THE EFFECT OF HYPOTHESIS GENERATION AND VERBALIZATION ON CERTAIN ASPECTS OF MEDICAL PROBLEM SOLVING

Dissertation for the Degree of Ph. D.
MICHIGAN STATE UNIVERSITY
SARAH A. SPRAFKA
1973

8-10-73 jk

pd $30

ABSTRACT

THE EFFECT OF HYPOTHESIS GENERATION AND VERBALIZATION ON CERTAIN ASPECTS OF MEDICAL PROBLEM SOLVING

By

Sarah A. Sprafka

Thirty medical students going into their fourth year of medical school were asked to solve three modified Patient Management Problems. Subjects either were or were not constrained to think aloud during problem solving. They were instructed to generate diagnostic hypotheses early, to withold judgment about diagnosis until the end of the problem, or were given no instructions about hypothesis generation. Number of hypotheses generated, thoroughness of cue acquisition, efficiency of cue acquisition, and accuracy of solution were the dependent variables. Multivariate analysis of variance revealed that instructions concerning hypothesis generation had no effect on outcome. Subjects constrained to verbalize generated significantly more hypotheses for one problem (Problem III) than subjects without that constraint. Further, there was a significant interaction effect of instructions concerning hypothesis generation and constraint to verbalize on number of hypotheses generated for that same problem.

In the interest of assessing the relation between certain aspects of performance and outcome, subjects were reassigned to groups using many ($\geq$ 10) or few ($<$ 10) hypotheses as one independent variable and early or late hypothesis generation as the other independent variable. Dependent measures were thoroughness and efficiency of cue acquisition, and accuracy of outcome. The interaction of early or late hypothesis generation and many or few hypotheses affected thoroughness on one problem (Problem I).

Subsequently two of the three problems were analyzed to determine whether a different problem solving process had been used by subjects who had high accuracy scores on those problems and subjects who had low accuracy scores. It was found that on the one problem which had a complex solution subjects who had low accuracy scores generated but did not retain the elements of the complex solution. On the other problem it was found that subjects who received low accuracy scores either did not generate or generated and dropped the correct hypothesis. These subjects also generated a larger variety of inaccurate hypotheses than subjects receiving higher accuracy scores.

The results of the three stages of analysis demonstrate that instructions about hypothesis generation and constraint to verbalize have little overall effect on performance. Furthermore, whether subjects generate the first hypothesis early or late, or generate many or few hypotheses makes little difference. Lastly, differences in accuracy of solution can be tentatively attributed to different causes for different problems.

THE EFFECT OF HYPOTHESIS GENERATION AND VERBALIZATION ON CERTAIN ASPECTS OF MEDICAL PROBLEM SOLVING

By

Sarah A. Sprafka

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Counseling,
Personnel Services and Educational Psychology

1973

ACKNOWLEDGMENTS

I would like to acknowledge the professional direction and encouragement given me by the members of my guidance committee: Dr. Stephen Yelon, Dr. Lee Shulman, Dr. Arthur Elstein, Dr. Gordon Wood and Dr. Howard Teitelbaum.

I owe thanks also to Professor Christine McGuire and Mr. Gordon Page for the supply of materials, information and advice.

The cooperation of members of the staffs of the University of Michigan Medical School, Sparrow Hospital in Lansing, Saint Mary's Hospital in Grand Rapids and Hurley Hospital in Flint who made space available for interviewing subjects is gratefully acknowledged.

DEDICATION

To Merlin and Falstaff who gave me solace in my weakest moments.

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

Page

Table

Figure

LIST OF APPENDICES

LIST OF DISPLAYS

CHAPTER I

INTRODUCTION

Medical care is made up of many facets, one of which is diagnosis. Medically defined, diagnosis is "the determination of the nature of the case of disease" (1). The association of diagnosis exclusively with disease is a questionable one. Therefore a looser, non-medical definition seems more appropriate. More loosely defined, diagnosis is "the investigation or analysis of the cause or nature of a condition ... or problem" (2). In order to diagnose a patient's condition or problem a physician must go through an information gathering and processing procedure, arriving finally at some sort of conclusion about his patient's condition or problem.

Diagnosis as a cognitive activity may be studied in a number of ways. Two major approaches are: 1) the study of diagnosis as a decision-making process, and 2) the study of diagnosis as a problem solving process. Although the two processes are closely linked in practice, the terms "decision-making" and "problem solving" refer to two distinct areas of research. Students of diagnosis as decision making use a methodology which differs markedly from those who look at it as problem solving. The emphasis for the decision-makers is more on outcome than on process. The method of study depends much more on mathematical models used as predictors or describers of outcome.

Correlation models, regression models and probability models are among the most popular.

Those who study diagnosis as problem solving, the present writer included, place more emphasis on the process of arriving at a diagnostic decision, whether tentative or not, than on the decision itself. The method of study usually involves conducting experiments much like the one reported below in the interest of ascertaining what the essential elements of diagnostic problem solving are and what role they play in the problem solving process.

The Problem

A number of ways of progressing from the beginning to the solution of a diagnostic problem have been proposed. One way advocated by many texts and teachers of medical diagnosis is to gather a large amount of data according to a prescribed format, record this information and then use it to arrive at a tentative differential diagnosis or list of problems which will help determine what diagnostic tests should be ordered and what management steps would be taken. This approach is advocated by Lawrence Weed, for example (3,4). He feels that a complete Data Base should be gathered on a patient before any further steps are taken, except of course in cases of emergency. Weed offers criteria for the establishment of a Data Base which would apply to almost all patients who came to a given health care facility. The Data Base can be elicited by paramedical personnel, thus saving the physician who will see the patient a good deal of time and trouble. When presented with the Data Base the physician can then select those cues which fit together into potential problems and, when he sees the patient, can perhaps gather more data to enable him to formulate those problems and make diagnostic and management decisions based on them.

Another way of approaching a diagnostic problem involves the constant evaluation of information as it comes in. This, to a certain degree is the approach advocated by Morgan and Engel (5). Especially in gathering information about the presenting complaint one should at times let one's ideas about possible causes guide one's questioning. Morgan and Engel emphasize the necessity of finding out about the bodily location, chronology, setting, etc. of a patient's complaints. Questions about these aspects can be triggered by considering possible causes. These considerations can, for example, cause the physician to elicit other symptoms the patient has not mentioned yet. They can also help direct specific questions about aggravating and alleviating factors which the patient might not have mentioned. Further, on-going use of information to suggest possible causes can help direct the physical examination.

Morgan and Engel do not place great emphasis on the use of incoming information to guide inquiry. They appear to assume that physicians do this during interviews as well as physical examinations, and give examples of where this activity can be put to good use.

Although both of these approaches to gathering and using patient information yield similar results, i.e. a preliminary differential diagnosis, they arrive at this end via rather different routes. How, then, does the physician use the information he gathers to solve a diagnostic problem? It is felt that the essential reasoning mechanism used to transform information into a diagnostic solution is the generation of diagnostic hypotheses and the evaluation of these hypotheses in the light of the information available. Thus the Weed procedure and that advocated by Morgan and Engel have an essential element in common: hypothesis generation. But this element plays a different role in each approach to problem solving. On the one hand establishing the Weed Data Base involves gathering information in a prescribed order. Hypotheses are then generated based on a large amount of data. The problem solver reasons from facts to hypotheses. On the other hand, conducting an interview and physical examination in the manner prescribed by Morgan and Engel may lead to the gathering of different kinds of information in varying orders depending on what possible causes the physician entertains and how he goes about evaluating them. The problem solver in this case may reason either from facts to hypotheses or vice-versa -- from hypotheses (possible causes) to facts (gathered via specific questions about those causes).

The importance of the difference between these two approaches to diagnostic problem solving is borne out by recent investigations (6) which are finding that despite training in the fact-to-hypothesis approach to diagnosis, many physicians seem to reason both from facts to hypotheses and vice-versa as acknowledged by Morgan and Engel. The result is a combination of the format procedure and an hypothesis generating and testing procedure. Information is gathered partly according to a memorized format and partly in the interest of testing hypotheses. The actual solution to the problem is usually arrived at by verification of one or more hypotheses generated along the way, backed up by the elimination of others.

Questions To Be Investigated

Certain questions arise out of these observations. Should a diagnostician be encouraged to generate hypotheses early in a problem; should he be encouraged to reserve judgment; or should he simply be permitted to find his own problem solving style? Would instruction to use a format versus an early hypothesis generation approach have any effect on the physician's performance, both as to his approach to the problem as well as to the quality of his diagnosis? Independent of instructions, does the early generation of a number of hypotheses lead to a more or less accurate diagnostic formulation than the late generation of as few as one hypothesis?

Recent investigations (6) have raised some questions about the appropriate methodology for investigating hypothesis generation and testing behavior. Those investigations used a thinking aloud procedure as well as a stimulated recall to assess the physician's thinking as he solved each problem. The authors of that study (6) felt that the influence of thinking aloud on a physician's reasoning should be investigated. The present study therefore included a thinking aloud (verbalization) condition.

Specifically, the study investigated the following questions:

1. Do instructions to use early vs. late hypotheses have any effect on the diagnostician's approach to a diagnostic problem?
2. Do instructions to use one or the other of these approaches have any effect on the quality of his diagnosis?

3. Independent of instruction, is there any relationship between the number of hypotheses, and the delay of their generation, and the quality of a diagnostic workup?
4. Do instructions to verbalize during a problem have any effect on the diagnostician's approach to the problem?
5. Do instructions to verbailze have any effect on the quality of the diagnosis?

CHAPTER II

REVIEW OF LITERATURE

Two major areas of investigation are relevant. One has to do with the difference in problem solving strategy being studied here; the difference between generating and testing hypotheses during problem solving, and generating hypotheses only after most of the information relevant to the problem has been gathered. The second area of investigation deals with the effect of verbalization on problem solving.

The issue of problem solving strategy has its roots in the philosophy of science [see Kessel, 1969 (7) for a stimulating discussion of this subject]. In its purest form the difference being studied here has to do with the nature of so-called factual information and how this information is treated by scientists. Francis Bacon, an early philosopher of science (8) believed that facts should be treated as strictly objective. Scientific investigation should proceed step-wise from the observation of particulars to the creation of elementary axioms from those observations. These axioms should then be used as the basis for further experiments which would lead to the establishment of more global axioms, thence to further experiments, and so on to the creation of general principles. Bacon cautioned strongly against using a few observed particulars as bases for creation of broad principles, and then using these principles to direct new observations and the creation of more elementary axioms. This approach to science implied the interpretation of observations in the light of previously established principles. In his opinion observations should in no way be colored by the observer's predispositions. For him observed facts were the ultimate arbiters of theory.

This strictly empirical approach to scientific investigation has engendered a good deal of criticism. Fundamental questions arise as to the relationship between facts which may be observed and the observer. Is it not true that what is observed and how observations are interpreted are strongly determined by the observer? Is it not also true that many theories have been created on the basis of comparatively little information, and perpetuated in the fact of contradictory observable evidence?

Kessel (7) points out that the autonomy of observable facts is highly questionable. Facts cannot be separated from the reasons for gathering them. Scientific investigators' choices of problems to study and their methods of gathering and interpreting information reflect their own predispositions as well as the general predisposition of their field of study. Kuhn notes (9) that the interpretation of facts may change following a change in paradigm, i.e. a change in the intuitive conception of the nature of things. It thus appears that scientific investigation as practiced is not strictly empirical. Beyond that, it may be that strict empiricism is impracticable.

The alternative to a strict facts-to-theory approach is one that involves the generation of tentative formulations or hypotheses based on small amounts of data and the subsequent testing of these hypotheses. Hypothesis testing leads to the gathering of more data and the possible generation of more hypotheses. As noted above, this seems to be the way scientific investigation is practiced. Two strong proponents of this hypothetico-deductive approach are Popper (10) and Medawar (11).

Popper proposes that "... the work of a scientist consists in putting forward and testing theories" (10, p. 31). He feels that the goal of science is not to string facts together into generalizations, but to justify certain generalizations with experience. Furthermore, theories are not methodically built around facts. A lot of creativity and luck goes into the creation of a theory. Only once it is created can it be tested. Medawar criticizes the empirical approach for its emphasis on facts without interpretation. A fact is not a fact unless it is interpreted relevant to something, and that something is a hypothesis or theory. Therefore, to gather facts without interpreting them is likely to be impossible. And if possible it is going to lead to the unsystematic gathering of a lot of irrelevant information. He proposes that scientific investigation should be a hypothetico-deductive process. Hypotheses should be generated by whatever means (including observations of phenomena, luck, and inventiveness) and then further data should be gathered systematically in the interest of testing those hypotheses. By testing, a hypothesis can be refuted or temporarily accepted.

Diagnostic problem solving can never be equated with scientific investigation. Yet fundamental elements are shared. In both domains information is gathered and interpreted by humans. Hypotheses are formed. The hypotheses may be tested and the results may lead to the generation of new hypotheses as well as to the further confirmation or disconfirmation of previously entertained hypotheses. Furthermore, something akin to the strictly empirical approach to scientific investigation is possible for the diagnostic problem solver. Information may be gathered and recorded without being interpreted. After a large amount of information has been gathered, sets of observations may be grouped together leading to the generation of diagnostic hypotheses.

The results of either of these approaches to diagnosis is usually the same -- a correct diagnosis. Is one better than the other? A better question would be, should one be used to the exclusion of the other? Both approaches have their advantages and disadvantages as is demonstrated by studies of problem solving discussed below.

Early investigation into approaches to problem solving was done by Luchins and Luchins (12) in their work on set or Einstellung. Subjects were given problems for which the discovery of a mathematical formula was necessary to avoid a trial and error approach. However once the appropriate formula had been discovered and found to work on a series of problems, the problems were modified by the experimenter with the result that the subject could not apply the formula he had discovered to the solution of one of the modified problems. Rather than backing off and reformulating the problem, the subjects tended largely to stick with their original equation and fail to solve that problem. That they held to the original equation was evidenced by their using it to solve two other modified problems which could have been solved more neatly using a simpler equation. One may find a lesson herein. These studies were designed specifically to test problem solving set. The set which was established involved encouraging subjects to generate a hypothesis about the appropriate formula for solution. Subsequently they had trouble changing this hypothesis, even though it made it more difficult, even impossible in one case, to solve subsequent problems.

Similar findings are reported as the result of extensive investigation by Wason and others (13). It was found that subjects given the opportunity to generate a rule for a progression of numbers did so quite readily. Unfortunately if they generated the wrong rule they were unwilling or unable to change it or even to think of examples of number progressions which might prove their rule incorrect. A number of means were used to encourage subjects to invent counterexamples to their rule or to invent a totally different rule. Some worked to a certain extent, but none was totally successful. Wason concludes that people in these and similar circumstances will tend to stick to a conclusion even though erroneous, and will not entertain alternatives or attempt to disprove their conclusion.

Medical diagnosis may be a circumstance similar to this. It is possible that by generating hypotheses early on, a physician may not truly test and attempt to falsify his hypotheses, but gather information only in the interest of confirming them. As a matter of fact, this phenomenon has been observed in recent investigations of diagnostic reasoning. Elstein and Shulman (14) state that having formed a hypothesis that a patient is hysterically ill rather than afflicted with an organic disease, one physician studied elicited cues which would disconfirm this hypothesis by asking a set of rather routine questions, but did not process these cues, i.e. did not apply them to the disconfirmation of that hypothesis.

Returning once again to non-medical problem solving -- some of the most extensive research in how subjects go about solving a restricted they had trouble changing this hypothesis, even though it made it more difficult, even impossible in one case, to solve subsequent problems.

Similar findings are reported as the result of extensive investigation by Wason and others (13). It was found that subjects given the opportunity to generate a rule for a progression of numbers did so quite readily. Unfortunately if they generated the wrong rule they were unwilling or unable to change it or even to think of examples of number progressions which might prove their rule incorrect. A number of means were used to encourage subjects to invent counterexamples to their rule or to invent a totally different rule. Some worked to a certain extent, but none was totally successful. Wason concludes that people in these and similar circumstances will tend to stick to a conclusion even though erroneous, and will not entertain alternatives or attempt to disprove their conclusion.

Medical diagnosis may be a circumstance similar to this. It is possible that by generating hypotheses early on, a physician may not truly test and attempt to falsify his hypotheses, but gather information only in the interest of confirming them. As a matter of fact, this phenomenon has been observed in recent investigations of diagnostic reasoning. Elstein and Shulman (14) state that having formed a hypothesis that a patient is hysterically ill rather than afflicted with an organic disease, one physician studied elicited cues which would disconfirm this hypothesis by asking a set of rather routine questions, but did not process these cues, i.e. did not apply them to the disconfirmation of that hypothesis.

Returning once again to non-medical problem solving -- some of the most extensive research in how subjects go about solving a restricted set of problems (the discovery of figural concepts) is that done by Bruner and colleagues (15). Bruner's studies found that although subjects used a number of strategies to discover a concept, the strategy which worked best, i.e. produced the correct solution after the fewest trials was labeled the "conservative focusing strategy" by Bruner. This strategy involves finding an object which is a positive instance of the concept in question, identifying all the elements of that instance, and then picking successive objects which differ from the positive instance in only one way and establishing which of these are also positive instances of the concept. When all the necessary positive instances have been identified (all the necessary information has been gathered) the concept itself can be formulated. This type of problem is then best solved by reasoning from individual pieces of information to a higher order statement. The advantages of using this approach for solving a restrictive problem of this type are obvious -- once all of the necessary positive instances are identified, the formation of the concept itself follows naturally. All elements of the problem are mutually exclusive and exhaustive. Further, the problems are content-free. The selective application of previously learned information is not required to solve them. All information necessary for solution is contained in the problem. The application of this approach to the gathering and interpretation of data for a diagnostic decision however, is inappropriate for three reasons. All of the relevant symptoms which might serve to incontrovertibly establish a diagnosis can never be found. The concept problem is amenable to an exhaustive search, the diagnostic problem is not. Secondly, hypotheses to which symptoms may be applied are not mutually exclusive. Lastly, the selective use of previously learned information plays a large part in diagnostic problem solving.

Evidence presented thus far would lead one to believe that although a hypothetico-deductive approach to diagnostic problem solving may be attractive, it may also be a risky approach to take. Perhaps in reality a more thorough approach is called for. However, investigations described below will tend to indicate that human information processors may have trouble proceeding in that manner. These studies of problem solving and information processing have led to theories explaining the organization of and restrictions on the human problem solving mechanism.

Miller, Galanter and Pribram (16) propose that although algorithms for problem solving are thorough and always lead to the correct solution, people tend not to use them. Use of an algorithm usually involves the recall of large amounts of relevant information, and the human short term memory tends to become overloaded. Instead of algorithms, people employ heuristics, or what the authors call Plans for problem solving. The Plan used by a given individual to solve a given problem is also somewhat restricted by human factors, since it is largely determined by that individual's Image of the problem. The Image consists of what the problem solver knows about the problem situation, e.g. the boundaries of its solution, its essential elements, possible time factors, how good he is at problems of this type, and so on. Different problem solvers may have different Images of a problem and for that reason choose different Plans for its solution. In all events, if problem solving is going to take place, the problem solver must have an Image of the problem as well as a heuristic or Plan for solving it.

A more concise expression of the Miller, Galanter and Pribram idea is given by Simon and Newell (17). Their theory of problem solving posits that although very few characteristics of the human as an information processer are constant over task and person, those few are enough to determine that a task environment (problem) is represented by the solver as a problem space. Problem solving then, takes place in a problem space. The structure of the problem space is determined by the nature of the task. And the set of steps, heuristics, etc. used to solve the problem are determined by the structure of the problem space. Once a problem is detected the problem solver creates a space for that problem which contains some representation of his goal, the relevant information he has, as well as some strategies for gathering and using further information that will get him to his goal without over loading short term memory. The problem space then is a type of orientation of the problem solver. It is not as specific as a theory to be tested but it does preclude the random gathering of information without a referent.

A type of strategy might be that proposed by Bartlett in his studies of sectional map reading (18). He calls solving sectional map reading problems (how to get from point A to point B quite a ways away by the shortest route possible?) problem solving in an open system. He has found that map readers tend to explore along the line of greater possibilities since the more possibilities you have to work with the more probable you are to find the right one, i.e. the shortest route to your destination. Although sectional map reading does not present a short term memory problem, still some sort of scheme for search must be established at the outset. The preferred scheme seems to be a flexible exploratory one rather than a binary search or some other more rigid strategy.

Another scheme or strategy is the one proposed by Miller (19) which we use to avoid over loading short term memory. He proposes that as more and more information is gathered concerning the solution of a problem we need to group it together. The grouping is necessary because the capacity of short term memory is on the order of 7 ± 2 elements. To avoid exceeding that limit a problem solver needs to "chunk" (Miller's term) elements together which go together thus reducing the memory load.

It appears that not only is the search for and use of information for problem solving not random, it is highly organized and can be quite selective. Facts, or particulars, are selected for observation based on the problem solver's perception of the task.

The discussion thus far has focused on approaches to non-medical problems. How do subjects approach problems concerned with medical diagnosis? In his experiments with physicians, Kleinmuntz (20) found that his subjects used one very useful strategy for gathering and storing information. In a study where the subject was required to gather pieces of information one at a time from a data bank, Kleinmuntz found that subjects with more medical experience tended to start with general questions and converge on a diagnosis using progressively more and more specific questions. The types of questions asked were those which yielded the greatest amount of information (reduced ambiguity to the greatest extent). The information which was stored was relevant to a specific diagnostic hypothesis. Although a definite strategy is being used here, it is a binary search strategy rather than a more flexible one such as a hypothetico-deductive approach. If medical diagnosis can be considered a search in an open system, using a rigid binary search strategy to solve diagnostic problems tends to contradict Bartlett's findings. The strategy used was very likely an artifcat of the experimental setting. Subjects were constrained to asking "yes-no" questions, so the most efficient strategy available was obviously the binary search.

Other authors have found strong evidence for hypothetico-deductive reasoning in physicians. Price and Vlahcevic (21), both physicians, speak from their own experience. They make an excellent case for hypothetico-deductive reasoning in diagnosis. Their claim is that physicians choose hypotheses and interpret the data they gather in the light of those hypotheses. A diagnostic decision is arrived at by combining the elimination of erroneous hypotheses with the tentative confirmation of one or more appropriate hypotheses. They insist that both of these elements must be present. An effort should be made to find <u>the</u> diagnostic formulation which fits the greatest number of symptoms. Not only must that diagnosis fit that set of symptoms, but conversely all symptoms necessary to confirm that diagnosis must be present. Furthermore, all other hypotheses which could fit that set of symptoms must have been rejected.

In a similar vein Dudley (22) has observed that experienced physicians are much more selective about their data gathering than medical students. He sees the data gathered by physicians as being processed as it is gathered and lumped together into Boolean type nets or lattices relevant to one or more hypotheses. The establishment of these nets enables a physician's search to become more and more specific and the possible hypotheses to be honed down to only one.

Studies to investigate these aspects of diagnostic reasoning have been done by Sprosty (23) and Elstein et al. (6). Sprosty studied history taking and accuracy of diagnosis in medical students. He observed that those students who obtained the correct diagnosis seemed to ask more and shorter questions. Furthermore, their questions were more specific and it was apparent that more hypothesis testing was being done by these students than by those who did not obtain the correct diagnosis. Although Sprosty does not have a measure for hypothesis generation, it is obvious that those students whom he found gave good performances did generate and test hypotheses.

A recent study by Elstein et al. (6) has led to a preliminary theory of medical inquiry. The focus of this theory is hypothesis generation. The authors found that physicians tend to generate specific diagnostic hypotheses early in the workup, usually as the result of some discrepant finding interpreted by the physician as problematical. These hypotheses may be systematically tested and/or further data may be gathered according to some routine the physician has memorized. The data is then applied to the various hypotheses in the interest of disconfirming some and tentatively confirming others.

Another recent study by Elstein et al. (30) demonstrates that hypothetico-deductive thinking occurs in diagnostic problem solving and that different strategies are used for different problems. This study is a process analysis of four Patient Management Problems (35) completed by 15 physicians who participated in a larger study (6). In each problem subjects were asked to generate diagnostic hypotheses at various points throughout the problem. All subjects, as instructed, generated diagnostic hypotheses early in each problem. It was not possible to assess whether specific items of information were selected to test out those hypotheses. Thoroughness and efficiency of data acquisition were evaluated, however, and some inferences may be made from those scores about the effect of hypotheses on data acquisition and utilization. In one problem for example, those subjects who generated the correct hypothesis early in the problem needed less total data to arrive at the correct solution than did those subjects who did not generate the correct hypothesis until later. One may infer from this that the correct hypothesis, generated early, was guiding an efficient course of data acquisition. This phenomenon is not uniform across problems, however. On another problem an attractive but erroneous diagnostic hypothesis was presented at the outset of the problem. This clearly had an effect on subjects' hypothesis generation and data acquisition. In this problem all subjects who arrived at a correct solution generated the correct hypotheses late in the problem, perhaps due to the influence of the hypothesis suggested at the outset. Those subjects who did not reach an accurate solution, did not ever generate the correct solution, and restricted the scope of their data acquisition. This appears to be an example of Luchins' Einstellung effect (12) in which the early suggestion of one attractive but erroneous hypothesis led to restricted data collection, failure to generate the correct hypothesis, and failure to arrive at the correct solution.

These findings demonstrate that diagnostic problem solving is not a straight-forward uniform process. Different approaches may be taken to different problems, and success or failure may be achieved in different ways.

In conclusion then, it appears that hypothetico-deductive reasoning is not only justifiable as an approach to solving scientific problems, it may also be justified as an approach to more practical problems such as medical diagnosis. As information processors humans tend to employ heuristics for gathering, storing and using information relevant to different problems. One of the most productive of these heuristics is the hypothetico-deductive one. However, use of this heuristic exclusively may also be risky in that it restricts the amount and type of data gathered. It therefore should be, and usually is supplemented with other data gathering heuristics. The combination of heuristics used may vary depending on the problem. The present study did not investigate the other heuristics. They and other elements of diagnostic problem solving are being studied elsewhere (e.g. 24, 25).

The second major element of the study concerns use of the "thinking aloud" procedure to gather data and the resultant effects of verbalization on problem solving. Thinking aloud was used as early as 1917 when Claparede (26) studied the origin of hypotheses about problem solutions. He describes the procedure as being useful since it is neither retrospective nor introspective but gives a running count of the problem solver's process. He found the drawbacks of thinking aloud were that first of all it needs training, and even then some subjects do not talk during the most interesting moments of their problem solving. One reason for this may be that when one is thinking very hard he is not prone to verbalize. Furthermore, one thinks much faster than he talks. We might infer from this that constraints to think aloud may alter the course of thought as well as perhaps slow it down. More recently Newell and Simon (27, 28), and Simon and Newell (17) have relied heavily on the thinking-aloud technique to obtain information about specific heuristics being used by subjects for solving problems. Neisser (29) has criticized this procedure on the grounds that any complex or multiple processing being done by a subject may be made to appear sequential by the thinking aloud. Worse yet, thinking aloud may cause a subject to use a sequential process where he would not do so otherwise.

Elstein et al. (6) in their study of diagnostic reasoning used thinking aloud to determine what processes physicians were using. Since then McGuire (32) has expressed concern that thinking aloud may not only make diagnosticians' thought processes appear more orderly than they are, but may also cause their processes to be more orderly than they would otherwise have been.

Little concrete evidence is available on the effects of thinking aloud on problem solving. Gagne and Smith (33) built a thinking-aloud condition into their study of children's ability to solve a problem and to formulate a general rule for its solution. They found that differences in accuracy of solution were attributable to verbalization. Moreover those subjects who verbalized were able, after the fact, to generate acceptable general rules for solution.

It appears then, that verbalization does have an affect on problem solving, although perhaps not a detrimental one as Neisser fears.

On the contrary the effect may be to enhance problem solving ability. If this is so, the phenomonon clearly deserves further attention.

CHAPTER III

DESIGN OF THE STUDY

Hypotheses

The hypotheses relevant to this study fall into two major areas: one concerning the effect of verbalization on problem solving; the other concerning the effect of instructions on approach to problem, as well as the effect of instructions and approach to the problem on solution. A brief explanation of how the relevant hypotheses were investigated (expanded in Procedures section below) will clarify the meaning of each specific hypothesis.

Hypotheses relevant to verbalization were investigated by either constraining subjects to verbalize during problem solution, or omitting that constraint.

Hypotheses relevant to instructions and approach to problem were investigated as follows. Subjects were first divided into three groups. One (Group E) was instructed to generate hypotheses early in the problem; another (Group L) was instructed to withold judgment about diagnoses until all the information is in; a third (Group C) was given no instructions about hypothesis generation. All subjects were given identical problems to solve. The amount of information gathered before generation of the first hypothesis and the number of hypotheses generated by each subject were tabulated. All subjects were scored on the efficiency, thoroughness, and accuracy of their solution.

Subsequently subjects were reassigned to groups based on their performance. Those subjects who asked at least one question before generating their first hypothesis were assigned to one group, those who asked no questions before generating the first hypothesis to another group. Similarly, those subjects who generated comparatively many hypotheses were assigned to one group, those generating comparatively few hypotheses to another. The degrees of cross-over from the original Instructions groups to the after-the-fact performance groups were assessed, and efficiency, thoroughness, and accuracy were recomputed for the performance groups.

The reason for re-assignments was to assess the degree to which subjects in the various Instructions groups followed the instructions they were given. As stated earlier, humans as information processors tend to select information based on their perception of the task, and to store information temporarily in such a way as to avoid overloading short term memory.

A non-medical example of this occurred in a well-known study done by Frase in the area of prose learning (40). He organized sentences about chess men and their attributes in three ways: according to names, according to attributes and randomly. Subjects were asked to read the group of sentences and try to recall as much as possible in a given time. Amount and organization of the subjects' written efforts at recall were evaluated. It was found that although subjects in the group reading randomly organized sentences recalled less than those in the other two groups, they tended to use the names of the chessmen as the basis for organization of their recall. As a matter of fact, 30% of those subjects whose reading passages were organized around attributes reorganized the sentences in their recall to a names organization. The experimenter hypothesized that the tendency to organize recall around the names of the chessmen might have been due to the fact that such an organization required less memory than the attribute organization or no organization at all.

The major heuristic used for medical problem solving which helps one to avoid a memory overload is the generation of hypotheses to which cues may be related as they emerge. Use of this heuristic is shown in the reports of Price and Vlahcevic (21), Sprosty (23), and Elstein et al. (6). These studies led one to believe that in spite of instructions to the contrary, subjects might generate hypotheses before all the information was in. As concerns other hypotheses stated specifically below, studies cited earlier suggested that instructions about hypothesis generation would have a selective effect on how early in the problem hypotheses were generated as well as on how many hypotheses were generated. Furthermore, these instructions would have an effect on the thoroughness and efficiency of problem solution as well as on accuracy. There is no concrete evidence as to the effect of instructions to verbalize on earliness of hypothesis generation or number of hypotheses generated. However, the experimenter felt that concerns that verbalization helped to guide thinking were legitimate. This was investigated. There is concrete evidence (33) showing a positive effect of verbalization on accuracy of solution. The specific hypotheses tested in this study are listed below.

Effect of instructions on efficiency, thoroughness, and accuracy of performance

1. Subjects instructed to generate hypotheses early will give a more efficient, less thorough, and more accurate performance than those instructed to withold judgment.
2. Subjects given no instructions about hypothesis generation will show the same pattern as in #1 above.
3. Subjects constrained to verbalize will give a more accurate solution than those without that constraint.
4. Instructions to verbalize will have no effect on the efficiency or thoroughness of performance.

Effect of instructions on earliness of hypothesis generation

5. Subjects instructed to generate hypotheses early will generate hypotheses earlier than those instructed to withold judgment.
6. Subjects given no instructions to generate hypotheses will show a pattern similar to #5 above.
7. Instructions to verbalize will have no effect on how early hypotheses are generated.

Effect of instructions on number of hypotheses generated

8. Subjects instructed to generate hypotheses early will generate more hypotheses than those instructed to withold judgment.
9. Subjects given no instructions about hypothesis generation will show a similar pattern to #8 above.
10. Instructions to verbalize will have no effect on the number of hypotheses generated.

Effect of performance (re: hypothesis generation) on efficiency, thoroughness and accuracy

11. Subjects who generate hypotheses early will give more efficient, less thorough, and more accurate performance than those who generate hypotheses later.
12. Subjects who generate comparatively many hypotheses will give a more efficient, less thorough, and more accurate performance than those who generate comparatively few hypotheses.
13. Independence of earliness of hypothesis generation and number of hypotheses generated: Earliness of hypothesis generation and number of hypotheses generated will be statistically independent.

Subjects

Subjects were 30 medical students going into their fourth year of medical school. Fifteen of these subjects were randomly selected from the Michigan State University College of Human Medicine, and 15 came from the University of Michigan Medical School. This sample was chosen because:

1. The materials used were deemed difficult enough to challenge this group, but were not perceived to be too difficult for them.
2. These medical students were more accessible than, for example, a group of physicians would be.
3. This group of students had similar backgrounds. They had similar amounts of medical knowledge and had had similar practical experiences. A more sophisticated group such as physicians would have had a more divergent set of experiences.

Since it has been found that medical students' information gathering skills vary as they progress through a training program (34), a group of students at different levels in medical school would also be inappropriate.

4. This group was judged to be interested in the materials to be used in the study since the problems resemble Part III of the examination of the National Board of Medical Examiners which some of these students would be taking within a year of the study.

Procedures

All subjects were given a modified version of three Patient Management Problems (PMP) developed by the Interdepartmental Appraisal Committee of the University of Illinois College of Medicine. The experimenter administered each problem individually to each subject. After reading and checking the common instructions as well as experimental group-specific instructions (see below), each subject was asked to request information from that available. The information was made available to the subject by a cue sheet containing numbered items. As the subject requested information he recorded the number identifying each item. The experimenter then handed that information, printed on a file card, to the subject. All subjects, regardless of group assignment were requested to record a differential diagnosis at the end of each problem. All subjects did the three problems in the same order.

The independent variables manipulated were instructions concerning hypothesis generation and verbalization. Each subject was assigned to one of six groups as follows:

Group E-V (instructions to generate hypotheses early and verbalize) -- Subjects in this group were instructed to generate diagnostic hypotheses as early in the problem as possible and, if they wish, to use these hypotheses to guide their data gathering. They were also stopped periodically during the problem and asked to write down any diagnostic hypotheses they had at that point as well as describe how and when those hypotheses were generated.

Group E-NV (instructions to generate hypotheses early but no verbalization) -- Subjects in this group were instructed to generate diagnostic hypotheses as early in the problem as possible, and, if they wished, to use them to guide the course of their data gathering. At the end of each problem each subject was asked to go back over the problem orally with the experimenter and indicate what hypotheses were generated and at what point in the problem generation occurred.

Group L-V (instructions to generate hypotheses only after all the data are in and to verbalize) -- Subjects in this group were admonished to withold judgment about diagnostic hypotheses until most or all of the data were in. They were stopped periodically during the problem and asked how they were coming along and what their thoughts were about the data they had gathered up to that point.

Group L-NV (late generation, no verbalization) -- Subjects in this group were similarly admonished to reserve judgment about diagnostic hypotheses until the end of the problem. They were then asked to review the problem orally and comment on what their thoughts had been about the progress of the problem at certain points.

Group C-V (no hypothesis generation instructions, verbalize) -- Subjects in this group were given no instructions concerning hypothesis generation. They were simply stopped periodically and asked if they had any idea as to where the problem was going or any other comments on the problem up to that point.

Group C-NV (no hypothesis generation instructions, no verbalization) -- As in Group C-V above, subjects were given no instructions about hypothesis generation. At the end of each problem subjects were asked to review the problem orally with the experimenter. Review questions emphasized where the subject thought that problem was going at certain points.

The instructions which were read to each subject are contained in Appendix A.

Description Of The Patient Management Problems And Modifications Made For The Present Study

The Patient Management Problems (PMP) were developed over a number of years by the Interdepartmental Appraisal Committee of the University of Illinois College of Medicine. A recently published book, Clinical Simulations (35) contains a large selection of the problems developed to date. Each problem begins with a brief introduction containing some information about the "patient", including the chief complaint. All problems deal with a patient with some kind of ailment requiring a physician's care. None are of the healthy recruit or insurance physical variety. Having obtained the initial information, the task of the examinee (problem solver) is to gather more information in the interest of diagnosing and/or managing the patient. The particular subset of problems to be used in modified form for this experiment consists of a booklet containing four problems (three will be used). In this booklet is the introductory information and a list of further types of information available. Accompanying the booklet is a set of answer sheets and figures. The answer sheets contain the answers to the "questions" the examinee may choose from the booklet. The figures (blood smears, x-rays, etc.) are non-verbal answers to questions in the booklet. To obtain an answer to a question the examinee must rub out an opaque overlay covering that section of the answer sheet corresponding to the question he asked. Correspondence is achieved by numbering each possible question and each response. No track is kept of the order in which an examinee requests information, except for a record of the order in which certain sections of a problem are done. What hypotheses may be governing his search is not determined in the original version of the booklet. A special version developed for the MSU Medical Inquiry Project asks the problem solver to list the hypotheses he is considering at the end of each problem section.

The PMP's were developed as a result of early work done by Rimoldi (36). Rimoldi's Test of Diagnostic Skills gives the subject some initial information about the patient and asks the subject to obtain further information to solve the problem. The solution is a diagnosis of the patient's illness. Additional information is provided on cards contained in a problem folder. On one side of each card is printed the question the subject may wish to ask. History questions, possible physical exam manipulations, and laboratory and other studies are included. On the reverse side of each card is printed the answer to each question. All questions available for any one problem are displayed for the subject at one time. The sequence of each subject's choices is recorded by the experimenter. Scores on the Test of Diagnostic Skills related to the number and usefulness of items chosen; the order in which choices were made; and the accuracy of the final diagnosis. Primary emphasis is placed on the utility of each choice. Utility is defined in terms of the frequency with which each item is chosen by the group taking the test. An item chosen by many subjects has high utility. A high score is gained by choosing as few high utility items as possible. Rimoldi claims (36) that the reasoning of the subject as he reaches a diagnosis may be assessed via these problems.

The PMP's differ from the Test of Diagnostic Skills in a number of ways. First, PMP's are diagnosis and management problems, not just diagnostic problems. Secondly, the format of the PMP offers the subject a number of alternative,, and equally good routes through a problem. Only one optimal route through a problem will result in a high score on the Test of Diagnostic Skills. Thirdly, the number of options available to a subject doing a PMP is usually greater than the Test of Diagnostic Skills, thus reducing the effect of cueing. Lastly, and most important, the PMP's are scored in such a way as to measure something quite different from what is measured by the Test of Diagnostic Skills. Items are weighted as to their value to the problem solver. Strongly positive weights are given to those items which help the subject to diagnose and manage his patient. In addition positive weights are given to items which should be included in a thorough workup and conscientious management plan. Negative weights are assigned to items which should not be chosen (e.g. because they may be costly to the patient) and zero weights are assigned to items which are non-contributory or simply distractors. The decision on assignment of weights is made by a group of criterion physicians.

Subjects are scored for over-all competence in working up and managing a patient (see [37]). A proficiency score as well as an efficiency score is calculated. In contrast to the Test of Diagnostic Skills, strong emphasis is placed on proficiency, a kind of selective thoroughness, and less emphasis is placed on efficiency, or reaching a solution in the fewest possible steps. A high score is obtained by choosing a reasonably large number of positively weighted items or by doing a thorough workup (diagnosis as well as management) of the patient. Distortions in performance due to cueing are greatly reduced by offering a large number of options and making it difficult for the subject to scan all available options at once.

PMP scores are calculated as follows:

Definition of terms:

H_s = Positively weighted items selected by S

h_s = Negative and zero weighted items selected by S

MAX = Sum of all positive weights possible

$$\text{Efficiency} = \frac{\Sigma Hs}{\Sigma Hs + \Sigma hs}$$

$$\text{Proficiency} = \frac{\Sigma \text{ weights of } H_s + \Sigma \text{ weights of } h_s}{\text{MAX}}$$

$$\text{Errors of Omission} = \frac{\text{MAX} - \Sigma \text{ weights of } H_s}{\text{MAX}}$$

$$\text{Errors of Commission} = \frac{-\Sigma \text{ weights of } hs}{\text{MAX}}$$

The subject receives all of the above scores for his diagnostic performance and his management performance. He also receives a score for attack strategy by which he is rewarded for following an appropriate sequence of sections and penalized for doing certain sections out of order.

Reliability Of The PMP

Since the PMP is an unconventional (i.e. not multiple choice or true-false) achievement test, estimates of its reliability cannot be made by using standard methods such as the Spearman-Brown or the Kuder-Richardson formulas. Thus alternative approaches must be developed (38). For these purposes reliability cannot be considered strictly as the accuracy with which a test measures something, reflecting a judgment about how closely a subject's score on the test approximates his true score. One must instead consider the purpose which estimations of reliability serve, namely to show with what consistency a test measures what it purports to measure.

Calculations of measurement consistency for the PMP have been made in two areas: 1) consistency across different ways of scoring the test, and 2) consistency across different but similar tests (38). In the first instance two methods were used. First the standard weights of items were changed to increase penalties for errors as well as to increase reward for correctness. Spearman's rho was computed on tests of the six medical specialties in the battery for the two systems of rating. Rho ranged between .95 and .97. A more significant change in the scoring procedure was to ask two groups of judges to independently assign weights to each item (previous weights had been assigned by consensus of one group of judges). Subjects' scores were then recomputed using the two new sets of weights. Correlations between scores using the two sets of weights ranged from .92 to .95.

Consistency across different tests was assessed in a number of ways, two of which were: 1) consistency across subtests, and 2) consistency across problems in different disciplines. One estimation of consistency across subtests was obtained by dividing a given problem into two problems, each of which tested the same competence factors as were tested in the original whole test. Correlations were then computed between scores on the subtest and on the whole test for all disciplines included in the battery. Average correlations ranged between .445 and .912, a result which does not differ greatly from that obtained on multiple choice tests. Consistency across problems in different disciplines was obtained by correlating scores on the whole battery of twelve problems as they occur in one discipline (e.g. Surgery) with scores on those problems as they occur in another discipline (e.g. Medicine). A coefficient alpha of .56 was obtained.

The investigators feel that these estimations of consistency of measurement show the PMP to be a reliable instrument.

Validity Of The PMP

The validity of the PMP, or how well it measures what it purports to measure has been considered from four points of view (39).

1. Predictive validity or how well performance on the test predicts subsequent performance of a similar type. Thus far follow-up studies which would evaluate the predictive validity of the PMP have not been done.

2. Concurrent validity or how well performance on this test resembles performance on another test that is considered a true measure of the competence in question. Performance of physicians on the PMP was compared to their performance of data gathering in an actual clinical setting as revealed by chart audits. It was found that physicians recorded considerably less data on their charts than they requested on the PMP. However when six items considered critical to solution of the PMP were singled out, a rather high correspondence was found between physicians' recording of that item on the chart and their requesting that item on the PMP. It can be stated then that performance on the chart and performance on the PMP are highly similar for certain specific relevant items. No correlations were reported (39, p. 9).
3. Content validity or how closely intellectual processes used to solve the PMP resemble those used in a clinical setting. The correspondence has been assessed by asking subjects to comment on their thought processes while solving the problems and compare them to the processes used in practice. Anecdotal data indicate that the process of responding to PMP's closely simulates the thinking process one goes through in a clinical setting (39, p. 4). Again, no statistical data are reported.
4. Construct validity or how closely differences in performance of different groups on this test corresponds to reasonable hypotheses about how the groups should differ. Two reasonable hypotheses considered were: 1) As a subject grows older and more experienced, he will tend to make a decision based on

less information. It was found that experienced physicians gave a less thorough performance on the PMP than medical students. No data are reported. 2) As a subject grows older and more experienced, he will be more willing to take prompt, even radical action. In a study using residents, candidates and examiners as subjects and a problem calling for amputation, 36% of the residents, 40% of the candidates, and 50% of the examiners chose amputation as a course of action.

Since the PMP is an effort to use a paper and pencil simulation to measure a highly complex kind of competence, a high degree of correspondence between performance on this simulation and performance in a clinical situation cannot be expected. The authors (39) feel that the degree of correspondence that has been found demonstrates that the PMP is an adequate simulation of clinical problem solving.

Modifications To The PMP For The Present Study

General

The present study used the PMP's more as an observation instrument than an evaluation instrument. Subjects were not scored relative to any criterion. Judgment was not passed as to the quality of a subject's work-up, only as to the accuracy of his diagnosis. This use of the PMP justifies certain changes in the format as well as in the scoring of the problems.

Format modifications

1. The PMP presents available information in a booklet format. The booklet also includes instructions for progressing through

the problem, i.e. instructions about what section of the problem the subject should do next. The booklet format and instructions may restrict the subject's freedom of choice in requesting information. In the interest of observing the sequence which a subject would naturally choose in this situation the presentation was changed to a set of cue sheets and the instructions were omitted. The cue sheets for each problem are contained in Appendix B.

2. The PMP makes available the information requested by having the subject rub out an opaque overlay, thus revealing the "answer". This format enables the subject to obtain more than one item of information at a time and denies the experimenter the opportunity of keeping track of the order in which information was requested. In the present study information was presented in printed form on cards one item at a time. This procedure assured that the subject would receive no more information than he requested, and it facilitated the recording of the order in which items were presented.

Scoring modifications

1. Thoroughness -- The PMP Proficiency (or selective thoroughness) score is a calculation of the percentage of positive points (or total weights of positive items) chosen by the subject (see p. 32 for formula). This way of calculating thoroughness rewards subjects as much for choosing a smaller number of heavily weighted items as for choosing a larger number of less heavily weighted items. Perhaps this is why it is called a Proficiency

rather than a thoroughness score. In the opinion of this experimenter a more thorough solution is given by choosing more positively weighted items regardless of the magnitude of their weight, as well as by choosing zero-weighted items. For that reason thoroughness for this experiment was the percentage of positively and zero-weighted items chosen by the subject (see formula p. 39). In addition to fulfilling the criteria stated above, this method of calculating thoroughness yields a score having the same metric as the efficiency score which is the percentage of all the items the subject chose which were positively weighted.

2. In the light of the purpose for which the present version of the PMP was used, no errors of omission or of commission were calculated. Furthermore, no separate efficiency and thoroughness score for diagnosis and management were computed, nor was an overall competency score determined.
3. Accuracy, which is not calculated for the PMP was computed on a five-point scale (0, 1, 2, 3, 4). The accuracy of each subject's definitive diagnosis was determined by comparing his diagnostic formulation with a set of often-entertained diagnostic hypotheses which had been weighted for their appropriateness by the developers of the PMP.

In summary, problem scores calculated for the study reported here are:

$$\text{Efficiency} = \frac{\text{Number of positively weighted items chosen by subject}}{\text{Total number of items chosen by subject}}$$

$$\text{Thoroughness} = \frac{\text{Number of positively and zero-weighted items chosen by subject}}{\text{Total number of positive and zero items in problem}}$$

Accuracy = 0, 1, 2, 3, or 4

Procedures for determining Thoroughness, Efficiency and Accuracy scores are contained in Appendix C.

Validity And Reliability Of The Modified Problems

To make judgments about content and concurrent validity of the modified problems, performance on two of these problems by the subjects in the present study was compared to performance of physicians on the same two problems in the "original PMP" form as well as to those physicians' performance on two similar problems presented as high fidelity simulations. Since the same subjects did not complete all three types of simulation (modified PMP, original PMP, and high fidelity simulation) the comparisons are at best crude. Generally, it was hoped that students' thought processes as well as other aspects of their performance on the modified PMP's would closely resemble physicians' performance on the high fidelity simulations. Where comparisons between all three types of problems were possible, it was hoped that the comparison of student performance on the modified PMP's with physician performance on the high fidelity simulations would be more favorable than comparison of physicians' performance on the original PMP's with their own performance on the high fidelity simulations.

The original PMP's used for comparison here are contained in the booklet of four problems mentioned earlier. The only difference between this booklet and the booklet originally created by the University of Illinois group is that the former contains opportunities for subjects to list the diagnostic hypotheses they were entertaining at various points in the problem.

The high fidelity simulations used as criterion for judgments about concurrent validity were two cases presented by programmed patients in a realistic office setting (for further description of these simulations see reference 6). These simulations were chosen for this purpose because they closely resemble the actual clinical setting, and because measures were available from these simulations which could be compared to similar measures on the two types of PMP.

Regarding content validity, the intellectual processes used by students to solve the modified PMP's closely resembled those used by physicians on the high-fidelity simulations in at least two ways. First, subjects doing both types of simulation spontaneously generated diagnostic hypotheses early in the problem after having obtained very few cues. Secondly, subjects doing both types of simulation used the cues they obtained to help them evaluate these hypotheses. Furthermore there were instances where a cue was elicited specifically in the interest of testing out a hypothesis. Information is not available on what specific strategies were used to solve the original PMP's so a comparison cannot be made at present between them and the high-fidelity simulations.

Certain tentative conclusions can be drawn concerning concurrent validity by comparing means and standard deviations of Thoroughness, Efficiency, and Accuracy scores on the three types of simulations. These means and standard deviations are presented in Table 1.

The GI problem presented in the high fidelity simulations completed by physicians was a case of ulcerative colitis in a 22-year-old male. The content of the original PMP completed by physicians was identical to that of the modified PMP completed by students. The hematological problem presented in the high fidelity simulation was a case of hereditary spherocytosis and infectious mononucleosis in a 21-year-old female. The content of the original and modified PMP was identical.

The thoroughness score calculated for the high fidelity simulation and the original PMP was the percentage of cues available in the case which was acquired by the subject. The thoroughness score calculated for the modified PMP was the percent of positive and zero-weighted items (using the weighting system described on p. 31 and 32) chosen by the subject. The efficiency score calculated for the high-fidelity simulations and the original PMP was the percent of cues acquired by the subject which were critical for one or more of his diagnostic hypotheses. Efficiency on the modified PMP was the percent of cues acquired by the subject which were positively weighted. Accuracy on the high-fidelity simulations and the original PMP ranged from zero to two. Accuracy on the modified PMP ranged from zero to four.

Thoroughness and efficiency scores on the three types of simulation may be compared by inspection of Table 1. For the GI problem average thoroughness on the modified PMP is more similar to thoroughness on the high-fidelity simulation than is thoroughness on the original PMP. However, efficiency on the modified PMP is less

Table 1
Means and Standard Deviations of Thoroughness, Efficiency, and Accuracy on Two Problems for Three Types of Simulations

		Thoroughness	Efficiency	Accuracy
GI Problem	Hi fi Simulation	47.14 (7.99)*	32.36 (9.09)	1.63 (.79)
	Original PMP	32.67 (11.44)	44.60 (7.77)	1.62 (.51)
	Modified PMP	40.79 (13.08)	62.79 (9.24)	2.92 (.78)
Hematological Problem	Hi fi Simulation	57.43 (6.40)	46.81 (7.02)	1.80 (.40)
	Original PMP	41.07 (22.94)	67.53 (17.65)	1.23 (1.01)
	Modified PMP	40.17 (12.88)	65.50 (13.32)	2.04 (1.73)

* Standard deviations are in parentheses

comparable to efficiency on the high fidelity simulation than is efficiency on the original PMP. Thus, although average thoroughness on the modified PMP compares most favorably with that on the high fidelity simulation, this is not the case for average efficiency. For the hematological problem thoroughness on the modified PMP compares less favorably with that on the high fidelity simulation than does that on the original PMP. The opposite is true of the average efficiency scores. It should be noted here that the thoroughness and efficiency scores on the original and modified PMP's are quite similar to each other for this problem and neither one compares particularly favorably with its counterpart in the high fidelity simulations.

In summary, the overall comparison of modified PMP scores with high fidelity simulation scores is not more favorable than the comparison of original PMP and high fidelity simulation scores. The author feels that any comparisons favoring the modified PMP might have been serendipitous. The data are taken from two different samples; the scores are obtained by different means; and the content of the problems is not identical for all three types of simulation; all of which renders conclusions about the meaning of these comparisons somewhat doubtful.

It may be appropriate to establish the validity of a paper simulation by comparing performance on it with that on a high fidelity simulation. However, the same subjects should participate in both simulations, the content of both simulations should be identical, and the scores on outcome variables should be arrived at in the same way for both types of simulation. Until this is done, no definite conclusions can be drawn about the validity of these PMP's.

Reliability

In the interest of determing whether the three problems could be considered a three-item test which measured thoroughness, efficiency and accuracy of diagnostic problem solving independent of experimental group assignments, these scores were correlated (Pearson r). The results were as follows:

Thoroughness

Prob 2	.32	
Prob 3	.77	.55
	Prob 1	Prob 2

Efficiency

Prob 2	.09	
Prob 3	.10	.32
	Prob 1	Prob 2

Accuracy

Prob 2	.19	
Prob 3	.09	-.15
	Prob 1	Prob 2

Problem 1:	Acute Abdomen
Problem 2:	Pale Lethargic Child
Problem 3:	Pale Confused Patient

The strong amount of variability in these correlations indicates that there is considerable lack of consistency in Thoroughness, Efficiency and Accuracy across problems and that therefore each problem should be considered a separate test. All subsequent analyses are done on each problem separately.

As stated earlier (page 33) the type of reliability estimates appropriate for these problems are those estimates which reflect the consistency with which a test measures what it purports to measure. With this in mind, the internal consistency of each problem was calculated using one of the procedures described by Lewy and McGuire (38). Each problem was divided into sections as follows:

Problem I -- Credited introductory items (i.e. items 1a and b, 2, 3b, 3c, 4a-c, 9a-c on the cue sheets, see Appendix C)

Physical exam

Laboratory

Non-surgical intervention

Surgical intervention

Problem II -- Diagnosis, Prognosis

Problem III -- Credited introductory items (i.e. items 1a-d, 2a-d, 3-7 on cue sheets, see Appendix C)

History

Physical examination

Laboratory

Therapy

A subproblem was then constructed from each problem by picking every third item, beginning with the first item in each section. Ten randomly selected subjects were then given a score on the whole problem and on the subproblem consisting of the sum of the weights of the items he chose in the total and subproblem. The weights used were those assigned the items by the University of Illinois criterion group. These scores served as the basis for estimation of the reliability (internal consistency) of each problem.

Internal consistency was calculated using the formula derived by Angoff (42) and his correction for spuriousness (43). The resulting internal consistencies were as follows:

Problem I:	.80
Problem II:	.34
Problem III:	.87

These coefficients indicate that Problems I (Acute Abdomen) and III (Pale Confused Patient) were internally consistent. Internal consistency is interpreted here as meaning that the set of items which was chosen for the subtest was a representative sample of the items used in the whole test. This apparently was not the case for Problem II (Pale Lethargic Child). The items which made up the subtest were not a representative sample of the frequency distribution for subjects' choices of items on the whole test. On the whole test the ratio of number of items chosen by more than half of the subjects to number chosen by less than half of the subjects was approximately two to one. On the subtest this ratio was closer to four to one.

In the opinion of this writer, these coefficients do not reflect the generalizability of the results on this test to other tests dealing with the same type of patient and containing a similar factor (defined as sectional divisions) structure. It is felt that to accomplish this goal a set of parallel tests containing items selected from a pool representative of the universe of items appropriate for this type of patient would have to be constructed, and a test-retest reliability would have to be calculated.

Additional comments are in order concerning the use on these tests of any reliability coefficient based on part-whole or part-part correlations. The choice of an item in these problems may depend in one of two ways on the choice of another item in the problem, First, some items are redundant as discussed above. If one of a pair is chosen in any problem then the other member of the pair logically cannot be chosen. When calculating the reliability of the test one item cannot be automatically declared the redundant one and its weight dropped from the calculation. Thus all redundant items are included in the reliability. It is sometimes possible to create a subtest that is equivalent in its redundancy to the whole test, but this possiblity is not always available. Secondly, irrespective of redundancy the items are interdependent from the point of view of the problem solver. What he has learned by a certain point in the problem may determine his choice of subsequent items. Most formulas for calculating reliability, especially those single-administration procedures such as the split-half coefficient or the Angoff formula used here depend heavily on the assumption that the test items are independent of one another. This assumption is violated by the problems discussed here.

CHAPTER IV

ANALYSIS OF RESULTS, SUMMARY AND DISCUSSION

The results of the study were analyzed in two ways. First, the experimental hypotheses were tested using the multivariate analysis of variance program developed by Jeremy Finn (44). This procedure was chosen for the bulk of the analysis for a number of reasons. First all dependent variables were interrelated. Not only was there a close relationship among some independent measures within a problem, but it could not be assumed that performance on any one problem was independent of that on any other. Secondly, the hypotheses were phrased in such a way as to require the analysis of clusters of variables rather than isolated ones. The multivariate procedure permits this type of analysis.

The second type of analysis was a process or clinical analysis of certain subjects' performances. This second analysis was done in the interest of identifying other elements not included among the initial set of variables which might be contributing to variations in performance.

Tests Of Hypotheses Concerning Hypothesis Generation And Verbalization

The first hypothesis to be tested was Hypothesis 13 to establish the statistical independence of earliness of hypothesis generation and number of hypotheses generated. The hypothesis of statistical independence should be rejected if $\chi^2 > \chi^2_1 (.95) = 3.84$. A statistic of $\chi^2_1 = 1.87$ was obtained thus confirming the independence of the two variables.

The next hypotheses to be tested were 1-4 and 8-10. Specifically these hypotheses are:

1. Subjects instructed to generate hypotheses early will give a more efficient, less thorough, and more accurate performance than those instructed to withold judgment.

2. Subjects given no instructions about hypothesis generation will show the same pattern as in #1 above.

3. Subjects constrained to verbalize will give a more accurate solution than those without that constraint.

4. Instructions to verbalize will have no effect on the efficiency or thoroughness of performance.

8. Subjects instructed to generate hypotheses early will generate more hypotheses than those instructed to withold judgment.

9. Subjects given no instructions about hypothesis generation will show a similar pattern to #8 above.

10. Instructions to verbalize will have no effect on the number of hypotheses generated.

Table 2 shows the results of the multivariate analysis of variance relevant to those hypotheses. Note that this analysis could not evaluate the effects of independent variables on Problem III Accuracy since this variable had no variance. Although this occurrence is unfortunate from an experimental viewpoint, it points to possible problem specific differences in the study. Only for this problem did all subjects reach the correct solution. Based on the results shown in Table 2 the hypotheses that instructions about hypothesis generation would have a differential effect on thoroughness, efficiency and accuracy of performance as well as on the number of hypotheses

Table 2
Manova of Effects of Instructions on Number of Hypotheses, Thoroughness, Efficiency, and Accuracy

	# Hypotheses			Thoroughness			Efficiency			Accuracy		
Problem	I	II	III	I	II	III	I	II	III	I	II	III
Cell Means												
Generate early												
Verbalize	8.6	9.0	13.6	46.0	51.6	57.6	61.0	63.0	88.4	2.6	1.0	4.0
No verb.	10.6	6.8	6.0	40.6	49.4	48.8	63.0	65.8	83.8	3.0	1.8	4.0
Generate late												
Verbalize	11.4	6.0	8.6	35.8	45.2	45.6	66.8	68.4	84.4	3.1	2.4	4.0
No verb.	11.6	9.4	10.8	42.0	48.0	54.8	63.0	63.0	84.8	3.0	1.8	4.0
Control												
Verbalize	11.6	9.8	12.0	42.0	53.6	52.4	63.0	68.4	89.4	2.6	2.8	4.0
No verb.	9.6	9.0	7.8	48.2	49.2	55.8	59.0	58.0	89.6	3.4	1.0	4.0
MANOVA												
Effect of Instructions on hypothesis generation (F = 1.04; df = 22, 28; p < .45)												
Effect of Verbalization instructions (F = 2.42; df = 11, 14; p < .06)												
Univariate												
F(df = 1, 24)	.004	.01	4.61	.96	.06	.04	.22	.79	.34	.89	.78	
P less than	.95	.91	.04	.34	.81	.84	.65	.38	.56	.18	.39	
Interaction (F = 2.35; df = 22, 28; p < .002)												
Univariate												
F(df = 2.24)	1.30	2.31	3.71	1.25	.17	.70	.25	.62	.52	.95	1.55	
P less than	.29	.12	.04	.30	.84	.51	.78	.55	.60	.40	.23	

generated must be rejected. Specifically hypotheses 1, 2, 8 and 9 must be rejected.

The multivariate analysis of hypotheses concerning the effect of verbalization constraints on outcome shows that these hypotheses probably should not be rejected. An investigation of the univariate effects of verbalization shows that verbalization probably did not affect the accuracy of performance, thus leading to the rejection of hypothesis 3. Further, there was no apparent effect of verbalization on the thoroughness or efficiency of performance. Hence, hypothesis 4 cannot be rejected. On the other hand verbalization did have an apparent effect on the number of hypotheses produced, especially on Problem III. Subjects constrained to verbalize produced more hypotheses for Problem III than those without that constraint, which points again to possible problem specific differences.

In addition to the verbalization differences, there were differences in number of hypotheses generated due to the interaction of hypothesis generation instructions and verbalization. These differences were particularly evident again on Problem III. This interaction is plotted in Figure 1. Both the vertical and horizontal dimensions of the interaction are of interest. On the vertical dimension subjects instructed to generate hypotheses early and subjects given no hypothesis generation instructions (control) behaved similarly. Both groups generated more hypotheses when constrained to verbalize than when not under that constraint. Subjects instructed to generate hypotheses late showed the opposite pattern. They generated fewer hypotheses when constrained to verbalize than when not constrained to do so. On the horizontal dimension subjects

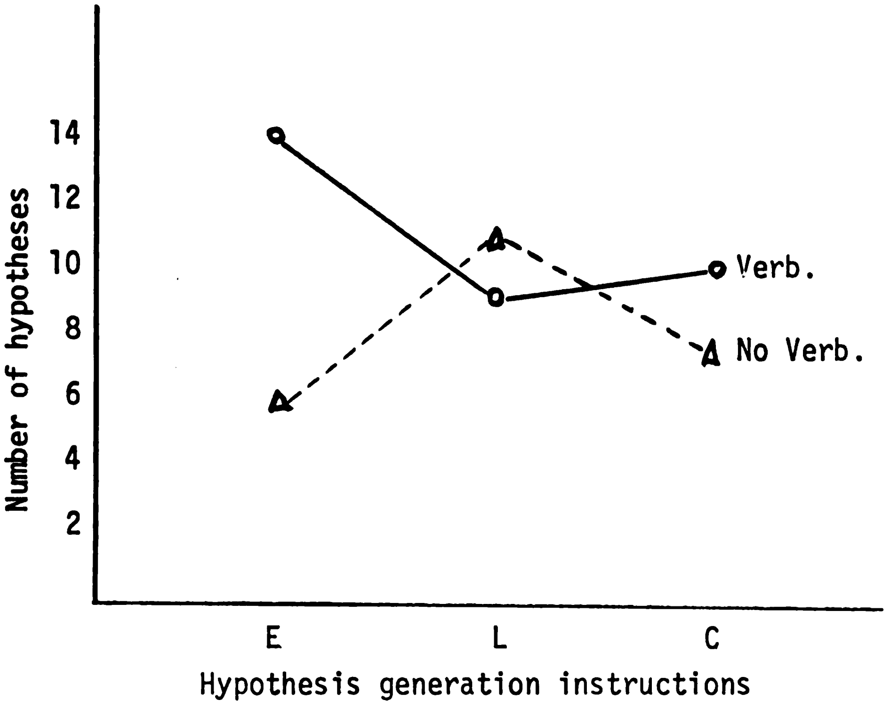

Figure 1: Instructions X Verbalization interaction, Problem III

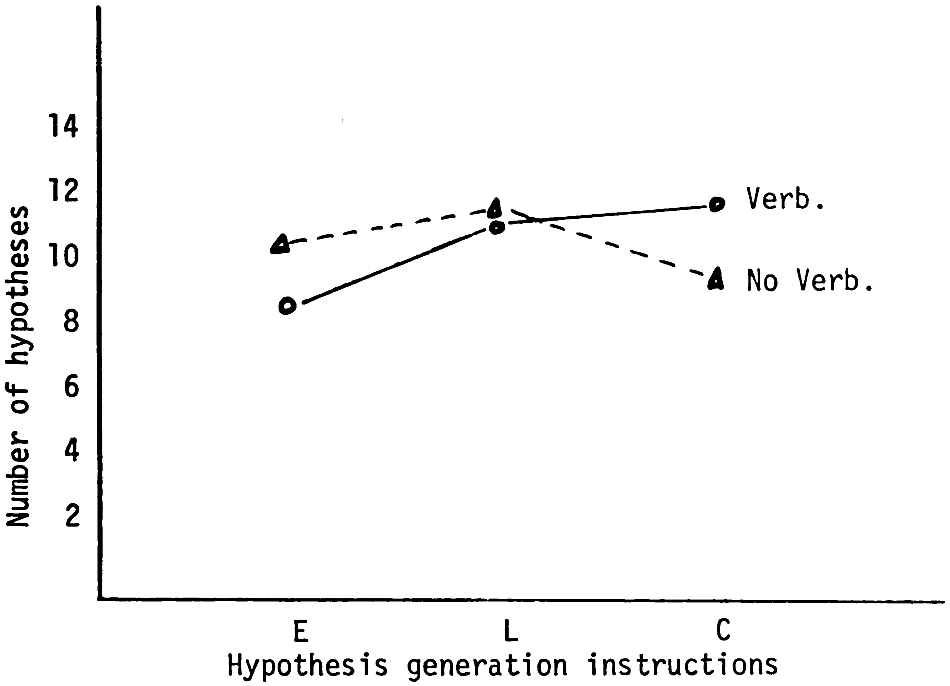

Figure 2: Instructions X Verbalization interaction, Problem I

who verbalized produced more hypotheses when instructed to generate hypotheses early or given no instructions, than they did when instructed to generate hypotheses late. Conversely subjects who were not constrained to verbalize produced fewer hypotheses when instructed to generate hypotheses early or given no instructions than when told to generate hypotheses late.

In the interest of determining whether this was a problem specific effect, the relation of instructions and verbalization to number of hypotheses was plotted for the other two problems as well as for the average number of hypotheses generated by each subject over the three problems. As shown in Figures 2, 3 and 4 these are also interactions. It appears then that this is not a problem specific phenomenon. Except for Problem I the interactions are disordinal. The interaction pattern for Problem I is dissimilar in that 1) subjects in the "early" group who verbalized produced fewer hypotheses than those who did not verbalize, and 2) verbalizers produced fewer hypotheses in the "early" group than in either the "late" or control group.

A statistical analysis was not conducted to evaluate the following hypotheses:

5. Subjects instructed to generate hypotheses early will generate hypotheses earlier than those instructed to withold judgment.

6. Subjects given no instructions to generate hypotheses will show a pattern similar to #5 above.

7. Instructions to verbalize will have no effect on how early hypotheses are generated.

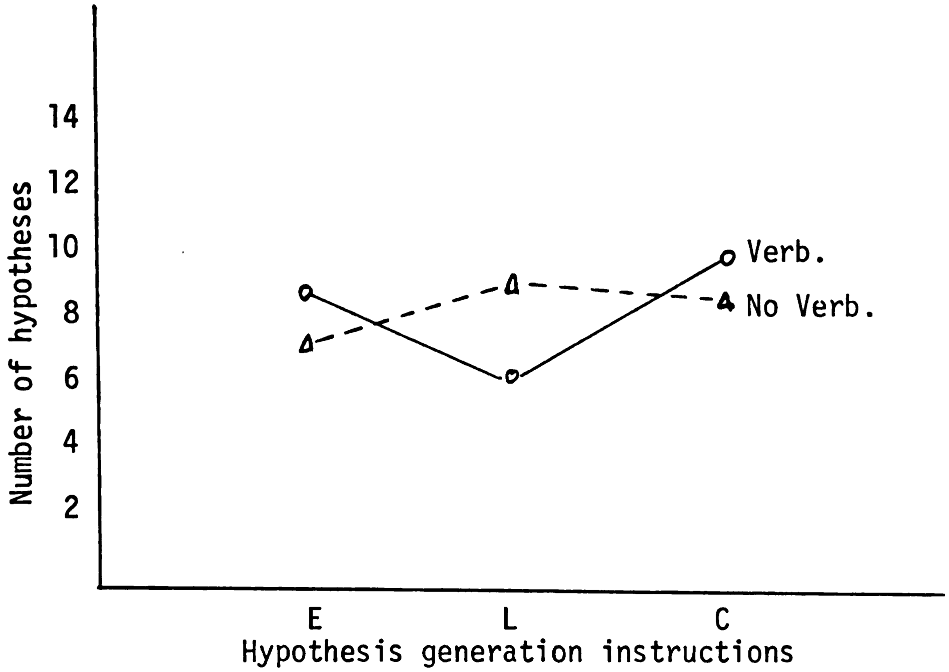

Figure 3: Instructions X Verbalization interaction, Problem II

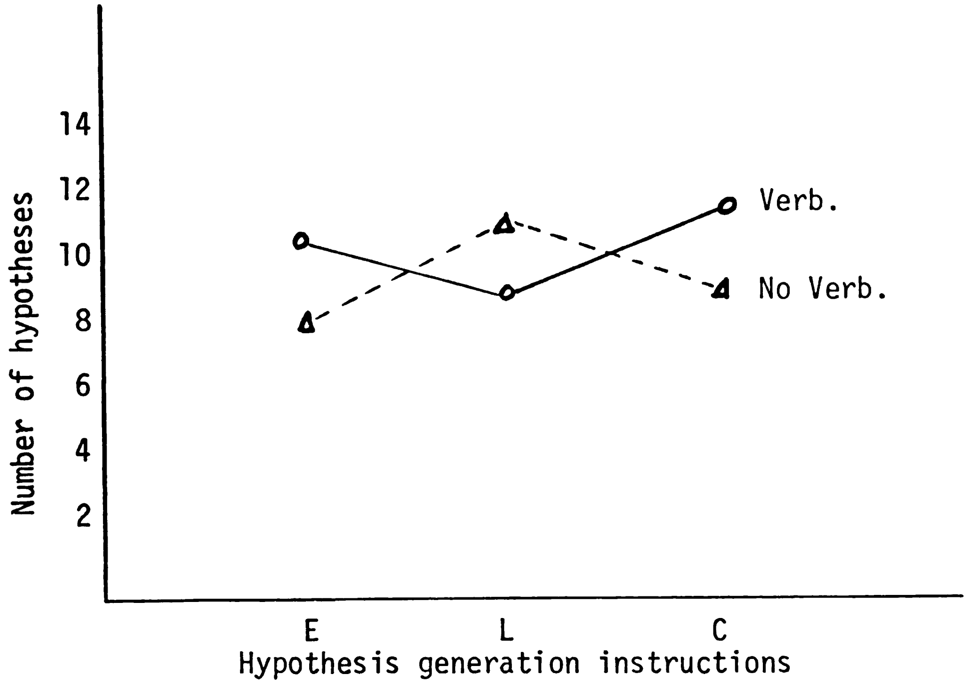

Figure 4: Instructions X Verbalization interaction, Average

Inspection of Table 3 will help clarify why this was not deemed necessary. As stated earlier, early generation was defined as generation of the first hypothesis occurring directly after the reading of the problem introduction. As can be seen from Table 3 this occurred in 75 cases (85% of the time). With this great difference it was felt that a statistical analysis would contribute little. Instructions had no apparent effect on subjects' early or late hypothesis generation behavior. Instructions to generate early produced only one more early generation than instructions to generate late (withold judgment). Subjects in control groups produced one fewer early generation than subjects instructed to withold judgment. Subjects constrained to verbalize produced one fewer early generation than those without that constraint. Instructions to generate late and verbalize produced the fewest early generations, but only one fewer than did instructions to generate early and not verbalize.

It was noted that Problem I produced the most late generations, a total of 10. Problems I and III combined produced only five late generations. A χ^2 Test for homogeneity across problems was conducted to evaluate the significance of this occurrence. The hypothesis of homogeneity should be rejected if $\chi^2 > \chi^2_2$ (.95) = 5.99. A statistic of χ^2_2 = 10.22 was obtained which confirms a lack of homogeneity of hypothesis generation across problems. Problem I is producing significantly more late generations than Problems II and III, further evidence of problem specific differences.

Before pursuing the analysis of hypotheses dealing with the effect of hypothesis generation performance on the dependent variables, a discussion of significant findings obtained thus far is in order.

Table 3
Hypothesis generation X Groups X Problems
(cells contain frequencies)

		Problem	Early generation	Late generation
Verbalize	Generate early	I	5	0
		II	5	0
		III	5	0
	Generate late	I	1	4
		II	5	0
		III	4	1
	Control	I	3	2
		II	5	0
		III	4	1
No Verbalize	Generate early	I	3	2
		II	3	2
		III	5	0
	Generate late	I	5	0
		II	5	0
		III	5	0
	Control	I	3	2
		II	5	0
		III	4	1

Summary And Discussion Of Verbalization And Interaction Effects

Based on results obtained thus far it would appear that instructing subjects in hypothesis generation and constraining them to verbalize has a combined effect on the number of hypotheses generated, but not on the thoroughness, efficiency, or accuracy of performance.

There was a significant effect for verbalization on Problem III only. This effect got its primary contribution from the "early" group where more than twice as many hypotheses were produced by verbalizers as by non-verbalizers (an average of 13.6 vs. 6.0). Control verbalizers produced more hypotheses than non-verbalizers, but late verbalizers produced fewer hypotheses than non-verbalizers. For this reason primary attention must be given to the effect of interactions between hypothesis generation instructions and verbalization on number of hypotheses generated.

It should be noted first of all that the interactions for all but Problem I show a similar pattern in the "early" and control groups. Combining this observation with the results presented in Table 3, one might infer that under instructions to generate hypotheses early, subjects behave as they would naturally. In other words, instructing a subject to generate hypotheses early is simply telling him to do what he would do anyway. It should also be noted that the average performance simply reflects a general trend across the problems. This general trend demonstrates that on the average subjects in the "early" group performed similarly to those in the control group. What needs explanation is why subjects in the "late" group performed differently and what there is about Problem I that induces a different hypothesis generation pattern than Problems II and III.

On the "early-late" dimension "early" instructions led to more hypotheses being produced by verbalizers than by non-verbalizers whereas "late" instructions led to more hypotheses being produced by non-verbalizers than by verbalizers. It is possible that this occurred due to a combination of the subjects' efforts to follow instructions and a change in their perception of the problem once the decision about the problem solution had been reached. Both Problems II and III were of the type for which a single diagnosis was appropriate, i.e. all the data fit together under a single disease rubric. However, these two problems presented some ambiguity at the outset. In each case the patient's presenting complaint did not lead subjects to think of the solution immediately. Most subjects (20 in Problem II, 19 in Problem III) had collected more than half of their data before generating their solution hypothesis. This ambiguity combined with instructions to generate hypotheses and to verbalize may have produced more hypotheses by this group on these problems. On the other hand the non-verbalizers in the "early" group did not enumerate their hypotheses until they had completed the problem, i.e. decided on a solution. In Problem III this was always the correct solution, and subjects were convinced it was correct. The solution (Pernicious anemia) was not an unusual one and there was a good deal of confirmatory data for it. These subjects then, in looking back on the problem may have suppressed many of the hypotheses generated during the problem once the solution became obvious to them. The solution to Problem II was not as obvious, so that although a similar form of suppression probably occurred the difference between verbalizers and non-verbalizers was not as great.

This explanation is supported in part by Miller, Galanter and Pribram's (16) concept of the Image of a problem and a Plan for solving it. Once a problem is solved one's Image of it may change and, retrospectively, the Plan used for solving it may be more congruent with the post-solution Image than with the pre-solution one. Similarly the explanation may be further supported by Newell and Simon's concept of a problem space. While working through the problem the subject may be operating in one space, whereas after solving the problem, the space he constructs for that problem may be different.

One component of the problem spaces -- pre- and post-solution -- is storage of information. Retrospectively it is possible that a greater number of cues appeared to belong in the Pernicious anemia category, therefore no new hypothesis had to be generated to accomodate them. Thus, retrospectively the number of hypotheses needed was fewer than during solution of the problem.

The solution to Problem I is not as clear cut as those for Problems II and III. For this reason the subjects' Image of the problem may not have changed perceptibly after solution. This may have led them to generate more hypotheses retrospectively than during solution. This still does not explain the difference in pattern between "early" and control groups. The interaction of instructions and verbalization on Problem I is not significant. A more significant interaction should be obtained before lengthy steps are taken to explain this phenomenon.

The facts that subjects in the "late" group generated fewer hypotheses during verbalization than they did retrospectively as well as generating fewer hypotheses during verbalization than the "early" group may both be due to the effect of the "late" instructions. This group (see Appendix A, L-V and L-NV instructions) was asked to gather enough data to do a good workup and group the data together into formulations. They were discouraged from leaping to conclusions based on small amounts of data. Subjects may have followed these instructions, not by generating the first hypothesis later in the problem, but by generating fewer hypotheses. Further, during verbalization they gave more evidence of following the instructions than retrospectively.

It is difficult to conceive of a reason why subjects in the "late" group generated more hypotheses retrospectively than subjects in the "early" group. Why, under any circumstances would the "late" group, who had been encouraged to be conservative, generate more hypotheses than the "early" group who had been encouraged to speculate? This phenomenon is particularly noticeable on Problem III. It was noted that subjects in the "early" group on the average generated their final hypothesis during retrospection earlier than those in the "late" group. Specifically the "early" non-verbalizers on the average had produced the solution hypothesis by the time they were almost 50% into the workup, whereas the "late" non-verbalizers did not produce that hypothesis until they were almost 63% into the workup. It may be then that the "late" instructions led subjects to do a more thorough evaluation in retrospect and generate more hypotheses, whereas subjects in the "early" group allowed their knowledge of the final solution to influence their retrospective hypothesis generation causing them to generate fewer hypotheses.

In summary the interactions of instructions and verbalization, noticeable only in Problems II and III are probably due in part to the structure of these problems as well as to subjects' interpretations of the instructions combined with a changed perception of the problem after it has been solved.

Summary And Discussion Of Early/Late Hypothesis Generation

It was noted earlier that except for Problem I, subjects in the "early" generation group behaved similarly to the control group regarding the number of hypotheses generated during verbalization or retrospectively. The statement was made earlier (p. 56) that telling subjects to generate hypotheses early encourages them to do what they would do anyway. This conclusion is further borne out by the results of the effects of instructions on early or late hypothesis generation. The instructions themselves had no apparent differential effect on early or late hypothesis generation. Thus not only does telling a subject to generate hypotheses early encourage him to do what he would do anyway, but also telling him not to generate hypotheses early does not discourage him from doing so. Early hypothesis generation and therefore hypothetico-deductive thinking seem to be an integral part of any diagnostic problem solving strategy. This conclusion derived from the data obtained is so obvious as to seem trivial. To the layman and to modern philosophers of science such as Kessel (7), Popper (10), and Medawar (11), a hypothetico-deductive approach to the solution of scientific problems seems essential. In the medical domain physicians such as Price and Vlahcevic (21) as well as investigators such as Elstein et al. (6) and teachers of medical students such as Morgan and Engel (5) contend that to solve a diagnostic problem hypotheses are generated and evaluated throughout the problem. Although this point is blatantly obvious to some, the present investigator included, it is not always obvious to others. One of those who does not espouse this viewpoint is Lawrence Weed and, by association, those who teach the Problem-Oriented Record (POR) using Weed's approach. This is not a criticism of the POR. As a record keeping device it is excellent, and when implemented in a health care system it provides the uniformity and conciseness missing in non-problem-oriented systems. Further, it allows data to be gathered from patients by non-physicians thus saving the physician time and effort.

If the present findings and those discussed above are indeed true, use of the POR as a device for organizing thought about diagnostic problems goes counter to the problem solver's intuitive approach which starts with the generation of diagnostic hypotheses. Further investigation of the POR and its role in the training of medical students is necessary before this issue can be resolved.

The finding that Problem I produced significantly more late generations again points up the possibility of problem specific differences. Earlier, in the Instructions X Verbalization interaction results, Problem I stood out as having a different pattern than Problems II and III. The reasons behind this may be in a number of areas. One is the structure of the experiment. Problem I was the first problem subjects did and without instructions they were hesitant to generate hypotheses early since many had been trained not to. Another reason may be the structure of the simulation. The introduction to Problem I (see Appendix B, Problem I) presented a comparatively small amount of information proportional to the total amount of data needed to solve the problem. Some sort of critical mass concept may be at work here whereby certain subjects needed more data than was presented at the outset to generate their first hypothesis. This critical mass may have been exceeded in the other two problems. Lastly the structure and/or content of the problem may have induced more subjects to put off generating their first hypothesis until more data had been obtained.

A combination of the above two explanations is plausible. The information presented at the outset is non-definitive enough that the patient's complaints could stem from any number of sources. The problem is presented as an emergency, yet the emergent nature of the situation is not immediately obvious. No subset of elements fits together immediately to suggest a hypothesis that one might pursue. These aspects of Problem I can be contrasted with the introduction to Problem II (see Appendix B, Problem II) for example in which a great deal of information is presented. The subject's primary task is to obtain laboratory tests to evaluate his hypotheses. Problem III (see Appendix B) also contrasts with Problem I in that the introduction to the former presents a female patient with the types of complaints which not infrequently occur in female patients of her age. Thus the range of possible initial hypotheses on Problem III is probably less than that suggested by the introduction to Problem I.

To use Newell and Simon's terminology, it may be more difficult to establish a problem space based on the introduction to Problem I. The concept of problem space has not been defined for medical problems as yet. It is felt that medical problem spaces at the very least consist of classes of hypotheses, aggregations of cues and heuristics for relating cues to hypotheses. Based on the introduction to Problem I it may be impossible for some subjects to establish this space in which to operate. Further investigation is needed into the nature of the problem space for various problems, and the amount of data that is needed to establish a problem space for different types of problems.

Results Of Performance Hypotheses

As could be predicted from the results of the previous hypotheses discussed (Hypotheses 5, 6 and 7), it was only possible to evaluate the effects of Early *vs*. Late and Many *vs*. Few hypotheses for Problem I. Attempts to make these evaluations for Problems II and III resulted in at least one cell with an N of less than 2.

Table 4 shows the results of a multivariate analysis of variance on the performance variables. No overall significant differences were obtained for early or late hypothesis generation, or for the generation of many (10 or more) or few (less than 10) hypotheses. Thus Hypotheses 11 and 12 must be rejected. Further, there was no interaction effect. Univariate analysis showed a significant effect for many *vs*. few hypotheses with subjects generating many hypotheses being more thorough than subjects generating few hypotheses. The interaction between Early/Late and Many/Few was disordinal and significant. This interaction is plotted in Figure 5. Subjects who generated the first hypothesis early and used few hypotheses were somewhat less thorough than those who started early and used many hypotheses. Those who generated the first hypothesis late and used few hypotheses were considerably less

Table 4
Manova of Effects of Many/Few and Early/Late Hypothesis Generation on Three Outcome Variables, Problem I (Acute Abdomen)

	Thoroughness	Efficiency	Accuracy
Cell Means			
Few (<10) Hyp			
Early	37.1	66.2	3.0
Late	26.3	66.0	2.5
Many (≥10) Hyp			
Early	42.0	63.0	2.9
Late	51.3	55.6	3.1
MANOVA			
Effect of Early or Late (F = .98; df = 3, 24; p < .42)			
Effect of Number of Hypotheses (F = 1.99; df = 3, 24; p < .14)			
Univariate			
F(df = 1, 26)	6.19	3.15	.21
P less than	.02	.09	.65
Interaction (F = 1.8; df = 3, 24; p < .17)			
Univariate			
F(df = 1, 26)	4.07	.91	1.45
P less than	.05	.35	.24

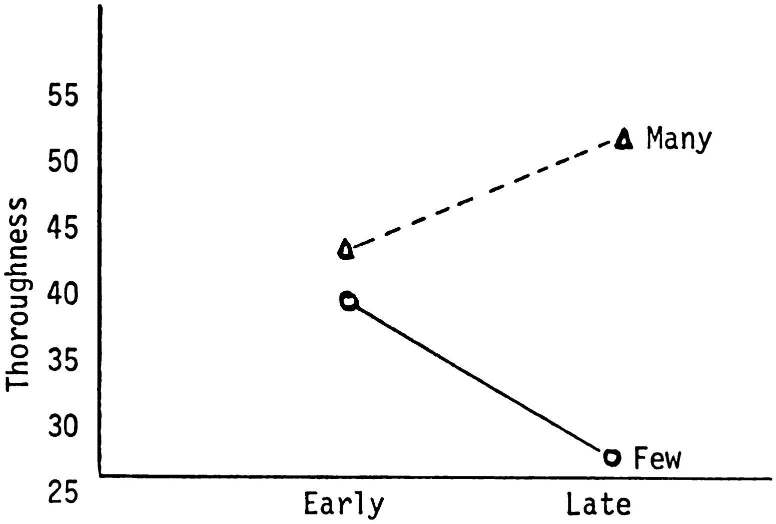

Figure 5: Many/Few X Early/Late interaction, Problem I

thorough than those who started late and used many hypotheses. Those who started early and used few hypotheses were considerably more thorough than those who started late and used few hypotheses.

Summary And Discussion Of Performance Results

The results obtained from the analysis of performance hypotheses are inconclusive. They show generally that thoroughness correlates with number of hypotheses and that there is some kind of combined effect on thoroughness of earliness of hypothesis generation and number of hypotheses used.

It was hoped that first of all there would be enough subjects who generated hypotheses late that this analysis could be done for all three problems. This was not possible. Secondly, it was hoped that the analysis that was done would show some significant differences on the Early/Late dimension. Since early hypothesis generation plays such an important role in diagnostic problem solving, one might think that those subjects who do not generate hypotheses early would go about problem solving differently than those who do generate early hypotheses. Although no effect of early hypothesis generation was observed for any of the outcome variables there was an interaction effect. The possible behavioral implications of this are discussed below. The results of the Many/Few dimension and the interaction were somewhat more rewarding. Generating many hypotheses caused subjects to be more thorough on Problem I. Inspection of the correlation matrix for these data reveals that this phenomenon is somewhat uniform across problems. Number of hypotheses correlates with thoroughness .41 on Problem II and .18 on Problem III.

Generating the first hypothesis early results in a similar amount of thoroughness whether few or many hypotheses are used. However generating the first hypotheses late results in considerably more thoroughness when many hypotheses are used and considerably less thoroughness when few hypotheses are used. Further evaluation of this interaction showed that only 3 of the 10 subjects who generated the first hypothesis late used less than 10 hypotheses. Seven subjects who started late also used many hypotheses. This, it is felt, further supports the earlier comment about Problem I (p. 63). This seems to be a problem for which it is difficult to establish a problem space which leads to a large number of late first hypothesis generations. Subjects who do have trouble establishing a problem space seem to find this problem difficult to work through. For this reason 7 of the 10 late generators use a comparatively large number of hypotheses and do comparatively thorough workups.

It would be most interesting to know whether late generation combined with the generation of a large number of hypotheses and a thorough workup is a problem specific phenomenon. It would also be most interesting to know whether the fact that the three subjects who started late also used few hypotheses and were noticeably less thorough is a chance occurrence. If not one might have to investigate two types of late starters; those who have trouble establishing a problem space on this type of problem and continue to have trouble clarifying it, and those who have trouble establishing one, but once it is established can arrive at a solution with fewer hypotheses and a less thorough workup than those who were able to establish a problem space early. Under those circumstances problem type by subject interactions could be rather involved.

Results And Discussion Of Problem By Problem Analysis

Having observed a number of apparent problem specific differences, a problem-by-problem multivariate analysis of variance was deemed appropriate. Although this procedure could not clarify problem specific differences in accuracy (Problem III appeared to stand out) it might clarify problem specific differences on the other dependent variables. The results of this analysis reaffirmed some of the results found earlier but, except in one possible case did not reveal any overall problem specific differences. As is shown in Table 5, the interaction between instructions and verbalization helped to generally differentiate between groups on Problem I.

There were two significant univariate F values:

Effect of verbalization on Problem III hypotheses
Univariate F = 4.61; df = 1, 24; $p < .04$

Interaction effect on Problem III hypotheses
Univariate F = 3.71; df = 2, 24; $p < .04$

These results tend to confirm some of those found earlier. The finding of differences between groups on Problem III hypotheses due to the effect of verbalization and due to an interaction of verbalization and instructions has been reported and discussed. It would appear that this may not be a problem specific phenomenon sincc when all three problems were analyzed as a group, there was an overall significant effect for interaction whereas when Problem III was separated out, the overall

Table 5
Problem-By-Problem Manova on Dependent Measures

Overall Effects

	Hyp. Gen. Instr.	Verbalization	Interaction
Problem I	F = 1.19; df = 8, 42; p < .33	F = .58; df = 4, 21; p < .68	F = 1.90; df = 8, 42; p < .08
Problem II	F = .49; df = 8, 42; p < .85	F = .33; df = 4, 21; p < .85	F = 1.06; df = 8, 42; p < .41
Problem III	F = .49; df = 6, 44; p < .62	F = 1.55; df = 3, 22; p < .22	F = 1.24; df = 6, 44; p < .30

effect disappeared. In the analysis of the combined problems, Problem II probably made a contribution to the interaction effect on the hypotheses variable.

The finding of an interaction effect on Problem I is a new occurrence. Inspection of a step down F shows that once the variance contributed by hypotheses is removed, thoroughness is still significant. Thus the differences in number of hypotheses generated and the differences in thoroughness are making primary contributions to the interaction. These two interactions are plotted in Figures 2 and 6. Figures 7 and 8 show the contributions of efficiency and accuracy to that interaction. It is felt that before going to great lengths to explain the interaction, it should be further investigated to assure that it did not occur by chance and to ascertain exactly what factors might be at work here.

Results And Discussion Of Process Analysis

As was stated earlier, both a statistical and a clinical analysis were done on the results of this study. This clinical or process analysis, as will be seen below, is not of the type that tries to describe the sequential strategy used by certain subjects to solve a problem, such as Simon and Newell (17), for example, have done for their cryptarithmetic tasks. The process analysis done here was simply an effort to extract certain elements from the students' performances after the fact, and to try to make inferences from these about the differences between high scorers and low scorers.

The process analysis was felt to be highly justified since the statistical analysis revealed certain possible problem specific

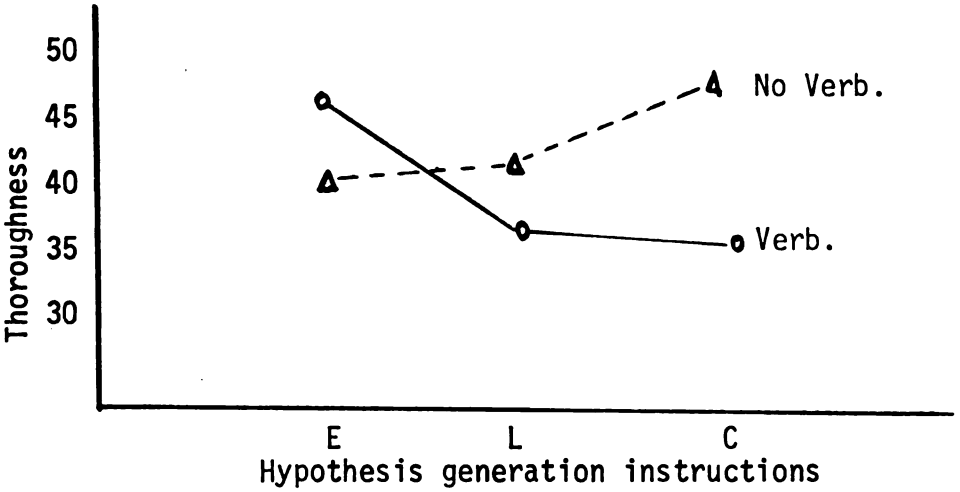

Figure 6: Instructions X Verbalization interaction, Problem I

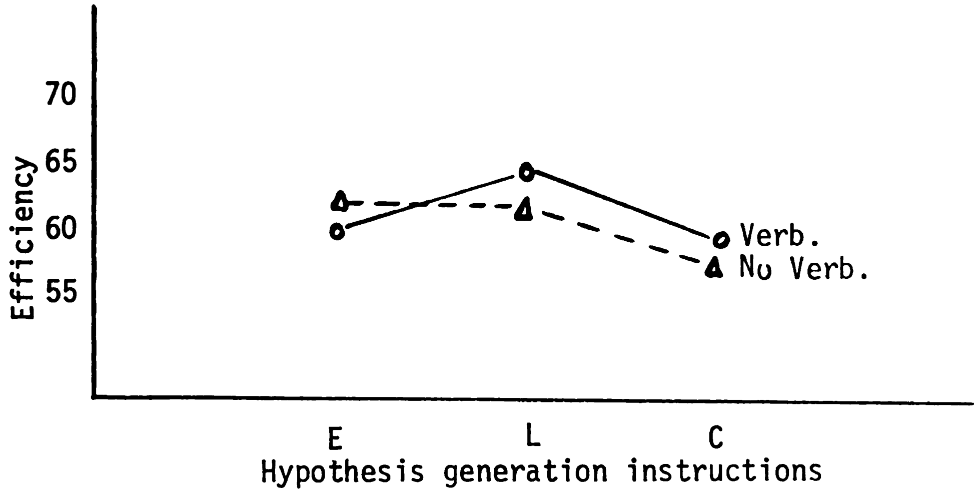

Figure 7: Instructions X Verbalization interaction, Problem I

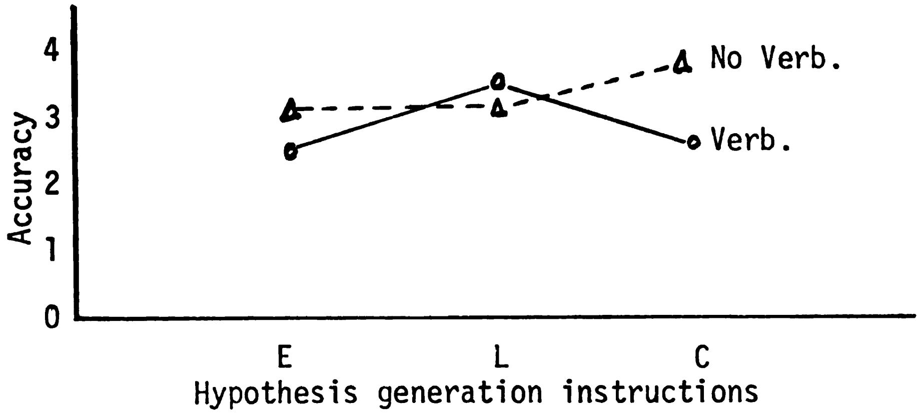

Figure 8: Instructions X Verbalization interaction, Problem I

differences but did not clarify them particularly. Further, the statistical analysis was not able to discriminate groups on the accuracy of their performance. Since accuracy is an important aspect of medical problem solving the process analysis tries to differentiate between high and low scorers on accuracy. The analysis was done on Problems I and II only since there were no low scorers on Problem III. This fact points up a potential problem specific difference. Problem I had five low scorers, Problem II had 20 low scorers and Problem III had none. Furthermore, Problem II was the only problem on which any subject received a score of "0". This may be due to the content and/or structure of the problem, or it may be the result of more stringent scoring being unknowingly applied. The criteria for assigning scores, as discussed earlier was derived from a criterion developed by experts. Thus the suggestion that it was somehow a more stringent or less fair scoring scheme is probably unjustified. The imbalanced distribution of accuracy on the three problems suggests that before simulations are more widely used -- especially low fidelity simulations such as these -- scoring schemes should be established which in some way assure that solutions with parallel degrees of accuracy receive equivalent scores.

A number of new elements were found which seem to point up differences between those subjects who received high and low scores. As can be seen in Table 6 these elements deal with the acquisition of cues which were positive for a high scoring (called generally "correct" in Table 6) solution. The high scoring solutions are shown in Display I. The elements also deal with the interpretation of these cues. The criterion used for determining which cues were positive for the correct

Table 6
Data for Process Analysis, Problem I (Surgical Abdomen)

	High Scorers (n = 25)	Low Scorers (n = 5)
Cue acquisition		
Thoroughness	43.7	37.4
Efficiency	61.1	66.4
% of cues + for correct hyp.	41.8	40.1
*Avg. # of cues per subject + for correct hyp.	24.7	20.5
Hypothesis generation		
# of hypotheses	10.5	11.0
*Avg. # of cues + for correct hyp. obtained		
before generation	16.8 (n=10)	9.2
after generation	.2 (n=10)	7.6
*Avg. # of cues + for solution hyp. obtained		
before generation		6.5 (n=4)
after generation		12.5 (n=4)
Early:Late	7:3	4:1
Many:Few	5:5	2:3
Cue utilization		
Avg. percent of cue misinterpreted	25	21

*GI hypothesis only

solutions is one developed at Michigan State University (30). Briefly, the criterion consists of a grid which relates cues to hypotheses. The cells of the grid contain weights from -3 to +3. A positive weight means that cue tends to confirm that hypothesis; a negative weight means that cue tends to disconfirm or rule out that hypothesis; and a 0 weight means that cue is non-contributory to that hypothesis. The grids used for the process analysis of Problems I and II are contained in Appendix D.

Process analysis of Problem I

The variables used in the process analysis of this problem are shown in Table 6. Thoroughness and efficiency are as defined earlier using the weighting system developed at the University of Illinois. The percent of cues positive for the correct (high scoring) hypothesis answers the question: overall of the diagnostic (as opposed to intervention) cues obtained by subjects, what percentage of them was positive for the correct hypothesis? This element gives a feeling for how closely the subjects' problem solving centered around cues positive for a high-scoring solution. It should be noted that calculation of the average numbers of cues (per subject) positive for correct and other outcome solutions was calculated for the GI hypothesis only. This figure was calculated on a sample of 10 high scorers in the one case. In the other case four of the five low scorers gave solutions of pancreatitis and/or cholecystitis rather than solutions involving ulcer and complications. The latter average is for the number of cues positive for those outcome solutions obtained by subjects before and after generation of the hypothesis.

An earlier discussion centered around the different thoroughness scores obtained by subjects who generated the first hypothesis late and who used many or few hypotheses. The discussion proposed that those who generated the first hypotheses late were having trouble establishing a problem space for this problem and since 7 of the 10 subjects also used many hypotheses perhaps these subjects continued to have trouble. It was thought that this possible difficulty establishing a problem space might have an effect on the accuracy of subjects' performance. As shown in Table 6 apparently this is not the case. High and low scorers are equivalently unequally divided on the early/late dimension, and equivalently equally divided on the many/few dimension.

Table 6 shows that in general subjects who received high scores were more thorough and less efficient than those who received low scores. The statistical analysis bears this out to some degree. A correlation of .29 was obtained between thoroughness and accuracy, and a correlation of -.13 was obtained between efficiency and accuracy on this problem. These correlations are not significantly different from zero and thus do not point to clear differences between the two groups. The same is true of the percent of cues positive for a correct hypothesis and the average number of cues per subject positive for a correct hypothesis. In sum, the differential ability to arrive at a correct solution is probably not due to any differential cue acquisition abilities. A similar conclusion may be drawn about cue utilization. In fact high scorers made slightly more errors in their cue interpretation than did low scorers.

To analyze the elements of hypothesis generation the problem was separated into its two components: Diabetic ketoacidosis and ulcer with complication. Concerning the diabetic ketoacidosis hypothesis, low scorers either did not generate a diabetes hypothesis at all (there was only one case of this) or did not include both diabetes and acidosis or ketoacidosis in their solution. Three of these subjects generated either diabetic ketoacidosis, or diabetes and acidosis or ketoacidosis separately, but did not include both in their solution. This was probably either due to negligence in recording the solution or because they did not perceive the fact that diabetes and acidosis were two parts of a separate solution component and should be combined. One subject generated the diabetes portion but not the acidosis portion although he obtained three cues that were strongly positive for it. The one subject who did not generate a diabetic hypothesis at all obtained four cues which were strongly positive for diabetes and/or ketoacidosis. One of these cues was correctly interpreted relative to another hypothesis (elevated blood sugar was seen as positive for pancreatitis) but was not used to generate a diabetic hypothesis. This subject may have been unwilling to deal with the ambiguity of a two solution problem and was able to interpret cues only relevant to one type of hypothesis, i.e. GI hypotheses.

Concerning the GI hypothesis, it should first be noted that all low scorers generated a GI hypothesis which would have received a score of 3 had it been retained as the final solution. These subjects apparently dropped that hypothesis before the end of the problem. It has been suggested by another analysis of this problem (31) that failure to retain this hypothesis could be due to errors in cue interpretation. However, low scorers on this problem did not do particularly poorly at cue interpretation. All but one subject accurately interpreted at least half of the cues they used. Two of these subjects had no inaccurate interpretations for a correct hypothesis. One subject accurately interpreted 10 out of 21 cues and inaccurately interpreted 7 out of 21 cues for a correct hypothesis. Generally, then it does not seem that inaccurate cue interpretation underlies rejection of a correct hypothesis combined with acceptance of an incorrect one.

It should be observed that because of the structure and content of this problem not only are there two correct solutions (diabetic ketoacidosis and a GI solution) but a correct GI solution is complex. As can be seen from Display I (pp. 139, 140) a score of 3 is given to solutions which involve 1) ulcer + localization, 2) obstruction + localization, or 3) ulcer + obstruction. Three low scorers generated an ulcer hypothesis as well as an obstruction hypothesis but did not ever combine the two. Failure to do this may in some way be due to an inability to combine individual elements of a complex hypothesis. This, however, does not explain why the other two subjects who generated ulcer + localization hypotheses failed to retain them.

The ability to combine elements of a complex hypothesis and to use an appropriate hypothesis generation and cue utilization pattern may contribute to being successful on this problem. As is shown in Table 6 high scorers did not generate the accurate hypothesis which they used as their solution until they had obtained an average of 16.8 cues which were positive for that hypothesis. Furthermore, they obtained an average of only .2 cues positive for that hypothesis after its generation. This means that most high scorers generated the accurate hypothesis which they used as their solution very near the end of their data gathering. In contrast low scorers generated a correct hypothesis (or the components of a correct hypothesis) much earlier in their data gathering. To establish a parallel between judgments about the generation of components of a correct hypothesis, high scorers' performances were analyzed to determine by what point they had generated the essential components of the accurate hypothesis they used as their solution. It was found that 6 of the 10 subjects did not generate these essential components until they had obtained _all_ of the cues positive for that hypothesis. Two subjects generated the essential components before obtaining any positive cues and then used the positive cues to refine (increase the specificity of) the hypothesis. Of the remaining two high scorers, one generated the essential components after obtaining 7 out of 21 positive cues. His final solution was a refinement of the essential elements. The other subject behaved much like the low scorers in that he generated the elements of the solution after obtaining 6 out of 13 positive cues and did not refine the solution beyond that point. He differed from the low scorers in that he retained that solution whereas they did not.

These analyses seem to show that high scorers are able to perceive that this is a complex problem. They perceive that not only are there two accurate outcomes but one of them is complex in itself. They tend to respond then, particularly vis-a-vis the GI hypothesis, by either waiting until they have gathered a large number of cues before making a decision about a solution or by generating the general elements of the solution very early and using data subsequently obtained to refine the solution. What distinguishes these subjects from the low scorers is primarily the delay in decision about a correct solution. Rather than dropping any essential elements generated along the way in favor of a less complex hypothesis as do low scorers, high scorers retain essential elements and use data to generate a more complex solution. A simple analysis of accuracy of cue interpretation does not distinguish these groups from one another. Therefore the ability to generate and build on a potentially complex hypothesis may stem from an ability to deal with ambiguity combined with an ability to accurately use incoming information to generate accurate alternatives and reject inaccurate ones.

Subjects' behavior on this problem appears to be a classic case of the use and misuse of the strategy outlined by Price and Vlahcevic (21). Low scorers are not able to use cues to rule in an accurate hypotheses, possibly because they are not able to generate, retain, and relate the elements of a complex hypothesis in a complex problem; and they are not able to rule out a less complex but inaccurate hypothesis. High scorers are able to use cues to establish and/or refine a complex hypothesis whether they decide early or late on its components. They are also able to use cues to rule out inaccurate hypotheses. More generally, in Miller, Galanter, and Pribram's (16) terms, high scorers seem to have an accurate Image of the problem and are able to select an appropriate Plan for solving it. They are aware of the essential elements of the problem as well as of the solution. They are also aware of the boundary conditions for its solution, i.e. that it is probably a complex problem and a good deal of information needs to be gathered before deciding on a solution.

In contrast, poor scorers although apparently aware of the essential elements of the problem may not select a good Plan for its solution due to a lack of awareness of the boundary conditions.

Process analysis of Problem II

The variables used in the process analysis of this problem are shown in Table 7. Certain of the variables are self-explanatory. Those deserving further definition are explained below:

Thoroughness and Efficiency, as defined earlier, using University of Illinois weighting system

Percent of cues positive for a correct (high scoring) hypothesis defined as the average percent of cues obtained by subjects which were positive for a correct hypothesis

Hi = subjects receiving high scores

Lo = subjects who received low scores and did not generate a correct hypothesis

Lo & Drop = subjects who generated a correct hypothesis but dropped it and received low scores

Table 7 shows that subjects who received high scores differed from those with low scores primarily in the areas of cue acquisition and hypothesis generation (the number as well as type of hypotheses generated). Cue utilization did not appear to differentiate among these subjects.

Subjects who received high scores were generally more efficient in their data gathering and particularly more so in their gathering of cues positive for a correct hypothesis. However they did not gather

Table 7
Data for Process Analysis, Problem II (Pale Lethargic Child)

	High scorers (n=10)	Low scorers (n=16)	Low scorers who dropped correct hypothesis (n=4)
Cue acquisition			
Thoroughness	47.6	50.9	
Efficiency	73.9	59.1	
% of cues + for correct hyp.	41.9	29.0	
Avg. # of cues per subject + for correct hyp.	6.4 (4-8)	6.1 (3-8)	
Hypothesis generation			
# of hypotheses	6.7 (3-11)	9.2 (3-15)	
% of workup where final hyp. was generated	68.1	50.9	
Avg. # of cues + for correct hyp. obtained			
before generation	4.3		.8
after generation	2.2		5.0
Cue utilization			
Number of subjects who misinterpreted any cues	4(out of 10)	9(out of 16)	3(out of 4)
Number of cues misinterpreted for correct hypothesis	0		1

any more positive cues than did subjects with low scores. This seems to indicate that high scoring subjects were better able to establish a problem space for this problem. Contrary to the findings for Problem I discussed earlier, subjects did not differ concerning the time at which the first hypothesis was generated. Only two subjects generated the first hypothesis late. One received a high score, the other received a low score. The establishment of a problem space for this problem appears to hinge in part on the knowledge of what cues must be collected to solve the problem and what cues are not necessary for its solution. High-scoring subjects were better able to focus their cue acquisition on those cues which were helpful in arriving at a correct solution.

The ability to use a focused problem solving strategy is further borne out by a crude analysis of the types of alternative hypotheses generated by subjects in the various groups. These hypotheses are listed in Display II (p. 144). The list is accompanied by the frequency with which each hypothesis was entertained by each group. One notes that almost all subjects in all three groups entertained the hypothesis of sickle cell anemia and a large number of these subjects entertained the hypothesis of G6PD deficiency. These hypotheses were strongly suggested by the introduction to the problem. However there is an apparent difference between the other types of hypotheses entertained by high and low scorers. High scorers entertained primarily the alternative hypotheses of hemolytic anemia, autoimmune and iron deficiency anemia. Leukemia was considered by five high scorers. On the other hand, low scorers primarily entertained alternative hypotheses of blood loss anemia, some sort of hereditary cell problem, hypersplenism, bone marrow repression, leukemia, and lead poisoning. Although there was a good deal of overlap -- for example in the consideration of leukemia and autoimmune anemia -- there was also a good deal of divergence. For example, blood loss anemia was entertained by four low scorers and only one high scorer, hypersplenism by five low scorers and no high scorers, and lead poisoning by 10 low scorers and only one high scorer. Furthermore, a greater number and a wider variety of hypotheses were considered by low than by high scorers.

Clearer definition of the role played by the alternative hypotheses in diagnostic problem solving is definitely needed before any conclusions can be drawn about the implications of these hypothesis generation patterns. For the time being it appears that subjects who obtained high scores on this problem were able to focus their cue acquisition as well as their hypothesis generation strategies better than low scorers.

Concerning the number of hypotheses generated (see Table 7), one notes that subjects who received high scores generated fewer hypotheses than low scorers. The statistical analysis bears this out. A correlation of -.41 was obtained between number of hypotheses and accuracy on Problem II. This correlation is significantly different from zero ($p < .05$). Not only did high scorers generate fewer hypotheses than low scorers, they generated their final hypothesis later and based it on more supportive cues. For 8 out of the 10 high scorers the last hypothesis generated was a correct one and the high scorers had obtained an average of 4.3 cues positive for that hypothesis before generating it. They obtained an average of only 2.2 positive cues after its generation. This may be contrasted with low scoring subjects who generated but dropped a correct hypothesis. This group generated the correct hypothesis quite early (an average of 8% into the workup) and had very little supportive data (an average of .8 cues) when they did so.

These results indicate that for this problem, generation of the correct hypothesis should come rather late in the problem if one is going to obtain a correct outcome. Although this indication does not go contrary to the findings of Elstein et al. (6), their findings may need some clarification. They noted that subjects generate diagnostic hypotheses early, but did not differentiate between the generation of a correct hypothesis and other incorrect alternatives. For this particular problem, although hypothesis generation occurs early for almost all subjects, generation of the last and correct hypothesis occurred later for high scorers than for low scorers.

An attempt was made to determine whether the discovery of a particular cue triggered the generation of the correct hypothesis and whether this cue by chance was obtained late in the problem. Attention was focused on the osmotic fragility test (cue #16), a test which is supposed to confirm the presence of spherocytosis. Although it appears that subjects related this cue to the correct hypothesis, one cannot conclude that obtaining that cue triggered generation of the hypothesis. In five instances cue #16 was obtained one, two, or three cues before hypothesis generation, and in five cases this cue was obtained one, two, or three cues after hypothesis generation.

In conclusion, these results show that for this problem, subjects who obtain high accuracy scores are able to use a focused problem solving strategy in the sense that they are apparently able to sense the problem and use a small number of steps (hypotheses as well as cues) to arrive at the correct solution. Although they generate hypotheses early, they need fewer and a smaller variety of hypotheses than low scorers. In addition the high scorers are able to gather a higher proportion of cues positive for the correct hypothesis than low scorers.

An additional somewhat anomalous aspect of the high scorers' strategy is that they do not generate the final and accurate solution to the problem until rather late, and until they have obtained a reasonably large number of cues positive for that hypothesis. But no one single cue seems to trigger generation of the hypothesis. One cannot conclude from this, however, that these subjects are using a strategy which involves the elimination of all alternative hypotheses before generation of the accurate one. The subjects' patterns of positive and negative cue associations indicates a strategy which combines elimination and confirmation of hypotheses.

In contrast, some subjects who receive low scores are not able to generate an accurate hypothesis even though they 1) generate more and a greater variety of alternatives than high scorers, 2) are generally less efficient than high scorers, and 3) obtain the same number of cues positive for the correct hypothesis as high scorers. Other low scorers are able to generate a correct hypothesis but do not retain it. The latter subjects generate the correct hypothesis early in the problem having obtained less than one cue which is positive for it.

The concept of "focused strategy" is as yet only grossly defined in the diagnostic problem solving context. It differs from the conservative focusing strategy discussed by Bruner and colleagues (15) in that it involves both the generation of alternative hypotheses and the acquisition of cues. This concept seems to fit more closely with Simon and Newell's (17) definition of a problem space and with Miller, Galanter and Pribram's (16) concept of the Image of a problem. High scorers seem to have a more closely defined or smaller space for this problem than low scorers. They seem to be using heuristics for hypothesis generation and for the relation of cues to hypotheses that differ from those used by low scorers. Possible strategies have been suggested such as reasoning from general to specific (Kleinmuntz, 20) or honing down involving the collection of more and more specific cues (Dudley, 22). The high scorers' hypothesis generation and cue association patterns do not suggest that they are using these strategies. Although a careful analysis of these patterns was not done, the hypotheses generated by high scorers seem to vary in specificity throughout the problem. Furthermore, the high scorers' hypothesis generation and cue association patterns do not differ noticeably from those used by low scorers. A possible strategy used by high scorers may involve the elimination of certain hypotheses combined with the confirmation of others as suggested by Price and Vlahcevic. A more careful analysis of the subjects' cue association patterns would have to be done to establish this fact.

Summary And Discussion Of Process Analyses

The process analyses conducted on Problems I and II were able to identify aspects of diagnostic problem solving which differentiated between subjects who received high scores and those who received low scores. These aspects were at once similar and different for each problem. Relevant to Problem I differentiating elements such as the ability to perceive and deal with a complex problem, and to delay final decision about a solution until a comparatively large number of cues had been gathered. Elements which differentiated high scorers from low scorers on Problem II were related to the ability to delineate a closely circumscribed problem space and to put off deciding on a solution until a large amount of information positive for that solution had been acquired.

Two brief points should be made relevant to these process analyses. First, although it is important to know about elements of problem solving which differentiate high scorers, one cannot infer that people who possess these abilities will do well on problems for which they are appropriate. Much investigation is needed before conclusions about the abilities of subjects who score high on these problems can be turned into conclusions about the performance of subjects who possess these abilities.

Secondly, although it has not been mentioned up to now, the role played by knowledge and experience in solving these problems cannot be overlooked. Thus far it has been suggested that subjects did well or poorly on these problems because of certain general, though ill-defined abilities they possessed. These apparent abilities may be the result of increased knowledge of that type of problem and/or of that content area. Similarly, subjects who received high scores may have had more and/or more recent experience with that type of problem. Again before conclusions are made about the performance of subjects who possess some of the abilities discussed here, the relation of these abilities to knowledge and experience must be clarified.

CHAPTER V

SUMMARY, CONCLUSIONS, AND IMPLICATIONS

Fourth year medical students completed three modified Patient Management Problems. Subjects were randomly assigned to six groups in a two by three design. On one dimension of the design, instructions concerning hypothesis generation were manipulated. Subjects were encouraged to generate diagnostic hypotheses early in the problem, to withold judgment about diagnostic hypotheses, or were given no instructions about hypothesis generation. On the other dimension subjects were either constrained to think aloud during problem solving or were asked to discuss the problem after they had solved it. Their performance on each problem was scored for thoroughness and efficiency of cue acquisition as well as for accuracy of outcome, and number of hypotheses generated.

Before analyzing the results of the study, efforts were made to estimate the internal consistency and concurrent validity of the modified Patient Management Problems. These efforts revealed that two of the three problems were internally consistent and one (Pale Lethargic child) was not. Furthermore, students' thoroughness and efficiency scores on one of the problems more closely resembled physicians' scores on the original PMP than those on the high fidelity simulation. Students' scores on the other problem showed the opposite trend, resembling more closely physicians' scores on the high fidelity simulation than their scores on the original PMP.

After effect of instructions on number of hypotheses, thoroughness, efficiency, and accuracy had been determined, each subject was reassigned in a two by two matrix according to whether he had generated comparatively many or few hypotheses, and whether he had generated his first hypothesis comparatively early or late in the problem. The effect of these two variables on thoroughness, efficiency and accuracy was analyzed. Lastly, a process analysis, relating certain aspects of performance to outcome, was done on two problems.

Results were as follows:

1. Instructions had no effect on thoroughness, efficiency, accuracy, or number of hypothesis generated.
2. For one problem the constraint to verbalize prompted subjects to produce significantly ($p < .04$) more hypotheses than the absence of that constraint.
3. For that same problem the interaction of instructions and constraint to verbalize had a significant effect on the number of hypotheses generated ($p < .04$).
4. Regardless of instructions early hypothesis generation occurred 85% of the time.
5. One problem prompted significantly ($p < .05$) more late hypothesis generations than did the other two.
6. On one problem subjects who generated comparatively many hypotheses were significantly more thorough in their acquisition of cues than those who generated comparatively few hypotheses ($p < .02$).

7. For this same problem the interaction between generating comparatively many or few hypotheses and generating the first hypotheses early or late had a significant effect on thoroughness of cue acquisition.

8. Obtaining a high score on one problem (Acute Abdomen) is associated with the ability to generate and retain the elements of a complex solution.

9. Obtaining a high score on another problem (Pale Lethargic Child) is associated with the ability to conduct a focused inquiry by gathering a small number of high yield cues, and generating a relatively small number of diagnostic hypotheses.

10. On both the Pale Lethargic Child problem and the Acute Abdomen problem obtaining a high score is associated with postponing generation of the solution until a relatively large number of cues positive for it have been gathered.

Conclusions

1. Concerning the Patient Management Problems as an instrument: Due to the variability of performance across problems (discussed in conclusion #3), the Patient Management Problems should not be judged as a general technique. Each problem presents a different situation to the problem solver. At this time it appears that success on one Patient Management Problem (PMP) is determined by different factors than is success on another PMP. For that reason judgments about the validity and reliability of PMP's should be made about

individual problems rather than about PMP's as representing a general measurement technique.

a. Validity of PMP's: The two most recent estimates of the validity of PMP's are that done for the present study and that done by Goran et al. (41). The present study found that physicians performed differently on the two PMP's which were analyzed than on the two high-fidelity simulations with which they were compared. The differences may have occurred because the PMP situations and the high-fidelity situations were not analogous to each other. The study done by Goran et al. (41) also found that physicians and fourth year medical students performed differently on a real patient (observed via chart audit) than they did on an analogous PMP. These differences may be due in part to the lack of thoroughness of the charts which were audited.

   Studies of validity of PMP's have thus far demonstrated that the validity of these tests is difficult to ascertain and that based on the results of these experiments the validity of the PMP's studied is questionable.

b. Reliability of PMP's: The only forms of reliability which have been estimated for PMP's is the consistency of scoring systems and the internal consistency of problems. The latter was estimated in the present study. Although the problems sampled appear to be internally consistent, this measure of reliability is weak. A measure of test-retest reliability or of generalizability of responses is needed before any conclusions can be drawn about the reliability of any PMP.

2. Concerning the methodology of the present study

   a. Instructions to the subject can play a number of roles in an experiment (see Sutcliffe [45] for a discussion of the role of instructions in psychological experiments). Instructions are usually used to simply acquaint the subject with his task. The clearer these instructions are, the more predictable is the effect they will have on the outcome of the experiment. The instructions concerning hypothesis generation used in this study went beyond subject orientation to exhort the subject to approach a problem in a certain way. The specific effect of these instructions is not known. The results of the study do indicate that the instructions had no measurable effect on any of the dependent variables. It can be concluded from these results that an experimenter should never assume that a certain set of exhortations is going to have the desired effect on a group of subjects.

   b. Verbalization during problem solving provides the experimenter with varied and reliable information about a subject's thought processes without apparently interfering with those processes. On the other hand, at least for certain problems, verbalization about thought processes after the problem has been solved can introduce retrospective distortion into the experiment. For some problems the information about processes given retrospectively may be unreliable.

3. Concerning the results of the study:

   a. Early hypothesis generation is an integral part of diagnostic problem solving. Moreover, how early in a problem hypotheses are generated is probably determined more by the problem and the way in which it is perceived by the problem solver than it is by training or exhortations from his teachers. Although early hypothesis generation has no apparent relation to the ability to gather data about a case or to the ability to accurately solve a problem it is an essential element of clinical inquiry.

   b. One of the strongest influences on how a subject approaches a diagnostic problem is the nature of the problem itself. The amount of information presented at the outset, whether or not the situation is emergent, and how complex the solution is are only a few of the elements which combine to influence a subject's performance. There is great variability across the three problems used in this study. One presents a large amount of information at the outset and leaves few options for the subject to choose; the other two present less initial information and contain more options. One has a complex solution, the other two have simple solutions (i.e. the patient has only one disease). Of the last two one is apparently quite easy since all subjects arrive at the correct solution, the other apparently quite difficult since the fewest subjects solve it. Lastly different behaviors seem to characterize successful solvers of one problem than characterize successful solvers of another. Diagnostic problems are often differentiated based on organ system and clinical specialty area. These two factors are not sufficient to categorize problems. The other elements discussed above play an equally if not more important role in the characterization of clinical problems than do organ system and clinical specialty.

c. Although no conclusions can be drawn about the kinds of abilities which lead to success on Patient Management Problems, tentative conclusions can be made concerning certain abilities associated with success (arriving at an accurate solution) on the two problems which were clinically analyzed. In both cases successful problem solvers did not generate the hypothesis they used as a solution until near the end of the problem after a large number of cues had been acquired. This seems to imply that being able to delay arriving at a decision about a solution until a large number of cues are known, and being able to rule out other hypotheses generated early in the problem may be helpful in reaching an accurate solution.

Implications

1. Concerning the Patient Management Problems as an instrument:

a. Patient Management Problems, particularly the more complex variety such as the Acute Abdomen and the Pale Lethargic Child are an attractive way of simulating the clinical setting. They involve sequential information gathering such as is practiced when working up a patient and they give the problem solver the opportunity to use thought processes similar to those he would use in working up a real patient. However it has been demonstrated on a number of occasions that problem solvers perform differently on Patient Management Problems than they do on real patients or even on higher fidelity clinical simulations. Until this descrepancy is corrected, use of Patient Management Problems as a replacement for a clinical encounter is unacceptable. In this writer's opinion wide spread use of Patient Management Problems for the evaluation of general clinical competence is inappropriate until the validity and reliability of the problems is more clearly established.

b. The reliability of the problems could be established using a test-retest technique or by clearly establishing the universe of skills to which performance on a PMP can be generalized. The former approach is straight-forward if done in the appropriate context. The two problems used must be essentially identical since there are so many hidden variables which can cause apparently parallel problems to differ. Time lag between the administration of the two problems must be great enough so that performance on the one will minimally influence performance on the other. Further, the subjects used for the test should be at a stage in their career where a large amount of learning is no longer taking place. In this manner the time lag between tests will not be accompanied by a noticeable change in the subject caused by new experiences and acquisition of large amounts of new knowledge. The establishment of a universe of generalizability is a more complex task which is being dealt with elsewhere (46) and will not be pursued here.

2. Concerning the methodology:

   a. The ineffectiveness of instructions as they were used in the present study should serve as a caveat for future investigators of problem solving process. Instructions should be clear. If possible they should go no further than to acquaint the student with his task. It is very difficult to ascertain a subject's response to exhortations. Instructions should not require a subject to do something in a given situation which conflicts with what he usually does in that situation unless a specific goal of the experiment is to create such conflict. In the present study for example, instructions to withold judgment in hypothesis generation conflicted with the subjects' apparent natural propensity to generate hypotheses early. Although the conflict did not seem to hamper the subjects' performance it did render instructions useless as an experimental variable. Lastly, subjects should be constrained to give a behavioral manifestation of the fact that they are following instructions. In the present study the verbalization instructions were successful at least in part because subjects were not only instructed but constrained to verbalize during or after problem solving.

b. Verbalization, particularly thinking aloud has a bright future as an experimental as well as training tool. Experimentally, thinking aloud can give the experimenter information about the logic and sequence with which the problem is progressing for the subject. Thinking aloud can also help the experimenter to identify factors which may influence problem solving. In the medical context what hypotheses a subject is entertaining at any one time; how new cues are related to those hypotheses; what goal led the subject to seek a specific item of information; and what, if any, informal decision rules the subject is using are all items of information which can be obtained by asking the subject to verbalize during or after problem solving. The present writer feels that verbalization during problem solving, i.e. thinking aloud not only yields more reliable information about the subjects' thought processes but does not noticeably interfere with those processes.

The use of thinking aloud as a training tool allows the teacher or other members of a peer instruction group to monitor the problem solver's progress. Information from the problem solver about his progress can help the monitor(s) to give supportive or corrective feedback where appropriate. This especially helps the problem solver avoid compounding his errors by constantly recycling him onto the main track of the problem.

3. Concerning the results of the study:

   a. The extremely important role that early hypothesis generation plays in diagnostic problem solving makes it an essential component of medical training and evaluation programs. Traditionally early hypothesis generation is not emphasized in medical training. As stated earlier (page 4, 5) certain training and record keeping procedures (specifically Weed's Problem Oriented Record (3, 4) discourage that activity. Others (e.g. Morgan and Engel [5]) do not discourage that activity but do not advocate that the acquisition of skill in early hypotheses generation be an important aspect of a physician's training. It may be that this activity is so fundamental to any kind of problem solving that it need not be specifically taught to medical students. Whether or not specific training in that skill is needed, a medical student's possession of that skill should be assured before he takes responsibility for the care of patients. The possession of the skill should be included in the objectives of any medical curriculum and all students should be evaluated on it.

The fundamental role of early hypotheses generation in medical problem solving has implications for use of the Problem Oriented Record (POR) as a training and record keeping tool. It will be some time before the technology is available for the automation of patient data acquisition and storage. For this reason among others, data gathering and data interpretation skills will continue to be an important part of a physician's training. If early hypothesis generation is indeed as unusual as it appears to be then the POR format which discourages this activity might create the type of conflict described earlier if used as a training tool. The POR has established its value as a record keeping tool but how useful is it as a training device? A simple bit of applied research could probably answer this question. A group of first year medical students might be started out with training in the use of the POR for record keeping and training in the generation and evaluation of hypotheses to guide their thinking. One would want to keep track of the kinds of conflicts which arose in both groups, what the sources of confusion were and how positive the students attitudes were and award each training program. At the end of the second year, each student could be evaluated on how often and in what manner he used the POR as well as be given a simulated problem to solve. During the problem he would be asked to think aloud which would tell the experimenter whether he was thinking hypothetico-deductively or like a POR.

b. The content and structure of any simulated diagnostic problem a subject solves have a strong influence on his problem solving behaviors. This observation has strong implications for anyone wishing to use diagnostic problem simulations in an experiment, or in a training and/or evaluation program.

The knowledge that different people solve different problems differently should discourage an experimenter or teacher/evaluator from trying to use the results of a performance on one or several diagnostic problems to make judgments or predictions about general clinical competence. Even if a subject could be given a "score" for clinical competence this score would not be justifiable if based only on solutions of a few diagnostic problems.

Particularly in a problem oriented curriculum, attempts should be made to identify aspects of problems used for training which care be associated with different problem solving approaches. Which problems tend to converge on a single solution? Which tend to be ambiguous or inconclusive? Can it easily be determined which problems are relatively easy (i.e. the solution is rather obvious from the outset) and which problems are more difficult (e.g. a special type of knowledge is needed to solve them)?

Which problems are common, which are uncommon? If these elements are identified, then problems can be found or created which best address certain problem-solving objectives in a medical curriculum.

BIBLIOGRAPHY

BIBLIOGRAPHY

1. Dorland's Illustrated Medical Distionary, 24th Edition, Philadelphia and London, W.B. Saunders Co., 1965

2. Webster's Seventh New Collegiate Dictionary, Springfield, Mass., G. and C. Merriam Co. Publishers, 1970

3. Weed, L.L., Medical Records, Medical Education, and Patient Care, Cleveland, The Press of Case-Western Univ., 1971

4. Hurst, J.W. and Walker, H.K., The Problem-Oriented System, New York, Medcom Press, 1972

5. Morgan, W.L. Jr. and Engel, G.L., The Clinical Approach to the Patient, Philadelphia, W.B. Saunders Co., 1969

6. Elstein, A.S., Kagan, N., Shulman, L.S., Jason, H., and Loupe, M.J., Methods and theory in the study of medical inquiry, Journal of Medical Education, 47, 1972, p. 85-92

7. Kessel, F.S., The philosophy of science as proclaimed and science as practiced: "identity" or "dualism"?, American Psychologist, Nov. 1969, p. 999-1005

8. Bacon, Francis, Novum Organum (First Part), in The English Philosophers from Bacon to Mill, Ed, E.A. Burtt, New York Modern Library, 1931

9. Kuhn, T.S., The Structure of Scientific Revolutions, Chicago, University of Chicago Press, 1962

10. Popper, K.R., The Logic of Scientific Discovery, New York, Basic Books, Inc., 1959

11. Medawar, P.B., Induction and Intuition in Scientific Thought, London, Methuen and Co., Ltd., 1969

12. Luchins, A.S. and Luchins, E.H., New experimental attempts at preventing mechanization in problem solving, Journal of General Psychology, 42, 1950, p. 279-297

13. Wason, P.C., 'On the failure to eliminate hypotheses ...' -- a second look, in Thinking and Reasoning, Ed, P.C. Wason and P.N. Johnson-Laird, Baltimore, Penguin Modern Psychology Readings (U.S.), 1968, p. 165-174

14. Elstein, A.S. and Shulman, L.S., Scoring and analysis of medical inquiry protocols, Unpublished manuscript, 1971

15. Bruner, J.S., Goodnow, J.J., and Austin, G.A., A Study of Thinking, New York, Wiley, 1956

16. Miller, G.A., Galanter, E., and Pribram, K.H., Plans and the Structure of Behavior, New York, Holt, 1960

17. Simon, H.A. and Newell, A., Human problem solving: the state of the theory in 1970, American Psychologist, December, 1970

18. Bartlett, F.S., Thinking, New York, Basic Books, 1958

19. Miller, G.A., The magical number 7 ± 2: some limits of our capacity for processing information, Psychological Review, 63, 1956, p. 81-97

20. Kleinmuntz, B., The processing of clinical information by man and machine, in Formal Representation of Human Judgment, Ed, B. Kleinmuntz, New York, Wiley, 1968

21. Price, R. and Vlahcevic, Z.R., Logical principles in differential diagnosis, Annals of Internal Medicine, 75, 1971, p. 89-95

22. Dudley, H.A.F., Clinical method, The Lancet, Jan. 2, 1971

23. Sprosty, P.J., The use of questions in the diagnostic problem solving process, in the Diagnostic Process, Ed, J.A. Jacquez, Ann Arbor, 1964

24. Elstein, A.S., Current status of the medical inquiry project, paper presented at AAMC Invitational Workshop, Washington, D.C., June 6-8, 1972

25. Gordon, M.J., Training in the use of heuristics in diagnostic problem solving among advanced medical students, Tentative title of doctoral dissertation in progress, ETC, May, 1973

26. Claparede, E., La genese de l'hypotheses, Archives de Psychologie, 24, 1934, p. 1-154

27. Newell, A., Simon, H.A., and Shaw, J.C., Elements of a theory of human problem solving, Psychological Review, 65, 1958, p. 151-166

28. Newell, A., On the analysis of human problem solving protocols, Paper given at the International Symposium on Mathematical and Computational Methods in the Social Sciences, Rome, 1966

29. Neisser, U., The multiplicity of thought, British Journal of Psychology, 54, 1963, p. 1-14

30. Elstein, A.S., Sprafka, S.A., and Shulman, L.S., Analyzing Medical Inquiry Processes, Paper read at 1973 Annual Meeting of American Educational Research Association, New Orleans, La., Feb. 25 - Mar. 1, 1973

31. Elstein, A.S., Kagan, N., Shulman, L.S., and Jason, H., Final Report of Medical Inquiry Project, Chapter 4, Office of Medical Education Research and Development, Michigan State University, ETC, July, 1973

32. McGuire, C.H., Personal Communication, Feb. 25, 1972

33. Gagne, R.M. and Smith, E.C., Jr., A study of the effects of verbalization on problem solving, Journal of Experimental Psychology, 63, 1962, p. 12-18

34. Helfer, R.E. and Ealy, J.M., Observations of pediatric interviewing skills, Paper read at the 1971 Annual Meeting of AAMC

35. Interdepartmental Appraisal Committee of University of Illinois College of Medicine, Clinical Simulations, Christine H. McGuire and Lawrence M. Solomon, Eds., New York, Appleton, Century, Crofts, 1971

36. Rimoldi, H.J.A., The Test of Diagnostic Skills, Journal of Medical Education, 36, 1961, p. 73-79

37. McGuire, C.H. and Babbott, J.M., Simulation techniques in the measurement of problem solving skills, Journal of Medical Education, Spring, 1967

38. Lewy, A., and McGuire, C.H., A study of alternative approaches in estimating the reliability of unconventional tests, Paper read at Annual Meeting of AERA, February, 1966

39. McGuire, C.H., A summary of the evidence regarding the technical characteristics of Patient Management Problems, A special report prepared for the Committee on Examinations of the Americal Academy of Orthopedic Surgery, Fall, 1970

40. Frase, L.T., Paragraph organization of written materials: the influence of conceptual clustering upon the level and organization of recall, Journal of Educational Psychology, 60, 1969, p. 394-401

41. Goran, M.J., Williamson, J.W., and Gonnella, J.S., The validity of Patient Management Problems, Journal of Medical Education, 48, 1973, p. 171-177

42. Angoff, W.H., Test reliability and effective test length, Psychometrika, 18, 1, 1953, p. 1-14

43. Angoff, W.H., A note on the estimation of nonspurious correlations, Psychometrika, 21, 3, 1956, p. 295-297

44. Finn, J., Multivariance -- Univariate and Multivariate Analysis of Variance and Covariance, a FORTRAN IV Program, State University of New York at Buffalo, 1970

45. Sutcliffe, J.P., On the role of "instructions to the subject" in psychological experiments, American Psychologist, 27, 8, 1972, p. 755-758

46. Levine, H., University of Texas at Austin, February, 1973, Personal Communication

APPENDICES

APPENDIX A

Instructions Read to Subjects

INSTRUCTIONS

During this session you will be asked to solve three simulated clinical problems. Your task will be to arrive at a definitive diagnosis for each problem with the aid of the information available. The procedure used for solving all three problems is the same. As you can see from the sheet entitled Problem I which I have given you, a written introduction is given at the beginning of each problem. The introduction is followed by numbered options which you may request. These options give you information which will help you solve the problem, i.e. reach a definitive diagnosis. The options fall into the general categories of history, physical,laboratory and x-ray studies, and management. As you request the options, I will hand you the appropriate information printed on a card. Please request information one item at a time by stating the number of the item. As you request an option please list its number in the column labeled "options" on the sheet provided.

As you work on the problem, I would like you to use any useful information you get to help you think of possible problem formulations. I encourage you to speculate somewhat and base formulations on relatively small amounts of information. Furthermore, if you wish to, use these formulations as guides in the selection of more information.

While you are doing the problem, I will interrupt you periodically and ask you to do two things. First I will ask you to write down any problem formulations you may have thought of at that point. Secondly, I will ask you to quickly describe to me what caused you to entertain those formulations. For example, did a certain item of information cause you to think of a formulation; had you seen a case before which presented in this manner, and so on. Each time I ask you to list problem formulations, please list as many as you are thinking of up to five (5). If you have no formulations in mind you need not list any. If you are still entertaining the same formulations you listed previously, you may simply list those same ones again. At the end of each problem please state your definitive diagnosis for that problem.

You will be evaluated on the efficiency, thoroughness, and accuracy of your work. The accuracy score refers only to the definitive diagnosis. Efficiency and thoroughness refer to how much information you gather to reach a solution. These two scores are closely related. They tend to balance each other. A good performance is given by choosing that amount of information which results in an adequate workup. You should not try to solve the problem by choosing as few items as possible, nor should you request information which is useless to you or may be harmful to your patient.

I will keep this tape recorder going while you are working. Do not pay any attention to it. It is harmless. Now, are there any questions before we begin?

If you are ready, please read the introduction to Problem I aloud.

INSTRUCTIONS

During this session you will be asked to solve three simulated clinical problems. Your task will be to arrive at a definitive diagnosis for each problem with aid of the information available. The procedure used for solving all three problems is the same. As you can see from the sheet entitled Problem I which I have given you, a written introduction is given at the beginning of each problem. The introduction is followed by numbered options which you may request. These options give you information which will help you solve the problem, i.e. reach a definitive diagnosis. You may request these options in any order you wish. The options fall into the general categories of history, physical, laboratory and x-ray studies, and management. As you request the options, I will hand you the appropriate information printed on a card. Please request information one item at a time by stating the number of the item. As you request an option please list its number on the sheet provided.

I would like you to gather as much information as you feel you need to do a generally good workup and to make a definitive diagnosis. When you have the information you need, please group together those cues which fit together into problem formulations, and indicate which formulation may be considered as the definitive diagnosis for that problem. I caution you against leaping to conclusions about possible formulations based on relatively small amounts of data. Drawing early conclusions can adversely bias a workup and may lead you to overlook some important information. Throughout the problem you will have available all the cards I have given you. Thus when you have the data you need, it will be an easy task for you to aggregate groups of data together into problem formulations.

While you are doing the problem, I will interrupt you periodically and ask you to "think aloud" for me about how you are coming along. I may ask you a few questions about how helpful you find the data I am giving you, and what types of data you still need in order to arrive at a solution for the problem. My main interest in doing this is to get an idea from you about how you approach a problem of this type, and what goes through your mind while you are solving it.

You will be evaluated on the efficiency, thoroughness and accuracy of your work. The accuracy score refers only to the definitive diagnosis. Efficiency and thoroughness refer to how much information you gather to reach a solution. These two scores are closely related. They tend to balance each other. A good performance is given by choosing that amount of information which results in an adequate workup. You should not try to solve the problem by choosing as few items as possible, nor should you request information which is useless to you or may be harmful to your patient.

I will keep this tape recorder going while you are working. Do not pay any attention to it. It is harmless. Now, are there any questions before we begin?

INSTRUCTIONS

During this session you will be asked to solve three simulated clinical problems. Your task will be to arrive at a definitive diagnosis for each problem with the aid of the information available. The procedure used for solving all three problems is the same. As you can see from the sheet entitled Problem I which I have given you, a written introduction is given at the beginning of each problem. The introduction is followed by numbered options which you may request. These options give you information which will help you solve the problem, i.e. reach a definitive diagnosis. You may request these options in any order you wish. The options fall into the general categories of history, physical, laboratory and x-ray studies, and management. As you request the options, I will hand you the appropriate information printed on a card. Please request information one item at a time by stating the number of the item. As you request an option please list its number in the column labeled "options" on the sheet provided.

As you work through the problem you may make notes on that same sheet about interesting items of information, possible things you might look for, and so on. You do not have to make notes if you do not wish to. They are purely for your convenience.

While you are doing the problem I will interrupt you periodically and ask you to "think aloud" for me about how you are attacking this problem and how helpful you find the information I am giving you. My main interest in doing this is to get an idea from you about how you go about solving a problem like this and what kinds of things you think about as you work through it.

You will be evaluated on the efficiency, thoroughness, and accuracy of your work. The accuracy score refers only to the definitive diagnosis. Efficiency and thoroughness refer to how much information you gather to reach a solution. These two scores are closely related. They tend to balance each other. A good performance is given by choosing that amount of information which results in an adequate workup. You should not try to solve the problem by choosing as few items as possible, nor should you request information which is useless to you or may be harmful to your patient.

I will keep this tape recorder going while you are working. Do not pay any attention to it. It is harmless. Now are there any questions before we begin?

INSTRUCTIONS

During this session you will be asked to solve three simulated clinical problems. Your task will be to arrive at a definitive diagnosis for each problem with the aid of the information available. The procedure used for solving all three problems is the same. As you can see from the sheet entitled Problem I which I have given you, a written introduction is given at the beginning of each problem. The introduction is followed by numbered options which you may request. These options give you information which will help you solve the problem, i.e. reach a definitive diagnosis. You may request these options in any order you wish. The options fall into the general categories of history, physical, laboratory and x-ray studies, and management. As you request the options, I will hand you the appropriate information printed on a card. Please request information one item at a time by stating the number of the item. As you request an option please list its number on the sheet provided.

As you work on the problem, I would like you to use any useful information you get to help you think of possible problem formulations. I encourage you to speculate somewhat and base formulations on relatively small amounts of information. Furthermore, if you wish to, use these formulations as guides in the selection of more information. Please try to remember any formulations you entertain along the way. If you can, also try to recall what events caused you to think of these formulations. When you have finished the problem, I will go back over it with you and ask you to review your thoughts as you worked through the problem.

Although this is a simulation, I would like you to treat me as much like a patient as possible. Particularly, I request that you not reveal to me what you are thinking about while you are doing the problem, by, for example "talking to yourself" about it.

You will be evaluated on the efficiency, thoroughness and accuracy of your work. The accuracy score refers only to the definitive diagnosis. Efficiency and thoroughness refer to how much information you gather to reach a solution. These two scores are closely related. They tend to balance each other. A good performance is given by choosing that amount of information which results in an adequate workup. You should not try to solve the problem by choosing as few items as possible, nor should you request information which is useless to you or may be harmful to your patient.

Now are there any questions before we begin?

INSTRUCTIONS

During this session you will be asked to solve three simulated clinical problems. Your task will be to arrive at a definitive diagnosis for each problem with the aid of the information available. The procedure used for solving all three problems is the same. As you can see from the sheet entitled Problem I which I have given you, a written introduction is given at the beginning of each problem. The introduction is followed by numbered options which you may request. These options give you information which will help you solve the problem, i.e. reach a definitive diagnosis. You may request these options in any order you wish. The options fall into the general categories of history, physical, laboratory and x-ray studies, and management. As you request the options, I will hand you the appropriate information printed on a card. Please request information _one_ item at a time by stating the number of the item. As you request an option please list its number on the sheet provided.

I would like you to gather as much information as you feel you need to do a generally good workup and to make a definitive diagnosis. When you have the information you need, please group together those cues which fit together into problem formulations, and indicate which formulation may be considered as the definitive diagnosis for that problem. I caution you against leaping to conclusions about possible formulations based on relatively small amounts of data. Drawing early conclusions can adversely bias a workup and may lead you to overlook some important information. Throughout the problem you will have available all the cards I have given you. Thus when you have the data you need, it will be an easy task for you to aggregate groups of data together into problem formulations. I would like to find out more about how you solve these problems than your written responses can tell me. For that reason, when you have finished the problem, I will go back over it with you. At that time I would like you to review for me any interesting thoughts you may have had while solving the problem.

Although this is a simulation, I would like you to treat me as much like a patient as possible. Particularly, I request that you not reveal to me what you are thinking about while you are doing the problem, by, for example "talking to yourself" about it.

You will be evaluated on the efficiency, thoroughness and accuracy of your work. The accuracy score refers only to the definitive diagnosis. Efficiency and thoroughness refer to how much information you gather to reach a solution. These two scores are closely related. They tend to balance each other. A good performance is given by choosing that amount of information which results in an adequate workup. You should not try to solve the problem by choosing as few items as possible, nor should you request information which is useless to you or may be harmful to your patient.

Now, are there any questions before we begin?

INSTRUCTIONS

During this session you will be asked to solve three simulated clinical problems. Your task will be to arrive at a definitive diagnosis for each problem with the aid of the information available. The procedure used for solving all three problems is the same. As you can see from the sheet entitled Problem I which I have given you, a written introduction is given at the beginning of each problem. The introduction is followed by numbered options which you may request. These options give you information which will help you solve the problem, i.e. reach a definitive diagnosis. You may request these options in any order you wish. The options fall into the general categories of history, physical, laboratory and x-ray studies, and management. As you request the options, I will hand you the appropriate information on a card. Please request information one item at a time by stating the number of the item. As you request an option please list its number on the sheet provided.

As you work through the problem you may note particularly interesting bits of information. I am sure that many things will flash through your mind which you will not write down. I would like to find out more about how you solve these problems than your written responses can tell me. For that reason, when you have finished the problem, I will go back over it with you. At that time I would like you to review for me any interesting thoughts you may have had while solving the problem.

Although this is a simulation, I would like you to treat me as much like a patient as possible. Particularly, I request that you not reveal to me what you are thinking about while you are doing the problem, by, for example "talking to yourself" about it.

You will be evaluated on the efficiency, thoroughness and accuracy of your work. The accuracy score refers only to the definitive diagnosis. Efficiency and thoroughness refer to how much information you gather to reach a solution. These two scores are closely related. They tend to balance each other. A good performance is given by choosing that amount of information which results in an adequate workup. You should not try to solve the problem by choosing as few items as possible, nor should you request information which is useless to you or may be harmful to your patient.

Now, are there any questions before we begin?

Problem I

A SURGICAL ABDOMEN

Assume you are a young general practitioner, a member of the staff of your modern 300-bed community hospital. You are called by the intern at 10:30 p.m. to see a patient in the Emergency Room.

When you arrive, you find a forty-seven year old man who complains of abdominal pain and vomiting. The pain began 3 weeks ago; the patient took Bromo-Seltzer^R with some relief. He continued to work until 1 week ago when he stopped working because of pain. After 2 days at home the pain began to improve, but he began to vomit small amounts. Similar, though less severe, episodes of pain have occurred off and on for the past three years.

In working up this patient you would be interested in doing or finding out which of the following (select as many items as you consider pertinent in the order you feel is appropriate):

1 Admit patient to hospital
2 Give antispasmodics, analgesics, and antidotes; reassure the patient; send him home and plan to see him at home early the next morning
3 Call the operating room and schedule the patient for surgery
4 Observe the patient closely for the next few hours
5 Obtain history information
6 Examine the patient
7 Obtain laboratory, x-ray, and other diagnostic information
8 Start appropriate therapy
9 Continue management without surgical intervention

Laboratory, X-ray and other diagnostic tests:

10 Hemoglobin and hematocrit
11 White blood cell count
12 Red cell smear, morphology
13 Differential white count
14 Urinalysis
15 (Report)
16 Urine culture
17 Stool guaiac
18 Erythrocyte sedimentation rate
19 Serum electrolytes
20 Arterial pH and pCO_2 determinations
21 Venous pH
22 Blood urea nitrogen
23 Serum creatinine
24 Total protein
25 Albumin/globulin ratio
26 Serum protein electrophoresis
27 Blood ammonia
28 Total and direct bilirubin
29 Cholesterol
30 Bromsulphalein retention
31 Cephalin flocculation
32 Thymol turbidity
33 Serum glutamic oxalecetic transaminase
34 Alkaline phosphatase
35 Lactic dehydrogenase
36 Acid phosphatase
37 Serum amylase
38 Urine amylase
39 Random blood sugar
40 Fasting blood sugar

41 2 hour postprandial blood sugar
42 Serum calcium and inorganic phosphorus
43 VDRL
44 Urine electrolytes
45 (Report)
46 Purified protein derivative skin test
47 Blood volume
48 Gastric analysis
49 Chest X-ray
50 Upright film of abdomen
51 Flat film of abdomen
52 Barium enema
53 Upper gastrointestinal series
54 Intravenous pyelogram
55 Oral cholecystogram
56 Intravenous cholangiogram
57 Electrocardiogram
58 Venous pressure
59 Circulation time (arm to tongue)
60 Pulmonary scan
61 Pulmonary function studies
62 Hepatic scan
63 Lumbar puncture
64 Echoencephalogram

History:

65 Headache
66 Epistaxis
67 Hemoptysis
68 Chest pain
69 Cough
70 Previous hypertension
71 Appetite
72 Dysphagia
73 Nausea and vomiting
74 Bowel habits
75 Type of pain
76 Location of pain
77 Nature of vomiting
78 Nature of stools
79 Nature of diet
80 Weight loss
81 Alcohol intake
82 Jaundice in past
83 Bleeding tendency
84 Belching
85 Radiation of pain
86 Chills and fever
87 Pruritus
88 Steatorrhea
89 Hematemesis
90 Food intolerance
91 Fatigue
92 Dizziness, vertigo, fainting
93 Character of urine
94 Family history
95 Angina
96 Dyspnea
97 Dysuria
98 Edema
99 Allergies
100 Drug history
101 Smoking history
102 Previous hospitalization
103 Previous operations
104 Previous X-ray studies
105 Trauma history

Physical Examination:

106 Pupils
107 Eyegrounds
108 External ear
109 Scalp
110 Nose
111 Mouth
112 Pharynx
113 Neck
114 Chest and lungs
115 Breasts
116 Peripheral pulses
117 Blood pressure
118 Valsalva maneuver
119 Pulse rate
120 Respiratory rate
121 Temperature
123 Abdominal wall
124 Bowel sounds
125 Liver
126 Spleen
127 Abdominal mass
128 Abdominal tenderness
129 Inguinal area
130 External genitalia

131 Right lower quadrant tenderness to palpation
132 Rebound tenderness
133 Referred rebound
134 Costovertebral angle
135 Back tenderness
136 Range of motion of spine
137 Striaght leg raising
138 Rectal examination
139 Sigmoidoscopy
140 Skin
141 Heel-to-knee test
142 Serial sevens
143 General appearance of patient
144 Chvostek's sign
145 Axilla
146 Visible peristalsis

Intervention (nonsurgical):

147 Nothing by mouth
148 Clear liquid diet
149 Bland, low residue diet
150 Force fluids
151 Nosogastric suction
152 Long intestinal tube for suction
153 Gastric lavage
154 Tap water enema
155 Magnesium sulfate 15 gm by mouth
156 Catheterized urine for urinalysis
157 Condom drainage
158 Record intake and output
159 Record urine output every 2 hours
160 Maalox^{R}, 30 ml every hour
161 Type and Crossmatch 1500 cc whole blood
162 Oxygen by nasal catheter at 6 liters per minute flow
163 Irrigate nasogastric tube every hour
164 Oxygen tent with high humidity
165 Atropine 0.4 mg intravenously
166 Morphine 5 mg intramuscularly
167 Meperidine 100 mg intramuscularly every 4 hours
168 Neomycin by mouth in therapeutic doses
169 Procaine penicillin G in therapeutic doses
170 Ampicillin in therapeutic doses
171 Kanamycin and penicillin in therapeutic doses
172 Chloramphenicol and penicillin in therapeutic doses
173 Lanatoside C 0.5 mg by mouth
174 Digitoxin 0.4 mg intravenously
175 Hydrocortisone intravenously
176 Lente insulin 40 units subcutaneously
177 Protamine zinc insulin 80 units subcutaneously
178 Tolbutamide
179 Calcium chloride intravenously
180 Packed RBC slowly intravenously
181 5% dextrose in water, 74 ml per hour
182 Hypotonic saline (0.5%) 1000 ml intravenously in next 4 hours
183 1/6 molar sodium lactate, 1000 ml intravenous in next 4 hours
184 Potassium 20 mEq for each hour of intravenous fluids
185 Add potassium chloride, 60 mg to each bottle of intravenous fluid
186 Central venous pressure catheter placed
187 Crystaline insulin 70 units stat intravenously and 50 units subcutaneously
188 Crystaline insulin every 1 to 2 hours depending on blood sugar determination
189 Phosphate 14 millimoles for each hour of intravenous therapy
190 Intravenous fluids containing ammonium chloride
191 Intravenous antibiotics and hydrocortisone in appropriate doses
192 Tracheostomy and respirator assisted ventilation

193 Order intravenous fluids for next 24 hours: 1000 ml 5% dextrose in water, 500 ml Ringer's lactate, 40 mEq KCl given slowly; leave nasogastric tube in place but clamped
194 Start liquid diet
195 Order intravenous fluids for next 24 hours: 2500 ml 5% dextrose in water, 1500 ml normal saline, 120 mEq KCl; continue nasogastric suction
196 Pull nasogastric tube, start fluids cautiously by mouth
197 Order intravenous fluids for next 24 hours: 1500 ml 5% dextrose in water, 1000 ml normal saline; continue nasogastric suction

Intervention (surgical):

198 Emergency laparotomy
199 Cholecystectomy
200 Cholecystostomy
201 Exploration of common bile duct
202 Vagotomy, antrectomy and gastroduodenostomy
203 Transverse colostomy
204 Wide gastrotomy and repair of bleeders
205 Vagotomy and gastrojejunostomy
206 50% gastric resection and gastrojejunostomy
207 Drainage of pancreatic pseudocyst
208 Pancreaticojejunostomy
209 Appendectomy
210 Lysis of obstructing adhesions
211 Repair of hiatus hernia

Problem II

THE PALE, LETHARGIC CHILD

The patient is an eight year old Negro boy who is brought to the Emergency Room because of fatigue and pallor which has developed over a one week period. He also complains of occasional abdominal pain, but denies nausea or vomiting. Prior to this episode the boy had been in apparent good health, but occasionally was thought to be pale compared to his sister, especially following infections (colds, sore throats). Three weeks prior to admission he had tonsillitis and received an antibiotic for one week. The symptoms and fever associated with this latter episode improved within 3 to 4 days after starting the drug.

Physical examination reveals a well developed moderately pale Negro boy in no acute distress. Oral temperature - 99.6°F, Pulse - 98 per minute, Respiration - 26 per minute, Blood pressure - 108/66 mm Hg. The spleen is palpable 3 cm below the left costal margin; the liver is not felt. Slight diffuse abdominal tenderness without rebound is present. The remainder of the examination is entirely within normal limits.

The initial blood count was recorded in the Emergency Room as follows: hemoglobin - 3.0 gm/100 ml; hematocrit - 10%; white blood cell count - 13,000/cu mm; differential: bands - 2%, segmented neutrophils - 58%, lymphocytes - 40%.

The patient is admitted to the hospital for further evaluation.

In working up this patient you would be interested in which of the following (select as many items as you consider pertinent in the order you feel is appropriate):

1 More detailed family history
2 More detailed dietary history
3 More detailed past history
4 Chest film
5 Tuberculin skin test
6 Urinalysis
7 Serum protein electrophoresis
8 Hemoglobin electrophoresis
9 Blood smear for morphology
10 Platelet count
11 Sickle cell preparation
12 Bone marrow aspiration
13 Serum iron and iron binding capacity
14 Serum B_{12}
15 Serum folic acid
16 Osmotic fragility of red cells
17 Prothrombin time
18 Partial thromboplastin time
19 Bleeding time
20 Reticulocyte count
21 Serum bilirubin
22 Red cell survival study with radioactive chromium (patient's cells)
23 Bone marrow biopsy (surgical)
24 Blood urea nitrogen

25 Serum uric acid
26 Urine coproporphyrin
27 Blood lead level
28 Coombs test
29 Lupus erythematosus cell preparation
30 Erythrocyte glucose - 6 phosphate dehydrogenase screening test

Your prognosis for this patient is:

31 Excellent with proper medical therapy
32 Excellent with proper surgical therapy
33 Should respond to proper management of acute episode, but is unlikely to have normal longevity
34 May respond to medical management but is most likely to die of disease within months to a few years
35 Unlikely to survive present episode

Problem III

A PALE, CONFUSED PATIENT

A young married woman brings her fifty-seven year old gray-haired mother to your office for a medical checkup. The daughter tells you that her mother's appetite, weight, strength and well-being have progressively deteriorated over the last 8 months. Lately she has become more confused and mildly disoriented and recently has exhibited slight memory loss. The patient added that for the past 4 weeks she has tired easily, especially when walking.

In working up this patient you would be interested in doing or finding out which of the following (select as many items as you consider pertinent in the order you feel is appropriate):

1 Defer further investigation for the purpose of observation and ask the patient to return in:
  a) 1 week
  b) 2 weeks
  c) 3 weeks
2 Order chest, cervical spine, and skull films and ask the patient to return in:
  a) 1 week
  b) 2 weeks
  c) 3 weeks
3 Arrange for an urgent psychiatric examination
4 Request a neurosurgical consultation
5 Hospitalize the patient and arrange for further diagnostic procedures
6 Hospitalize the patient in a mental institution
7 Transfer the patient to a psychiatric ward
8 Obtain history information
9 Examine the patient
10 Arrange for laboratory evaluation on an inpatient basis
11 Arrange for laboratory tests on an outpatient basis
12 Initiate therapy

History:

13 The family history
14 Episodes of febrile illness
15 Nausea and vomiting
16 Frequency or urgency
17 Dietary intake
18 Food fads
19 Soreness in mouth or tongue
20 Alcohol intake
21 History of diabetes mellitus
22 History of hypertension
23 History of heart disease
24 History of arthritis
25 Shortness of breath on exertion
26 Any abnormalities in sensation or changes in perception
27 Motor weakness
28 Pain on urination
29 Diarrhea
30 Involuntary loss of urine
31 Coordination
32 Chest pain
33 Cough
34 Allergies
35 Gait
36 Changes in intellectual capacity
37 Orientation to time, place and people
38 Uncontrollable crying and laughing
39 Difficulty speaking
40 Double vision
41 Headaches

42 Insomnia
43 Black stools
44 Vaginal bleeding
45 Leg pains
46 Dyspepsia
47 History of abdominal surgery

Physical Examination:

48 Extraocular movements
49 Fundi
50 Pupils
51 Mucosa and nail beds
52 Heart and lungs
53 Abdomen
54 Lymph nodes
55 Rectal
56 Pelvis
57 Sensory examination
58 Motor examination
59 Deep tendon reflexes
60 Babinski sign
61 Mental status
62 Cerebellar system
63 Romberg test
64 Skin
65 Sclera
66 Blood pressure, pulse, respirations temperature
67 Tongue
68 Visual fields
69 Muscle atrophy
70 Breasts
71 Tenderness over spine

Laboratory, X-ray, and Other Diagnostic Tests:

72 Hemoglobin, hematocrit, white blood cell count, differential
73 Urinalysis
74 Tri-iodothyronine uptake (T_3) (Resin)
75 Fasting blood sugar and 2 hour post-prandial blood sugar
76 Blood urea nitrogen
77 Electrocardiogram
78 Chest X-ray
79 Stools for occult blood
80 Gastric analysis
81 Serum sodium, potassium, carbon dioxide, chloride
82 Skull X-ray
83 Liver function studies
84 Blood smear
85 Red cell indices
86 Sickle cell preparation
87 Lupus erythematosus cell preparation
88 Sedimentation rate
89 Serum creatinine
90 Serum uric acid
91 Serum total protein
92 VDRL
93 Stool for fat
94 Bone marrow
95 Lumbar puncture
96 Electroencephalogram
97 Serum calcium, phosphorus, alkaline phosphatase
98 Gastric analysis after subcutaneous histamine
99 Schilling test
100 Myelogram
101 Liver biopsy
102 Reticulocyte count

APPENDIX C

Procedures for Scoring Modified PMP's

Therapy:

103 Multivitamins, 2 tabs three times a day by mouth
104 Prednisone, 5 mg three times a day by mouth
105 Adrenocorticotropic hormone (ACTH) 30 units intravenous infusion
106 Iron
107 Vitamin B_{12} 1000 mg intramuscularly every 3 days
108 Laminectomy for decompression of the spinal cord
109 Whole blood transfusion (3 units)
110 Multivitamin capsules, one capsule three times a day
111 Aristocort 4 mg twice a day
112 Vitamin B_{12} 100 microgm intramuscularly for two weeks
113 Folic acid 10 mg per day by mouth
114 Physiotherapy

APPENDIX C

Procedures for Scoring Modified PMP's

1. Thoroughness

Before calculating any scores, the total number of possible points had to be calculated for each problem. Total possible points was defined as the total number of non-redundant positively and zero-weighted items in each problem. A redundant choice is one which logically cannot be ordered if another is ordered. For example, in one problem the opportunity to hospitalize the patient occurs twice. Depending on what point in the problem the subject chooses to hospitalize the patient, the response to his choice of that item is different. Logically the subject can choose that item only once. Therefore one hospitalization choice is redundant.

To obtain the total possible points for a problem all the response cards were counted and certain cards were subtracted from that total. First cards for which the subject received no credit were subtracted. These are not zero-weighted cards, but cards which simply guide the subject through the problem. They give him no information about the status of the patient. An example of one of these is: the student chooses the option which states "Obtain history information". He is given a card which says "See History section below". Secondly, all redundant cards are subtracted. Thirdly, all non-redundant negatively weighted cards are subtracted. The remaining total non-redundant positively and zero-weighted cards are the total possible points for that problem. One point is assigned to each card. Although this is a somewhat complex method for arriving at a possible total, it was thought necessary. The scores (thoroughness particularly) are proportions and inclusion of redundant and no-credit items in a total would artificially and non-uniformly reduce subjects' scores.

Thoroughness, then is the proportion of non-redundant positive and zero-weighted items chosen by the subject. The formula for the calculation is given in the text on page

2. Efficiency

This score is the most straightforward. It is simply the proportion of items chosen by the subject which were positively weighted. The formula is given on page

3. Accuracy

The accuracy of a subject's final response to a problem was scored in two ways: maximum accuracy and mean accuracy. These two scores are explained below. Responses on each problem were scaled from 0 to 4 using as criterion a set of weights provided by Christine McGuire. Display I gives a breakdown of final responses and the accuracy scores assigned them. The responses listed in the Display are all of those given by subjects in the present study. They are variants of those found on the criterion provided by McGuire.

The maximum accuracy score a subject could obtain was determined by the presence or absence in his final diagnosis of one of the possibilities listed in Display I. If one of these solutions was present, the subject's maximum accuracy was simply the accuracy score assigned to that solution. There were two accurate solutions to Problem I. Maximum accuracy was the mean of those two solutions as listed by the subject. If the subject listed only one solution, his maximum accuracy was half of the value assigned that solution. For example:

Final Solution	Max Accuracy
Peptic ulcer with pyloric obstruction and Diabetic ketoacidosis	$\frac{4 + 4}{2} = 4$
Acute pancreatitis and Diabetic ketoacidosis	$\frac{2 + 4}{2} = 3$
Diabetic ketoacidosis	$\frac{4 + 0}{2} = 2$

The mean accuracy score was the mean of the accuracy scores of all the solutions he listed. Examples of these scores on Problems I, II and III are shown below.

Problem I

Solutions	Max Accuracy	Mean Accuracy
Peptic ulcer with pyloric obstruction Diabetic ketoacidosis Dehydration	4	$\frac{4 + 4 + 1}{3} = 3$

Problem II

Solutions	Max Accuracy	Mean Accuracy
Familial hemolytic anemia Bone marrow failure	3	$\frac{3 + 1}{2} = 2$

Problem III

Solutions	Max Accuracy	Mean Accuracy
Pernicious anemia Peripheral neuropathy Hypertension	4	$\frac{4 + 2 + 1}{3} = 2.3$

When the two accuracy scores were calculated it was thought that the mean accuracy score would be used in the analysis. This was later deemed unwise since it tended to penalize students for being thorough in their final listing of solutions or problems as those trained in the problem-oriented tradition have been taught to do. For this reason primarily, the maximum accuracy score was used in analysis.

4. Hypothesis Generation and Number of Hypotheses

Before determining when hypotheses were generated and how many were entertained by a subject the entity, hypothesis, had to be defined.

A hypothesis is any disease entity or problem (in the problem-oriented sense) mentioned by the subject. It must be mentioned in a positive context. The following are examples of hypotheses:

Ulcer

Diabetes

Anemia mentioned <u>before</u> CBC is obtained. In Problem II, anemia is <u>not</u> a hypothesis unless qualified, e.g. Hemolytic anemia

Bleeding

Gall bladder problem

GI problem

Depression

Psychological problem

The above hypotheses would be credited if mentioned in the following types of context:

"The patient may have HYP"

"This makes me think of HYP"

"I'm going to check for HYP"

Those hypotheses would not be credited if mentioned in the following types of context:

"This is probably not HYP"

"This cue does not go along with HYP"

Generation of a hypothesis was credited at the time the hypothesis was first mentioned. If the hypothesis was first mentioned in association with a cue (e.g. "When I got CUE it made me think of HYP"), generation was credited at the time the cue was obtained. For subjects in the Verbalization condition this procedure was easily applied since the subject stopped periodically during the workup and commented. His comments often contained hypothesis generations and cue associations. Subjects in the Non-Verbalization condition were asked to recall the problem by going back through it card by card. This procedure was used to reduce retrospective distortion to a minimum. Hypothesis generation was credited during the recall using the same rules as were used for the Verbalization group.

To test one of the research hypotheses (see page) subjects were divided on two hypothesis generation dimensions: early or late, and many or few. Early hypothesis generators were those who generated the first hypothesis after having read the introduction to the problem, but before choosing any options. Late generators were subjects who generated the first hypothesis after choosing one or more options. The total number of hypotheses generated were plotted in a separate histogram for each problem. A natural break fell at about ten hypotheses in each problem. Therefore subjects were divided into two groups at that point with up to nine hypotheses being "few", and ten or more hypotheses being "many".

APPENDIX D

Criteria Used for Process Analyses

DISPLAYS

DISPLAY I

Accuracy Scores

PROBLEM I

Cues	Hypotheses										
	Pep. Ulcer[1]		GI Malig.	Gall Bladder	Pancrea-titis	Diabetes Mellitus[2]	Gastritis	Reg. Enter.	Bowel Obst.	Myo. Infarc.	Append.
	a	b									
10 Hemoglobin and hematocrit: Hb. -- 16.8 gms/100 ml, hct. -- 48%	+1	-1	-1								
11 White blood cell count: 12,300/cu mm.		+1		+1	+1			+1	+1	+1	+2
12 Red cell smear, morphology: The red cells are normochromic and normocytic		-1	+1					-1			
13 Differential white count: Bands -- 3%, neutrophils -- 70%, lymphocytes -- 24%, monocytes -- 3%		-1		-1	-1			-1	-1	-1	-2
14 Patient unable to void				+1	+1					+1	
15 Urinalysis, catheterized sample: Color -- deep yellow, specific gravity -- 1.035, pH -- 5.2, protein -- trace, glucose -- 4+, acetone -- 4+, bile -- negative, microscopic -- normal					+3	+3				+1	

[1] Peptic ulcer: a) uncomplicated b) complicated

[2] Diabetes mellitus with ketoacidosis

PROBLEM I (continued)

Cues	Hypotheses										
	Pep. Ulcer[1]		GI Malig.	Gall Bladder	Pancrea-titis	Diabetes Mellitus[2]	Gastritis	Reg. Enter.	Bowel Obst.	Myo. Infarc.	Append.
	a	b									
16 Urine culture: a) Patient unable to void spontaneously, and must be cather-ized. Catheter specimen sent to lab.					+1	+1				+1	
b) No growth.						-1					
17 Stool guaiac: Positive 2+.		+1	+1				+1	+1	+1		
18 Erythrocyte sediment-ation rate: 32 mm in 1 hour (Westergren).	+1	+1	+1	+1	+1			+1	+1	+1	+1
19 Serum electrolytes: a) Na -- 129 mEq/l; K -- 2.8 mEq/l; Cl -- 80 mEq/l; CO_2 -- 18 mEq/l.					+2	+2					
b) Na -- 129 mEq/l; K -- 2.8 mEq/l; Cl -- 80 mEq/l; CO_2 -- 16 mEq/l.					+2	+2					

[1] Peptic ulcer: a) uncomplicated b) complicated

[2] Diabetes mellitus with ketoacidosis

PROBLEM II

Cues	L-L	Blood Loss	Red Cell Aplasia		Infec. Mono	Mes. Lymph.	Inad. Diet	Neph.	Lead Intox.	Porph.	SCA	CSA	AI
			Con	Acq									
7 Serum protein electrophoresis: Total protein 7.4 gm/100 ml; Albumin-60%; Alpha$_1$ globulin-7%, Alpha$_2$-10%, Beta-10%, Gamma-13%	-1							-1					
8 Hemoglobin electrophoresis: Hemoglobin AA.		+1	+1	+1							-3		
9 Blood smear for morphology: See Figure 10.	-1	-1	-3	-3	-1		-2	-1	-1		-2	+3	
10 Platelet count: 300,000/cu mm.	-2												
11 Sickle cell preparation: See Figure 11.											-3		
12 Bone marrow aspiration: See Figure 12.	-3	-2	+1	+1	-1		-2			-2		+2	
13 Serum iron and iron binding capacity: Serum iron-130 microgm/100 ml; iron binding capacity-175 microgm/100 ml	-1	-2			-1	-1	-1		-1		+1	+1	

L-L = Leukemia-Lymphoma
Blood Loss = Blood Loss GI Tract
Red Cell Aplasia = Red Cell Aplasia
Infec. Mono = Infectious mononucleosis

Mes. Lymph. = Mesenteric Lymphadenitis
Inad. Diet = Inadequate Diet
Neph. = Nephritis
Lead Intox. = Lead Intoxication

Porph. = Porphyria
SCA = Sickle Cell Anemia
CSA = Congenital Spherocytic Anemia
AI = Autoimmune Hemolytic Anemia

PROBLEM II (continued)

Cues	L-L	Blood Loss	Red Cell Aplasia		Infec. Mono	Mes. Lymph.	Inad. Diet	Neph.	Lead Intox.	Porph.	SCA	CSA	AI
			Con	Acq									
14 Serum B_{12}: 260 micro microgm/ml							-1						
15 Serum folic acid: 13 milli microgm/ml							-2						
16 Osmotic fragility of red cells: Hemolysis beginning at 0.70% saline; complete at 0.45% saline		-3	-2	0			-2				-3	+3	+3
17 Prothrombin time: 12 seconds (control 13 seconds)													

L-L = Leukemia-Lymphoma
Blood Loss = Blood Loss GI Tract
Red Cell Aplasia = Red Cell Aplasia
Infec. Mono = Infectious Mononucleosis

Mes. Lymph. = Mesenteric Lymphadenitis
Inad. Diet = Inadequate Diet
Neph. = Nephritis
Lead Intox. = Lead Intoxication

Porph. = Porphyria
SCA = Sickle Cell Anemia
CSA = Congenital Spherocytic Anemia
AI = Autoimmune Hemolytic Anemia

DISPLAY I

Accuracy Scores

Problem I -- GI solutions

Score of 4

Peptic ulcer with pyloric obstruction

Duodenal peptic ulcer with pyloric spasm

Peptic ulcer with scarring and narrowing of the proximal duodenum

Duodenal ulcer with obstruction

Duodenal bulb ulcer with gastric outlet obstruction

Pyloric outlet obstruction secondary to old duodenal ulcer

High partial obstruction with or without ulcer

Score of 3

Right sided obstruction of ascending colon

Partial obstruction at duodenum

Active chronic duodenal ulcer

Acute GI bleed with duodenal ulcer

Gastric outlet obstruction

Obstruction secondary to ulcer

Duodenal peptic ulcer

Zollinger-Ellison Syndrome

Score of 2

Acute pancreatitis

Cancer of the head of the pancreas causing obstruction

Obstruction secondary to pancreatitis

Alcoholic pancreatitis

Cholecystitis

Stroke

Score of 2 (continued)

Pancreatic pseudocyst

Chronic pancreatitis

Ulcer

Small bowel infarct secondary to obstruction

Score of 1

Alcoholism

Barium peritonitis

Alcoholic gastritis

Dehydration

GI hemorrhage with question of obstruction

Score of 0

Gall bladder problem

Problem I -- Diabetes solutions

Score of 4

Diabetic ketoacidosis

Score of 2

Diabetes

Diabetes mellitus

Acidosis

Problem II

Score of 4

Congenital spherocytosis

Hereditary spherocytosis

Score of 3

Familial hemolytic anemia

Score of 2

Hemolytic anemia

Hemolytic process

Score of 1

Ideopathic hypersplenism

Chronic disease of hematopoetic and immune systems

Aplastic anemia

Hereditary cell defect

Bone marrow failure

Hereditary RBC disease

Score of 0

Iron deficiency anemia

Chronic lymphocytic leukemia

Leukemia

Lead poisoning

Coagulopathy

Autoimmune hemolytic anemia

Problem II -- Other identified problems

Score of 1

Porphyrinuria

Decreased production of RBC's

Severe anemia

Problem III

Score of 4

Pernicious anemia with neurologic complications

Pernicious anemia

Decreased production of Intrinsic Factor with decreased absorption of Vitamin B_{12}

Pernicious anemia with achlorhydria

Vitamin B_{12} deficiency

Vitamin B_{12} deficiency with subacute combined degeneration

Score of 3

Subacute combined degeneration

B_{12} or Folic dependent macrocytic anemia

Megaloblastic anemia

Score of 2

Peripheral neuropathy

Score of 1

Brain stem problem

Score of 0

Gastric carcinoma

Strobe

Score of 0 (continued)

Plummer-Vinson Syndrome

Organic brain syndrome

Spinal cord compression

Cerebral metasteses

Cerebral dimentia

Problem III -- Other identified problems

Score of 1

Hypertension

DISPLAY II

Representative Hypotheses

Hypothesis	Hi (n=10)	Lo (n=10)	Lo & Drop (n=4)
Sickle cell anemia	10	10	3
G6PD deficiency	8	6	4
Thallassemia	1	1	1
Hemolytic anemia	6	5	3
Autoimmune anemia	4	6	3
Blood loss anemia	1	4	
Hereditary cell problem		5	
Hypersplenism		2	3
Bone marrow repression	1	4	2
Leukemia	5	7	2
Lymphoma	1	1	1
Iron deficiency anemia	5	6	1
Vit. B_{12} deficiency	3	4	
Folic acid deficiency	2	3	1
Lead poisoning	1	7	3
Infection	1	1	1