DOCUMENT RESUME

ED 050 164 TM 000 557

AUTHOR Elstein, Arthur S.; Shulman, Lee S.
TITLE A Method for the Study of Medical Thinking and Problem Solving.
INSTITUTION Michigan State Univ., East Lansing.
PUB DATE Feb 71
NOTE 31p.; Paper presented at the Annual Meeting of the American Educational Research Association, New York, New York, February 1971

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS *Clinical Diagnosis, Cognitive Measurement, *Critical Thinking, Data Analysis, Decision Making Skills, Hypothesis Testing, Medical Education, Medical Evaluation, *Physicians, Problem Solving, Rating Scales, Research Design, Scientific Attitudes, *Scientific Methodology, *Simulation, Video Tape Recordings

ABSTRACT

A method for studying medical reasoning in a life-like setting is reported. Simulated medical problems, amplified by concurrent thinking aloud, episodic retrospection during the work-up, and videotape-stimulated retrospection, are used to obtain records of the behavior and reasoning physicians use to solve diagnostic problems. The fundamental units of analysis are questions, critical findings, and hypotheses. Eight categories of questions relate the information seeking behavior of the inquiring physician to a widely accepted outline for medical history taking. Critical findings in a case are elicited by questions and are assigned weights depending upon their relation to any conceivable diagnostic hypothesis. Hypotheses tested by an inquirer are identified from his thinking aloud and retrospection. Findings elicited are evaluated in relation to inquirer's hypotheses or to those he might have considered but did not. Medical diagnosis is thus analyzed as a special case of hypothesis testing. The method is illustrated by application to two work-ups of the same problem; one globally rated substantially better than the other. The method effectively distinguishes between the two in psychologically relevant ways. Discussion relates the findings to current work in problem solving. (Author)

A METHOD FOR THE STUDY OF MEDICAL THINKING AND PROBLEM SOLVING

Arthur S. Elstein and Lee S. Shulman

Michigan State University

Abstract

A method for studying medical reasoning in a life-like setting is reported. Simulated medical problems, amplified by concurrent thinking aloud, episodic retrospection during the work-up, and videotape-stimulated retrospection, are used to obtain records of the behavior and reasoning physicians use to solve diagnostic problems. The fundamental units of analysis are questions, critical findings, and hypotheses. Eight categories of questions relate the information seeking behavior of the inquiring physician to a widely accepted outline for medical history taking. Critical findings in a case are elicited by questions and are assigned weights depending upon their relation to any conceivable diagnostic hypothesis. Hypotheses tested by an inquirer are identified from his thinking aloud and retrospection. Findings elicited are evaluated in relation to inquirer's hypotheses or to those he might have considered but did not. Medical diagnosis is thus analyzed as a special case of hypothesis testing. The method is illustrated by application to two work-ups of the same problem; one globally rated substantially better than the other. The method effectively distinguishes between the two in psychologically relevant ways. Discussion relates the findings to current work in problem solving.

A METHOD FOR THE STUDY OF MEDICAL THINKING AND PROBLEM SOLVING*

Arthur S. Elstein and Lee S. Shulman
Michigan State University

We are investigating medical thinking as a paradigm of reasoning and problem solving in a practical domain. We have chosen not to employ an experimental setting that is devoid of "life-like" elements, as have been the characteristic investigations in cognitive psychology, with their concept attainment boards (Bruner, Goodnow and Austin, 1956), memory drums or Towers of Hanoi (Simon, 1970). Instead, we are focusing on studies of an actual cognitive activity, medical problem solving, conducted by experienced practitioners in settings as natural as the requirements of disciplined inquiry permit. We then propose to generalize from the findings of these studies to other similar domains, arguing that they possess characteristics analogous to those of the medical problem solving situation. Schwab (1969), for example, argues that medicine is a far more appropriate analogue to curriculum development and educational decision-making than are the theoretical disciplines most often looked to for guidance in that field.

This paper describes the development and initial testing of a method for scoring and evaluating the problem solving of experienced physicians as they perform diagnostic work in a simulated medical setting. The paper discusses the aims which directed our choice of a particular scoring system; specifies the precise manner in which the data are collected and subsequently scored; and reports the results of a pilot attempt to investigate the success of the scoring system in meeting the criteria enunciated.

---

* This research is supported in part by NIH Grant PM-00041 to the Office of Medical Education Research and Development, Michigan State University.

In weighting alternative methods of scoring medical problem solving protocols, four major criteria were used:

1. The method must be objective and reliable. That is, given formal statements of the rule for each scoring category independent judges ought to reach at least 85% agreement on the specific categories to which any particular unit of behavior is assigned.
2. The method must reflect the critical and relevant characteristics of the particular mode of cognitive functioning under study. Thus, in assigning scores or weights to aspects of the observed behavior, the scoring system should draw attention to the clearly more relevant aspects of the functioning and pay less or no heed to the irrelevant aspects.
3. The scoring system should measure aspects of the activity under study that can be related to parallel variables in other theories of problem solving and/or studies of similar processes in other content domains. That is, the scoring system should not only be a description of subject performance in medically relevant terms, but should also afford a way of describing the cognitive functioning of the subject that is meaningful in the light of broader theories of cognitive functioning and problem solving.

4. The specific scores or assessments generated by the scoring procedure ought to result in scores which distinguish effectively between clearly different levels of competence in medical functioning. That is, the scoring system will demonstrate its validity through effectively discriminating between levels of competent performance.

These then are the four desiderata which directed the formulation of the present scoring system. We now move to a description of the specific characteristics of the scoring procedures themselves.

COLLECTING THE BASIC DATA

Simulated medical cases based on actual clinical records are used to observe, in moderately controlled circumstances, the procedures by which physicians gather data and reason clinically (Kagan, Elstein, Jason, Shulman and Loupe, 1970). A room has been designed to resemble a physician's office; two television cameras are mounted near the ceiling and the entire interaction between the doctor and the simulated patient is videotaped. Actors have been carefully trained to simulate patients in these problems. The information potentially available to each physician-subject is thus known, so that different physicians may be observed while solving the same diagnostic problem. Historical data, physical findings, and laboratory examinations are all available upon request. It is stressed to the physician-subject that he is free to elicit as much or as little data as he feels is necessary for adequate solution of the diagnostic problem, and that he may

elicit these data in any order that he chooses. He is asked to work in his customary manner and to do whatever he feels is appropriate for the case at hand.

Whenever a natural break occurs in the diagnostic work-up, the physician is asked to review and consolidate his findings and hypotheses aloud so as to provide an ongoing record of his reasoning at intervals. The points at which this review is most usually obtained are between the history and the physical examination and at the conclusion of the physical examination before ordering any laboratory tests. After the full work-up has been completed, the "stimulated recall" section of the experiment begins. The videotape of the physician's work-up is replayed for him. He is given a stop-start switch with which he can control the playback and he is asked to stop it whenever there is an even on which he can elaborate. He is encouraged to use the tape as a vehicle to stimulate his memory and to relate his thoughts during the original encounter. Thus, a record of his thinking and reasoning supplements the videotaped record of his overt behavior during the work-up. Generally, scrutiny of the first fifteen to twenty minutes of an encounter, a procedure ordinarily requiring one to one-and-a-half hours, provides effective clarification of the basic hypotheses generated by the physician and his proposed strategies for testing them.

THE UNITS OF ANALYSIS

Once the full record of the interaction between doctor and patient has been transcribed into a typewritten protocol the process of scoring can begin. This

process involves transcriptions both of the original doctor-patient interaction and of the subsequent review of the videotape by the physician.

As in so many parallel domains, the first problem is that of identifying the fundamental units of analysis. In this research the units of analysis will be: 1) questions, 2) critical elements or findings, and 3) hypotheses. It will be seen that the question serves to parse the protocol into the smallest constituent elements of surface structure, playing much the same role as the morpheme in grammatical analysis. When we discuss what the physician is doing we will be discussing his questioning behavior. Questions will take many forms, ranging from the explicit interrogation regarding past medical history to the shining of a light to examine a patient's eye grounds. Findings may be volunteered by a patient or elicited via the physician's inquiry. Subsequently, they may be sensed as critical by the physician or ignored. We will observe that both the elicitation and sensing of critical problem elements are crucial variables in the analysis of physician inquiry.

Both questions and findings can be said to lie on the surface of the observable medical inquiry. Below that surface lie the mental operations which lead the physician to ask the questions he does and to process them in the way he chooses. We have found that the hypothesis is a most powerful way of characterizing one important aspect of these mental operations. In an earlier paper (Elstein, Shulman, Kagan, and Jason, 1970) we argued that most medical inquiries are characterized by the relatively early generation of working hypotheses. These in turn appear to direct the subsequent patterns of data collection and evaluation. In the analysis

of the present protocols, it can be seen that a hypothesis can be generated alone or in the company of competitors. A hypothesis has not only a moment of birth but also an ultimate fate. It may be entertained and then rejected. It may never be explicitly rejected, but allowed simply to fade away as better alternatives move into place. It may be confirmed, in which case it moves from the status of hypothesis to that of tentative or ultimately final diagnosis. The purpose of asking specific questions to elicit critical findings is to manipulate the status of these hypotheses in order to achieve a correct diagnosis.

Questions (Q)

A question is defined as any statement or act of the physician which either: 1) seeks information from the patient, 2) instructs the patient concerning a procedure in the examination, or 3) establishes rapport between the physician and the patient. To provide a link between these questions and more typical classifications of physician activity, eight content categories were identified into which any question could be further assigned. The first six categories are minor modifications of an outline for examining patients which is widely accepted by physicians and taught to medical students (Harvey, et al., 1968). These categories and their explanations are described in Table 1. Problems of ambiguous questions, that is, questions that could justifiably be included in more than one category because they were simultaneously serving multiple functions or because the function intended by the physician was not made clear either in the course of the original work-up of the subsequent stimulated review, will not be discussed in

TABLE 1. CATEGORIES OF QUESTIONS

1. Present Illness

   Patient's account of onset, duration and course of illness. Chief complaints and associated symptoms.

2. Personal and Social History

   Personal status, habits, home conditions, occupation, environmental factors, military medical records.

3. Family History

   State of health and cause of death of parents and siblings. History of tuberculosis, diabetes, heart trouble, cancer and other disease with hereditary components.

4. Previous Medical History

   History of illnesses, operations, injuries and allergies, review of functioning of organ systems (neurological, endocrine, respiratory, ect.).

5. Physical Examination

   Search for signs of illness. Examination of skin, heart, lungs, abdomen, etc.

6. Laboratory Data

   Tests performed on various bodily fluids, products or functions. Examination of blood, urine, sputum, cerebro-spinal fluid. Diagnostic x-rays, electrocardiogram, electroencephalogram.

7. Rapport

   Statements or questions dealing with the doctor-patient relationship or the patient's anxiety about illness.

8. Instructions

   Statements telling the patient what is about to occur or asking the patient to do something.

the body of the paper. These more technical matters are reviewed in full in a scoring manual currently under development.

Since the analysis of these protocols depends upon the reliability with which they can be objectively parsed into question-components, the rules for such divisions were given to two judges who proceeded independently to divide two protocols into their respective questions. A very high percentage of agreement (91%) was achieved for identifying the actual number and identity of questions. The agreement for assigning questions to specific content categories was only slightly lower (86%).

Critical Findings (CF)

For each case a list of critical findings is compiled. These findings are: (a) answers to possible questions that might be asked during a history, (b) specific physical findings that would be observed in a physical examination, or (c) results of laboratory tests that might be ordered. Thus, some findings are critical because they are positive while others are equally critical although negative. The questions asked serve as the milestones of the inquiry, indicating how far the individual has moved in his investigation. The critical findings are potentially problematic elements with which he must come to terms to solve the problem. If significant numbers of critical findings are missed, this is likely to preclude the inquirer from reaching his intended destination. On the other hand, no single finding is indispensable because the interaction of both psychological and physiological systems in the human organism creates redundancy among cues. The critical

findings list plays much the same role in studies of medical inquiry as the manual of potentially problematic elements played in studies of teacher inquiry (Shulman, 1965; Shulman, Loupe and Piper, 1968) or pupil inquiry in local politics (Allender, 1969).

Each critical finding is important in different ways, depending upon the particular hypotheses that the physician is entertaining at the moment the finding is elicited. We judged that it was important to assign a weight or valence to each critical finding in relation to any of the hypotheses that might be entertained by the inquirer. Therefore, each critical finding was assigned a weight from -2 to +2, with regard to its impact upon any hypothesis that might be held in a particular case. For example, if the physician is simultaneously entertaining the hypotheses of hysterical paralysis and multiple sclerosis, the finding of a positive plantar reflex (Babinski's sign) has a weight of -2 for the hypothesis of hysteria and +1 for multiple sclerosis. This is because a positive finding unequivocally rules out hysteria while, though positive for something like multiple sclerosis, it does not rule out a number of other disorders which could also produce spinal cord lesions. We are currently in the process of determining the degree of objectivity with which independent medical judges will assign such weights.

Hypotheses (H)

The hypotheses generated, entertained, explicitly rejected, forgotten, simply ignored, or ultimately accepted are identified through analysis of the physician's thinking aloud, both during the inquiry and in the natural breaks

between phases, as well as from his reflections during the stimulated recall period. Hence, we know that a physician is entertaining a particular hypothesis because he tells us so. Usually, he volunteers such information without the necessity for probing. On some occasions, the existence of an underlying hypothesis emerges when the physician is questioned regarding his choice of a particular question or test.

To summarize the discussion to this point, it is possible to take the transcribed protocol of a doctor-patient interaction and divide it into basic components called questions. These questions can be assigned reliably to medically relevant content categories. The consequences of having asked those questions can further be reflected in the elicitation of critical findings. The order in which these findings emerge is a consequence of the particular questions asked and can clearly be indicated in a chart. This chart can be said to map the surface structure of the medical inquiry session that is being analyzed.

The "deep structure" of a particular inquiry makes use of the findings elicited and the hypotheses generated. The findings are evaluated in relation to any particular hypothesis. Second, they are scored to reflect whether or not the physician sensed the importance of the finding if elicited.

Clearly, the major constituent of this deeper level of analysis is the hypothesis. Charts at this deeper level reflect the relations among findings elicited and sensed (or not sensed) and the natural history of hypotheses as they are created and consigned to some particular fate. Questions are used to index the particular points at which these events occur, serving much the same purpose as

here as page numbers in a book.

By carefully examining the relationships among hypotheses and findings we can compare the number of positive and negative findings which he elicited for any considered hypotheses. One measure of the strength of a physician's subjective probability estimate for a particular hypothesis, for example, may be reflected in the degree to which findings inconsistent with that hypothesis are elicited or volunteered, but fail to be sensed.

A SPECIFIC EXAMPLE

Let us now apply these analytic tools to a specific example, a comparison of two work-ups of the same simulated medical problem. These work-ups are of interest because one (Dr. X), has been uniformly rated by viewers of the videotape record as one of the poorest in our pilot series of work-ups, while the second (Dr. Y) has been equally uniformly rated as an excellent example of clinical work. Can the analytic scheme outlined distinguish between two work-ups that impress clinicians so differently? And are the identifiable differences (if found) comprehended within a more general psychology of problem solving? Can we then begin to analyze medical diagnosis as a specialized form of problem solving, not simply as an art sui generis?

Table 1 presented the categories used for classifying questions in the medical work-up. Table 2 presents the same categories, showing the numbers and proportions of questions asked by Dr. X and Dr. Y. Dr. Y asks many more questions than Dr. X, but both ask about the same proportion of questions in Category 1,

TABLE 2. QUESTIONS ASKED BY DR. X AND DR. Y

	Doctor X (Poor Work-Up)		Doctor Y (Good Work-Up)	
	Number	% Total	Number	% Total
1. Present Illness	62	45	143	43
2. Personal & Social History	20	14	12	4
3. Family History	0	0	14	4
4. Previous Medical History	19	14	18	5
5. Physical Examination	27	19	99	30
6. Laboratory Data	0	0	0	0
7. Rapport	9	6	22	7
8. Instructions	2	1	24	7
TOTAL	139	99	332	100

present illness. They differ in number of questions about personal and social history (Category 2), previous medical history (Category 4), and physical examination (Category 5). This surface analysis alone may suggest that Dr. X is searching for data that will relate to a historical or psychogenic basis for the present disorder while Dr. Y is testing hypotheses about organic etiology. We suspect, and shall soon demonstrate, that these surface differences in the data sought reflect different hypotheses about the nature of the problem. The work-up is not an invariant routine. Rather, it is structured to answer certain questions, and different diagnostic hypotheses lead physicians to consult different sources of information (Harvey, et al., 1968).

Table 3 presents the list of critical findings for the case in question. The patient is a 21-year-old female who is brought to the emergency room early one morning paralyzed in both legs. Having gone to bed the night before believing herself well, she is quite upset and agitated over the sudden appearance of severe motor loss. These facts are given to each physician at the start of the problem, as he walks into the examining room to meet the "patient" for the first time. The initial facts are consistent with a wide range of diagnoses and hence there is a diagnostic problem to be solved.

The table also shows the weights that are assigned to the critical findings for two common diagnostic hypotheses, hysteria and multiple sclerosis. The patient in question is single, a college student, has a boyfriend with whom she is not contemplating marriage; she may possibly be pregnant. These facts tend to support a diagnosis of hysteria and for this reason a "+" has been indicated opposite each

TABLE 3. CRITICAL FINDINGS WITH WEIGHTS FOR TWO HYPOTHESES: HYSTERIA AND MULTIPLE SCLEROSIS

FINDING	HY	MS
Given at Start of Problem:		
1. 21-year-old female		
2. Paralysis of both legs (chief complaint)		
3. Brought in by ambulance		
4. No fever (99°F oral temperature)		
5. Upset and agitated		
Personal and Social Data:		
6. Single	+	
7. College student	+	
8. Has boyfriend	+	
9. Marriage not contemplated	+	
Medical History and Systems Review:		
10. Acute onset of paralysis (overnight)	+	
11. No previous history of paralysis or similar disturbance		-
12. Visual disturbance (4 weeks ago)	-	++
13. Peculiar sensation on right side of body (starting 3 weeks ago and continuing)		++
14. Urinary urgency (started 2 days ago)		+
15. No history of exposure to infection		
16. No history of recent injury or trauma		
17. Menses 2 weeks overdue	+	
18. Denial of recent stress	-	
19. Knowledge of possible pregnancy	+	
20. No toxic exposure		
21. No difficulty with practiced movements of hands	+	
22. No difficulty with speech	+	
23. No significant headaches		
Physical Findings:		
24. Positive Babinski's sign bilaterally	--	+
25. Blind in one eye	--	++
26. No stiffness of neck		
27. Weakness of left arm, hand and fingers	-	+
28. Paralysis of left triceps	-	+
29. Temperature lost to T2 (collar bone) bilaterally	-	+
30. Deep pain - Lost to T3 on right (2" above nipple) and Lost to T4 on left (nipple line)	-	+
31. Vibration lost to T9 bilaterally (base of rib cage)	-	+
32. Touch lost to T10 (umbilicus)	-	+
33. Sensation OK in saddle area	-	+

FINDING	HY	MS
34. Complete loss of voluntary motion from waist down:		
a. leg extensors		
b. leg flexors		
c. calf extensors		
d. calf flexors		
e. adductors		
f. gluteal		
35. Complete loss of sensation from waist down:		
a. touch		
b. deep pressure		
c. temperature		
d. pain		
e. Proprioception		
f. vibration		
36. No limitation in range of motion of joints	-	+
37. Palpable bladder		
38. Deep tendon reflexes increased	-	
39. Abdominal reflexes decreased	-	+
40. Abdominal muscles weak (can't sit up)	-	+
Lab Data:		
41. CBC		
42. Urinalysis		
43. Electrophoresis of CSF		
44. Skull X-Rays		
45. Spinal X-Rays		
46. Cervical myelogram		
47. Colloidal Gold	-	+

TOTALS	HY	MS
+	9	14
++	0	2
-	14	1
--	2	0
	25	17

in Table 3 under the column "HY". The findings pointing most strongly toward multiple sclerosis (incidentally, the correct diagnosis) are: history of visual disturbance four weeks earlier, a peculiar sensation on the right side of the body starting three weeks ago and continuing, a positive Babinski's sign bilaterally and the fact that the patient is indeed blind in her right eye on the morning of the examination although she does not know it. Because these findings, taken as a whole, point so strongly to multiple sclerosis most have been weighted ++. Note that a sizable group of findings are + for one hypothesis and - for the other (e.g., 24 and 25), while others are + for one, and equivocal (no entry) for the other. Some findings do not aid in differentiating between hysteria and multiple sclerosis (e.g., 34a-f), since they are consistent with either alternative.

Table 3 could be extended to provide a set of weights for every conceivable diagnostic hypothesis, but in the interests of simplicity only two are presented. The table permits the investigator to analyze any work-up of this medical problem in terms of how many findings were elicited and sensed for any possible diagnosis, the ratio of confirming (+) and disconfirming (-) findings elicited to the numbers potentially available, and thus to compare work-ups to each other in terms of a common standard. (Those familiar with the Rorschach test may find a resemblance between our method here and Beck's approach of comparing any inkblot response to a published standard for evaluating F+% [Beck, et al., 1961].)

In Table 4, Dr. X's work-up of the case is summarized. At the far left, the question numbers serve as an index to the points at which critical findings were elicited. The columns in the body of the table indicate the diagnostic hypotheses

TABLE 4. DR. X'S WORK-UP: CRITICAL FINDINGS AND HYPOTHESES

QUESTION NUMBER	HY	H1	H2	H3	H4	MS
Given Findings 1-5	Gen.	Gen.				
5	10+		10+		10+	
10						11-
11			Gen.			
21	7+					
25	18-					
28				Gen.		
28				23-		
30	(12-)					(12++)
33	(27-)	(27+)		26-		(27+)
33	(40-)	(40+)		Rej.		(40+)
43	6+					
45			Rej.			
45			15-			
55	36-					
57	(38-)					
69	32-	32+				
73	8+					
75	9+					
89					Gen.	
89					16-	
96					Rej.	
112	17+					
116	19+					
126	(30-)					(30+)
127	34b					34b
130	34d					34d
137	(39-)					(39+)

Key:

HY = Hysteria
H1 = Organic Disease
H2 = Viral Infection
H3 = Meningitis
H4 = Trauma
MS = Multiple Sclerosis
( ) = Findings elicited but not sensed

which Dr. X, in fact, entertained in the course of the work-up, hysteria being on the far left. Multiple sclerosis is a diagnosis which he never considered although it is in fact correct. H1, H2, H3 and H4 refer to four other hypotheses which he generated and partially tested. The term "Gen." identifies the approximate point at which the hypothesis was generated while "Rej." indicates where it was terminated or rejected. Hypotheses H2, H3 and H4 were rejected after one or two pieces of negative evidence had been elicited. The hypothesis of organic disease (H1) was never formally terminated by Dr. X, but merely allowed to fade away. At the close of the inquiry, he is testing only one hypothesis, his early favorite, hysteria. Note that six findings which are negative for hysteria are marked in parentheses in the appropriate column in Table 4. This indicates that Dr. X elicited these findings but did not sense their significance for his work-up. This illustrates the effect of an early commitment to the diagnosis of hysteria upon his inquiry. He did not process disconfirming evidence. Ironically, the findings not sensed were strong evidence for multiple sclerosis. The elicitation of these findings did not lead him to generate this hypothesis and without the hypothesis as an organizing schema within which to evaluate these findings, they were not sensed.

Dr. Y's work-up is shown in Table 5. For simplicity and ease of presentation, a complete analysis is shown only for two hypotheses, hysteria and multiple sclerosis, although his other hypotheses could be similarly analyzed. Dr. Y's early hypothesis, generated on the basis of the evidence given to him at the start of the case, was hysteria. For the first part of the work-up, up to Q

TABLE 5. DR. Y'S WORK-UP: CRITICAL FINDINGS AND HYPOTHESES

QUESTION NUMBER	HY	H1	H2	MS
Given Findings 1-5		4-		
	Gen.			
10	10+			
22	18-			
27	7+			
29		11-		11-
59	6+			
67		Gen.		
71		Rej.	Gen.	
71			13	
96	21+			
98	22+			
116				12++
116				Gen.
147				23
212	34a			34a
213	34b			34b
215	36-			36+
215	34c			34c
221	35e			35e
225	34d			34d
231	35f			35f
233	35d			35d
233	30-			30+
234	35a			35a
234	32-			32+
235	29-			29+
261	27-			27+
281	39-			39+
283	40-			40+
284	34e			34e
286	38-			38
290	24--			24++
322				14+

KEY:

HY = Hysteria
H1 = Infection
H2 = Peculiar vascular condition
MS = Multiple Sclerosis
H3 = Neurofibroma

116, findings elicited are largely supportive of that diagnosis. At that point, he elicits a finding (#12, transient visual disturbance) which is strongly positive for multiple sclerosis. He immediately generates a new hypothesis, multiple sclerosis. Shortly thereafter, he proceeds into the physical exam where he quite exhaustively searches for findings which would enable him to differentiate multiple sclerosis and hysteria. In the sequence beginning at Q 212 and ending at Q 290 he elicits a range of findings about half of which are equivocal for the two diagnoses, the other half of which point toward multiple sclerosis and uniformly away from hysteria. Dr. Y sensed all the facts he elicited. Everything that he found contributes to his evaluation of the case. Cues imply hypotheses and subsequently evidence is marshalled leading to acceptance or rejection.

Table 6 presents a statistical summary of the two work-ups. There are 57 critical findings. Both physicians are given 5 at the start of the case. Dr. X elicited and sensed another 18, 32% of the available findings. Dr. Y elicited and sensed 30, 53% of the available findings. Dr. X elicited but did not sense 6 critical findings; all were negative for hysteria and 5 were positive for multiple sclerosis. Dr. Y sensed every finding which he elicited. To illustrate the impact of commitment to a hypothesis on the elicitation of facts, look at the percentage of critical findings positive and negative for the two diagnoses. Dr. X elicited 78% of the critical findings which are positive for hysteria and only 6% of the findings that are positive for multiple sclerosis. Dr. Y was much more evenhanded in his elicitation of positive findings. He elicited 55% of the critical findings positive for hysteria and 63% of the findings positive for

TABLE 6. STATISTICAL SUMMARY OF TWO WORK-UPS

	Dr. Y	Dr. X
Total Critical Findings	57	57
Total Problem Elements Given	5	5
Total Problem Elements Elicited	30	18
Total Problem Elements Elicited but not Sensed	0	6
Total Hypotheses Generated		
% Critical Findings (CF) Elicited and Sensed	30/57 = 53	18/57 = 32
% CF + for Hysteria	5/9 = 55	7/9 = 78
% CF - for Hysteria	10/14 = 71	3/14 = 21
% CF + for Multiple Sclerosis	10/16 = 63	1/16 = 6
% CF - for Multiple Sclerosis	1/1 = 100	1/1 = 100
Number of Critical Findings		
Hysteria	5+, 10-, 0 not sensed	7+, 3-, 6- not sensed
Multiple Sclerosis	10+, 1-, 0 not sensed	1+, 1-, 5+ not sensed

multiple sclerosis. Thus, we see that Dr. Y searched about equally for positive findings for two diagnoses. A commitment to one diagnosis did not cause him to overlook evidence that favored another. Having a clear contrasting alternative to hysteria in mind helped Dr. Y greatly in testing and weighing evidence. As Table 4 shows, Dr. X never did generate a strong alternative to hysteria and discarded most alternatives after minimal disconfirmation.

Even more striking are differences in their handling of negative evidence. Dr. X elicited and processed only 21% of the negative evidence for hysteria while Dr. Y found 71% of this evidence. It is perhaps tempting to conclude that it is the ability to utilize disconfirming evidence which distinguishes good from poor clinical work in this illustration. But the facts do not necessarily imply that Dr. Y is a more efficient processor of negative information. The structure of this medical problem itself dictates that a sizable body of findings are + for multiple sclerosis and - for hysteria. Thus, having generated both hypotheses, Dr. Y can search for positive findings for either. His strength as a problem solver may lie not in a relatively rare gift to draw inferences from negative information, but rather in his capacity to generate alternative hypotheses so that all the facts he finds are + for some concept. Then, they can be sensed, retained in memory, and utilized in solving the problem. Dr. X, in contrast, never generated the alternatives he needed for which his unsensed findings would have been + data. His repeated failure to do so, when presented with many of the same cues which Dr. Y observed, implies premature commitment to a single alternative.

Analysis of these cases thus suggests that, in this example, a good medical

work-up can be differentiated from a poor one in three ways:

1. The better work-up shows greater flexibility in generating alternative hypotheses based on minimal information. It is crucial for Dr. Y's success that he generate the hypothesis of multiple sclerosis the instant he encounters a strongly positive (++) finding for that disease. Having generated it, it implies for him a plan of search and a schema for organizing findings.
2. Therefore, the better work-up is characterized by greater sensitivity to critical findings. This feature, is in our opinion, contingent upon having a hypothesis available as an organizing framework for the data. Thus, early sensitivity to cues facilitates hypothesis generation which in turn facilitates sensitivity to findings emerging later.
3. Finally, the better work-up appears to exemplify a more comprehensive, efficient use of negative proof. But this too, is a consequence of having available for testing competing hypotheses so structured that data positive for one are negative for the other.

Thus, efficiency in diagnosis seems to be a function of not simply generating early hypotheses, but more specifically, of generating hypotheses which are strong conceptual competitors. Dr. X, in fact, generated and tested more hypotheses than Dr. Y, but none of his alternatives to hysteria were framed so as to be strong competitors. Perhaps his inability to generate strong alternatives was a function of

defects in his knowledge, perhaps a result of premature closure on the psychogenic hypothesis. Dr. Y seems to employ a method of multiple working hypotheses (Chamberlin, 1965). A question for further study is, what conditions of the problem setting or attributes of the problem solver increase the likelihood of using this method?

Finally, it should be stressed that we are not claiming that all, or even most, good medical inquiry is structurally similar to Dr. Y's approach. We are simply demonstrating here a method for the comparative study of different work-ups, so that common features of good work can be empirically determined. We are, however, encouraged with this analysis because it can be readily related to principles and findings in the psychology of non-medical problem solving, and it is to these conceptual links that we now turn.

DISCUSSION AND IMPLICATIONS

We will briefly discuss the theoretical implications which derive from the pilot study reported. Another paper (Elstein, Shulman, Kagan and Jason, 1970) more fully develops the theoretical model which directs this research.

This paper has reported on an analysis of medical inquiry which combined a variety of investigative methods. The methods used included direct observation of physician performance while dealing with a simulated patient; thinking aloud techniques; segmented retrospection, in which the physician was encouraged to reflect on what he had just done during natural "breaks" in the medical interviewing process; and stimulated recall retrospection, in which the interview was reviewed as a whole

by physician and investigators with the aid of an immediate videotape playback.

The findings generated using these methods can now be reviewed in the light of the four criteria enunciated at the beginning of this paper. There is an acceptable level of inter-rater reliability at those points where reliability has already been calculated. There are several other aspects of the scoring system whose reliability has not yet been systematically investigated but we have no reason to anticipate that there will be a great deal of difficulty in those areas. Examination of the scoring from the vantage point of clinical medicine reveals that the relevant aspects of the medical interview have been captured in the scoring procedures. We can examine the duration and character of the interaction between physician and patient. We can analyze the distributional breakdown of particular questions by medical content categories. These categories reflect the amount of effort that the physician is expending for both information-gathering and the establishment of interpersonal rapport. Analysis of the deeper structure of the interview begins to explain how the physician is using these questions in order to move toward a diagnosis. The points in the interview where diagnostic hypotheses are generated and the apparent reasons why some continue to develop while others are rejected can be studied and understood.

The language of the "deep structure" level of analysis is drawn to a great extent from the lexicon of cognitive psychology. The very structure of the analysis makes it readily amenable to comparison and contrast with existing positions on the psychology of thinking and problem solving. Since our purpose is not only to develop a deeper understanding of one particular domain of inquiry, namely medical

diagnosis, but also to use this understanding to augment general cognitive theory, the compatibility of these two language systems is an important and desirable characteristic.

We have also demonstrated that when applying this scoring system to two contrasting protocols, one of which can be judged globally as an example of successful inquiry and the other unsuccessful, our system meets the psychometric criterion of discrimination. That is, the scores effectively distinguish between levels of performance as rated independently.

Related Theories

Clearly, the division of medical inquiry into levels of surface and deep structure derives from the seminal work of Chomsky (1965) in linguistics. At this stage, we are merely using his constructs as a convenient descriptive language for emphasizing the contrast between observable performance and underlying operations. Whether the theory of medical inquiry which ultimately evolves from this research takes on the character of a "grammar", i.e., a set of generative rules of inquiry competence, remains questionable.

Analysis of the two inquiry protocols suggests that the two physicians differed markedly in their ability to process and to make use of the information they elicited, especially that information which we might call "negative instances". We know that, in general, negative instances are extremely difficult to process (Hovland and Weiss, 1953; Bruner, et al., 1956; Donaldson, 1959). We know further that it is characteristic of many problem solvers to ignore negative instances

if they can find sufficient positive instances to bolster a hypothesis which they are holding (Wason, 1968). This is clearly one way of accounting for the differences between the performances of Dr. X and Dr. Y. There is another way of accounting for those differences. What constituted a negative instance for Dr. X, because he had not already formulated a hypothesis within which to accomodate the finding, constituted a positive instance for Dr. Y, since the set of multiple working hypotheses with which he was operating included one for which any particular observation could constitute positive evidence. This argument is more fully developed in the previous section.

The descriptions of chess thinking by De Groot (1965) are also very suggestive. Examination of our protocols lends credence to De Groot's concept of progressive deepening. De Groot argued that chess masters develop several alternative lines of possible attack and explore them mentally in an alternating fashion, moving back and forth at continually deeper levels. We believe that one of the major virtues of progressive deepening is that it guarantees the operation of multiple working hypotheses. It may very well be that the coexistence of several working lines of inquiry is a necessary feature of any approach to problem solving which must combat the dangers of premature closure leading to inadequate handling of negative instances, Einstellung, and other psychological states which inhibit the effectiveness of the problem solver.

In their classical studies of concept attainment, Bruner, et al. (1956) argued that focusing strategies were much more efficient than scanning strategies. Scanning strategies, you will recall, are strategies in which the problem solver

begins with hypotheses, either single or multiple, and processes the information in the light of these hypotheses. Focusing strategies are more purely inductive in nature, and differ from each other only in their degree of conservatism in information processing. The reason, Bruner argued, for the relative inefficiency of scanning strategies, is that they lay far too great a burden of cognitive strain on the information processor. We have taken note of Bruner's observation and the marked contrast between his assertion and our reality. We have found that early generation of hypotheses in a scanning mode, rather than indcutive focusing strategies are the characteristic hallmarks of the experienced diagnostician. This observation has been further supported in recent articles by an Australian investigator (Dudley, 1970; 1971).

We believe that it is readily understandable why Bruner's observations and ours do not agree. Bruner and other psychologists have constructed experimental settings in which, for purposes of maintaining the control needed, they divorce the content of the experimental task from all previous experiences and systematic bodies of knowledge which the inquirer may bring into the research setting. For obvious reasons, it was neither possible nor desirable to do that in our studies of medical diagnosis. In fact, the world at large is a place where problem solving is rooted in and dependent upon systematic bodies of knowledge stored in various structured ways in the memories of problem solvers. What we need is a set of theoretical formulations which will account for how cognitive functioning occurs in the presence of such structured bodies of knowledge, not in the absence of them. We hope that the present series of studies will serve to make some small contribution to that yet infant body of investigation and theory.

REFERENCES

Allender, J.S. The teaching of inquiry skills using a learning center. AV Communication Review, 1969, 17, 399-409.

Beck, S.J., Beck, A.G., Levitt, E.E. and Molish, H.B. Rorschach's test. Vol. 1: Basic Processes. New York: Grune and Stratton, 1961.

Bruner, J.S., Goodnow, J.J., and Austin, G.A. A study of thinking. New York: Wiley, 1956.

Chamberlin, T.C. The method of multiple working hypotheses. Science, 1965, 148, 754-759.

Chomsky, N. Aspects of the theory of syntax. Cambridge, Mass.: M.I.T. Press, 1965.

De Groot, A.D. Thought and choice in chess. The Hague: Mouton, 1965.

Donaldson, M. Positive and negative information in matching problems. British Journal of Psychology, 1959, 50, 235-262.

Dudley, H.A.F. The clinical task. Lancet, 1970, 1352-1354.

_____. Clinical method. Lancet, 1971, 35-37.

Elstein, A., Shulman, L., Kagan, H. and Jason, H. A theory of medical inquiry. Proceedings of the Ninth Annual Conference on Research in Medical Education. Washington: Association of American Medical Colleges, 1970.

Harvey, A.M., Cluff, L.E., Johns, R.J., Owens, A.H., Rabinowitz, D. and Ross, R.S. Principles and practice of medicine. 17th Ed. New York: Appleton-Century-Crofts, 1968.

Hovland, C.I. and Weiss, W. Transmission of information concerning concepts through positive and negative instances. Journal of Experimental Psychology, 1953, 45, 175-182.

Kagan, N., Elstein, A., Jason, H., and Shulman, L. Methods for the study of medical inquiry. Proceedings of the Ninth Annual Conference on Research in Medical Education. Washington: Association of American Medical Colleges, 1970.

Schwab, J.J. The practical: A language for curriculum. School Review, 1969 78, 1-23.

Shulman, L.S., Seeking styles and individual differences in patterns of inquiry. School Review, 1965, 73, 258-266.

_____, Loupe, M., and Piper, R. Studies of the inquiry process. Department of Health Education and Welfare, Office of Education, Project No. 5-0597, Michigan State University, July, 1968.

Simon, H.A., Human problem solving: Current state of the theory. Address to the American Psychological Association, Miami Beach, 1970.

Wason, P.C. 'On the failure to eliminate hypotheses. . . . . .'--------a second look. In P.C. Wason and P.N. Johnson-Laird (Eds.), Thinking and reasoning, Baltimore: Penguin Books, 1968.