excluded, not because they are not of interest, but because they are more peripherally related to 'problem-solving'.

(3) A limited number of measures should be used. There were two reasons for this criterion: the first was to reduce the possibility of a type I error resulting from multiple comparisons, and the second was the constraint of feasibility in scoring the 240 encounters in the study.

(4) Measures should have the properties of interval or ratio scales in order to permit the application of analysis of variance methods, which are an essential approach to isolating the sources of variance.

It was decided to focus on measures of significant data gathered, diagnosis and investigations. Measures of data-gathering activity, such as the length of the encounter of the

number of questions asked, were excluded since at least one study of clinical reasoning demonstrated that these are unrelated to educational level, uncorrelated across encounters and unrelated to the appropriateness of diagnosis or management (Barrows *et al.*, 1978). Further, because a criterion group was available, it was decided to develop measures using the performance of this group as a standard rather than comparing performance to an arbitrary criterion.

The measures used in the assessment of performance are described in detail below.

(a) *Significant data gathered*

It was decided to focus on 'significant data'—information from the history and physical examination which was important to the resolution of clinical problems—rather than all possible data available. This then presented the challenge of defining 'significant'.

The approach used was first to develop a list of history and physical findings from the case protocols used to train the simulated patients. These lists consisted of twenty to thirty findings per case. The videotapes from the criterion physicians were then reviewed, and any additional data they elicited was appended to the initial list, resulting in an average of ten to fifteen additional findings. As a final step, the number of criterion physicians who elicited each finding was noted, and only those findings elicited by a majority (three out of five) of the subspecialists were identified as significant findings for the case.

(b) *'Critical' significant findings*

The list of significant findings identified contained information that was volunteered by the patient but that might not contribute to the resolution of the problem. In order to develop a subset of findings critical to the making of the diagnosis, criterion physicians were asked to weight each finding against their principal diagnosis and the correct diagnosis (if different). Each finding was rated as leading towards the diagnosis ('+'), leading away from it ('−') or neither ('0'). Only those significant findings that were weighted non-zero by a majority of physicians were designated as 'critical findings'.

Scores for critical findings and significant findings for each encounter were then derived by taking the number of significant findings elicited by each subject as a proportion of the number available.

(c) *Diagnosis*

The scoring of diagnosis used the performance of the criterion group to develop 'aggregate scores' for each encounter. The approach has been described in a recent paper, and has been shown to possess adequate construct and concurrent validity (Norman, 1985). The basic element in the aggregate score is a weight assigned to each diagnosis mentioned by the criterion group, equal to the proportion of all criterion physicians who mentioned the diagnosis. In the present study, each diagnosis mentioned by any subject in the criterion group received a weight of 0·2, 0·4, 0·6, 0·8 or 1·0 (1/5 to 5/5). A score was then assigned to each encounter, in the following manner. The numerator of the ratio consisted of the sum of the weights of diagnoses mentioned by the subject, with an additional weighting of 1·0, 0·8, 0·6, 0·4 or 0·2 to account for rank-ordering in differential diagnosis. Algebraically,

$$\begin{aligned} \text{NUM} = {} & 1{\cdot}0 \times \text{Wt (diagnosis 1)} \\ & +0{\cdot}8\ \text{Wt (diagnosis 2)} \\ & +0{\cdot}6\ \text{Wt (diagnosis 3)}. \\ & ++ \end{aligned}$$

A denominator was then developed consisting of a similar sum of the weights of those diagnoses ranked highest by the criterion group. The ratio of these two sums then is a number between 0 and 1, with a zero score obtained when none of the diagnoses of the subject were advanced by the criterion group, and a score of one obtained by mentioning those diagnoses most frequently mentioned by the criterion group.

(d) Investigation and management

Use of laboratory tests was assessed in a similar manner to diagnosis with an aggregate score approach. The ranking weights (1·0, 0·8,

etc.) were omitted from this calculation, since no ordering is usually implied in the laboratory requisition.

Management was not assessed because the patients were all complex cases seen on an initial visit, and management options were usually preliminary with definitive treatment awaiting the receipt of test results.

(e) *Formal knowledge*

The responses of subspecialists to each item on the MCQ test in their discipline were reviewed to determine if systematic differences were present between the correct response designated by NBME and the responses of the criterion group. Such differences were present for a few items in each subtest, and these were reviewed by the clinician investigators. In the majority of instances, both answers were viewed as acceptable and in scoring either was accepted as correct. For a few items, wording of questions was changed to remove ambiguity.

Results

(1) *Construct validity of measures*

In order to verify the usefulness of these measures, two forms of validity were examined using the cohorts from the study.

(a) *Divergent contruct validity*

It was hypothesized that the scores on each dimension should discriminate among subjects at different levels of experience. Analysis of variance was conducted using level of experience as a grouping factor.

The results are shown in Table 1. Significant differences among groups were demonstrated for all the measures, and nearly all measures showed monotonic trends in the expected direction. It was interesting that residents performed at the same level as cardiologists and respirologists on the cardiorespiratory part of the MCQ, but not so well as rheumatologists on the rheumatology part of the questionnaire. Given the difference in exposure to problems in these areas in a typical internal medicine residency, the results are not surprising.

(b) *Convergent construct validity*

To demonstrate convergent construct validity, it was hypothesized that there should be a strong positive correlation between the two measures of data-gathering, and weaker positive correlations between data-gathering and diagnosis and investigation scores.

Simple correlations among the measures, and partial correlations removing the effect of educational level, are shown in Table 2. From the Table, it is apparent that the construct was

TABLE 1. Mean values of variables by educational level (standard errors in brackets)

Performance	Clerk	Resident	Expert	*P*
% total significant findings elicited	67·2 (10·3)	64·7 (10·7)	83·4 (11·3)	<0·0001
% critical significant findings elicited	76·1 (15·4)	74·2 (13·3)	90·7 (10·5)	<0·0001
Diagnosis	60·1 (27·5)	67·5 (27·1)	87·6 (16·8)	<0·0001
Investigations scores	70·9 (15·5)	82·1 (14·2)	95·2 (5·8)	<0·0001
Knowledge				
% items correct, multiple choice questions rheumatology	58·8 (8·9)	65·8 (7·1)	80·6 (6·4)	<0·0001
multiple choice cardiorespiratory	51·8 (7·5)	74·2 (3·5)	74·5 (3·32)	<0·0001

TABLE 2. Convergent construct validity—correlations among performance measures

		Critical findings	Diagnosis	Investigation
Significant findings	Simple	0·668*	0·242*	0·258*
	Partial	(0·628)*	(0·123)	(0·057)
Critical findings	Simple	—	0·260*	0·212*
	Partial		(0·147)*	(0·059)
Diagnosis	Simple	—	—	0·325*
	Partial			(0·166)
Investigations	Simple	—	—	—
	Partial			

*$P < 0.05$

supported, with significant positive simple correlation among all measures and the strongest correlations between the two measures of data-gathering. The partial correlations, controlling for educational level, were lower, but remained positive, and three of the six were significant. Based on these results it was concluded that the measures used in the study possessed construct validity; therefore, it was appropriate to proceed to a test of the study questions.

(2) *Proportion of variance attributable to various factors*

The primary analysis of the study was directed at establishing the proportion of the variance in scores which could be attributed to the various factors manipulated in the study. In particular, it was wished to explore the degree to which a student or resident who achieved a high or low score on one problem would obtain a high or low score on a second problem which had a different complaint or diagnosis, or different subspecialty. The specific factors which could be examined in the study design are labelled and described below.

(a) Replication—each subject saw two presentations of the same problem, by different individuals, within each specialty area. These are problems A–A′, C–C′, etc. in Fig. 1.

(b) Same Complaint—Different Diagnosis (problems A–B, C–D, E–F, G–H in Fig. 1).

(c) Different Complaint—Same Diagnosis, (problems B–C, F–G in Fig. 1).

(d) Different Complaint—Different Diagnosis (clinical problems with a different complaint and diagnosis in the same specialty) (A–C, B–D, E–G, F–H).

(e) Different Specialty—rheumatology versus cardiology (ABCD versus EFGH).

The analysis was conducted using mixed model analysis of variance, and a standard statistical package (BMDP8V). The details of the analysis are found in Appendix 1.

The basic relationship between the research question and the analysis of variance relates to the components of variance; specifically, the components of variance due to a main effect of subjects, and the interaction between this factor and the remaining factors. If the problem-solving scores are evidence of a general skill which is independent of content, then one should find some subjects who performed consistently well across all problems and some who performed consistently less well. This would be identified as a large component of variance due to 'subject'; equivalently, a large main effect of subjects. Conversely, if there was no consistent difference in performance across subjects, even to two replications of the same problem, most of the variance would be unexplainable by the factors in the analysis, and this would result in a large residual or error variance component.

Content-specificity represents an interim situation between these two extremes. If the scores are influenced by content, then there should be large variance components due to the interaction between the other factors in the design, which are related to content, and the subject factor. For example, a large interaction between 'subjects' and 'specialty' would imply

that individual subjects do consistently well or poorly within a specialty but these skills do not transfer across the two specialties.

As outlined in the Appendix, the analysis grouped subjects by educational level; therefore, the variance components are based on differences between subjects within residency or clerkship groups, addressing the issue of the ability of the instruments to distinguish between good and poor residents or clerks.

The relevant variance components, expressed as a percentage of the total variance, with associated error of estimate, are shown in the first and second columns of Table 3. Although the errors in the estimates of percentage of variance components are of the order of 5 to 15%, the data are quite consistent. For three of the four variables, the residual variance, unexplained by any of the content factors, exceeded 60% of the total. Furthermore, there was little systematic difference between subjects in comparison with the residual variance.

These variance components were then used in the calculation of 'generalizability coefficients' according to the methods of Cronbach *et al.* (1963), described in Appendix 2. The generalizability coefficient is a ratio of variance components, and can be interpreted as a correlation between sets of scores. Different generalizability coefficients are developed corresponding to different degrees of generalization. In the present situation, the method permits examination of the correlation between scores derived from the presentations of the same problem, problems with the same complaint but different diagnoses, different complaint and same diagnosis, different complaint and diagnosis, and different specialty. If problem-solving was a general skill, one would expect the correlations to remain high and constant across all conditions. At the other extreme, if the measures were unreliable, or problem-solving was highly unstable, the correlations should be consistently low. Content-specificity would be evidenced by a high correlation across situations with similar content, and a gradual monotonic decrease in correlation as the content became less similar.

The results of this analysis are shown in Table 4. With the exception of the diagnosis score, measures showed only low generalizability, even to two presentations of the same problem, and little evidence of a relationship to content knowledge. The diagnosis score did follow the expected pattern of 'content specificity', with relatively high correlation on very similar problems, low correlation with dissimilar problems in the same specialty, and no

TABLE 3. Components of variance attributable to various sources

	Residual	Same complaint different diagnosis	Same diagnosis different component	Different complaint different diagnosis	Different specialty	Subjects
Sig. findings						
Variance	59·0	7·0	9·1	0·0	11·9	7·8
Error	9·8	9·9	11·4	8·3	8·3	7·4
%	62·0	7·0	9·6	0·0	12·5	8·2
Crit. findings						
Variance	163·9	0·0	0·0	0·0	1·8	23·2
Error	27·3	21·7	25·6	18·5	11·8	13·5
%	87·0	0·0	0·0	0·0	1·0	12·0
Diagnosis score						
Variance	339·8	36·1	193·1	289·5	0·0	0·0
Error	56·3	56·2	94·4	118·6	90·7	41·9
%	40·0	4·2	22·5	33·7	0·0	0·0
Investigations						
Variance	217·7	5·1	0·0	0·0	0·0	10·7
Error	36·3	32·4	32·3	26·4	14·4	12·0
%	93·0	2·2	0·0	0·0	0·0	4·8

TABLE 4. Generalizability coefficients

Problem	Same	Different	Different	Different	Different
Complaint	Same	Same	Different	Different	Different
Diagnosis	Same	Different	Same	Different	Different
Specialty	Same	Same	Same	Same	Different
Significant findings	0·37	0·30	0·27	0·20	0·08
Critical findings	0·13	0·13	0·13	0·13	0·12
Diagnosis score	0·60	0·56	0·37	0·33	0
Investigations score	0·07	0·05	0·05	0·05	0·05

correlation across specialties. However, the correlation of 0·60 across two presentations of the identical problem, accounting for 36% of the variance in scores, clearly demonstrates that the variation in scores cannot be attributed simply to variable content.

Therefore, from the present study, there was only weak evidence of content specificity, as reflected in the diagnosis score, and no evidence that problem-solving, as reflected in the four measures used, is a general skill.

This observation regarding the minor role of content knowledge in explaining the observed variation in scores was confirmed by an analysis of covariance, in which the scores on the multiple choice subtests were used as covariates. No significant positive relationship between this measure of content knowledge and the performance scores was present, thus content knowledge, as assessed by the MCQ tests, was unrelated to performance on any problem.

Some additional confirmation of these findings was obtained by supplementary analyses which simply focused on whether subjects obtained the correct diagnosis, either as principal diagnosis or on the differential. The first analysis examined the two presentations of the same problem seen by each resident and clerk. Each encounter was analysed to determine whether or not the correct diagnosis was listed as principal or differential diagnosis on the first and second presentations. As shown in Table 5, there was only a weak and non-significant association between the two presentations, yielding a Kappa coefficient of 0·12. The results therefore provided confirmation of the high variability of performance, even across two presentations of the same problem.

The second analysis examined the presence of the correct diagnosis as related to performance on the multiple choice test. Subjects were grouped above and below the median score on the multiple choice test in each specialty, and the number of correct diagnoses obtained out of a maximum possible of twenty-five (five subjects × five problems) was tabulated.

As shown in Table 6, in no instance was there a significantly higher proportion of cor-

TABLE 5. Correctness of diagnosis on first and second presentation of the same problem (residents and clerks)

		Second presentation			
		Principal	Differential	Absent	Total
First presentation	Principal	10	5	1	16
	Differential	4	6	1	11
	Absent	7	5	1	13
	Total	21	15	3	40

$\chi^2_4 = 1{\cdot}82$ $P = 0{\cdot}77$
Kappa (weighted) = 0·06

TABLE 6. Relationship between MCQ scores and presence of correct diagnosis as principal of differential diagnosis

	Number (%) of correct diagnoses			
	Rheumatology		Cardiology	
	Principal	Differential	Principal	Differential
Residents				
Above median	13 (52)	22 (88)	12 (48)	18 (72)
Below median	13 (52)	17 (68)	12 (48)	21 (84)
χ^2	0	2·91	0	1·04
P	–	0·10	–	–
Students				
Above median	11 (44)	18 (72)	11 (46)	21 (84)
Below median	10 (40)	17 (68)	8 (30)	20 (80)
χ^2	0·08	0·09	0·76	0·13
P	–	–	–	–

rect diagnoses for subjects with higher performance on the multiple choice test. Thus, content knowledge was not related to accuracy of diagnosis, again confirming the previous analysis.

Discussion and conclusions

The results of the study are consistent. Sparse evidence had been found for content or case-specificity; rather it is apparent that the measures used to assess problem-solving in this study contain a large component of variability which could not be explained by systematic changes in content or systematic differences between subjects.

One possible explanation for these results is simply that the variation attributable to factors controlled in the study—subjects and content—was small, leading to a high proportion of variance due to random variation. This circumstance could arise if the subjects in the study were relatively homogeneous in ability, so that there was no observable variation between subjects, or if the cases were chosen in such a way that the range of observed performance across cases was very similar. Examining the means and standard deviations of scores in Table 1, this does not appear to be the case, as no mean value in the clerk and resident sample approached 100%, which would indicate a 'ceiling effect', and standard deviations were of reasonable magnitude—10 to 27%, suggesting that considerable variability in individual scores was present.

Examining other possible biases in the study, the two disciplines chosen differed in terms of the amount of exposure residents might have—from high exposure in cardiorespirology to little exposure in rheumatology. With the single exception that resident performance on the cardiorespiratory MCQ was higher than on the rheumatology MCQ test, there is no evidence that the choice of specialties resulted in any bias. The study used a simulation format—live simulated patients, rather than actual clinical performances, in order to permit experimental control over the range of content. However, the validity of live simulated patients has been demonstrated, using measures similar to these of the present study (Norman & Tugwell, 1982), therefore it is unlikely that the use of this simulation format resulted in any bias which could lead to the low generalizability of performance across problems.

Perhaps the strongest argument in support of the generalizability of the present findings is that the correlations across problems fall into the range reported by other studies in which content was not controlled (Skakun *et al.*, 1979; Berner *et al.*, 1977; Norman & Tugwell, 1982; Elstein *et al.*, 1978; Donnely *et al.*, 1982; Norcini *et al.*, 1983).

If one accepts the findings of the study, there

are two possible explanations. Either the measures used in the study are unreliable indicators of clinical performance, and are thus not sensitive to true variations in problem-solving skill across individuals, or clinical performance really is as variable and unpredictable as indicated in the present results. Some evidence exists to support both.

Examining first the evidence regarding the validity of the measures, thoroughness of data-gathering is a traditional virtue associated with competent clinical practice, and many scoring methods, such as the 'significant findings' score in the present study or the 'proficiency' score of patient management problems, reflect this attribute. However, the evidence of the value of thoroughness is essentially negative. Barrows *et al.* (1978) and Norman & Tugwell (1982) both found that thoroughness of data-gathering was uncorrelated with obtaining the correct diagnosis. Marshall (1977) showed that a score which encouraged thoroughness was inversely related to experience, whereas a modification of the score, which rewarded efficiency, was positively related to experience and more highly correlated across problems. The present study lends support to this argument; correlations between data-gathering measures and diagnostic and laboratory outcomes were only in the range 0·05 to 0·15 (Table 2) when educational level was partialled out.

The evidence regarding measurement of diagnosis and use of investigations is more limited. Most work on correlations across cases for diagnosis and management is derived from patient management problems, and correlations are consistently lower for diagnosis and investigations than for history and physical sections (Donelly *et al.*, 1982; Norcini *et al.*, 1983. This finding may reflect the fewer number of alternative options available in this section of the PMP, but no investigator has examined this possibility. Certainly, the scoring of diagnoses is problematic. The number of reasonable alternative diagnoses in a particular problem is very limited, so that at the level of diagnosis, a single patient problem may be like a single multiple choice question. Conversely, although the alternatives are few, rarely is there a single right answer, and scoring must account for degrees of 'rightness'. The aggregate score approach used in this study is an attempt to account for a range of plausible alternatives, but the method is not, as yet, well developed, and the possibility remains that these scores do not optimally detect systematic differences between individuals.

What of the alternative possibility that clinical problem-solving is a highly variable activity, rather than a general skill?

Although this hypothesis runs counter to a prevailing view in medical education over the past two decades, there is recent evidence from a diversity of fields that the expert problem solver may be an expert, not because of any innate or learned advantage in problem-solving skill, but because he knows more in his domain than the novice. Work in artificial intelligence began in the 1950s with the development of general problem-solving programmes (Newell & Simon, 1972) but more recent, and far more successful, programmes operate only in highly circumscribed domains and operate on extensive rules and heuristics, which reflect, and are specific to the domain. A recent review article (Waldrop, 1984) summarized the state of the art recently as:

> 'The essence of intelligence seems to be less a matter of reasoning ability than of knowing a lot about the world' (p. 1279).

Similarly, studies of expert–novice differences in a wide variety of domains from physics and chess to medicine (Chi *et al.*, 1981; De Groot, 1965; Chase & Simon, 1973; Norman *et al.*, 1979; Muzzin *et al.*, 1982), have led from a focus on general problem-solving strategies towards an attempt to understand the characteristics and organization of knowledge structures acquired by experts. As Glaser (1984) comments:

> 'Our interpretation is that the problem-solving difficulties of novices can be attributed largely to the inadequacies of their knowledge bases and not to limitations in their problem-solving capabilities' (p. 99).

If knowledge is what distinguishes expert from novice, why, in the present study, did we see so little effect resulting from the systematic manipulation of knowledge? The reason may lie in the definition and organization of know-

ledge. Expert knowledge in medicine, like any professional domain, is highly complex and interwoven. The generation of appropriate diagnostic hypotheses early in the encounter (Elstein *et al.*, 1978; Barrows *et al.*, 1978) may result from a pattern-recognition process against prototypes or templates in memory (De Groot, 1965; Muzzin *et al.*, 1982), which are in turn a product of extensive experience with patients in addition to formal educational experiences. If this is the case, it suggests that expert knowledge is unlikely to be organized in a logical hierarchical form, as was implicitly assumed in the experimental manipulations of the present study. The solution of a single patient problem, would derive not from a general problem-solving process utilizing a logically consistent knowledge base, but from a pattern-matching process against experiences in memory. If this is the case, the results of the present study may be anticipated.

Conclusions

The results of the study clearly indicate that patient-based evaluation of health professionals will not allow generalizations about competence based on one or two patient encounters. However, the picture is not as discouraging as that presented by Elstein *et al.* (1978) who suggested that, since transfer across problems is so limited, one could only certify competence based on those problems which the learner had actually encountered. Instead, it would appear that this issue is not one of limited transfer of knowledge, but is related to the inherent, and apparently random, variability present in performance on a single problem. At best, this variation may simply represent a measurement problem, and a better choice of measures may improve generalizability. Alternatively, one may have to accept that this degree of variability is to be expected, and devise new approaches to assessment which more directly top into the extensive body of knowledge which is the hallmark of clinical expertise.

Acknowledgement

This study was supported by an Ontario Ministry of Health Research Grant.

References

American Board of Internal Medicine (1979) *Definition of Competence in Internal Medicine*. American Board of Internal Medicine, Philadelphia.

Barrows, H.S., Neufeld, V.R., Feightner, J.W. & Norman, G.R. (1978) *An analysis of the clinical methods of medical students and physicians*. Report to Ontario Ministry of Health.

Berner, E.S., Bligh, T.J. & Guerin, R.O. (1977) An indication for a process dimension in medical problem solving. *Medical Education*, **11**, 324–8.

Case, S.M. (1981) A new examination for the evaluation of diagnostic problem solving. *Proceedings of the 20th Conference on Research in Medical Education, Washington, DC*.

Chase, W.G. & Simon, H.A. (1973) Perception in chess. *Cognitive Psychology*, **4**, 55–81.

Chi, M.T., Feltovich, P.J. & Glaser, R. (1981) Categorization and representation of physics problems by experts and novices. *Cognitive Science*, **5**, 121–52.

Cronbach, L.J., Rajaratnam, N. & Gleser, G.C. (1963) Theory of generalizability: a liberation of reliability theory. *British Journal of Statistical Psychology*, **16**, 137–63.

De Groot, A. (1965) *Thought and Choice in Chess*. Mouton, The Hague.

Donnely, M.B., Fleisher, D.S., Schvenker, J. & Chen, C.Y. (1982) Problem solving within a limited content area. *Proceedings of the 21st Conference on Research in Medical Education, Washington, DC*.

Elstein, A.S., Shulman, L.S. & Sprafka, S.A. (1978) *Medical Problem Solving: An Analysis of Clinical Reasoning*. Harvard University Press, Cambridge.

Finkel, A. & Norman, G.R. (1973) The validity of direct observation. *Proceedings of the 12th Conference on Research in Medical Education, Washington, DC*.

Glaser, R. (1984) Education and thinking: the role of knowledge. *American Psychologist*, **39**, 93–103.

Helfer, R.E. & Slater, C.H. (1971) Measuring the process of solving clinical diagnostic problems. *British Journal of Medical Education*, **5**, 48–52.

Hubbard, J.P. (1978) *Measuring Medical Education*. Lea and Febiger, Philadelphia.

Marshall, J. (1977) Assessment of problem solving ability. *Medical Education*, **11**, 329–34.

Mazzuca, S.A., Cohen, S.J. & Clark, C.M. (1981) Evaluating clinical knowledge across years of medical training. *Journal of Medical Education*, **56**, 83–90.

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

Muzzin, L.J., Norman, G.R., Feightner, J.W. & Tugwell, P. (1982) Manifestations of expertise in recall of clinical protocols. *Proceedings of the 21st Conference on Research in Medical Education, Washington, DC*.

Newell, A. & Simon, H.A. (1972) *Human Problem Solving*. Prentice Hall, Englewood Cliffs.

Norcini, J.J., Swanson, D.B., Grosso, L.J. & Webster, G.D. (1983) A comparison of several methods for

scoring patient management problems. *Proceedings of the 22nd Conference on Research in Medical Education, Washington, DC.*

Norman, G.R. (1985) Objective measurement of clinical performance. *Medical Education,* **19,** 43–47.

Norman, G.R. & Tugwell, P. (1982) A comparison of resident performance on real and simulated patients. *Journal of Medical Education,* **57,** 708–15.

Norman, G.R., Feightner, J.W., Jacoby, L.L. & Campbell, E. (1979) Clinical experience and the structure of memory. *Proceedings of the 18th Conference on Research in Medical Education, Washington, DC.*

Robinson, S.A. & Dinham, S.M. (1977) Reliability and validity of simulated problems as measures of change in problem-solving skills. *Proceedings of the 16th Conference on Research in Medical Education, San Francisco, CA.*

Schumacher, C.F. (1983) Validation of the American Board of Internal Medicine written exam. A study of the examination as a measure of achievement in graduate medical education. *Annals of Internal Medicine,* **78,** 131–5.

Senior, J.R. (1976) *Toward the Measurement of Competence in Medicine.* American Board of Internal Medicine, Philadelphia.

Skakun, E.N., Taylor, W.C., Wilson, D., Taylor, T., Grace, M. & Fincham, S.M. (1979) Preliminary investigation of computerized patient management problems in relation to other examinations. *Educational and Psychological Measurement,* **39,** 303–10.

Smith, P.L. (1981) Gaining accuracy in generalizability theory using multiple designs. *Journal of Educational Measurement,* **18,** 147–54.

Waldrop, M.M. (1984) The necessity of knowledge. *Science,* **223,** 1279–82.

Appendix 1: Analysis of Variance

The primary analysis reported in the paper used the scores generated by residents and students on eight clinical problems—four each in rheumatology and cardiology, distributed as shown in Fig. 1. The criterion clinicians were not included in this analysis, nor were the replications of the identical cases.

As a result there were five factors identified in the analysis:

(1) Educational level—resident or clerk (two levels);

(2) Subjects—ten subjects (residents or clerks) 'nested' within each educational level;

(3) Discipline—cardiology or rheumatology—a repeated measure on each subject;

(4) Complaint—a repeated measure on each subject, with two levels, nested within each discipline;

(5) Diagnosis—a repeated measure with two levels nested within complaint and discipline.

The analysis used a mixed model ANOVA with educational level as a fixed factor and the remaining as random factors.

Components of variance were determined from the expected mean squares by the package programme (BMDP8V). Only those components of variance including the 'subject' factor are reported in the text, as the remaining components are of peripheral interest to the study question.

Estimates of errors associated with each variance components were determined using the methods described by Smith (1981), where for a particular estimated variance component σ_k^2, the error variance is

$$\text{var}(\hat{\sigma}_k^2) = \frac{2}{C_k^2} \sum \frac{(EMS_i)^2}{df_i}$$

where C_k^2 is the coefficient of the mean square used in the estimation of variance, EMS is the calculated mean square, with df its associated degrees of freedom. Thus, for example, the variance due to subjects in the present design is based on differences between mean squares as:

$$\hat{\sigma}_S^2 = (MS_S - MS_{SD})/N_R N_C N_D$$

where

MS_s = mean square (subjects) df = 18
MSSD = mean square (subjects × discipline)
df = 18
NR = no. of levels of 'replication' = 2
NC = no. of levels of 'complaint' = 2
ND = no. of levels of 'discipline' = 2

Therefore

$$\text{var}(\hat{\sigma}_S^2) = \frac{2}{(N_R N_C N_D)^2}\left(\frac{MS_S^2}{df_S} + \frac{MS_{SD}^2}{df_{SD}}\right)$$

$$= \frac{2}{(2\times 2\times 2)^2}\left(\frac{MS_S^2}{18} + \frac{MS_{SD}^2}{18}\right)$$

Standard errors were then determined as the square root of the estimated variance.

Separate estimates of the residual variance were determined by analysis of the cases which were seen by each resident. Since the primary analysis used an estimate of residual variance based on two different problems with the same chief/complaint, this overestimated the true residual variance.

The secondary analysis included four repeated measures on each subject—the two replications of the case within each discipline, and the two disciplines. Variance estimates were conducted as in the primary analysis, and are shown in the line labelled 'Replication' in Table 3. Variance due to interaction between subjects and complaints was then obtained by subtracting the two residual variances. In addition, a separate analysis was conducted for those problems in each specialty which had the same diagnosis but different complaints (B–C, F–G). Variance attributed to the interaction of subjects and diagnoses was then obtained by subtracting the residual variance from the ANOVA of replications.

Appendix 2

The estimation of generalizability coefficients involves a ratio of variance components. In its simplest form, the numerator contains the variance due to subjects, σ^2_s, and the denominator contains the sum of all other sources of variance, σ^2ERR, and it is identical to the classical reliability coefficient.

$$G = R = \frac{\sigma^2_s}{\sigma^2_s + \sigma^2_{ERR}}$$

In more complex designs, the error variance is comprised of the variance due to the interaction between subjects and the other 'facets' of design, (in this case problem [P], complaint [C], diagnosis [D], and specialty [S]). Different generalizability coefficients are constructed depending on the degree of generalization required.

Thus, for example, the coefficient for generalizing to the second presentation of the same problem is:

$$G = \frac{\sigma^2_s + \sigma^2_{sC:} + \sigma^2_{sD} + \sigma^2_{sS}}{\sigma^2_s + \sigma^2_{sP} + \sigma^2_{sC:} + \sigma^2_{sD} + \sigma^2_{sS}}$$

and to a different specialty is:

$$G = \frac{\sigma^2_s}{\sigma^2_s + \sigma^2_{sP} + \sigma^2_{sC:} + \sigma^2_{sD} + \sigma^2_{sS}}$$

The G coefficient, then, is a number between 0 and 1, where 0 implies that all variance is a result of other factors in the design, and 1 implies that all variance is due to true variance between subjects.

Received 21 *December* 1985; *accepted for publication* 4 *March* 1985