Title
Teaching clinical problem-solving strategies to medical students

Author
Beck, Arne L.

Publication Date
1985

Peer reviewed|Thesis/dissertation

Teaching Clinical Problem-Solving Strategies to Medical Students: An Information-Processing Intervention

by

Arne L. Beck

Health Psychology Program
University of California, San Francisco

San Francisco, California, 1985

ABSTRACT

Diagnostic expertise is based on the organization of clinical information in the expert's long-term memory and the use of efficient strategies for rapidly accessing this knowledge.

This study assessed the use of expert-like knowledge organization and problem-solving strategies to enhance the clinical problem-solving skills of medical students.

Thirty-five preclinical medical students were randomly assigned to an experimental or control group and given written material on congenital heart diseases to study for four hours. Experimental subjects received material in a format that grouped together logically competing sets of diseases (LCSs), based on the similarity of their clinical presentation. They were also given a brief lecture on general clinical reasoning strategies. Control subjects received the same material but in a "classical" text book format. They did not receive the clinical reasoning lecture. Between two and five days later, subjects' clinical reasoning was assessed with three simulated cases of congenital heart disease which were presented to them on a microcomputer. Cases varied in typicality and, hence, diagnostic difficulty. The first case was prototypic and the easiest to diagnose; the second was typical, that is, relatively common and of moderate diagnostic difficulty; and the third was atypical and the most difficult to diagnose.

The results showed: 1) experimental subjects acquired a higher ratio of diagnostic to nondiagnostic clinical information than controls, across all cases.

2) experimental subjects mentioned the correct diagnosis sooner in their workups than controls for the atypical case, but not for the prototypic and typical cases; 3) experimental and control subjects were equally extensive in their evaluation of LCSs for the salient information in each case; 4) although no group differences in diagnostic accuracy were found for the prototypic case, more control than experimental subjects correctly diagnosed the typical case, and more experimental than control subjects correctly diagnosed the atypical case; 5) experimental subjects incurred slightly lower workup costs than controls; and 6) subjects' learning styles interacted with the intervention.

Additional research was suggested to explore the generalizability of the intervention and the role of individual learning styles in clinical problem-solving.

ACKNOWLEDGMENTS

I am grateful for the support and guidance provided by many people during my dissertation research. I thank my wife, Kathie, for her love, humor, and support during this difficult process. I appreciate the substantive contributions and the encouragement offered by my dissertation chair, Nancy Adler, Ph.D., and commmittee member Catherine Lewis, Ph.D. I am especially thankful for the support and intellectual stimulation provided by my committee member, David Bergman, M.D., whose ideas helped shape and focus this research. I would like to thank the UCSF computer center staff for their invaluable assistance, which ranged from text processing and statistical package consulting to providing money for my computer account on a continual basis. I would also like to express my appreciation to the UCSF Graduate Division for both the Graduate Research Fellowship and the Patent Fund award given to me. This research would have been difficult if not impossible without the substantial contributions of these funds. Finally, my thanks to the medical students from UCSF, the joint UCB-UCSF program, and Stanford University who participated in this study. It is to the betterment of their clinical training that this research is dedicated.

CONTENTS

INTRODUCTION

Significance

Research on clinical reasoning has increased dramatically within the last decade. One contributing factor has been the application of cognitive science research on complex problem-solving to the field of medicine. Psychologists have directed their efforts toward understanding the cognitive processes involved in medical problem-solving and clinical expertise using the paradigms of information-processing, judgment, and decision theory (Elstein & Bordage, 1979).

Research from the information-processing paradigm has been particularly relevant to medicine by suggesting that the way in which information is organized in memory (knowledge representation) and the strategies used to rapidly access this information, are crucial to efficient problem-solving. In light of the information explosion in medicine and the concommitant increase in demands on medical students to assimilate this information, the issues of knowledge representation and clinical problem-solving strategies have assumed increasing importance.

A second factor has involved the rapid growth of the artificial intelligence field. Using increasingly sophisticated programs and computer hardware, artificial intelligence researchers have constructed computerized expert medical systems that model clinical reasoning and serve as decision support systems or tutors to physicians.

A third and equally important research stimulus has come from medicine itself. Physicians are becoming sensitized to the information explosion in medicine and the attendant demands on their ability to make efficient and accurate diagnostic and treatment decisions. Moreover, many physicians are becoming aware of the growing body of research which suggests that the process of medical decision making is subject to numerous errors and biases (Bergman & Pantell, 1984; Detmer, Fryback, & Gassner, 1978; Elstein, 1976). For example, physicians often ignore rates of disease prevalence when considering diagnostic possibilities (Casscells, Schoenberge, & Grayboys, 1978). Instead, they may base their considerations on the salience of the disease in memory, or the degree to which the patient's symptoms appear highly representative of that disease, regardless of it's rarity (Eddy, 1982). Other systematic errors include the subjective distortion of information later in the medical workup to support initial diagnostic hunches (Wallsten, 1981), and excessive reliance on noncontributory diagnostic information (Bergman & Beck, 1983), or on laboratory tests with only moderate predictive value for the disease (Balla, Elstein, & Gates, 1983).

Pressure to avoid these documented errors, as well as reduce the costs associated with them, has led to the collaboration of physicians with cognitive psychologists and artificial intelligence researchers. The goal of this research collaboration has been to better understand and work with the cognitive limitations of the medical decision maker and to develop techniques to improve the decision maker's effectiveness.

As a result of the research on clinical reasoning, much is now known about the components of clinical expertise and the manner in which it is achieved. However, there is a paucity of research on the application of such knowledge to the improvement of clinical reasoning. In particular, one logical and potentially

important application of this knowledge is to the teaching of clinical problem-solving. The question which arises is, can such expert-like clinical reasoning skills be facilitated through teaching approaches which are based on information-processing concepts? The purpose of the present study is to investigate this question.

Review of Literature

The following review is divided into three separate but interdependent sections. The first section provides an overview of the following major theoretical approaches to medical decision making: Judgment, decision theory, artificial intelligence in medicine (AIM), and cognitive psychology-information-processing. Because the primary theoretical orientation of the present study derives from the cognitive psychology-information-processing approach, this approach will be emphasized while the other approaches will be presented more briefly.

The second section will examine clinical problem-solving from the standpoint of assessment issues and methodology. Various assessment methods will be evaluated, with an emphasis on simulated processes, particularly patient management problems. This evaluation will focus on issues of validity, reliability, fidelity, and ease of development, implementation and scoring of these assessment approaches.

The last section will review literature on problem-solving instruction. Examples of research in problem-solving instruction within medical education, educational psychology, and mathematics will be presented.

Overview of Theoretical Approaches to Medical Decision Making

The theoretical distinctions between the various approaches to medical decision making are largely due to differences in the methodologies each uses to investigate this phenomenon. One general but useful distinction between the four approaches is the degree to which they use "process-tracing" or "black box" methodologies in the investigation of clinical decision making (Elstein, Shulman, & Sprafka, 1978). The former attempts to describe the actual cognitive process of clinical problem-solving by using various measures of the problem-solving process, such as thinking aloud protocols. The latter attempts to model the inference process without speculating on the actual cognitive processes behind the judgments.

Judgment and decision theory might be considered black box methodologies because they rely on techniques for measuring decisions with accuracy but they lack a focus on the cognitive processes involved in problem-solving. These methodologies attempt to predict decisions through the use weighted information entered into a multiple regression equation, or through the use of bayesian statistics.

AIM research might be considered as consisting of both black box and process-tracing methodologies. Early AIM programs represented the former, being diagnostically accurate but not psychologically valid in their approach to problem-solving. However, recent work on expert systems attempts to model cognitive processes in clinical problem-solving more closely and therefore might be classified as following a process-tracing approach.

The last theoretical approach to be discussed is the information-processing paradigm. This paradigm is grounded in cognitive psychology and relies primarily on process-tracing methodologies to investigate problem-solving.

Each of the theoretical approaches to medical decision making will be discussed in more detail next.

Judgment. The judgment paradigm has long been used in experimental psychological studies of psychophysics and attitude scaling, as well as in more applied settings where judgments are used for personnel selection, placement and evaluation decisions. In clinical medicine, the judgment paradigm can be used to examine the optimal combination and weighing of clinical data—signs, symptoms, or laboratory test results, for the purpose of making a judgment about the diagnosis or treatment.

Much of the work from the judgment paradigm is based on Brunswik's lens model and the adaptation of it to judgment studies (Hammond, Hursch, & Todd, 1964). The model describes the relationship between the judgment, based on observable cues, of the unobservable criterion, or true state. The relationship between the judgment, cues, and true criterion may all be stated in correlational terms. The correlation between the judgments and the cues represents the degree to which various cues are used in the judgment process. The correlation between the cues and the true criterion (ra) is a measure of ecological validity, or the degree to which the cues accurately represent the true criterion. In addition to the accuracy of the judgments, there are three other components of the lens model. The first component is the degree to which the task or environment is predictable, represented by the correlation between the actual and predicted criterion value. This variable is labelled Re. The second component, G, measures the knowledge of the properties of the task, and is expressed as the correlation between subjects' predicted judgments and the predicted criterion values. The third component in the lens model, Rs, represents the cognitive control over the utilization of the knowledge. It is expressed as the correlation between subjects' actual and predicted judgments.

These four components of the lens model have been expressed in the equation ra=GReRs, that is, judgmental accuracy (ra) is based upon the predictability of the task (Re), knowledge of the task properties (G), and cognitive control over the use of that knowledge (Rs).

The lens model can be easily applied to clinical judgment tasks where the primary concern is the evaluation and categorization of information in the form of patient cues to arrive at a diagnostic or treatment decision. Hoffman (1960) was one of the first investigators in the field of psychology to take such an approach, proposing that multiple regression equations be used to model the clinical judgments of clinical psychologists. This technique involves presenting cues separately to judges who rate their relative importance. A regression equation is then computed based on a linear combination of the weighted cues to estimate the criterion judgment. The technique of developing regression equations to model judges' weighting and combining of cues, called policy capturing, has been successfully applied to a variety of judgment tasks. Within the medical realm, policy capturing has been used to determine expert cue utilization for several diagnostic and treatment decisions. For example, Slovic, Rorer, & Hoffman (1971) studied nine radiologists' use of seven signs to judge the malignancy of gastric ulcers. They were able to precisely describe the individual physicians' use of the signs and point out the disagreements between them in the relative weights they assigned to these signs. Further, Slovic et al compared the physicians' ratings with the Bayesian probabilities for the likelihood of the malignancy, given the various signs. This analysis indicated an encouraging degree of agreement, with a correlation between the average ratings and the Bayesian probabilities of .80.

Einhorn (1974) used a model of expert judgment to examine three pathologists' ratings of the amount of nine histological signs present on biopsy slides taken from patients with Hodgkin's disease. He used a multitrait-multimethod matrix to assess 1) intrajudge reliability, 2) convergent and discriminant validity of ratings, and, 3) the degree of similarity between the judges' weighting schemes. Einhorn was able to accurately describe individual judges' policies and demonstrate their reliability and convergent and discriminant validity. However, as with Slovic et al, the results showed a fair amount of interjudge disagreement on the relative weighting of the signs.

In a study of the use of judgments in thyroid treatment decisions, Moore, Aitchison, & Parker (1974) asked six clinicians to choose one of three treatments based on five items of patient information. Again, they found differences in clinicians' use of information. In addition, all six clinicians used less than the five items available to them in selecting their choice of treatments.

Several consistent findings have emerged from research on clinical judgment. For example, clinicians often use fewer cues for judgments than are typically available to them (Moore et al, 1974; Elstein & Bordage, 1979). Further, these cues are usually combined summatively, despite the belief that more sophisticated patterns of cue combinations are necessary for accurate judgments (Dawes and Corrigan, 1974). In addition, it appears that even the weighing of cues is a simple matter of assigning +1 or -1 (Dawes & Corrigan, 1974).

Another related finding from studies of policy capturing is that the multiple regression equations developed frequently surpass the accuracy of the judges upon which they were based. This phenomenon, known as bootstrapping (Goldberg, 1970), occurs because the derived regression equation represents the judge's policy in an optimal way, based as it is upon repeated (and therefore reliable) estimates that are immune to potential sources of error once developed.

It is important to note that although regression models of judgment are usually quite accurate, they must be considered paramorphic in that they do not duplicate the actual cognitive processes used in making the judgment (Hoffman, 1960).

The judgment paradigm is of value in prescribing the use of regression techniques, albeit paramorphic, to improve the accuracy and reliability of clinical judgments.

Decision theory.. Decision theory, like the judgment approach, attempts to describe how information is used in clinical decision making. However, its focus is on the process of rational decision making under conditions of uncertainty. Whereas the judgment approach examines the use of fairly stable estimates in which all of the information is known or available, decision theory is concerned with changes in estimates as a function of new information (Elstein et al, 1978). In addition, while the judgment approach is based on a regression model, decision theory is based on a Bayesian model that views its various components in probabilistic terms.

The primary concern of decision theory is the selection of the most appropriate course of action based on an assessment of prior probabilities of the symptoms and diseases in question, and the utility, or value, associated with the outcome of each course of action. Determination of the appropriate action therefore depends on the calculation of several pieces of information, including the probability of the disease given the symptom, P(D/S), the probability of the symptom given the disease, P(S/D), and the distribution of both symptom and disease in the patient population. This information may be aggregated using Bayes' theorem, a formula for optimizing the revision of probabilities in the light of new evidence. The formula may be stated as follows:

$$P(D|S)=P(S|D) \times P(D)/P(S)$$

It can be seen that the probability of a disease given a symptom is a function of both the degree to which the symptom and disease are associated, and their distribution in the population.

Bayes' theorem is frequently used to calculate the probability that a diagnostic screening test is truly indicative of the disease in question. For this case, then, the screening test may be substituted for the symptom in the equation. For example, a screening test may correctly classify 90% of the cases of a particular disease and incorrectly classify 5% of the normal population as having the disease. Furthermore, if the disease occurs in 5% of the population, then the probability that an individual has the disease given a positive screening test is 49%. To many, this probability estimate might seem low, given the relatively high hit rate of the screening test to detect the disease. This is because of a tendency to place greater weight on the power of the screening test to detect the disease than on the relatively low base rate of the disease, which accounts for the low probability estimate (Bar-Hillel, 1980; Lyon & Slovic, 1976).

Although Bayesian analysis provides a useful means for estimating revised probabilities on the basis of new information, it is subject to distortion when 1) the prevalences of the signs, symptoms, and diseases are not well known, 2) when several diseases or signs and symptoms are not independent, 3) when two or more diagnoses are considered simultaneously, and 4) when the probability estimates used do not accurately represent the population from which they are derived (Bergman & Pantell, 1984). These criticisms reflect the application of a precise mathematical model to the imprecise domain of clinical medicine.

Fortunately, not all of these drawbacks are insurmountable. For example, nonindependence of data may be dealt with by complex mathematical models

(Fryback, 1978; Ludwig & Heilbron, 1983). Perhaps more important is that in the absence of objective probability estimates, which is a frequent occurrence in clinical medicine, subjective probability estimates may be used. When this is done, however, an important consideration must be kept in mind. Subjective estimates may follow the same laws of probability as objective estimates in a mathematical sense, but there is a clear difference between the two in the manner in which each is generated. While objective probability estimates are based on observed frequencies of events, subjective probability estimates are based on intuitions about the frequencies of events which may or may not be valid. (Elstein and Bordage, 1979). In fact, there is a substantial amount of research indicating that in the process of generating subjective probability estimates, individuals may be prone to several biases and heuristics. Typically, these biases occur as a result of our inability to process probabilistic information accurately.

Kahneman, Slovic, and Tversky (1982) discuss three major heuristics used by decision makers when making probability estimates. The first is availability. Availability occurs when the salience of a particular event in memory increases its subjective probability, and when less salient events are forgotten and thus estimated to occur less frequently. An example of the availability bias in medicine is when serious illnesses or rare cases may cause the physician to overestimate their actual prevalence.

The second bias is representativeness, which occurs when one's subjective probability estimate is based on the degree to which the event is prototypic of that class of events, regardless of the actual frequency of the event. For example, a physician may see a patient with "classic" signs and symptoms of a disease. However, if the disease has a low prevalence rate in the patient

population, then the patient's signs and symptoms may be manifestations of a more common disease. This inattention to base rates when assessing the likelihood of an event or outcome is a frequently observed bias (Bar-Hillel, 1980; Lyon & Slovic, 1976; Eddy, 1982; Casscells, Schoenberge, & Grayboys, 1978).

The third bias, Anchoring and adjustment, refers to the "tendency of people to give unduly tight distributions when assessing uncertain quantities" (Lichtenstein, Fischhoff, & Phillips, 1982, p. 333). Uncertainty of the quantity or event leads to an initial subjective probability estimate, which then serves as a dominant anchor around which subsequent adjustments are made. This bias leads to insufficient adjustments and overconfidence in estimates.

Another medical decision making bias sometimes seen involves the subjective distortion of information later in the medical workup to support initial diagnostic hunches. This bias might be considered a form of the primacy effect, which occurs when "early information is assigned undue weight in deciding among hypotheses" (Wallsten, 1981, p. 150).

These medical decision making biases are not always maladaptive. For example, the overestimation of the subjective probability of a serious disease may reflect the clinician's concern about the consequences of missing the diagnosis. This bias involves the combination of an assessment of the utility of the outcome with the actual subjective probability estimate, and is called a value-induced bias (Wallsten, 1978).

Two questions arising from this discussion are, to what extent are physicians' subjective probability estimates accurate and how can they be improved? The studies pertaining to the first question yield mixed results. For example, clinicians tend to be overconfident in their subjective probability estimates (Oskamp, 1965; Gilbert, McPeek, & Mosteller, 1977; Christensen-

Szalanski & Bushyhead, 1981), and to demonstrate the operation of heuristics and biases =n their use of probabilistic information (Balla, Elstein, & Gates, 1983; Detmer, Fryback, & Gassner, 1978). However, other studies indicate that physicians' subjective probability estimates are reasonably accurate (Thornbury, Fryback & Edwards, 1975; Gustafson, Kestly, Greist, & Jansen, 1971)

The second question regarding the improvement of subjective probability estimates reveals some interesting techniques for calibrating subjective probability estimates (Elstein & Bordage, 1979; Lichtenstein, Fischhoff, & Phillips, 1982). For example, one approach suggests a four step process for calibrating estimates. The first involves relating the unknown events to events more familiar to the rater. The second step involves a ranking of the events. Third, the relative likelihood of events should be estimated, and finally, probability estimates for these events should be compared to events with known likelihoods. Other calibration approaches require the assigning of numerical weights to feelings of confidence about the event in question (Koriat, Lichtenstein, & Fischhoff, 1980; Ferrell & McGoey, 1980). Koriat et al's method requires that the assessor emphasize contradictory evidence in order to avoid overconfidence.

One method of assessing the accuracy of subjective probability assessments is to construct a calibration curve (Lichtenstein et al, 1982). The calibration curve is a plot of the actual versus predicted occurrences of the event, thus providing a numerical assessment of the degree of over- or underconfidence of the probability estimates.

Decision theory has made valuable contributions to medical decision making, particularly in the development of decision aids, such as algorithms, and in the elucidation of common errors in subjective probabiliy estimates that occur

in the process of medical decision making. However, decision theory is more concerned with improving the accuracy of decisions through statistical models than it is in describing and understanding the actual cognitive processes underlying clinical reasoning.

Artificial intelligence in medicine (AIM).. Research in AIM has been steadily increasing and expanding in scope since its inception some 25 years ago. The early work in AI focused on the use of powerful, high speed computers to solve a variety of problems ranging from chess to mathematical proofs. Early work in AIM involved the use of statistical techniques for diagnostic problem-solving in well defined problem domains. These techniques included the use of probabilistic information on diseases and symptoms, modeling of physiological processes, and the development of algorithms to aid physicians with clinical decisions (Duda & Shortliffe, 1983). Such simulations of expert decision making yielded accurate and reliable systems capable of generating diagnostic hypotheses and providing a relatively complete list of courses of action requiring consideration.

Ironically, an important by-product of research in AIM has been the realization that there are areas of human reasoning inadequately simulated by these programs. For example, knowledge domains which contain imprecision in facts and problems are not readily amenable to simulation. In addition, the early hopes of developing powerful, widely applicable problem-solving procedures have been dashed by repeated findings that problem-solving heuristics are to a lage extent dependent on the knowledge base from which they are derived (Duda & Shortliffe, 1983). This last revelation has generated a sigificant shift in the approach of AIM research from the use of powerful general problem-solving mechanisms to the structuring of a particular knowledge base in an optimal manner for efficient problem-solving within that domain.

The deficiencies in early AIM programs and the shift to knowledgebased approaches have led to the development of expert systems. Expert systems are typically characterized by their formalization and organization of large, though usually incomplete, knowledge bases, and their use of rule-based reasoning strategies for arriving at diagnostic or treatment alternatives. The more well-developed expert systems have understandable explanations of their reasoning strategies for the benefit of the user. Expert systems are developed by trial and error process whereby the initial programs are run on test cases and problems in the knowledge base or problem-solving strategy are revealed and corrected. Eventually, the validation of an expert system includes a formal comparison of its performance to that of several experts in the field. Examples of medical expert systems include:

MYCIN (Shortliffe, 1976), a consulting program that selects appropriate antimicrobial therapy for treatment of infectious diseases

GUIDON and NEOMYCIN (Clancey, 1983), refined versions of MYCIN developed for teaching and characterized by more human-like and transparent reasoning strategies

ONCOCIN (Shortliffe, Bischoff, Campbell, van Melle, & Jacobs, 1981), used for oncology treatment

INTERNIST and its successor, CADUCEUS (Pople, 1983), used for a wide variety of problems in internal medicine

CASNET (Weiss, Kulikowski, & Safir, 1978), an expert system based on a causal model which provides consultation for glaucoma patients

PIP (Szolovits & Pauker, 1976), a system which uses both categorical and probabilistic reasoning mechanisms to associate patient findings with hypotheses

Expert systems in medicine have made considerable progress toward producing accurate and reliable medical decisions. Equally important, they have taken a significant step in the direction of modeling human expert reasoning. However, a number of issues remain to be addressed, as outlined by Duda & Shortliffe (1983). First, the acquisition and encoding of a relatively complete knowledge base remains a difficult and time consuming task that presenty has not been mastered in an efficient manner. This step requires both comprehensible domain knowledge and workable knowledge representation strategies for effective problem-solving.

A second issue which lies at the heart of expert systems research is the development of adequate knowledge representation strategies. A variety of knowldege formalisms have been proposed, each with various strengths and weaknesses. For example, mathematical logic is a popular scheme because of its flexibility and precision. Many expert systems, such as MYCIN, are rule-based, allowing for excellent representation of empirical associations but lacking in the elucidation of pathophysiology or temporal trends in the disease process. Other representation methods include semantic networks and frames (structures for organizing knowledge) (Barr & Feigenbaum, 1981). To be effective, knowledge representation strategies must closely simulate the way in which the knowledge is structured in the expert's memory. Yet, it is equally important that this strategy be amenable to description for understanding and use by clinicians. The complexity of the expert's knowledge representation scheme and the need for understandable explanations of it place heavy demands on expert systems researchers. An additional problem faced by these researchers is the task of altering the complex computer programming when updating the knowledge base and the inference process.

A third issue in expert systems research invovles the use of inference systems to come to reasonable conclusions from the data. Inference systems may be goal driven, reasoning backward from goals to data, or data-driven, reasoning forward from data to conclusions. More advanced systems, such as ABEL (Acid-Base and ELectrolyte program) (Patil, Szolovits, & Schwartz, 1983), have been developed which use data represented in a multilevel causal network which is based on pathophysiological knowledge. Inferences are drawn through the aggregration and elaboration of these causal concepts at multiple levels of detail. However, ABEL is limited to the fairly circumscribed problem of acid-base and electrolyte disturbances.

It is clear that for expert systems to accurately simulate human experts, they will need to contain a of combination data-driven inference systems and organization of information within a causal network. This is a difficult task which increases in complexity when the inferences are characterized by uncertainty, a frequent occurrence in clinical medicine. Strategies for drawing inferences under uncertainty must then involve some form of probabilistic model, such as Bayes' theorem. In fact, Szolovits & Pauker (1983) assert that a truly expert system must demonstrate both categorical and probabilistic reasoning, "the former to estabish a sufficiently narrow context and the latter to make comparisons among hypotheses and eventually to recommend therapy" (p. 210).

Although research in AIM has developed to the point of attempting to formalize human reasoning, the issues discussed above provide considerable challenge and demonstrate the incomplete development of the field. In comparison to the previously discussed approaches of judgment and decision theory, AIM is more comprehensive in its ability to simulate reasoning processes as well as make accurate diagnostic or treatment decisions. However, expert

systems are costly, time consuming to develop, and fall short of capturing the subtleties of expert reasoning. In addition, their ability to generate hypotheses and recommend alternatives is subject to the same biases and errors as human decision makers. Finally, expert systems are useful only in well circumscribed areas which are rich in knowledge and have a narrow, specific focus.

Cognitive psychology-information-processing. The efforts of AIM researchers to formalize the knowledge representation systems of experts has brought them closer to the field of cognitive psychology, where the analysis of problem-solving expertise is a major topic of investigation. Cognitive psychology in general and the information-processing paradigm in particular has had a long-standing interest in the way in which information is structured in memory and its relationship to problem-solving expertise.

A central assertion of the information-processing approach is that man has rational limitations on his cognitive capacity to receive, store, process, and retrieve information for problem-solving (Newell & Simon, 1972). In particular, it is widely accepted that working memory is limited in its capacity to process more than seven plus or minus two bits of information at a given time (Miller, 1956). In contrast, long-term memory is seen as unlimited in its storage ability. Thus, it is necessary to adopt problem-solving strategies to use the maximum amount of information available in the most efficient manner possible. One such strategy involves combining large amounts of information into a smaller number of more manageable chunks which can be stored in long-term memory and retrieved as complete sets. This strategy circumvents the information-processing limitations of working memory, thereby allowing the processing of large amounts of information which are actually stored and retrieved as a much smaller number of discrete categories.

A second problem-solving strategy follows naturally from the first and involves the hierarchical organization of information in long-term memory (Mandler, 1967) and the development of rich interconnections between these chunks of information. The actual form in which the information is stored in long-term memory is the subject of debate in cognitive psychology (Anderson, 1978; Wood, 1983), and several forms have been proposed, such as propositions, schemata, scripts, and pictorial representations (Wood, 1983). However, the idea that problem-solving ability is in large part a function of the way in which information is organized in long-term memory is much less controversial.

The third strategy for efficient problem-solving is to develop strong associations between certain important cues presented by a problem-solving task and the chunks of information in long-term memory which are relevant to the problem solution. These cues then become triggers for a large amount of information in long-term memory and a plan for its application to the problem-solving task.

Problem-solving strategies are a naturally occurring response for coping with information-processing demands. Because of the complex reasoning tasks inherent in some fields and the consistent application of problem-solving strategies to these tasks over time, some individuals become exceptionally proficient at problem-solving. These individuals have stored chunks of information in long-term memory which are relevant to the tasks they perform; this information is hierarchically organized with extensive interconnections; and strong associations between information chunks and problem cues have been developed which facilitate rapid and efficient use of the information for problem-solving. Such proficiency in information-processing forms the basis of expertise in many complex fields, including medicine.

Perhaps the best illustration of the use of problem-solving strategies is the research on expertise in diverse problem domains. Studies of problem-solving in chess (deGroot, 1965; Chase & Simon, 1973;, Simon & Chase, 1973). revealed that when briefly presented with typical chess board configurations, grand masters remembered significantly more positions of pieces on the board than novices. Further, it was apparent that experts chunked whole board configurations in memory while novices chunked single pieces. Not only did experts have denser and more meaningful chunks, they had more programmed strategies for a series of moves based on the board configuration presented to them. Similar findings have been reported for physics and math experts. Although introductory physics students only appear to process the surface features of physics problems, experts are able to produce the physics laws upon which the problem solutions are based (Chi, Feltovich, & Glaser, 1981). Similarly, math experts use a variety of heuristics, or rules of thumb, to simplify and reduce complex equations for quicker solution, while novices often perceive the same problems as unfathomable (Schoenfeld & Hermann, 1982). In these examples, it is clear that experts have a great deal of highly organized knowledge for more efficient problem-solving, use well-developed production systems for rapid retrieval and application of relevant information to a problem, and use a variety of heuristics to simplify problems.

It is noteworthy that expert knowledge is organized specifically for application to a given problem domain. Therefore its usefulness decreases when applied outside of that domain or when applied to a problem within the domain that is out of context. Thus, although Chase and Simon (1973b) estimated that chess experts have some 50,000 chess board configurations stored in memory, these experts were no better able to recall a random configuration presented to them than were novices.

A growing body of literature on clinical expertise in medicine concurs with research on expertise in other fields. Wortman (1966, 1971, 1972) was one the first psychologists to investigate medical diagnosis from an information-processing perspective. He proposed that medical diagnosis involves a memory search of information organized into a hierarchy of categories which are based on the location of the disorder, anatomical syndromes and groupings, etiology and etiological syndromes and groupings, and, at the lowest level, diseases. On the basis of a verbal protocol of a neurologist's decision rules and diagnostic problem-solving strategies, Wortman was able to construct a computer program that accurately performed the same tasks as the physician.

In 1978, Elstein, Shulman and Sprafka completed a series of studies analyzing clinical reasoning through the use of several methods, including simulated patients and stimulated recall of the problem-solving process. This methodology represents a long tradition of process-tracing studies in psychology, popularized by DeGroot (1965) in the U.S. but dating back to the Wurzburg group of "thought psychologists" (Elstein et al, 1978). The emphasis of this approach is that the study of the problem-solving process is as important as the outcome of that process.

Elstein et al found that diagnostic problem-solving resembled hypothetico-deductive reasoning in that clinicians rapidly generated and tested hypotheses based on the patient's information. Four components of the process were identified:

1) cue acquisition--information gathering via the history, physical exam, and laboratory tests. Thoroughness of cue acquisition was found by Elstein et al to be associated with diagnostic accuracy. However, excessive information collection was seen as potentially maladaptive in that it tended to overload the information-

processing capabilities of the physician. The typical strategy used by physicians to avoid information overload involved the collection of clinical data within a fairly limited context which was generated by the patient's problem.

2) hypothesis generation--the formation of hypotheses which are triggered by the information. During this process, physicians rapidly produced a set of about four to seven hypotheses. Hypotheses were often triggered almost immediately in response to patient information, even when the information was scant. This suggests "the existence of strong links in memory between salient cues and certain hypotheses triggered by these cues" (Elstein and Bordage, 1979, p. 338). In fact, this research indicates that physicians use their knowledge of pathophysiology less than their well-developed associations in long-term memory between cues and hypotheses and between competing hypotheses. Further, these associations enable the physician to 1) automatically retrieve problem formulations from memory, 2) arrive at correct conclusions quickly by using questions which maximize information concerning the problem formulation and, 3) rapidly narrow their number of hypotheses under consideration to a workable set.

In addition to the automatic use of associative reasoning, physicians may consider deliberately competing hypotheses so that data that is negative for one hypothesis is positive for another. Such a strategy allows the most efficient use of the diagnostic data by reducing the amount of information under consideration at any one time. In addition, the consideration of deliberately competing hypotheses facilitates the normally difficult task of processing negative information (Elstein & Bordage, 1979).

Thus, it becomes apparent that the process of hypothesis generation involves the use of automatic as well as conscious strategies for overcoming the limitations of working memory when processing large amounts of information.

These strategies are based on experience and thus are not often used by novice clinicians. As Barrows & Tamblyn (1980) state, "It is the absence of a problem formulation that causes students to recite endless amounts of data about findings on the history and physical examination of a patient when asked for a summary of the patient's problem." (p. 28).

3) cue interpretation--data are interpreted within the context of the hypotheses under consideration. The weighing of diagnostic information takes the simplified but fairly accurate form of a three-point scale, that is, cues are considered as positive, negative, or noncontributory. As with thoroughness of cue acquisition, Elstein et al found that accuracy of cue interpretation was associated with diagnostic accuracy. However, thoroughness of cue acquisition was not shown to compensate for errors in cue interpretation.

4) hypothesis evaluation--an effort to aggregate data to confirm one of the competing hypotheses as the most likely. This usually involved simply combining the weights for the cues or clusters of cues in order to assess diagnostic likelihood.

Elstein et al found two general types of errors that occured during the clinical reasoning process. The first involved the tendency to overemphasize positive findings, while paying less attention to disconfirming data. This error also took the form of assessing noncontributory data as positive for the diagnosis.

A second type of error seen was excessive data collection. This error was often committed in an effort to increase diagnostic accuracy through the use of additional confirmatory data. However, excessive data collection did not in reality increase diagnostic accuracy, because the additional information was

typically redundant. Despite its relative uselessness, this redundant information served to increase the physician's confidence in his or her decision.

Two negative outcomes of excessive data collection are the tendency to have more information than is manageable for decision making and the generation of excessive costs in the diagnostic workup through the use of additional laboratory tests.

Kassirer and Gorry (1978) also examined physicians' clinical problem-solving using a protocol analysis combined with introspection to elicit the problem-solving strategies used by six clinicians. On the basis of the physicians' verbal reports, Kassirer and Gorry classified the problem-solving process into three phases: 1) hypothesis activation, 2) hypothesis evaluation, and 3) information gathering.

During hypothesis activation, physicians were observed to rapidly generate one or more working hypotheses using relatively little information. The average number of active hypotheses under consideration at any one time was slightly less than seven, although many more hypotheses were considered at different points in the workup. These hypotheses provided a framework within which to organize existing information and seek additional confirmatory or discriminatory evidence.

The hypothesis evaluation process consisted of further refining and reducing the hypotheses under consideration, and employing several case building strategies to test them. One such strategy was confirmation, or the accumulation of several pieces of evidence that, taken together, strongly suggested a diagnosis. Another strategy was elimination, which inolved the use of an absent finding to discard a hypothesis normally associated with the presence of that finding. Elimination strategies were also used to distinguish one hypothesis from another by the presence of a certain cue. Exploration was the third

strategy used. It involved searching for additional expected information to substantiate the most likely hypothesis, as well as to check for complications associated with the condition.

The information gathering phase was seen by Kassirer and Gorry as an important part of the problem-solving process. They found that 60%-80% of the physicians' questions were centered around temporal relations, organ systems, disease severity, predispositions or complications, and the need for actions. The use of these questions, rather than a more general review of systems, suggests a problem-solving approach by experts that is hypothesis-driven.

The findings that have emerged from these seminal studies have been corroborated by several other investigations (Barrows & Bennet, 1972; Norman, 1983; Feltovich, Moller, & Swanson, 1983; Patel, 1983; Bordage, 1983).

An additional finding from this research is that the process of clinical reasoning does not change from medical student to expert (Neufeld, Norman, Feightner, & Barrows, 1981) and that experts do not have superior memories per se. Rather, expertise is based on the "availability of a broad interconnecting network of relevant knowledge and experience which can be efficiently and rapidly accessed in the solution of a problem" (Norman, 1983, p. 280). Thus, the organization of knowledge in memory is one of the distinguishing features of clinical expertise. One example of this expertise is the finding that experts' memory chunks contain three times more information than those of novices (Norman, Jacoby, Feightner, & Campbell, 1979). Moreover, experienced clinicians categorize information on the basis of its relationship to similar signs and symptoms for a given diagnosis, rather than on the basis of superficial similarities which are not related to diagnostic problem-solving. This knowledge representation scheme enables experts to better handle atypical cases and pick out their most critical cues.

A study by Feltovich, Johnson, Moller, and Swanson (1983) elegantly illustrates this relationship between knowledge structure and problem-solving ability. Feltovich et al presented four pediatric cardiology case summaries to 12 subjects at four levels of training, ranging from fourth year in medical school to 20 years of clinical experience. Each segment of case information was presented to subjects in a sequential, fixed order fashion. Information was grouped into the four sections of history, physical examination, x-ray, and EKG. Subjects were taped while they thought aloud during the problem-solving process. At the end of each of the four information sections, subjects were asked for any diagnostic hunches and alternative hypotheses under consideration.

Analysis of subjects' protocols was centered on the concept of the logical competitor set (LCS), which is a set of diseases that "share major underlying physiology with the operative or true disease in the case and hence have similar clinical presentation" (Feltovich et al, 1983). The use of LCSs allows an examination of 1) subjects' disease knowledge and the precision of their disease models; 2) subjects' ability to differentiate diseases into subtypes; 3) subjects' use of disease clusters corresponding to disease categories; and 4) the precision of subjects' knowledge of the variations of the diseases within the LCS.

Feltovich et al were interested in the degree to which subjects would consider all of the diseases in the LCSs as well as use cues to discriminate among competing diseases, thus enabling a narrowing of the hypothesis space to the correct diagnosis. The full use of the LCS and cues for its evaluation were seen by these investigators as evidence that these diseases are stored together as a single chunk in long term memory.

They found that the experienced clinicians tended to consider all of the diseases within the LCS for each case and, within each LCS, were able to

correctly evaluate cues in order to select the appropriate disease as the primary diagnosis. In contrast, the medical students typically did not consider the full range of diseases within the LCS nor did they demonstrate the precision in their disease knowledge necessary for using critical cues to arrive at the correct diagnosis. Subjects with a moderate amount of experience (residents and fellows) demonstrated problem-solving that at times resembled the experts while at other times resembled that of the students. The findings of this study demonstrate the relationship between expertise, knowledge organization, and problem-solving skill.

In sum, literature on expertise in medicine and other fields illustrates several characteristics of the human problem solver: "A limited capacity of short-term memory, a use of heuristic strategies to examine promising avenues, a tendency to search for information sequentially, and the importance of the problem solver's conceptualization of the problem at hand" (Kassirer & Gorry, 1978, p. 254).

The conceptualization of the problem, a product of the problem solver's knowledge representation scheme, is perhaps the most critical factor in efficient, accurate problem-solving (Posner, 1973; Wickelgren, 1974). Through experience, experts store knowledge in memory based on the features of the problems they encounter. Over time, they develop interconnections between these chunks of information, strengthen associations between salient cues and hypotheses, and refine the information in chunks in order to accurately discriminate competing hypotheses.

Assessment of Clinical Problem-Solving.

Because clinical problem-solving skill is the critical component of clinical expertise, a great deal of work has been done on its assessment from a variety of

perspectives. The major approaches to the assessment of clinical problem-solving to be reviewed here are simulated patients and patient management problems (PMPs). Both forms of assessment involve simulations of the clinical interaction but offer different degrees of fidelity to the actual doctor-patient encounter. Although other assessment forms exist, such as chart review, written and oral examinations, and case presentations (Barro, 1973), simulated processes are perhaps the most complete forms for measuring many of the variables involved in clinical problem-solving and are among the most widely used methods for research.

Simulated patients.. Simulated patient approaches involve the use of trained actors to portray patients with whom the physician interacts. This interaction is videotaped and coded for several process and outcome measures thought to comprise clinical problem-solving skills. Physicians may be asked to think aloud during the workup so that a better understanding of the cognitive processes involved in clinical problem-solving may be gained. In addition, a second technique, called stimulated recall, may be used, which involves the physician's review of the videotape of the workup immediately after the interaction with the patient.

The simulated patient method with stimulated recall was used in Elstein et al's (1978) classic analysis of clinical reasoning within a cognitive psychological framework. The use of simulated patients allowed Elstein to measure several critical components clinical problem-solving, such as data perception, problem formulation, hypothesis generation, and diagnostic and treatment decisions.

Kassirer & Gorry (1978) took an approach similar to that of Elstein's group. They used a protocol analysis of problem-solving behavior combined with requests for physicians to think aloud during the task. Responses of six clinicians were

audiotaped as they queried a simulated patient. Kassirer and Gorry were able to categorize physicians' behavior as involving hypothesis activation, hypothesis evaluation, and information gathering. Kassirer coined the term "clinical cognition" to refer to the use of verbal transcripts to study problem-solving process.

The primary advantage of the simulated patient method is that it is a high fidelity representation of the clinical problem-solving process, and yet provides sufficient standardization for comparative assessment of physicians. In addition, the simulated patient method measures a large number of the components of problem-solving. The fidelity and extensive sampling of the domain of clinical reasoning provide evidence for the construct and content validity of simulated patient methods (Elstein et al, 1978).

The drawbacks of simulated patient methods are that they take considerable time to prepare, are costly, and difficult to score. As a result, relatively few cases may be used and these may not be readily generalizable to the larger sample of cases typically encountered by clinicians. In addition, Elstein et al were unable to demonstrate their reliability or discriminant validity, because physician performance varied as a function of the case.

A more fundamental critique of the simulated patient method centers on the validity of verbal reports as data. The act of speaking may alter the normal sequence of thoughts during problem-solving. Also, verbal reports may omit critical intermediate steps in the reasoning process, or worse, be inaccurate (Kassirer et al, 1982; Nisbett & Wilson, 1977; Ericsson & Simon, 1980). The potential for error is greatest with the use of retrospective verbal reports (Fischhoff, 1975). This finding particularly calls into question the validity of the stimulated recall method.

Patient management problems.. Developed during the early 1970's, PMPs have gained rapidly in popularity and have undergone numerous revisions during their use in research, teaching, and certification. The basic purpose of the PMP is to assess clinical skills using a standardized presentation format and scoring system. Although PMPs have lower fidelity than simulated patients, they offer more control over administration and scoring. A PMP typically consists of a limited amount of patient information which is presented to the subject in written or computerized format. The subject then selects additional history, physical examination, or laboratory test data which he or she thinks is of diagnostic relevance. This process continues until a suitable diagnosis and/or course of managing the patient is chosen.

PMPs were originally conceived as a clinical assessment method that would measure aspects of clinical competence not addressed by objective examinations. The first PMPs were developed by the National Board of Medical Examiners to assess nine areas of clinical competence (Vu, 1979): History, physical examination, tests to be used, diagnostic acumen, treatment, care implemented, continuing care, doctor-patient relationship, and responsibilities as a physician. The PMP was presented in written format. After receiving initial patient information, students selected bits of information about the patient in a sequential fashion. Students' choices were indicated by removing the ink covering that particular response. Scoring was based on the number of correct choices made plus the number of incorrect choices avoided.

Since the development of the original PMP, several revised formats have been employed. McGuire & Babbott (1977) constructed a revised PMP which contains less initial patient information, places greater emphasis on interdependent patient management decisions, offers several alternative

diagnostic pathways, and uses a different scoring system. Scoring for the revised PMP is based on efficiency, profficiency, errors of omission, errors of commission, and a composite index of overall competence. Another form of the PMP was developed by Elstein, Shulman, & Sprafka (1978) for observational purposes. It focuses on the natural sequence of patient management decisions and the order of information collected. Three scores are derived from the modified PMP: Efficiency, thoroughness, and diagnostic accuracy. A third form of the PMP, called the Diagnostic Management Problem (DMP), was developed by Helfer & Slater (1971). The DMP, like the modified PMP, measures the process of clinical problem-solving. Individual items of patient information are presented on cards which subjects can select in any order and number they desire. The order, number, and usefulness of the selected cards form the basis of scores on process, efficiency, competence, and diagnosis.

The Sequential Management Problem (SMP) (Martin, 1975) was developed as another form of PMP with the specific goal of avoiding the problem of cueing associated with the previously discussed methods. Cueing occurs when the options made available to subjects bias their diagnostic reasoning and help them arrive at the correct diagnosis. To avoid the problem of cueing, the SMP was designed to require subjects to ask for each sequential item of patient information without being given a list of possible choices. Another unique aspect of the SMP is that subjects are provided with feedback immediately following their choices, thus preventing the accumulation of errors associated with poor initial choices. Scoring for the SMP is based on averages of positive, negative, or zero points for correct, incorrect, or equivocal choices, respectively. Separate scores are assigned to each section of the PMP.

One of the most recent problem-based assessment methods is the Portable Problem Patient Pack (P4) (Barrows & Tamblyn, 1980). In this format, different colored decks of cards are used to contain history, physical examination, laboratory, consulting, and treatment data. The front of each card contains questions to guide the student, and the back of the card provides answers to the questions. Each action is assigned weights from -2 (dangerous or inappropriate) to +2 (appropriate) and combined into the following scores: Clinical skills economy, proficiency, and the extent to which the workup is on-target or off-target.

In addition to paper and pencil formats, PMPs have been developed for use in a computer format. Two versions of this format are the Computer-Based Examination (CBX) and Computerized Patient Management Problem (CPMP). The CBX measures the efficiency of patient management and the sequence and efficiency of the tests ordered. The CPMP measures general components of problem-solving (Shakun, Taylor, & Osbaldeston, 1976). CPMPs have been used successfully in the certifying examination by the Royal College of Physicians and Surgeons of Canada.

Overall, PMPs have demonstrated their usefulness in teaching and evaluation situations. They offer a variety of formats and scoring systems, and provide for a high degree of control over the assessment process, although they have relatively low fidelity and take time to construct.

The most serious criticisms of PMPs involve issues of reliability and validity. Although most of the PMP formats discussed here have demonstrated reliability by the stability of scores across problems, or by the internal consistency of items or problems, neither measure of reliability may be totally appropriate. For example, because PMP items are interdependent, the use of

internal consistency violates the assumption of independence of items which is necessary for the use of this measure. In addition, reliability assessments based on the stability of scores across problems are not altogether appropriate because problem-solving performance has often been found to be case specific.

The validity of PMPs is also open to criticism, despite the findings that most versions of the PMP have demonstrated content validity, that is, they represent the domain of behavior which comprises the clinical problem-solving process. This is because neither the construct nor criterion-related validity of PMPs have been effectively demonstrated. For example, Some research has indicated that PMPs do not yield the exact results as actual real-life performance on the same diagnostic task (Goran, Williamson, & Gonnella, 1973; Newble, Hoare, & Baxter, 1982). Goran et al (1973) found that physicians ordered significantly more history and physical data, as well as lab tests, in response to a PMP versus real patients with the same diagnosis. These researchers conclude from their results that the concurrent validity of PMPs is questionable. However, Marshall (1983) argues that performance on PMPs and in real diagnostic settings is not necessarily comparable because PMPs measure only problem-solving, not attitudes and skills. Marshall also questions the accuracy of Goran et al's assessment of actual clinical performance, based as it was on record review. At best, it may be concluded that PMPs may be used for pure problem-solving assessment and for comparison of different groups on simulated problem-solving, as proposed in the present study. However, they should not be seen as substitutions for the complete assessment of actual performance in a clinical setting because such performance is under the influence of other factors not measured by the PMPs.

It should be clear that the controversy surrounding the validity of PMPs results not only from the nature of the PMPs themselves, but also from the complex, multidimensional nature of clinical problem-solving, and our presently inadequate attempts to define or measure it in any complete sense.

Bashook (1976) has proposed new conceptualizations of reliability and validity for application to clinical problem-solving assessment. He suggests that clinical problem-solving be categorized according to the problem-solving process (e.g, sensing, defining, resolving), the clinical discipline involved, and the context of care (for example, chronic versus acute). The issue of reliability becomes one of the number of clinical problems in one clinical situation that are required to determine performance. Validity may be defined as the degree of sampling across the breadth of the domain which is necessary for the generalization of performance to that whole domain.

Teaching Clinical Problem-Solving Strategies

Although the basic components of clinical expertise have been articulated, the application of such knowledge to the improvement of medical education has been slow. However, medical educators are becoming increasingly concerned about the relevance of teaching and evaluation methods. This concern has resulted in part from the research findings discussed earlier, which clearly show that experienced clinicians approach diagnostic problems quite differently than classical medical training dictates. As Kassirer and Gorry (1978) note, clinical instruction focuses on the

"personal interaction with the patient, the need to avoid biased questions, the necessity of assessing the patient's reliability in recalling the history, and the value of thoroughly characterizing the patient's symptoms...Critical elements

such as how diagnostic possibilities are first introduced and evaluated, how competing diagnostic possibilities are eliminated, and what strategies should be used to obtain data with the greatest diagnostic information content are typically ignored" (p. 253).

Research in medical education.. In response to the need for educational improvement, medical educators are developing new teaching and evaluation techniques which focus on clinical problem-solving (Helfer & Slater, 1971; Wright, Stanley, & Webster, 1983; Marshall, 1983; McGuire, 1980; Bashook, 1976). Allal and Shulman (1974) attempted to train 16 second year medical students to generate diagnostic problem formulations early in the clinical encounter. Allal and Shulman used films to simulate the early stages of the clinical encounter. Training involved generating initial problem formulations for this simulated data, followed by feedback on the outcome of this process based on an expert's problem formulation for the same task. One group of students received feedback on the outcome of the problem formulation and another group received both outcome feedback and feedback on the processes used by the expert in his problem formulation. A control group which received neither the training nor feedback was also used. Following three weeks of training, students were evaluated on measures of problem formulation, cue utilization and classification, and the degree of relationships among problem formulations. The two trained groups were found to have significantly different problem formulations than the control group, but differences were not seen on any of the other variables. Interestingly, the addition of process feedback to one of the training groups did not significantly increase their problem-solving skills compared to the outcome feedback only group.

In another effort to evaluate the teaching of clinical problem-solving, Gordon (1974) taught problem-solving heuristics to half of a group of 32 medical students. The other half was allowed to use their own problem-solving strategies. In addition, half of the students in each group were asked to apply the heuristics systematically, while the other half was not. The results of a posttest indicated that none of the groups was superior on a measure of diagnostic accuracy.

Several teaching programs have been initiated to teach clinical problem-solving through a combination of learning techniques. Ways, Loftus, & Jones (1973) developed the Focal Problems course, which uses study cases, presented sequentially, small group discussions, and an emphasis on problem-oriented patient workups, for example, problem formulation, cue interpretation, and hypothesis generation. Barrows and Tamblyn (1980) have also developed a comprehensive teaching program that is problem-based. This method involves simulated patient management problems with realistic choices for the student and immediate feedback in the form of comparison with previously developed standards. Barrows and Tamblyn propose three major components of teaching programs aimed at facilitating clinical problem-solving skills: 1) continual exposure to patients, 2) the use of simulation experiences, such as their Portable Patient Problem Pack (P4), and 3) complementary printed materials to aid students with the practice and evaluation of their learning experiences. In another comprehensive effort, Taylor, Harasym, and Laurenson (1978) taught the generation of early diagnostic hypotheses to 61 first year medical students. A five week course included lectures, small group discussions, independent learning, and practice with simulated patient problems. Posttest analyses of students' performance showed gains in factual recall, and identification and integration of information for hypothesis testing, but less change in the actual formation of

hypotheses. Finally, Kassirer (1983) proposed the use of iterative hypothesis testing to teach problem-solving skills. In this method, one student acts as the repository for the patient data, initially presenting only the patient description and chief complaint. Students then query the "patient", who provides answers to the question and no more. Participating students must justify their questions, describe the diagnostic hypotheses they are entertaining, discuss their expectation for the patient information, and interpret the information once it is given.

Although the primary work on teaching clinical problem-solving has emerged from the field of medical education, there are an increasing number of AIM researchers who are developing intelligent tutoring systems for medical education. One of the best examples of this work is NEOMYCIN (Clancey, 1983), an tutorial program for application to the problem domain of infectious disease. NEOMYCIN was developed to provide a psychologically valid model of expert diagnostic behavior that is relatively transparent to the student. It is characterized by "focused, forward-directed use of data (including trigger associations that suggest diagnoses); follow-up questions that establish the disease process (part of what a physician calls 'forming a picture of the patient'); and management of a changing 'working' memory of hypotheses under consideration" (p. 364). Knowledge is organized hierarchically and several diagnostic rules are employed that correspond to the use of disease process knowledge, data-hypothesis associations, and confirmation and discrimination heuristics. These include causal rules, trigger rules (to associate data with etiologies), data/hypotheses rules (to associate data with diseases only within the differential), and screening rules, to restrict the data under consideration.

The potential of expert tutorial systems such as NEOMYCIN is great, but at present there are no formal evaluation studies which demonstrate the extent to which they can improve clinical problem-solving skills. Moreover, the cost-effectiveness of expert tutorial systems needs careful evaluation because they are costly to develop and can only be applied to fairly circumscribed problem domains at present.

The teaching techniques just described do address the critical areas of clinical problem-solving which are in need of improvement. Yet, the use different measures of problem-solving ability and the relative lack of rigorous evaluation data on teaching efforts limits the conclusions that can be drawn from these studies. Unfortunately, these studies represent the "state of the art" in the field of medical education. More typically, "problem-solving techniques are...transmitted implicitly, with the expectation that the student will assimilate them by mimicking the observable practices of experts at work" (Kassirer, Kuipers, & Gorry, 1982, p. 257).

Application of information-processing paradigm to medical education. A fundamental problem with investigations of medical problem-solving instruction is that they are not firmly grounded in relevant theories of learning and problem-solving. These efforts might benefit from the fields of cognitive and educational psychology.

For example, Langley and Simon (1980, p. 368) suggest the following ways in which learning techniques might be employed to improve one's information-processing capacity:

1. Additions to or reorganization of the knowledge base.
2. Augmentation of the recognition mechanism, or index, for the knowledge base.
3. Augmentation of search strategies: Organized as production systems.

4. Modification of evaluation functions stored in memory and used to guide search.
5. (Apparent) augmentation of short-term memory capacity by storing new chunks in long-term memory.
6. Augmentation of lexical, syntactic, and semantic knowledge in language processing systems.
7. Enrichment of the representations of information (ways of organizing information) in memory.

Several of these modification strategies might be applicable to the learning of clinical information. Given the earlier discussion about the central role of knowledge representation in clinical problem-solving, the use of modifications to enrich knowledge representations seems a useful avenue to pursue. This can involve a restructuring of the material to be learned so as to improve the memorization of categories which resemble those used by experts. Information can be presented in the form of expert-like chunks with the important trigger cues and associations explicitly presented.

Another fruitful approach to teaching expert-like problem-solving would involve explicit instruction on general problem-solving strategies and the use of heuristics to simplify problems.

Although intervention studies in medical education have generally not taken an information-processing approach to improving knowledge representation, several basic and applied studies from cognitive psychology, and math and physics instruction address this topic. Wortman and Greenberg (1971) have shown that information stored in long-term memory can be reorganized according to a hierarchy specified by the experimenter. Subjects were presented with names of

16 common objects grouped on the basis of location, shape, material, and color. He found that after several recall trials, subjects reorganized the information in memory according to the hierarchy of categories he presented. Wortman showed that recoding this information involved three stages: a) perceiving the category hierarchy, b) chunking within a subordinate category, and c) establishing links between superordinate and subordinate categories. In addition, subjects who participated in a problem-solving task which highlighted the categorical relationships developed the organizational structure more quickly than those who did not.

Ausubel (1960) took a similar approach to subjects' learning and retention of an unfamiliar passage of text. He coined the term "advance organizers" to refer to the "advance introduction of relevant subsuming concepts" (p. 267). The purpose of advance organizers was to provide a conceptual framework at a general, abstract level within which detailed factual information, obtained later, could be integrated. Ausubel's thesis was that the provision of "cognitive scaffolding" prior to a learning task would lead to more efficient integration and retention of the material. Ausubel's results supported his hypothesis: Subjects provided with advance organizers scored significantly higher on knowledge tests based on the written material than did control subjects.

Shavelson (1972) examined the issue of knowledge structure change further. He was interested in the degree to which students' knowledge structures correspond to the structure of the course material after learning. Shavelson used an experimental control group, pre-posttest design in which the experimental subjects received instruction in physics over five days while the control subjects did not. Both groups were given achievement and word association tests each day. Shavelson found that following instruction, experimental subjects' cognitive

structures (based on the word association data) corresponded more closely to the content of the presented material and that key concepts were more strongly interrelated. Further, experimental subjects' achievement scores increased significantly from pre-test to post-test. The control group did not demonstrate any of these changes.

This research suggests that the organization of stimulus material is critical to retention and problem-solving. Further, organizational strategies based on key concepts seem to aid the learner by providing a framework within which to assimilate information for rapid retrieval during future problem-solving tasks.

Two additional studies used a slightly different approach to improving problem-solving skills but arrived at conclusions similar to the studies just reviewed. Schoenfeld (1980) and Schoenfeld and Hermann (1982) examined the impact mathematics problem-solving heuristics on students' problem-solving skills. In the first study, Schoenfeld assessed students' problem-solving performance before and after explicit instruction in the use of heuristics which are similar to those used by mathematics experts. These heuristics were found to improve students' performance as measured by increases in post-test scores on the math test. In the second study, Schoenfeld and Hermann assessed experimental and control subjects' performance on a card sort task and math test, before and after a mathematics problem-solving or structured programming course, respectively. The card sort task required subjects to categorize 32 math problems according to the similarity of the approach to their solution. The experimental group was taught problem-solving heuristics and a "systematic, organized approach" to mathematics problem-solving, while the control group was taught a "structured, hierarchical, and orderly way" to solve non mathematical problems using the computer. Post-test analyses showed that experimental

subjects performed better than control subjects on the math test. More important, experimental subjects sorted the problems on the basis of deep structure rather than surface structure more often than control subjects. Schoenfeld and Hermann used the term deep structure to refer to "the mathematical principles necessary for solution" (for example, solution by analogy, contradiction), and surface structure as "a naive characterization of a problem, based on the most prominent mathematical objects that appear in it (polynomials, functions, whole numbers) or the general subject area it comes from (plane or solid geometry, limits)" (p. 486). Thus, experimental subjects perceived math problems more like the experts and became more proficient at their solutions. Changes in knowledge representation and concommitant performance increases were not seen in control subjects.

The literature on instruction and knowledge structure changes has demonstrated that a) the pre-existing structure of knowledge in memory can be altered to resemble the structure provided through instruction, and b) expert-like knowledge representation strategies can be taught which facilitate retention of material and improved problem-solving performance.

The implication of these findings for medical education is clear. If the experts' representation of clinical knowledge in memory can be described, then it may be possible to design clinical instruction which fosters such knowledge representation in medical students. Just as Ausubel (1960) used advance organizers for the integration of information, medical educators might teach the use of specific problem formulations from which to gather and weigh clinical information. Similarly, Schoenfeld's (1980) use of problem-solving heuristics might be comparable to medical experts' use of well-learned rules of thumb which could also be taught to improve clinical problem-solving.

Statement of the Problem

The present study proposes that instruction involving the use of medical information structured in an expert-like fashion, along with clinical problem-solving heuristics, can improve the clinical problem-solving performance of medical students. It is hypothesized that students receiving this type of instruction will solve clinical problems more effectively and efficiently than students receiving "classical" clinical instruction on the same subject. The expert-like instruction is specifically hypothesized to:

1. Increase subjects' use of highly diagnostic clinical information while decreasing their use of noncontributory information (that is, information which is not highly predictive of a given diagnosis).

2. Increase the rapidity with which subjects generate the correct diagnosis for a case.

3. Increase the extent to which subjects evaluate the most likely groups of diseases for a case when presented with salient clinical information for that case.

4. Increase subjects' diagnostic accuracy.

5. Decrease the cost of the workups incurred by subjects.

Three simulated cases varying in their difficult of diagnosis will be presented to subjects in order to assess these dimensions of clinical problem-solving. Because previous research has shown that differences in diagnostic skill between experts and less experienced clinicians are only elicited with problems representing a moderate to high level of difficulty (Feltovich, 1983; Chase & Simon, 1973), it is hypothesized that subjects in the present study will show superior clinical

problem-solving performance only for the cases of above average difficulty. The level of difficulty of the cases will be determined primarily by their typicality, that is, the degree to which they are common and straightforward, having a classical "textbook" presentation, or they are rare, presenting as an unusual variation of a more common disease.

Because several abbreviations will be used throughout this paper, they are presented next in Table 1, along with brief definitions where appropriate. Following Table 1 are the operational definitions of the major dependent variables and the primary and secondary hypotheses.

TABLE 1

COMMONLY USED ABBREVIATIONS

LCS-- Logical competitor set.
This is a set of diseases which share underlying pathophysiology and thus, present with similar clinical findings. The use of LCSs in the diagnostic process is a hallmark of good clinical reasoning.

CDP-- Computerized diagnostic problem.

NOTE: The following eight diseases comprised subject's knowledge base for the present study. Detailed descriptions of these diseases are found in Appendix D and schematic diagrams of the lesions are found at the end of Appendix C.

ASD-- Atrial Septal Defect.

ECD-- Endocardial Cushion Defect.

PAPVC-- Partial Anomalous Pulmonary Venous Connection.

TAPVC-- Total Anomalous Pulmonary Venous Connetion.

VSD-- Ventricular Septal Defect.

PDA-- Patent Ductus Arteriosus.

PTA-- Persistent Truncus Arteriosus.

CTGV-- Complete Transposition of the Great Vessels.

Operational Definitions

1. Prototypic case— This is a case where the actual signs and symptoms for the target diagnosis closely match the classic signs and symptoms whose descriptions are based on the pathophysiology of the lesion. Prototypic cases can be found in most medical texts and often serve as the starting point in subjects' knowledge bases to which more detailed disease variations are later added. Because they are straightforward, prototypic cases are relatively easy to diagnose. Prototypic

cases also tend to be common. In the present study, Atrial Septal Defect served as a prototypic case.

2. Typical case-- This is a case where most of the actual signs and symptoms for the the target diagnosis match the textbook description of the disease. However, there may be some signs and symptoms normally associated with the disease that are ambiguous or absent. Typical cases vary in their difficulty of diagnosis, depending on the degree to which their clinical presentation matches their textbook description. Typical cases tend to be relatively common. Patent Ductus Arteriosus served as the typical case for this study.

3. Atypical case-- This is a case whose clinical presentation is quite different than that described in textbooks. In other words, many of the classic signs and symptoms for the disease are either ambiguous or absent. Atypical cases tend to be uncommon variations of more common (prototypic) diseases. For these reasons, atypical diseases are often difficult to diagnose. Total Anomalous Pulmonary Venous Connection served as the atypical case for this study.

4. Proficiency of critical cue acquisition-- the ratio of critical to total cues acquired. Critical cues are signs, symptoms, or laboratory test results which are highly diagnostic for a disease or set of similar diseases. Noncritical or noncontributory cues are signs, symptoms, or laboratory test results which may be associated with a disease or set of diseases but which are only weakly predictive for the disease or disease set. The measure of proficiency of critical cue acquisition indicates the degree to which a focused, efficient diagnostic approach is taken.

5. Proficiency of early hypothesis generation-- the percentage of total cues acquired before the correct diagnosis is first mentioned. Note that the correct diagnosis need only be mentioned, not necessarily evaluated for a specific cue. The variable of early hypothesis generation measures the rapidity with which the correct initial clinical problem is formulated.

6. Critical cue evaluation-- Evaluation of LCS members with respect to critical cues--This variable refers to the number of times LCS members are evaluated as positive, negative, or noncontributory with respect to the critical cues in the case. It is an indication of the extent to which subjects explicitly use critical cues to rule in or out diseases in the LCS.

7. Diagnostic accuracy-- diagnostic accuracy is operationalized as the proportion of subjects in each group who are correct in their final diagnosis.

8. Cost of workup-- cost is operationalized as the total cost of laboratory tests ordered by the subjects.

The methods for scoring each of the variables will be described in the results section prior to the presentation of the results for that particular variable.

Primary Hypotheses

1. Proficiency of critical cue acquisition: Experimental and control subjects will not differ significantly in the proficiency with which they acquire critical cues for the prototypic case. Experimental subjects will demonstrate significantly greater proficiency of critical cue acquisition than control subjects for the typical and atypical cases.

2. Proficiency of early hypothesis generation: Experimental and control subjects will not differ significantly in the rapidity with which they generate the correct hypothesis for the prototypic case. Experimental subjects will demonstrate significantly greater rapidity of early hypothesis generation than control subjects for the typical and atypical cases.

3. Critical cue evaluation: Experimental subjects will not differ significantly in the degree to which they evaluate LCS members with respect to critical cues for the prototypic case, but experimental subjects will have higher critical cue evaluation scores than control subjects for the typical and atypical cases.

Although two measures of proficiency and one measure of critical cue evaluation for LCS members will be the primary measures used, there are several additional measures of proficiency and cue evaluation that will be explored to aide in the interpretation of the results pertaining to the primary hypotheses.

Secondary Hypotheses

1. Diagnostic accuracy: Experimental and control subjects will be equally accurate in their diagnoses of the prototypic case while experimental subjects will be more accurate than control subjects in their diagnoses of the typical and atypical cases.

2. Cost of workup: The cost of the workup, primarily due to the number of lab tests ordered, will not differ significantly for experimental and control subjects for the prototypic case, but experimental subjects will incur significantly lower cost than control subjects for the typical and atypical cases.

METHOD

Overview of Design and Analysis

The purpose of this study was to assess an intervention designed to facilitate the development of expert-like clinical problem-solving skills in preclinical medical students.

A posttest-only control group desgin was used and Subjects were randomly assigned to an experimental or control group. Experimental and control subjects were given written material on eight congenital heart diseases to read for approximately four hours. The experimental intervention entailed the presentation of the disease material in a format that grouped together logically competing sets of diseases (LCSs), based on the similarity of their clinical presentation. The experimental group was also given a brief lecture on the characteristics of good clinical reasoning. Control subjects received the same material but it was presented in a text book format. These subjects did not receive a lecture on clinical reasoning.

The posttest involved an assessment of subjects' clinical reasoning on three simulated cases of congenital heart disease presented to them on a microcomputer. The dimensions of clinical reasoning that were assessed included proficiency of critical cue acquisition, proficiency of early hypothesis generation, the extent to which critical cues were evaluated with respect to LCS members, diagnostic accuracy, and cost of the workup.

Sample

Sample Characteristics

Subjects were pre-clinical medical students from three programs: The University of California, Berkeley-San Francisco joint medical program (UCB-UCSF), the University of California, San Francisco medical school (UCSF), and the Stanford University medical school (SU).

Thirty-Five subjects were recruited for the study. All were paid volunteers who received $25 for their participation. Eighteen subjects were randomly assigned to the experimental group and 17 to the control group. Table 2 presents the sample characteristics.

Subjects' familiarity with the clinical information on congenital heart diseases varied. Second and third year preclinical students had already taken a cardiology course and thus had some knowledge of congenital heart disease. However, this knowledge was usually limited and in no case did a subject have clinical experience in pediatric cardiology, although some subjects had clinical experience in other areas of medicine, public health, or nursing.

The use of subjects without previous extensive knowledge of the material allowed an examination of the way in which newly acquired knowledge is structured in memory as a function of the format of presentation. Moreover, this method is preferable to expert-novice comparisons because differences in problem-solving performance between experts and novices may be due to aptitude differences as well as to other factors (Schoenfeld & Hermann, 1982).

TABLE 2

SAMPLE CHARACTERISTICS

School	N
UCB-UCSF	10
UCSF	15
SU	10
	35
Year	
1	23
2	8
3	4
	35
Sex	
M	19
F	16
	35
Clinical Experience	
<= 1 month	23
2-6 months	6
7 months- 2 years	3
> 2 years	3
	35

Subject Recruitment

Subjects from the UCB-UCSF program were recruited with a letter describing the study. An interested instructor announced the study and distributed the letter to the first and second year students in his classes. Twenty-five students were contacted and ten, or 40%, chose to participate. However, students who participated were asked to contact other potential subjects, so the actual study

population could be considered to be the entire preclinical cohort of 36 students, equally divided into three years. Taking this more conservative estimate, the response rate was 28%.

It should be noted that only in the UCB-UCSF program are third year students preclinical. This is because of an extra preclinical year required for the completion of an M.S. degree in health sciences. In the other two programs, only the first and second years are preclinical. Of the two third year subjects from the UCSF and SU programs, one was on a leave of absence during the third year and the other had just begun his clerkships when he participated in the study. He had not yet done a clerkship in cardiology or pediatrics.

Subjects from the UCSF medical school were recruited with announcements posted near their mailboxes and ads placed in the school newspaper once a week for three weeks. When it became apparent that subject recruitment was more effective when done by word of mouth rather than with written announcements, this latter approach was stressed. After subjects had completed the study, they were asked to contact two or three classmates who might also be interested in participating. Because, without exception, subjects found the study interesting, the word of mouth approach was particularly effective.

Subjects were also recruited from a medical problem-solving course that was being taught to first year students. A letter describing the study was given to these students by their instructor and he announced the study in class. The course content was generally related to some of the ideas in the present study, but there was no specific material on pediatric cardiology.

The problem-solving course was used as a blocking factor for random assignment to experimental or control groups so that there would be no differential effect of the class participation on one of the groups.

Six subjects were recruited by word of mouth, eight were recruited from the medical problem-solving class, and one subject was recruited from the posted announcement. The potential first and second year population was 290, hence the response rate was 14/290, or approximately 5%. One student was from the third year class, although he was on leave and had not taken part in any clerkships.

Subjects from the SU medical school were recruited with announcements posted near student mailboxes and on bulletin boards. Again, the response was uniformly low, with only two students responding to the posted announcement. The other eight students were recruited by word of mouth from subjects. The potential population for this school was 172, evenly divided between years one and two. Thus, nine out of 172, or five percent, participated out of those who were potentially reached. As mentioned before, one third year student asked to participate, even though the posted announcement was directed toward first and second year students.

Although a response rate of between 5% and 28% is quite low, the assumption that the entire three populations of preclinical medical students were in fact alerted to the study is questionable. It is reasonable to assume that many of the students were not aware of the study announcements (no matter how strategically placed) nor were they contacted by a classmate who had participated.

The main reason offered by those actually declining to participate was pressure to study for midterms, final exams, or, in the case of second year students, board exams. This latter fact accounted for the relatively low participation rate of second year compared to first year students.

It is possible that the sample was not representative of the study populations because participants may have had more interest in clinical problem-

solving than nonparticipants. If so, this may have decreased the efficacy of the intervention because, in the absence of any intervention, control subjects may have been aware of the characteristics of good clinical problem-solving through their prior interest in this topic. Therefore, control subjects' performance may have been more similar to that of experimental subjects than would have been the case for subjects who were more naive to the topic of clinical problem-solving.

Random Assignment Checks

In order to check that random assignment within school and within year was accomplished, two chi square analyses were computed (group X year and group X school). Both were nonsignificant, indicating relatively equal distributions of subjects from each year and school in the experimental and control groups. Tables 3 and 4 display the numbers of subjects in the experimental and control groups by year and school.

TABLE 3

JOINT DISTRIBUTION OF YEAR IN SCHOOL BY GROUP

	Year		
Group	1	2	3
Experimental	11	5	2
Control	12	3	2
			35

X^2 = = .515, df = 2, p = .773

TABLE 4

JOINT DISTRIBUTION OF SCHOOL BY GROUP

Group	School		
	UCB-UCSF	UCSF	SU
Experimental	5	9	4
Control	5	7	5
			35

X^2 = = .333, df = 2, p = .847

Because prior clinical experience was also thought to be a potential confounder of the intervention, a chi square analysis was computed to check whether subjects with various amounts of clinical experience were randomly assigned to the two groups. The chi square was nonsignificant, indicating equal distributions of clinical experience in the experimental and control groups. Table 5 displays the numbers of subjects in the two groups by their clinical experience.

TABLE 5

JOINT DISTRIBUTION OF CLINICAL EXPERIENCE BY GROUP

	Clinical Experience			
	0-1 month	2-6 months	7 mos.-2 yrs.	> 2 Yrs.
Group				
Experimental	13	3	1	1
Control	10	3	2	2
				35

$X^2 = 1.03$ df = 3, p = .794

Procedure

Testing took place over a period of approximately four months. When possible, small groups of two to five subjects were scheduled for the learning and instruction sessions. Scheduling was flexible to allow subjects a choice of times according to their school schedules. Subjects were asked not to discuss the nature of the information they received in order to prevent discussion of the different instructional strategies used in the two groups, and the potential confounding results. Also, subjects were asked not to consult other pediatric cardiology reference materials in order to ensure that their knowledge bases were comparable. Instructional materials on pediatric cardiology were then given to subjects for review. The experimenter briefly reviewed the material with subjects and answered any initial questions about the material and study procedure.

Experimental subjects received information arranged to emphasize the learning of diseases with similar clinical presentation and the learning of sign and symptom clusters that are strongly associated with the diseases.

In addition, the experimenter gave these subjects a 30 minute lecture on clinical reasoning. The lecture was read to subjects from an outline in order to standardize its presentation. The lecture was not specifically oriented to congenital heart disease but, rather, was a general overview of the characteristics of good clinical problem-solving. The lecture emphasized the rapidity with which experienced clinicians generate diagnostic hypotheses in a clinical workup, their efficient organization of clinical information in memory, their use of critical cues to rule out or to confirm diagnostic hunches, and their appreciation for the considerable amount of variability in clinical findings for the same disease. An outline of this lecture is in Appendix A.

Control subjects received the same material on pediatric cardiology as experimental subjects, but it was structured in a more typical text book format. This approach classifies diseases according to pathophysiology (e.g., cyanotic and acyanotic diseases) and emphasizes the prototypic diseases within these classification schemes. Control subjects were not given a lecture on clinical reasoning. Instead, the time was spent discussing congenital heart diseases and their pathophysiology in general terms. If control subjects asked specific questions on diagnostic strategy, they were told that they should devise their own for the computer diagnostic simulation.

Following the introductory session, which lasted approximately 30 minutes, subjects were told that it was important to understand the material well enough to make simulated diagnostic decisions during the problem-solving assessment. Therefore, they were asked to study the written material for four hours on their

own. When possible, subjects spent 45 minutes to one hour reviewing the material before taking it home with them. This allowed them the time to ask the experimenter questions about the material before studying it in depth. Those subjects who were unable to study the material immediately after the introductory session were told that their questions would be answered when they returned for the simulated diagnostic session, before they began to go through the cases. Subjects were sufficiently motivated that allowing them to take the material home with them to study during their own time appeared to be the most effective approach, given that the majority had severe time constraints.

The mean number of hours subjects reported studying the material was 3.00, sd=.60. The mean number of hours reported studied was not significantly different for experimental and control subjects (2.98 and 3.03, respectively).

The problem-solving assessment took place between two and five days following the learning and instruction sessions. The problem-solving asssessment was carried out with three computerized diagnostic problems (CDPs) on congenital heart defects in infants and children. The format for the CDPs was developed by Guenin and Schwartz (1982).

Congenital heart diseases were chosen for the assessment for three reasons. First, these diseases provide relatively clear illustrations of the relationships between symptoms, pathophysiology, and the diagnosis. Second, the preclinical medical students had sufficient background knowledge of physiology and anatomy to assimilate the material on congenital heart diseases. Third, previous research on diagnostic expertise has been conducted in this knowledge domain and therefore, a basis for comparison of findings was available. A more detailed discussion of the construction of the CDPs is provided in the section on assessment of clinical problem-solving.

Subjects were assured that their responses on the computer would be confidential and anonymous. Therefore, they were asked to type in an identification number instead of their name. Subjects were then given written instructions on completing the CDPs (see Appendix B) and briefly familiarized with the computer. The CDPs contained additional instructions on how to proceed. Subjects were told that they could take notes during the problem-solving sessions if they found it helpful.

Each CDP began with a brief description of the patient's presenting complaint. The video display terminal then presented subjects with the categories of history, physical examination, and laboratory findings. Under each of these general categories were several subcategories such as "history of present illness" and "review of systems". Under the subcategories were additional subcategories, for example "chief complaint" and "murmur". These last categories contained the actual data items, for example, the nature of the chief complaint or the type of murmur found.

Subjects were asked to follow a diagnostic path typically used in clinical medicine, that is, to select history items first, proceed to the physical examination, and then order appropriate laboratory tests. After following this general path, subjects could return to any previous categories they wished in order to review earlier findings or select additional information prior to giving their final diagnosis. Because subjects were not yet familiar with interpreting EKG and X-ray findings, the experimenter provided a simplified explanation of them that was consistent with the description provided in the instructional material.

During the diagnostic problem-solving process, subjects were also asked to think aloud. Thus, prior to a request for patient information, subjects were

encouraged to discuss their reasoning and/or hypotheses behind that request. After subjects chose each data item, the experimenter read it aloud. Subjects were then asked to discuss any hunches they had regarding the item. If subjects had no hunches following an item, they were to continue to the next item. If subjects were silent but appeared to be thinking about an item, the experimenter prompted them with the statement, "please report your thoughts". Subjects' verbal responses were tape recorded for coding at a later time.

Subjects were encouraged to approach the CDPs as they might an actual clinical workup, taking into consideration "real world" constraints, such as the cost of the workup and the time involved.

When subjects had collected and reviewed patient findings to their satisfaction, they were asked for a primary and as many as two secondary hypotheses for each case. For each diagnosis given, subjects also provided a likelihood estimate for that diagnosis on a 5 point scale, ranging from "a little" likely to "very likely". All subjects diagnosed each of the three cases in the same order.

The first case was Atrial Septal Defect (ASD), a prototypic case, and the most straightforward of the three. It was considered prototypic because all of the important patient data items were consistent with this diagnosis. In other words, the clinical presentation of ASD on the CDP closely resembled its description in the knowledge base. There was no ambiguous or disconfirming information to complicate the diagnostic reasoning process. ASD was also considered a prototypic case because it is relatively common, accounting for 5 to 10% of all congenital heart disease, and because it receives more emphasis in pediatric cardiology texts than most other congenital heart diseases. In other words, this was a "textbook case" of ASD.

The second case was Patent Ductus Arteriosus (PDA). The case was modified to have some data items strongly confirmatory for PDA while others that were normally associated with PDA were ambiguous or absent. Because of the imperfect match of the data items to the actual diagnosis, this case was considered to be typical. That is, PDA represented the kind of case physicians are often presented with in an actual clinical setting. A secondary basis for the consideration of PDA as typical was that it has a relatively high incidence (10% of all congenital heart disease). It should be noted here that the typicality of the cases was based more on the degree of difficulty of diagnosis than on its incidence.

The third case was Total Anomalous Pulmonary Venous Connection (TAPVC). The case of TAPVC was considered atypical for two reasons. First, the clinical presentation of TAPVC was less severe than is normally portrayed in texts, particularly for a four and one-half year old child. Second, it is rare, accounting for only 1% all congenital heart disease.

After completing all of the cases subjects were informed of the the correct diagnosis for each and briefly queried about the difficulty of the material on congenital heart diseases, the way in which they studied the material, and the relative usefulness of the different instructional formats, such as the schematic diagrams of the heart, tables, flow diagrams, and introductory text. Subjects were reminded not to discuss the cases with anybody else until the study was completed.

Completion of the three CDPs took an average of one and one-half hours per subject.

Instructional Design

Selection of Knowledge Base

The use of a specific knowledge base to assess problem-solving is consistent with the literature in cognitive psychology which suggests that problem-solving skills are to a large extent dependent on the knowledge base to which they are applied, that is, they are context dependent (Anderson, 1981). Therefore, the assessment of general (domain free) problem-solving skills was thought to be of limited value for the present study.

Because the research questions for the present study are an extension of those posed by Feltovich in his work on expert-novice differences in diagnostic reasoning, it was decided to use part of the same knowledge base of pediatric cardiology as well as similar cases of congenital heart disease to assess subjects' diagnostic reasoning. In this way, a comparison could be made between characteristics of experts' and medical students' diagnostic reasoning after the latter have undergone an intervention designed to foster an expert-like approach to diagnostic problem-solving.

The instructional materials on pediatric cardiology were taken from three sources: Moller (1978), Moss, Adams, and Emmanouilides (1977), and Feltovich (1981). The first two sources are well-known tetxtbooks on pediatric cardiology. The third source describes the use of cases of congenital heart disease in an experiment which was designed to assess differences in diagnostic knowledge of physicians at different levels of training. The cases used in Feltovich's study were selected and modified for experimental use in collaboration with Moller, a professor of pediatric cardiology at the University of Minnesota and author of the first textbook cited.

Information on congenital heart diseases was abstracted primarily from Moller's textbook and this material served as subjects' knowledge base in the present study.

Because of constraints on subjects' time to learn the material and their lack of familiarity with it, it was decided to use a more limited subset of diseases than those described in Moller. Eight diseases were chosen: Ventricular septal (VSD), Patent Ductus Arteriosus (PDA), Endocardial Cushion Defect (ECD), Atrial Septal Defect (ASD), Partial Anomalous Pulmonary Venous Connection (PAPVC), Complete Transposition of the Great Vessels (CTGV), Total Anomalous Pulmonary Venous Connection (TAPVC), and Persistent Truncus Arteriosus (PTA). All of these diseases are congenital heart defects that cause increased pulmonary blood flow, either through the shunting of blood from the left-sided cardiac chambers to the right-sided chambers (left-to-right shunt) or from the mixing of pulmonary and systemic blood (admixture lesion). Detailed descriptions of these diseases can be found in Appendix D

The eight diseases were chosen for several reasons. First, they represented a group of related heart defects, each of which had one or more distinguishing symptoms, signs, or laboratory test findings. Therefore, subjects' evaluation of cases for one of these diseases necessitated an understanding of certain critical cues that could distinguish between them. The extent to which subjects acquired these critical cues relative to their acquisition of less important cues could then serve as one measure of their diagnostic proficiency.

Second, the diseases could be divided into different groupings, and subjects' use of these groupings was taken as evidence of the way in which the diseases were represented in memory. For example, the first five diseases listed above may be characterized as acyanotic diseases, that is, the patient shows a normal

pink color on examination because a sufficient amount of oxygenated blood is being pumped to the body. The last three diseases are cyanotic diseases because they lead to a mixture of oxygenated and unoxygented blood which is returned to the systemic circulation. This condition causes the patient to have a bluish color in the extremities.

Grouping based on the presence or absence of cyanosis is considered a classic categorization, that is, it is often taught in introductory texts. Other groupings of the diseases are based on such similarities as the type of heart murmur they produce, or the chamber(s) of the heart most affected by the defect. However, these latter disease groupings are not as widely taught, but they are often adopted by physicians after years of clinical experience. An example of such a category used in the present study was the group of shunts at the atrial level that caused increased blood flow to the right side of the heart. This group included ASD, ECD, PAPVC, and TAPVC. In the classical categorization of heart defects, the first three would be considered acyanotic while the last one would be considered cyanotic. Hence, it would be unlikely for subjects to actively consider TAPVC along with the others if they were using a classical disease grouping. However, because TAPVC shares many clinical findings with the others it also is a good candidate to consider. Thus, it can be seen that the use of the category of atrial level diseases with increased blood flow to the right side prevents the clinician from excluding a competing disease, TAPVC, in his or her differential.

In addition to the acyanotic-cyanotic and atrial level shunt categories, another category could be used to group diseases in the present study: ventricular level shunts (VSD, PDA, and PTA).

The groupings of atrial and ventricular level shunts were considered logical competitor sets because, in the presence of certain critical cues, they represented groups of diseases that should be considered together in a differential diagnosis.

A third reason for selecting the eight diseases for study was that they represented a continuum from common, easily diagnosed diseases to more rare and difficult ones. Commonness was based on incidence rates among all congenital heart disease, as well as the amount of textbook space devoted to the disease relative to other diseases. In general, difficulty was based on both the number of critical findings in common between the target disease and its logical competitors, and on the degree of ambiguity of these critical findings. If a disease had most of the same clinical findings as other LCS members, then it would be considered difficult to diagnose because diagnosis would be based on, at most, one or two key distinguishing findings. Such a case would be even more difficult to diagnose if its key distinguishing finding was either absent or ambiguous, that is, if it was difficult to judge as clinically significant.

There were exceptions to this definition of diagnostic difficulty. Although a few of the diseases in the knowledge base did have only one or two distinguishing findings, they were not considered difficult to diagnose because these key findings were always present with the disease, always absent with competing diseases, and were unambiguous.

In sum, the diseases chosen for use as subjects' knowledge base allowed an assessment of subjects' acquisition of critical cues, their evaluation of these cues with respect to logical competitor sets, and the degree of precision in their disease knowledge for discriminating among competing diseases.

Organization of Knowledge Base

Information for subjects' knowledge base was abstracted from Moller's (1978) and Moss et al's (1977) pediatric cardiology textbooks. Although some of the information was copied directly from the textbooks, most of it was condensed and clarified as much as possible. This process often involved excluding details that were not critical to understanding the diseases or excluding less common variants of the diseases. The construction and organization of the knowledge base was done in collaboration with a pediatrician who was familiar with research in medical problem-solving.

In order to determine that the knowledge base was understandable and at an appropriate level of difficulty for the time given to assimilate it, pilot tests were conducted with two psychology graduate students and four first year medical students. This process resulted in several revisions.

The revised knowledge base consisted of two parts. The first was an eleven page introductory section which included a brief description of the circulation of the normal heart, heart sounds, murmurs, EKG, pathophysiology and hemodynamics of congenital heart diseases, and schematic diagrams of the eight congenital defects. This information provided subjects with the necessary background information to understand the eight diseases. The introductory information was given in the same format to both experimental and control subjects. A copy of this section is in Appendix C.

The second part contained abstracted material on the eight congenital heart diseases. This material was organized differently for experimental and control subjects. A copy of the material for the experimental and control subjects may be found in Appendices D and E, respectively. It should be noted that the table of diseases and the two flow diagrams used by the experimental subjects were

reduced in size for the the Appendix. The actual size of the table was 11.5" X 17" and the flow diagrams were 8.5" X 14".

A brief description of the organization of the material for the two groups is provided next.

Experimental group. Material for the experimental group included a large table detailing the characteristics of the eight congenital heart diseases, and two flow diagrams which suggested a particular diagnostic pathway.

The table illustrated the relationship between each of the diseases and their findings from the history, physical examination, EKG, and X-ray tests. The purpose of the table was to enable subjects to quickly determine similarities and differences between the diseases in order to form logical competitor sets (LCSs). In fact, the first four diseases in the table formed one LCS while the second three formed another. The last disease, CTGV, could not easily be placed into an LCS because its clinical presentation varied considerably. However, it was not considered a particularly difficult disease to diagnose because it presented with intense cyanosis, and most infants who had the defect would not be expected to live longer than six months in the absence of corrective surgery.

The two flow diagrams were constructed to indicate an efficient diagnostic pathway. Small clusters of critical cues were used in the diagrams in order to lessen information processing demands on subjects' working memories. The diagrams also illustrated how diagnoses could be reached with a relatively small amount of salient clinical information.

The two diagrams were designed to complement each other. The first emphasized data-driven hypothesis generation. In other words, for the history, physical examination, and laboratory tests, combinations of three clinical findings were provided which were strongly associated with single diseases or groups of

diseases. Thus, Diagram 1 indicated which disease or diseases to consider when certain combinations of findings were present.

The second flow diagram focused on hypothesis-driven data gathering. This diagram illustrated which disease or groups of diseases were strongly associated with findings from the physical examination, EKG, and X-ray. Thus, Diagram 2 indicated which findings to expect when certain diseases were being actively considered as diagnostic hunches.

Both flow diagrams also emphasized the use of LCSs. Diagram 1 grouped the diseases as acyanotic or cyanotic in the history section of the workup. Diagram 2 grouped the diseases as atrial level or ventricular level shunts, thus crossing over the cyanotic classification scheme. For both diagrams, critical findings from the physical exam and laboratory tests further served to group and differentiate the diseases.

In order for subjects to learn the two diagrams well, they were provided with copies of the diagrams that were incomplete, that is, the key findings for each diagram were omitted from the boxes. Subjects were instructed to use the table of diseases as a guide to complete the key findings for each diagram (sets of diseases for Diagram 1 and expected findings for Diagram 2), and then to check their work with the already completed diagrams which they had been given. The purpose of this process was to help subjects learn data-hypothesis associations through active practice.

As mentioned earlier, another part of the experimental intervention consisted of a brief lecture on general characteristics of good clinical reasoning. An outline of this lecture is in Appendix A.

Control group. Material for the control group was organized along the lines of Moller's pediatric cardiology text. The diseases were presented serially,

the left-to-right shunts first, followed by the admixture lesions. No other categorization for the diseases was provided other than this classical one, based on the presence or absence of cyanosis. Information for each disease was subsumed under the categories of history, physical examination, EKG, X-ray, and summary.

Assessment of Clinical Problem-Solving

Cases

Three cases of congenital heart disease assessed subjects' use of the knowledge base for clinical problem-solving. Two of the cases, TAPVC and PDA, were based on actual patient records modified for use in an experimental setting. A third case, ASD, was constructed based on the information in Moller (1978). The cases of TAPVC and PDA were first developed for use by Feltovich, in collaboration with Moller, and were used in a study of clinical expertise. The cases were extensively tested on pediatric cardiology experts prior to their use in the study.

A few of the findings from theses cases were modified for the present study in order to make the cases somewhat more difficult, and thus to provide more variability in the dependent measures. The TAPVC case was modified only slightly for the present study and, therefore, subjects' responses to it may be compared to the responses of physicians in the earlier study.

The case of PDA, however, was substantially modified. First, some of the LCS members mentioned by physicians in the earlier study were outside the knowledge base of the present study. Thus, subjects used a more limited LCS. Second, although Feltovich used PDA as a straightforward case, it was made more difficult in the present study by increasing the ambiguity of some of the

critical findings. In particular, the continuous or machinery type murmur that is highly diagnostic of PDA was re-worded to exclude the terms "continuous" or "machinery type", so as not to immediately cue subjects to the correct disease. The primary reason for this modification was the finding from Feltovich's study that this case was too straightforward to elicit expert novice differences in diagnostic reasoning.

The cases represented a hierarchy of typicality ranging from prototypic (ASD), to typical (PDA), to atypical (TAPVC). The level of difficulty of diagnosis directly corresponded to the typicality of the case. The prototypic case was the easiest to diagnose, the typical case was moderately difficult, and the atypical case was the most difficult.

Although the cases differed in their typicality, all were designed to assess 1) the proficiency of subjects' critical cue acquisition and early hypothesis generation, 2) the degree to which subjects used appropriate LCSs when evaluating critical cues for a case, and 3) subjects' ability to correctly interpret critical cues that were ambiguous, in other words, the degree of precision of subjects' disease knowledge coupled with their understanding of the natural range of variability of clinical findings.

The individual cases, their critical cues, and LCS members are described below. More detailed descriptions of the three diseases may be found in Appendix C.

It is important to note that subjects may have correctly diagnosed the cases without having acquired all of the critical cues for them. Thus, the specification of cues as critical for the cases is a rule of thumb and not invariant. Also, although the critical cues and corresponding LCS members for the cases will be presented in a certain sequence, there were several equally successful diagnostic pathways that subjects could have taken for each case.

Atrial Septal Defect. The first case presented, ASD, was considered a prototypic case. The LCS for ASD included ECD, PAPVC, and TAPVC, although TAPVC was not as likely as the other members for this case because of its more severe clinical presentation.

Table 6 presents the introduction and the six critical cues for the case of ASD.

TABLE 6

INTRODUCTION AND CRITICAL CUES FOR THE ASD CASE

INTRODUCTION--The patient is a 5 year old white girl who weighs 37 pounds and is 44 inches tall. Her presenting problem is a murmur heard by her pediatrician.

CRITICAL CUES

History

1. Childhood illness: Had flu at age 2 1/2.

Physical Examination

2. Skin: No skin lesions; normal skin coloration.

3. Murmurs: Grade 2-3/6 systolic ejection murmur along upper left sternal border. Grade 2/6 mid to late diastolic murmur along left sternal border.

4. Auscultation: First heart sound has very loud component. Second heart sound is widely split all the time and appears fixed. The pulmonary component is a little prominent.

Laboratory Tests

5. EKG: Right axis deviation of +120 degrees, right atrial enlargement, and right ventricular hypertrophy. rSR' pattern seen in lead V1.

6. Chest X-Ray: Slight enlargement of right side of heart, increased pulmonary vasculature. Aorta and left atrium are normal size.

The case introduction for ASD was designed to lead the subject toward a less serious disease such as an acyanotic left-to-right shunt. The child was past the life expectancy for someone with CTGV, and her normal growth led away from the other more serious diseases such as TAPVC and PTA. A negative history for upper respiratory infections and the absence of cyanosis on physical examination further ruled out the possibility of a cyanotic admixture lesion.

The second heart sound and the murmur should have served to alert subjects to the possibility of an atrial level defect, such as ASD, ECD, or PAPVC. These auscultatory findings were also perfectly consistent with TAPVC, but this disease should have been given low priority in the LCS because the child had normal growth and appeared acyanotic.

The EKG findings of right atrial enlargement and right ventricular hypertrophy were additional evidence of right heart enlargement and were quite consistent with an atrial level shunt. Most important, the right axis deviation ruled out ECD, because this disease has the unique finding of left axis deviation on EKG.

The X-ray finding of right heart enlargement was confirmatory for an atrial level defect and the absence of an anomalous vascular shadow or cardiac silhouette definitively ruled out PAPVC and TAPVC.

The case of ASD was designed to assess subjects' proficiency of critical cue acquistion and early hypothesis generation, as well as their use of LCSs in a situation where the diagnosis could be arrived at with minimal uncertainty. Because this case was relatively straightforward, no differences were expected between experimental and control subjects on these measures. This prediction follows the findings of Feltovich et al (1983) and Chase and Simon (1973), which suggest that expert/novice differences in problem-solving are only elicited in problems characterized by at least a moderate level of difficulty.

Because the ASD case was considered the easiest to diagnose, it was presented first so that subjects could gain a sense of efficacy with regard to the diagnostic simulation and could become more familiar with the computer format.

Patent Ductus Arteriosus. The second case presented to subjects was PDA, a typical case. The LCS for PDA included two other diseases, VSD and PTA.

Table 7 presents the introduction and critical cues for the case of PDA. The relatively normal growth suggested a less serious disease. However, the finding of a higher than usual number of past respiratory infections could have led subjects to keep the more serious cyanotic diseases under consideration also.

The history of the mothers' pregnancy revealed a "flu" in the first trimester, which, although vague, was intended to lead subjects to a consideration of rubella, a disease associated with PDA. Because many of the subjects who acquired this cue interpreted it literally as the flu, and not a possible rubella, they were told that this "flu" might be other things, such as rubella.

The findings of dyspnea and cyanosis with increased activity suggested a moderate degree of congestive heart failure, which could be found with most of the eight diseases, and in particular, a large VSD, PDA, and the three cyanotic diseases, TAPVC, PTA, and CTGV. The cyanosis cue was important to interpret accurately because if it was interpreted as too serious (in other words, as central cyanosis), this finding would lead subjects to overemphasize the cyanotic diseases.

The finding of no cyanosis on physical exam was critical for ruling out CTGV, because it presents with intense cyanosis. At this point, a good LCS might have included VSD, PDA, and PTA. PTA was included here because, although it is a cyanotic disease, it does not always present with cyanosis in the first six months, and the child was six months old.

TABLE 7

INTRODUCTION AND CRITICAL CUES FOR THE PDA CASE

Introduction--The patient is a 6 month old Hispanic girl weighing 14 pounds, 2 ounces, with a length of 24 inches. A murmur was heard during her 6 month well baby visit.

History

1. Skin appearance or exercise tolerance (findings same for both): On questioning, mother says occasionally baby gets breathless and blue around the mouth during exertion.

2. Childhood illness: Number of past respiratory infections somewhat higher than usual.

3. Pregnancy: Full term uncomplicated pregnancy. Mother remembers "flu" in second month.

Physical Examination

4. Skin: No skin lesions; normal skin coloration.

5. Murmurs: Harsh grade 3/6 systolic murmur coupled to a grade 1/6 diastolic murmur heard in left infraclavicular area. Reaches peak intensity at about second heart sound.

6. Auscultation: A systolic ejection click is present and there is a loud pulmonary component of the second sound.

Laboratory Tests

7. EKG: Left atrial enlargement, biventricular hypertrophy.

8. X-Ray: Cardiomegaly; increased pulmonary vasculature; left atrial and left ventricular enlargement. Enlarged aorta.

The harsh systolic murmur should have ruled out ASD, ECD, PAPVC, and TAPVC, because the murmur was uncharacteristic of these diseases. The murmur was vaguely characteristic of VSD, PDA, and PTA. However, it was not explicitly stated to be either a continuous murmur (positive for PDA) or a

pansystolic murmur (positive for VSD and possibly PTA). Here it was important for subjects to remember that the continuous murmur of PDA is not always heard in the first six months of life.

The auscultatory findings of a systolic ejection click and split second heart sound were the most critical to the case. The systolic ejection click, indicating the presence of a dilated aorta, was confirmatory for either PDA or PTA. Hence, these two diseases were the strongest competitors at this stage of the workup. The split second heart sound was strongly disconfirmatory for PTA because PTA is always associated with a single second sound.

The EKG and X-ray findings were generally consistent with the LCS of VSD, PDA, and PTA, with one exception: Aortic enlargement on X-ray is uncharacteristic of VSD, and therefore should have served to rule it out.

Unlike ASD, the case of PDA required correct interpretation of ambiguous cues and the integration of cues from different parts of the workup in order to arrive at the correct diagnosis. Despite the differences betweeen this case and the prior case, the analysis again focused on subjects' proficiency of critical cue acquisition and early hpothesis generation, and their use of LCSs. Perhaps most important to successful diagnosis of this case was subjects' ability to interpret ambiguous clinical information.

Total Anomalous Pulomnary Venous Connection. The case of TAPVC was considered atypical because of its rarity and its unusually mild clinical presentation on the CDP. Table 8 presents the introduction and critical cues for the case of TAPVC.

The LCS for TAPVC was the same as that for the ASD case, consisting of ASD, ECD, and PAPVC. Because of this, the true disease was likely to be confused with other less severe members of the LCS, such as ASD, ECD, and especially

TABLE 8

INTRODUCTION AND CRITICAL CUES FOR THE TAPVC CASE

INTRODUCTION--The patient is a 4 year old Asian girl. She weighs 33 pounds (slightly underweight) and is 41 inches tall. She was referred to you for evaluation of a murmur.

History

1. Skin appearance: Mother notes that in last 2 years, when child is cold her lips turn blue.

2. Childhood illnesses: Between age 2 and 3 had numerous infections, including flu, upper respiratory infections, and otitis; or Hospitalizations: Required several hospitalizations for upper respiratory infections between age 2 and 3.

Physical Examination

3. Skin: No skin lesions; slight bluish coloration to lips and fingernails, although this is hard to evaluate as abnormal since child's skin pigmentation is dark.

4. Murmurs: Grade 2-3/6 systolic ejection murmur along upper left sternal border. Grade 2/6 mid to late diastolic murmur along left sternal border.

5. Auscultation: First heart sound has very loud component. Second heart sound is widely split all the time and appears fixed. The pulmonary component is a little prominent.

Laboratory Tests

6. EKG: Right axis deviation of +135 degrees, right atrial enlargement, and right ventricular hypertrophy. rSR' pattern in lead V1.

7. Chest X-Ray: Moderate cardiomegaly; markedly increased pulmonary vasculature. Unable to evaluate size of left atrium. Unusual vascular shadow seen in right side.

PAPVC. However, the integration of a few pieces of information from the history provided clues to the patient's increased symptom burden. The

introduction to the case mentioned that the girl was underweight, a possible sign of a more serious defect. A history of frequent respiratory infections provided more evidence of a serious disease. The evidence of peripheral cyanosis on history was important only in light of the other data; by itself, it was not an especially diagnostic finding.

The physical exam and laboratory findings should have further enabled subjects to narrow their differential diagnosis to TAPVC. The appearance of mild cyanosis on physical exam, even in its ambiguous form, indicated a cyanotic disease. The characteristics of both the murmur and heart sounds pointed to the group of atrial level shunts, of which TAPVC was a member. As with the first case, the EKG finding of right axis deviation ruled out ECD. The presence of an unusual vascular shadow on X-ray ruled out ASD and provided confirmatory evidence for some type of anomalous pulmonary venous connection, either PAPVC or TAPVC. This cue was ambiguous and by itself did not provide sufficient evidence to differentiate between these two defects. It was therefore necessary to place sufficient weight on the historical findings of increased symptom burden, that is, the history of upper respiratory infections and mild cyanosis on physical examination. These findings were more likely to occur with TAPVC than PAPVC.

The case of TAPVC focused on subjects' ability to include an atypical disease in their LCS of atrial level shunts so that it could be actively considered. To the extent that subjects did not actively consider TAPVC, they were more likely to consider PAPVC, a plausible but incorrect alternative.

As with the case of PDA, the case of TAPVC assessed the accuracy of subjects' interpretation of ambiguous clinical information and their synthesis of other data from the workup to aid in their interpretation of such information.

Apparatus

The three cases were adapted for use on an Apple IIE computer with 128k memory. The software used to create the computer cases, called Computer Assisted Medical Problem Solving (CAMPS), was developed by Guenin and Schwartz (1982), and was written in Pascal.

CAMPS is a general software package that allows the creation of an unlimited number of simulated cases by altering relevant patient data items from normal to abnormal. The CAMPS sofware consists of three floppy diskettes that contain 500 patient data items for a normal infant or child, and the structure within which to create a simulated case. The data items are organized into the categories of history, physical examination, laboratory tests, treatment, and consultation. The creation of the cases for the present study entailed altering approximately 25 data items per case.

Because the purpose of the present study was to examine diagnostic thinking and not patient management, the categories of treatment and consultation were eliminated. In addition, certain highly diagnostic laboratory tests, such as the echocardiagram, were deleted from the laboratory section so that subjects would not be given the diagnoses automatically. Other highly diagnostic laboratory tests, such as cardiac catheterization, were altered to read "Results available upon diagnosis". This allowed an assessment of subjects' lab costs without actually giving them the test results.

It is recognized that making certain highly diagnostic laboratory tests unavailable to subjects limits the generalization of these results to actual laboratory test ordering for the types of cases under consideration. However, the primary purpose of the study was to assess diagnostic reasoning and this would have been impossible if the subjects were given the option of ordering one or two tests that automatically provided a definitive diagnosis for them.

The CAMPS program records the number of cues acquired in the history, physical examination, and laboratory sections of the workup. It also tallies the cost of the laboratory tests ordered. A scoring system was developed by Guenin and Schwartz, based on the averaged responses of five experts who work through the simulated cases. This scoring system was not used for two reasons. First, the dependent measures already proposed for the study were derived from research on diagnostic expertise, and second, the scoring system was not particularly relevant to the dependent measures in this study.

Data and Analysis

Data was obtained from the the CAMPS program itself and from cassette tapes which recorded subjects' "thinking aloud" about the cases.

The following data was recorded on the CAMPS data diskettes: The number of cues acquired in the history, physical exam, and laboratory sections, the cost of the workup, and the primary diagnosis given. These data were transcribed on to computer code sheets for statistical analyses.

A cassette recorder was used during the diagnostic problem-solving sessions to record subjects' "thinking aloud" about the cases. Data obtained from the cassette tapes included the following for each case:

1. Which cues were acquired in the workup, and thus, the critical cues acquired and the ratio of critical to total cues acquired.

2. The number of cues acquired before the first correct diagnostic hypothesis was mentioned, and the percentage of total cues acquired before this hypothesis was mentioned.

3. The evaluation of critical cues with respect to LCS members. This measure was computed as the sum of all the evaluations of LCS members made in response to the critical cues of the cases. The evaluations were coded as positive, negative, or noncontributory. Although a given LCS member was coded only once for the same critical cue, it was coded each time it was evaluated for a different critical cue.

4. The evaluation of the correct disease with respect to critical cues. This was computed as the number of times the correct disease was evaluated as positive, negative, or noncontributory with respect to the critical cues that were associated with it.

5. Errors of critical cue evaluation. Errors of critical cue evaluation included 1) statements that a critical cue was positive for an LCS member when it was negative or noncontributory, 2) statements that a critical cue was negative or noncontributory for an LCS member when it was positive, and 3) statements that a cue was more diagnostic of one LCS member than it really was.

6. The total number of LCS members mentioned.

7. The highest number of LCS members evaluated as positive, negative, or noncontributory for any one critical cue.

8. Secondary diagnoses given and likelihood estimates for both primary and secondary hypotheses.

RESULTS

Preliminary Analyses

The analyses done previously demonstrated that random assignment was accomplished within year in school, level of prior clinical experience, and type of medical school. Therefore, there is no reason to suspect that these variables would differentially affect the relationship of experimental or control group membership to the major dependent measures, and thus bias the results.

However, it was of interest to determine whether these variables were in fact related to the major dependent measures, though they were not of primary importance to the study. Of specific interest was whether more senior students, and those with more prior clinical experience, would show greater proficiency of critical cue acquisition and early hypothesis generation, and would evaluate more LCS members with respect to critical cues, than would first year students and those with less clinical experience.

In addition, the question of whether differences existed between the three medical schools on the major dependent variables was of interest, although there were no specific hypotheses for the presence or direction of such differences.

Pearson correlations were computed in order to assess the relationship of year in school and clinical experience to the three major dependent variables: the ratio of critical to total cues acquired, percentage of total cues acquired before the correct hypothesis was first mentioned, and the number of LCS members evaluated for critical cues.

Table 9 shows the correlations of year in school and clinical experience with the major dependent variables.

TABLE 9

PEARSON CORRELATIONS OF YEAR IN SCHOOL AND CLINICAL EXPERIENCE WITH PROFICIENCY OF CRITICAL CUE ACQUISITION, EARLY HYPOTHESIS GENERATION, AND CRITICAL CUE EVALUATION

	Critical cue Acquisition			Early Hypothesis Generation			Critical Cue Evaluation		
Year	ASD	PDA	TAPVC	ASD	PDA	TAPVC	ASD	PDA	TAPVC
	.070	-.061	-.088	-.094	-.510*	.035	-.082	-.188	.025
Clinical Experience									
	.038	-.009	-.072	-.488*	.075	.114	-.005	-.175	-.076

*p = .001

With two exceptions, there were no significant relationships between subjects' year in school or amount of prior clinical experience, and the three dependent measures.

The first exception is that for the typical case, more senior subjects tended to acquire a smaller percentage of their total cues before mentioning the correct diagnosis than did less senior subjects (p = .001).

The second exception was that for the prototypic case, subjects with more clinical experience tended to acquire a smaller percentage of their total cues

before mentioning the correct diagnosis than did subjects with less clinical experience ($p = .001$).

Although these findings are not part of any general pattern, they suggest that for some of the cases, year in school and clinical experience may be related to efficiency of early hypothesis generation. Caution should be used when interpreting these results because several tests were performed, thus increasing the occurrence of a significant result by chance.

Because the type of medical school was a nominal variable, its relationship to the three dependent variables was assessed with three analyses of variance. Each of the ANOVAs had two factors, school with three levels, and case with three levels. The case factor was a repeated measure.

The ANOVAs showed no significant main effects of the type of medical school on any of the major dependent variables for any of the cases. The p values for the main effects of school for the three ANOVAs were .824 for critical cue acquisition, .526 for early hypothesis generation, and .512 for critical cue evaluation.

Primary Hypotheses

Analyses for Primary Hypotheses

Three two factor ANOVAs (group with two levels and case with three levels) were used to analyze subjects' scores on the three dependent measures. The case factor was a repeated measure because all subjects diagnosed all three cases.

Orthogonal contrasts were set up for the case factor which best represented the hypotheses. The first contrast assessed possible differences between the prototypic case and a combination of the typical and atypical cases. The second contrast assessed possible differences between the typical and atypical cases.

The assumptions necessary for ANOVA with a between and a within subjects factor were met for all three analyses (Norusis, 1985): 1) The dependent variables were normally distributed, 2) non significant Box's M tests demonstrated the equality of variance-covariance matrices at all levels of the independent variables, and 3) Bartlett's tests of sphericity were nonsignificant, indicating that the variances of the transformed within subjects variables were equal and their covariances were 0.

The ANOVA results for the each of the major hypotheses will be presented in the following manner: For each ANOVA, average F tests, pooled over the specific contrasts, will be presented first for the main effects and interactions. This allows an assessment of the overall effect of the intervention, averaging across contrasts (Norusis, 1985).

For significant main effects of group, the simple effects of group within each level of case will be explored to determine where the actual group differences lie.

For significant group by case interactions or main effects of case, the specific orthogonal contrasts will be examined to determine where the actual significant differences can be found.

It should be noted that there were no specific predictions regarding any main effects of case, and that although these results will be reported, they are of secondary importance to the study hypotheses.

An alpha level of .05 was adopted as the criterion significance level for the analyses. Although an effort was made to limit the number of overall statistical tests performed to three (one ANOVA for each major hypothesis), the possibility of significant results due to chance remains. Therefore, the strength of significant findings as well as the logic of their interpretation should be weighed carefully.

Additional Measures

Although two measures of proficiency and one measure of critical cue evaluation for LCS members were analyzed, there were several additional measures of proficiency and cue evaluation that were explored to aide in the interpretation of the ANOVA results. In order to limit the experiment wise error rate, these measures were not used for hypothesis testing (and were not statistically analyzed), but rather, served to illuminate the findings for the major hypotheses. These additional measures were the following:

1. Number of cues acquired in the history, physical exam, and laboratory sections of the workup.
2. Total number of LCS members mentioned.
3. Highest number of LCS members evaluated for any one critical cue.
4. Number of critical cues acquired for each case.
5. Number of evaluations of the actual disease for each case with respect to the critical cues for that case.
6. Number of errors of critical cue evaluation.

Proficiency of Critical Cue Acquisition

The ratio of critical to total cues acquired for each case served as the the dependent measure for this analysis. This score was derived by dividing the number of critical cues that each subject acquired for a case by the total number of cues he or she acquired. Table 10 shows the means, standard deviations, and ranges of this variable for the experimental and control groups on the three cases.

Table 11 displays the ANOVA results for the measure of critical cue acquisition.

TABLE 10

PROFICIENCY OF CRITICAL CUE ACQUISITION FOR EXPERIMENTAL AND CONTROL SUBJECTS ON THE THREE CASES

	$\bar{X}$	sd	range
ASD			
Experimental	.536	.202	.200-.860
Control	.356	.192	.140-.710
PDA			
Experimental	.617	.237	.300-.990
Control	.403	.227	.180-.890
TAPVC			
Experimental	.672	.221	.210-.990
Control	.396	.239	.140-.990

It can be seen that the main effect of group was highly significant ($p = .003$), indicating that the ratio of critical to total cues acquired was greater for experimental than control subjects.

A further examination of the simple effects of group within case was done in order to determine for which cases the group differences were significant. These results showed significant group differences across all cases (prototypic, $p = .011$; typical, $p = .010$; atypical, $p = .001$). An examination of the mean group differences showed an increase from the prototypic to the atypical cases, suggesting more pronounced differences at increasing levels of case difficulty.

TABLE 11

ANALYSIS OF VARIANCE FOR PROFICIENCY OF CRITICAL CUE ACQUISITION

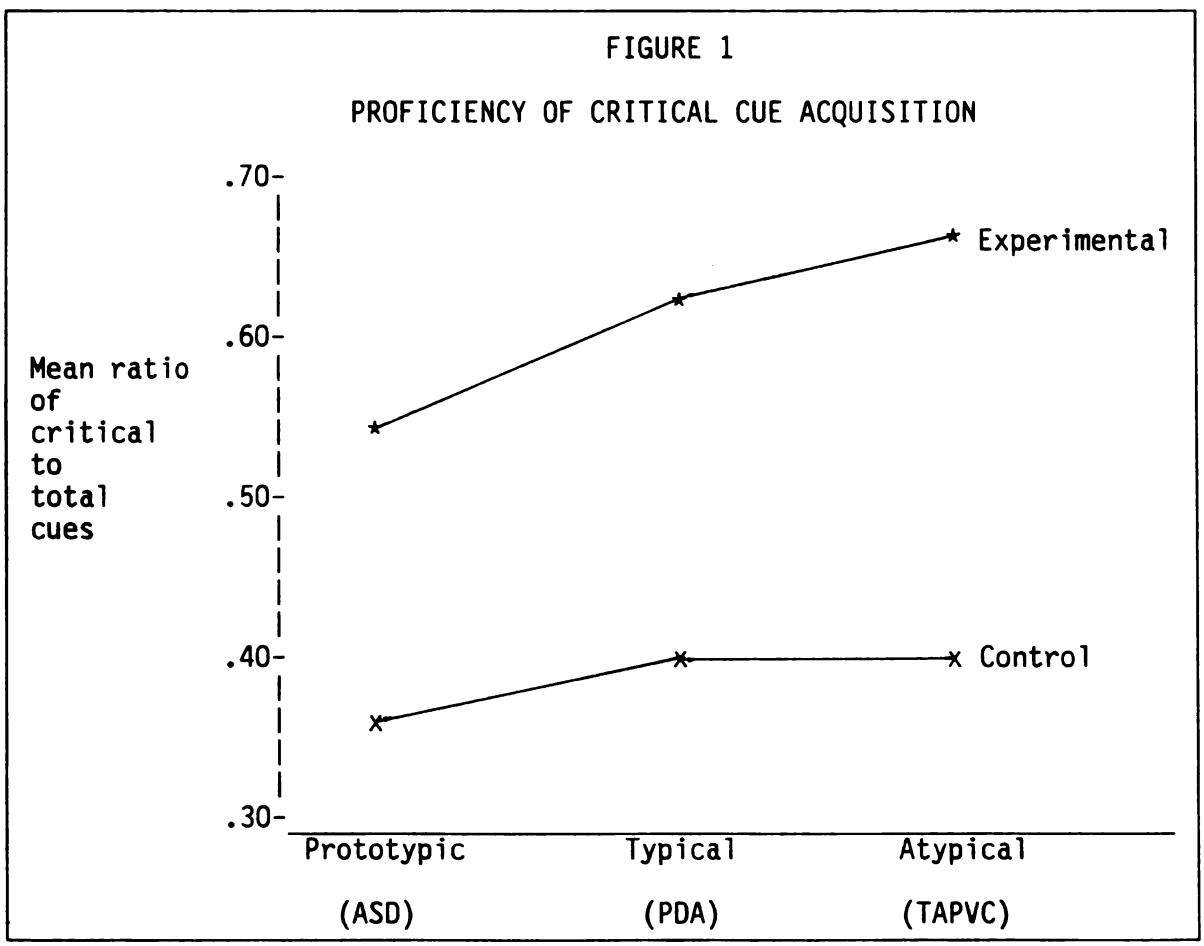
Source of Variance	Sum of Squares	df	Mean Square	F	p
Between Subjects					
Within Cells	4.29842	33	.13026		
Constant	25.89478	1	25.89478	198.80031	.0005
Group	1.30924	1	1.30924	10.05173	.003
Within Subjects					
Within Cells	.50443	66	.00764		
Case	.14392	2	.07196	9.41518	.0005
ASD vs. PDA, TAPVC				12.55299	.001
PDA vs. TAPVC				2.17409	.150
Group by Case	.04197	2	.02099	2.74580	.072
ASD vs. PDA, TAPVC				2.35696	.134
PDA vs. TAPVC				3.64310	.065

The results pertaining to the group by case interaction indicate that this interaction was not significant, although a trend was apparent ($p = .072$). The specific contrasts for the group by case interaction revealed that the increase in the ratio of critical to total cues acquired from the typical to the atypical cases was somewhat greater in the experimental than in the control group ($p = .065$). A look at the differences between experimental and control subjects' mean ratio scores on the three cases suggests that the group differences were in fact

greatest when comparing the prototypic and atypical cases. However, this specific contrast was not formally incorporated into the ANOVA.

The average F test for the main effect of case was also highly significant (p = .0005). An examination of the case contrasts indicates that the ratio of critical to total cues acquired was significantly lower for the prototypic case compared to the combination of the typical and atypical cases (p = .001). The contrast comparing the typical and atypical cases was not significant on this measure (p = .150).

Figure 1 provides a plot the group means across the cases.



To summarize these findings, experimental subjects were significantly more proficient than control subjects in their acquisition of critical cues relative to total cues, and this difference was evident across all three cases, although it was not predicted to occur for the prototypic case. Thus, the hypothesis of group differences on the proficiency of critical cue acquisition was supported.

The finding of a significant overall effect of case was not predicted. It points to an increase in the ratio of critical to total cues acquired for the prototypic case compared to the typical and atypical cases, over all subjects. These findings suggest a possible learning effect, that is, as subjects became more familiar with the structure and information of the initial case, they became more proficient in their critical cue acquisition for the subsequent cases.

The results thus far have shown group differences on the measure of proficiency of critical cue acquisition. However, several questions remain which can only be clarified through an examination of subjects' cue acquisition in more detail. First, did experimental subjects acquire more critical cues relative to total cues than control subjects, or did they acquire fewer noncritical cues relative to critical cues? Second, in what part of the workup were the differences in cue acquisition most pronounced? Third, was proficiency of critical cue acquisition related to diagnostic accuracy?

Table 12 presents a breakdown of experimental and control subjects' cue acquisition in different parts of their workups for the three cases.
In answer to the first question, the mean number of critical cues acquired for both groups was roughly comparable. Moreover, both groups acquired most of the critical cues that were available for each case: The total sample of subjects acquired 83% of the critical cues for the prototypic case, 79% for the typical case, and 84% for the atypical case. This finding is not surprising because the

TABLE 12

CUE ACQUISITION OF EXPERIMENTAL AND CONTROL SUBJECTS IN DIFFERENT PARTS OF THE WORKUP FOR THE THREE CASES

	ASD			PDA			TAPVC		
	$\bar{X}$	sd	Range	$\bar{X}$	sd	Range	$\bar{X}$	sd	Range
History									
Experimental	4.72	2.97	1-12	5.22	3.37	2-14	4.94	3.56	2-15
Control	8.76	6.62	2-28	12.00	8.82	1-28	10.59	8.21	1-31
Physical Exam									
Experimental	3.72	2.35	1-11	4.44	2.26	2-9	4.00	2.09	1-10
Control	7.65	3.97	1-14	7.65	3.26	2-12	7.35	3.55	1-12
Laboratory									
Experimental	1.89	.32	1-2	1.83	.62	1-3	1.89	.76	1-4
Control	2.41	.94	1-5	2.06	.97	1-4	2.24	.90	1-5
Total Cues									
Experimental	10.33	4.84	4-22	11.50	5.25	5-22	10.33	5.48	4-29
Control	18.82	10.16	6-42	21.71	12.05	5-43	20.18	11.32	5-44
Critical Cues									
Experimental	4.83	1.04	2-6	6.06	1.00	4-8	5.94	.87	4-7
Control	5.12	.86	4-6	6.53	1.28	4-8	5.76	.97	4-7

critical cues were straightforward and logical, such as the type of murmur, the

presence of cyanosis, or the EKG and chest X-Ray findings. The results thus indicate that experimental subjects acquired fewer noncritical cues relative to critical cues than did control subjects.

With regard to where in the workup the major group differences in cue acquisition were found, Table 12 indicates that for all three cases, they were in the history and physical exam, and not in the number of laboratory tests ordered. This finding might be expected because the intervention focused on efficient cue acquisition in the history and physical exam and not in the laboratory section of the workup.

An additional unexpected finding was that the variability in cue acquisition for the history was approximately twice as great for control subjects compared to experimental subjects. The meaning of this finding is not entirely clear. It is possible that control subjects followed their own idiosyncratic styles of cue acquisition in the history, many of them acquiring a large number of cues without any particular idea of what information they were looking for (in other words, being "data-driven"), while others spent relatively little time on the history, believing that the information it contained was not especially useful.

In contrast, exerimental subjects may have used the flow diagrams as a guide to what specific information, and how much of it, was important to acquire in the history.

A final question regarding proficiency of critical cue acquisition warrants investigation: Was there a loss of diagnostic accuracy among subjects who were more proficient in their critical cue acquisition? This question is important because more proficient critical cue acquisition is only useful when it does not lead to a reduction in diagnostic accuracy. T-tests were used to compare subjects who were correct versus those who were incorrect in their diagnosis for

the typical and atypical cases. The prototypic case case was not analyzed because only one subject misdiagnosed it. The results showed that for the typical case there were no significant differences in proficiency of critical cue acquisition between correct and incorrect subjects, $t = -.86$, $df = 33$, $p = .397$. For the typical case, then, proficiency in critical cue acquisition was unrelated to diagnostic accuracy. For the atypical case, correct subjects had a significantly higher mean ratio of critical to noncritical cues than incorrrect subjects, $t = 2.45$, $df = 33$, $p = .020$. Thus, at least for this case, acquiring additional noncritical cues to "get a better picture of the patient" (as several subjects stated), might have even led subjects away from the correct diagnosis.

These results provide reassurance that teaching medical students to be more proficient in their critical cue acquisition will not result in a sacrifice of diagnostic accuracy and, in fact, may enhance it for some cases.

Proficiency of Early Hypothesis Generation

The dependent measure of early hypothesis generation was operationalized as the percentage of total cues acquired in the workup before the subject mentioned the correct disease for a particular case. In order to receive a score on this measure, the subject needed only mention the disease, not evaluate it with regard to any specific cues.

With few exceptions, subjects generated the correct disease at some point their workups for each of the cases. This does not mean, however, that subjects always ended up specifying the correct disease as their primary diagnosis for a case.

The use of a dependent measure based on the percent of total cues was preferred to one based strictly on the number of cues acquired prior to hypothesis

generation, because the latter score is confounded with subjects' cue acquisition scores. In other words, if subjects' scores were based on the number of cues acquired prior to mentioning the correct hypothesis, then those who acquired the fewest cues in the workup would also tend to have acquired the fewest cues prior to their hypothesis generation. The use of a score based on the percent of total cues acquired removes the confounding effect of subjects' cue acquisition.

Table 13 displays the means, standard deviations, and ranges for the early hypothesis generation measure for experimental and control sbjects.

TABLE 13

PROFICIENCY OF EARLY HYPOTHESIS GENERATION FOR EXPERIMENTAL AND CONTROL SUBJECTS ON THE THREE CASES

ASD	$\bar{X}$	sd	range
Experimental	.484	.183	.048- .727
Control	.541	.259	.111- .905
PDA			
Experimental	.634	.296	.067-1.00
Control	.718	.290	.083-1.00
TAPVC			
Experimental	.469	.218	.077-1.00
Control	.670	.291	.034- .978

Table 14 shows the ANOVA results for the early hypothesis generation measure.

TABLE 14

ANALYSIS OF VARIANCE FOR PROFICIENCY OF EARLY HYPOTHESIS GENERATION

Source of Variance	Sum of Squares	df	Mean Square	F	p
Between Subjects					
Within Cells	2.62409	33	.07952		
Constant	36.03109	1	36.03109	453.11847	.0005
Group	.33774	1	.33774	4.24739	.047
Within Subjects					
Within Cells	4.01681	66	.06086		
Case	.47962	2	.23981	3.94034	.024
ASD vs. PDA, TAPVC				4.99509	.032
PDA vs. TAPVC				3.02673	.091
Group by Case	.10198	2	.05099	.837784	.437
ASD vs. PDA, TAPVC				.74660	.394
PDA vs. TAPVC				.91688	.345

The main effect of group was significant (p = .047), suggesting that experimental subjects mentioned the correct diagnosis earlier in their respective workups than control subjects. Further analysis of the simple effects of group within case showed that experimental subjects were quicker than controls in their early

hypothesis generation, and that these differences increased from the prototypic to the atypical cases. However, experimental-control differences in early hypothesis generation only attained significance for the atypical case ($p = .027$). The p values for the simple effects of group within the prototypic and atypical cases were .457 and .407, respectively.

The average F test for the group by case interaction was nonsignificant, suggesting that the group difference for the prototypic case compared to the combined typical and atypical cases nonsignificant, as was the group difference for the typical compared to the atypical case.

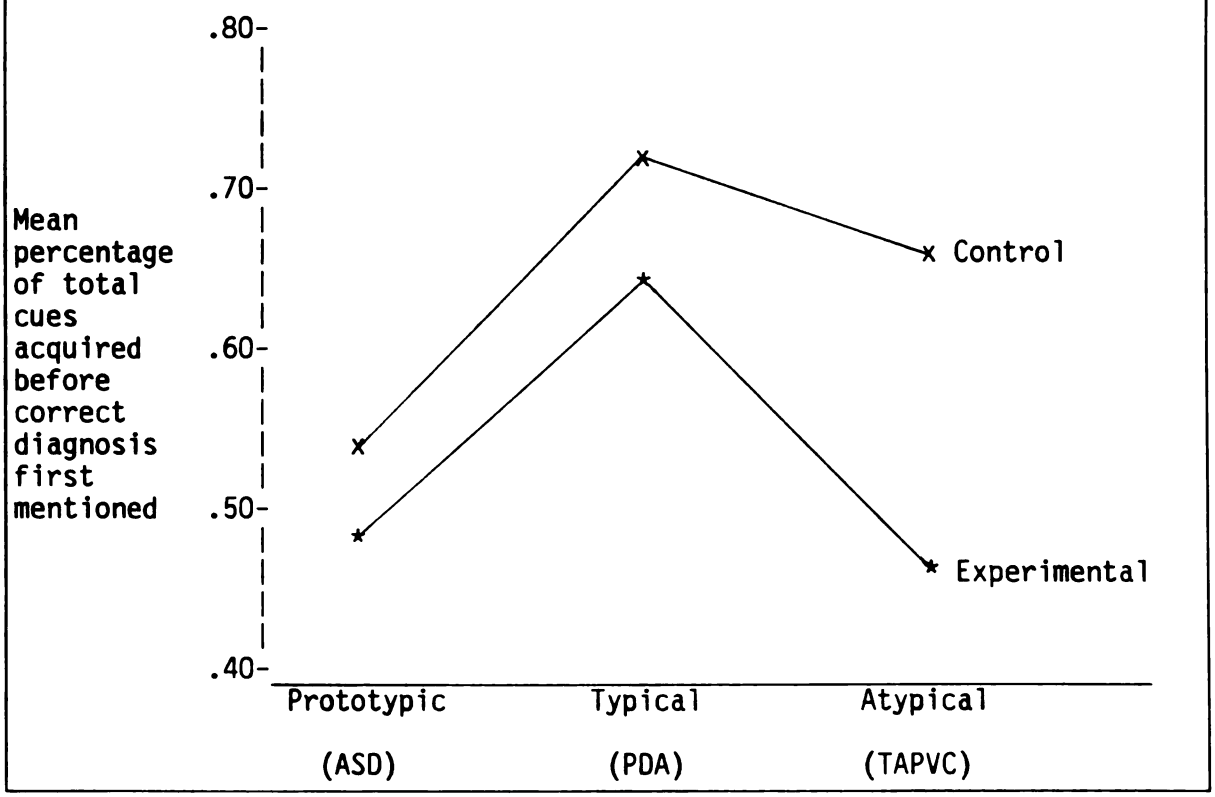
Finally, the average F test for the main effect of case was significant ($p = .024$). An examination of the specific case contrasts showed that the variance in case differences was accounted for primarily by the difference between the prototypic case and the combination of the typical and atypical cases ($p = .032$). The typical versus atypical case contrast was not statistically significant ($p = .091$).

Figure 2 shows a plot of the group means across the cases.

The results for the measure of early hypothesis generation may be stated as follows: experimental subjects generated the correct diagnosis in their workups on the basis of relatively fewer cues than control subjects for the atypical case. The group means for this case indicate that experimental subjects acquired 47% of their total cues before generating the correct hypothesis, while control subjects acquired 67% of their total cues before generating the correct hypothesis. For the prototypic and typical cases, experimental subjects generated 6% and 9% fewer cues, respectively, than control subjects before generating the correct diagnosis. However, these latter two group differences were not statistically significant.

FIGURE 2

PROFICIENCY OF EARLY HYPOTHESIS GENERATION



The significant group difference on the atypical case provides partial support for the second hypothesis, that is, that experimental subjects would be more proficient than controls in their early hypothesis generation for the atypical and typical cases. The hypothesis cannot be unequivocally supported in the absence of significant group differences on the typical case.

The significant difference in early hypothesis generation between the prototypic case and the combination of the typical and atypical cases was not predicted. However, because early hypothesis generation required the fewest number of cues for the prototypic case compared to the others, this finding is not entirely surprising. It is logical to expect that for the prototypic case, the

correct diagnosis would be generated earlier in the workup in comparison to the typical and atypical cases. This is because the prototypic case did not present any discrepant or ambiguous information to reduce subjects' confidence in their initial diagnostic hunches.

The reason that the typical case required more cues than the atypical case prior to early hypothesis generation is not clear. In theory, the atypical case, being the most difficult, should have required the most number of cues prior to the generation of the correct hypothesis. However, it is possible that the structure of the atypical case (TAPVC) allowed for the generation of the correct hypothesis earlier than for the typical case for the following reason. The finding of poor growth in the case introduction was an early cue for a possible cyanotic admixture lesion. On the basis of this information, several subjects entertained the three cyanotic diseases of CTGV, PTA, and TAPVC as an early LCS. Subjects taking this diagnostic pathway were quickly able to rule out CTGV based on the patient's age, and the auscultatory findings led them to strongly consider an atrial type disease. On this basis, astute subjects were able to conclude that the only cyanotic disease with auscultatory findings of ASD was TAPVC. TAPVC was further confirmed from the auscultatory findings because a split second heart sound definitively ruled out PTA.

In contrast, the typical case (PDA) had no such early "leads" for the diagnosis, with the exception of rubella, and this cue was not sufficiently diagnostic for an early, strong consideration of PDA.

Critical Cue Evaluation

The dependent measure of critical cue evaluation provided an assessment of the degree to which subjects 1) considered the "good" hypotheses for a case, that is,

the appropriate LCS, 2) generated these LCSs in response to the acquisition of critical cues, and 3) evaluated these LCS members in groups, which would suggest that they were part of a unit of information in long-term memory.

The scoring of critical cue evaluation was based on a method developed by Feltovich (1981) and is presented next. For each of the critical cues in a case, the subject's entire response to that cue was the unit of analysis. Any response to a critical cue that included an evaluation of an LCS member for that cue was coded. The evaluation of an LCS member for a given critical cue could be coded as positive (+), negative (-), or neutral (0). For example, the statement, "a systolic ejection click indicates the presence of PDA", represents a positive evaluation of an LCS member (PDA) with regard to a critical cue (systolic ejection click).

No score was recorded for the critical cue evaluation measure if the subject evaluated an LCS member with respect to a noncritical cue, or simply mentioned an LCS member without specifically evaluating it for a cue. This latter occurrence was coded separately and used in another measure that will be discussed later.

If the subject evaluated an LCS member more than once during a response to the same critical cue, the net valence was recorded. Responses that appeared to cancel each other out were recorded as 0. An example of such a response would be the statement that "a harsh systolic murmur heard in the left infraclavicular area seems to point to a PDA, but since the murmur is not continuous, it may not be PDA". Responses that cancelled each other out were very rare.

Subjects' total scores were then computed by summing the number of occurrences of pluses, minuses, or zeros across all the critical cues for a case.

Because the measure of critical cue evaluation involved some degree of subjective judgment, a reliability check was performed. Nine protocols were selected from both experimental and control subjects, representing a total of 27 cases. A second judge independently scored each of the cases in the same manner as the experimenter. All instances of agreement and disagreement on hypotheses evaluated for the critical cues were summed across subjects and cases. The coefficient Kappa was used as a measure of interrater agreement. This statistic calculates the percentage of agreement between raters and adjusts this figure by removing the percentage of agreement predicted by chance. The formula for Kappa and for the calculation of the proportion of chance agreement can be found in Fleiss (1973). The Kappa for the present study was .79, indicating an acceptable level of interrater agreement.

The measure of critical cue evaluation was used to assess the activeness of subjects' cue evaluation. In other words, the higher a subject's score, the more active he or she was in explicitly evaluating LCS members with respect to the critical cues of a case. However, this measure did not address the number of times that subjects evaluated the correct diagnosis with respect to the critical cues. This question will be discussed later.

Turning to the results for the critical cue evaluation measure, Table 15 shows the descriptive statistics for both groups.
Table 16 presents the results of the ANOVA.

The main effect of group was nonsignificant ($p = .548$), indicating that experimental and control subjects did not differ in the mean number of LCS members they evaluated with respect to the critical cues for the three cases. Because the overall effect of group was nonsignificant, no further explorations of the simple effects of group within case were warranted.

TABLE 15

CRITICAL CUE EVALUATION FOR EXPERIMENTAL AND CONTROL SUBJECTS ON THE THREE CASES

ASD	$\bar{X}$	sd	range
Experimental	7.72	3.71	2.00-13.00
Control	6.00	3.50	1.00-16.00
PDA			
Experimental	5.78	2.41	3.00-12.00
Control	4.76	2.73	1.00-13.00
TAPVC			
Experimental	5.17	3.20	1.00-12.00
Control	6.53	2.48	2.00-11.00

The average F test for the group by case interaction was significant ($p = .036$). An examination the specific contrasts for the interaction showed that the experimental-control differences were significantly different between the typical and the atypical case ($p = .029$). For the typical case, experimental subjects were slightly more active than control subjects in their critical cue evaluation, whereas the opposite occurred for the atypical case.

The average F test for the main effect of case was also significant ($p = .037$). The specific contrasts revealed that this difference was in the contrast of the prototypic case and the combination of the typical and atypical cases ($p = .036$). This finding indicates that subjects' mean cue evaluation scores were

TABLE 16

ANALYSIS OF VARIANCE FOR CRITICAL CUE EVALUATION

Source of Variance	Sum of Squares	df	Mean Square	F	p
Between Subjects					
Within Cells	490.50980	33	14.86393		
Constant	3768.69020	1	3768.69020	253.54595	.0005
Group	5.49020	1	5.49020	.36936	.548
Within Subjects					
Within Cells	430.00654	66	6.51525		
Case	45.30775	2	22.65387	3.47705	.037
ASD vs. PDA, TAPVC				4.76762	.036
PDA vs. TAPVC				1.22536	.276
Group by Case	45.65061	2	22.82530	3.50337	.036
ASD vs. PDA, TAPVC				2.53242	.121
PDA vs. TAPVC				5.19735	.029

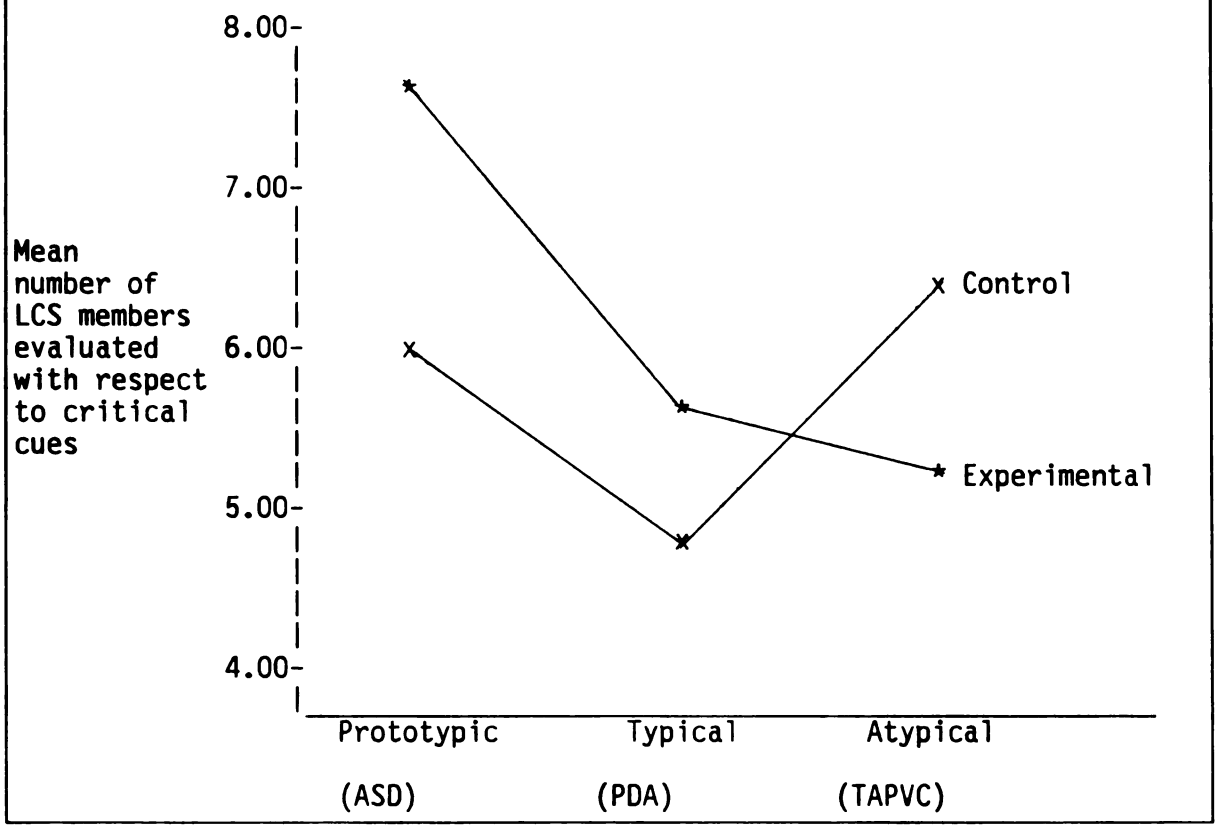
significantly higher for the prototypic case compared to the combination of the typical and atypical cases.

Figure 3 shows a plot of the group means across cases.

In summary, experimental subjects were not significantly more active than controls in their evaluation of critical cues with respect to LCS members for any of the cases. In light of these findings, the hypothesis of more active critical cue evaluation for experimental compared to control subjects on the typical and atypical cases was not supported.

FIGURE 3

CRITICAL CUE EVALUATION



The finding that critical cue evaluation decreased from the prototypic to the typical and atypical cases does not seem readily interpretable. One explanation is that as subjects worked through the cases, they may have become somewhat fatigued and thus, less verbal in their critical cue evaluation for the cases presented later. Other explanations for this finding that are based on the differences in case structure are not especially clear.

Comparison of findings with previous research. Although experimental and control subjects' critical cue evaluation scores were assessed relative to each other, it was of interest to compare these subjects to

experienced physicians on this measure, so that some criterion for performance could be established.

Such data were available from Feltovich's (1981) study, but only for the atypical (TAPVC) case, because this was the only case in common between both studies.

In order to compare the TAPVC cases from both studies, critical cue evaluation scores were computed for Feltovich's sample of 12 physicians and medical students. Four of the physicians were experts in pediatric cardiology, two with over 20 years of experience. Four physicians were either third year general pediatrics residents or first year pediatric cardiology fellows. The last four subjects were fourth year medical students who had just completed a six week course in pediatric cardiology prior to their participation in the study.

The four experts were divided into one group and the rest of the sample into another.

The mean critical cue evaluation score for the experts on the TAPVC case was 7.75, while for the trainees and fourth year students, it was 5.88.

In the present study, experimental and control subjects' mean scores on this measure were 5.17 and 6.53, respectively. Thus, it can be seen that the experimental and control groups were somewhat less active than the experts in their critical cue evaluation. However, their scores were comparable to those of the trainees and fourth year students.

The fact that the present study sample was comparable to advanced medical students and pediatric cardiology trainees suggests that, at least for the measure of critical cue evaluation, subjects' performance was good. However, as mentioned before, the experimental group was no better than the controls on this measure. It must also be remembered that this assessment is limited to only one case.

Additional measures of LCS use. The purpose of assessing critical cue evaluation was to examine the extent to which subjects used the appropriate logical competitor sets for the cases. However, this measure only addressed the issue of activeness of evaluation and therefore only provided partial information on subjects' use of LCSs.

Additional questions remain regarding 1) whether subjects actually used all of the critical cues for the cases, 2) the extent to which subjects used LCS members together when evaluating the same critical cue, 3) whether subjects' cue evaluations were accurate, and 4) the extent to which subjects used the most important member of the LCS: The actual disease in each case. Data will be presented next in order to address these issues.

The first question is, how many different LCS members were actually mentioned for the cases? In the previous measure of critical cue evaluation, it would have been possible for a subject to demonstrate relatively active evaluation of some LCS members with respect to critical cues, but not to have ever considered other LCS members. In this case, a failure to use one or more LCS members would not have been picked up by the critical cue evaluation measure because the subject would have received a high score based on his or her active evaluation of other LCS members.

Table 17 presents the mean number of different LCS members mentioned for both groups on each of the cases, along with the actual number of LCS members for each case for comparison purposes.
The Table clearly shows that both groups of subjects mentioned most of the appropriate LCS members for each case at some point in their workups.

It can be concluded that experimental subjects did not mention more of the LCS members for the cases than did control subjects. Thus, if any group

TABLE 17

MEAN NUMBER OF DIFFERENT LCS MEMBERS MENTIONED FOR EXPERIMENTAL AND CONTROL SUBJECTS ON THE THREE CASES

ASD	$\bar{X}$	sd	Range
Experimental	3.67	.59	2-4
Control	3.41	.80	1-4
Number of LCS members = 4			
PDA			
Experimental	2.78	.43	2-3
Control	2.47	.72	1-3
Number of LCS members = 3			
TAPVC			
Experimental	3.22	1.11	1-4
Control	3.47	.80	2-4
Number of LCS members = 4			

differences existed on subjects' use of LCSs, they cannot be attributed to a failure to consider the "good" hypotheses for the cases, however cursory this consideration was.

It must be remembered that this measure provides no information on whether subjects actively evaluated the LCS members with respect to any cues,

whether they evaluated most of the LCS members together, or whether their evaluation of these members was even accurate.

A second measure provides a rough indication of the degree to which subjects used LCS members together in their evaluation of critical cues. It was calculated as the highest number of different LCS members that a subject evaluated for the same critical cue in a case. Table 18 shows the group means for all cases on this measure.

TABLE 18

HIGHEST NUMBER OF DIFFERENT LCS MEMBERS EVALUATED FOR SAME CRITICAL CUE

	$\bar{X}$	sd	Range
ASD			
Experimental	3.00	.97	1-4
Control	2.65	1.12	1-4
PDA			
Experimental	2.78	.43	2-3
Control	2.47	.72	1-3
TAPVC			
Experimental	2.56	1.34	1-4
Control	2.94	.90	1-4

Again, it can be seen that the groups were quite comparable in their use of LCS members together for at least one critical cue. In addition, most of the LCS members for each case were evaluated together for at least one critical cue.

A third important dimension of subjects' use of LCSs was the extent to which they were inaccurate in their evaluation of critical cues for LCS members. It may have been possible for subjects to be scored as active in their critical cue evaluation even if their evaluations were often erroneous.

Errors in critical cue evaluation included 1) statements that a critical cue was positive for an LCS member when it was negative or noncontributory, 2) statements that a critical cue was negative or noncontributory for an LCS member when it was positive, and 3) statements that a cue was more diagnostic of one LCS member than it really was. For example, the statement, "Increased pulmonary vasculature on chest X-ray is confirmatory for PDA", was erroneous because this finding could have been present with any of the eight diseases in the knowledge base.

Table 19 presents the mean number of errors of critical cue evaluation for both groups of subjects on the three cases.

The most important finding here is that the mean number of errors for either group never exceeded one for any of the cases, indicating that subjects were generally error-free in their critical cue evaluation. For the typical and atypical cases, experimental subjects appeared to commit more errors than control subjects. This difference was significant for the atypical case ($t = -2.13$, $df = 33$, $p = .045$). The meaning of this finding is not readily apparent.

The fourth measure of LCS use involved the extent to which subjects were active in evaluating the correct disease with respect to the critical cues for a case (corrrect hypothesis evaluation). This measure does not address subjects'

TABLE 19

MEAN NUMBER OF ERRORS OF CRITICAL CUE EVALUATION FOR EXPERIMENTAL AND CONTROL SUBJECTS ON THE THREE CASES

ASD	X	sd	Range
Experimental	.11	.32	0-1
Control	.29	1.21	0-5
PDA			
Experimental	.72	.96	0-3
Control	.47	.87	0-3
TAPVC			
Experimental	.39	.61	0-2
Control	.06	.24	0-1

joint LCS use (this was discussed earlier) but focuses instead on whether subjects were actively weighing the best member of their LCS during the workup. Table 20 presents the mean number of times both groups of subjects evaluated the correct diseases for the cases with respect to the critical cues of these cases. It indicates that the groups were similar on this measure for all cases. Table 20 also shows a slight downward trend in the mean number of evaluations from the prototypic to the atypical cases. This finding may be due to the increased case difficulty which might have led subjects to more actively consider competing LCS members as the case difficulty increased.

TABLE 20

MEAN NUMBER OF EVALUATIONS OF CORRECT DISEASE WITH RESPECT TO CRITICAL CUES

ASD	$\bar{X}$	sd	Range
Experimental	3.11	1.13	1-5
Control	3.12	.70	2-5
PDA			
Experimental	2.11	1.41	0-5
Control	2.34	1.03	1-5
TAPVC			
Experimental	2.11	1.37	0-4
Control	1.47	1.28	0-4

To summarize the findings pertaining to the additional measures of LCS use, subjects 1) mentioned most of the appropriate LCS members for the cases, 2) evaluated most of the LCS members together at least once for any one critical cue, 3) committed relatively u35 errors of critical cue evaluation, 4) actively evaluated the correct disease with respect to the critical cues for a case, and 5) did not differ on these measures depending on whether they were in the experimental or control group, the one exception being significantly more critical cue evaluation errors for experimental subjects than controls on the atypical case.

Secondary Hypotheses

Diagnostic Accuracy

The measure of diagnostic accuracy was the proportion of subjects in each group who gave the correct primary diagnosis for each case. Table 21 shows these proportions for the three cases and the Chi square tests of significance for each. This table also shows what the incorrect diagnoses were for each of the cases. Only one subject (2.9%) misdiagnosed the prototypic case, whereas 12 subjects (34.4%) misdiagnosed the typical case, and 15 (42.9%) misdiagnosed the atypical case. These figures reflect the increasing level of difficulty from the prototypic to the atypical cases.

It is interesting to note that 25 of the 28 incorrect diagnoses given for the three cases involved members of the LCS for that particular case. Thus, subjects who misdiagnosed the cases were at least working with the appopriate LCS for each. The proportions of experimental and control subjects misdiagnosing the typical and atypical cases were significantly different. For the typical case, 12% of the control subjects made a misdiagnosis while 56% of the experimental subjects made a misdiagnosis ($p = .018$).

For the atypical case, these proportions were in the opposite direction. 65% of the control subjects and 22% of the experimental subjects misdiagnosed the case, respectively ($p = .028$).

The findings for diagnostic accuracy support the hypothesis that experimental subjects would not be more accurate than controls for the prototypic case, but would be more accurate for the atypical case.

However, the finding that control subjects were more accurate than experimental subjects for the typical case was opposite of what was

TABLE 21

PROPORTIONS OF EXPERIMENTAL AND CONTROL

SUBJECTS WHO CORRECTLY DIAGNOSED EACH CASE

ASD

Group	CORRECT	INCORRECT	INCORRECT DIAGNOSES
			PDA
Experimental	18	0	
Control	16	1	1
		35	

X^2 = = 0.00, df = 1, p = .977

PDA

	CORRECT	INCORRECT	INCORRECT DIAGNOSES		
			VSD	PTA	TAPVC
Experimental	8	10	4	6	
Control	15	2	1		1
		35			

X^2 = = 5.62, df = 1, p = .018

TAPVC

	CORRECT	INCORRECT	INCORRECT DIAGNOSES	
			PAPVC	CTGV
Experimental	14	4	4	
Control	6	11	10	1
		35		

X^2 = = 4.82, df = 1, p = .028

hypothesized. This discrepant finding will be explored in more detail in the discussion.

In sum, it may be stated that the hypothesis regarding diagnostic accuracy received only partial support.

Additional analyses. Because it was of interest to determine the possible sources of subjects' diagnostic errors, two additional exploratory analyses were done.

The first analysis was qualitative and involved reviewing subjects' protocols for occurrences of errors of critical cue evaluation. This was done to examine whether cue evaluation errors eventually led to the ruling out of the correct disease or the confirmation of an incorrect one. As mentioned earlier, errors of critical cue evaluation included 1) statements that a critical cue was positive for an LCS member when it was negative or noncontributory, 2) statements that a critical cue was negative or noncontributory for an LCS member when it was positive, and 3) statements that a critical cue was more diagnostic for one LCS member than it really was.

The review of critical cue evaluation errors yielded surprisingly little information pertaining to subjects' diagnostic errors. In fact, errors of cue evaluation appeared to be idiosyncratic, following no particular pattern. Also, as mentioned before, errors of critical cue evaluation occurred relatively infrequently.

Another possible reason for diagnostic inaccuracy might have been that subjects incorrectly weighed evidence for competing diseases. In this case, subjects could have been accurate in their critical cue evaluation, yet interpret the weight of evidence as favoring the incorrect diagnosis. A good example was the atypical case, where incorrect subjects often decided that the patient was not

quite sick enough (or cyanotic enough) to have TAPVC, so, therefore, she probably had PAPVC. These types of errors relate to "judgment calls", and do not necessarily reflect subjects' inacurrate interpretation of a specific cue. These errors might also be considered errors of cue synthesis.

The second analysis focused on the relationship between diagnostic accuracy and the extent to which subjects evaluated the correct hypothesis with respect to the critical cues for the cases. The measure of correct hypothesis evaluation was coded as the number of times the correct disease for the case was evaluated as positive, negative, or noncontributory for any of the critical cues for the case. T-tests were used to compare the accurate and inaccurate subjects on this measure of correct hypothesis evaluation. T-tests were only computed for the typical and atypical cases, because all but one subject correctly diagnosed the prototypic case.

Table 22 shows the t-test results.
It indicates that there were no differences between accurate and inaccurate subjects on correct hypothesis evaluation for the typical case.

However, for the atypical case, subjects who gave the correct diagnosis had significantly higher scores on correct hypothesis evaluation than those who were incorrect in their diagnosis ($p = .0005$). Thus, at least for the atypical case, diagnostic accuracy was strongly associated with the extent of subjects' correct hypothesis evaluation. This finding suggests that those subjects incorrectly diagnosing the atypical case may not have given much weight to the correct disease, TAPVC, in the first place. In fact, on the average, inaccurate subjects evaluated TAPVC with respect to the critical cues slightly less than once, whereas accurate subjects evaluated TAPVC an average of two and one-half times with respect to the critical cues. Perhaps this finding reflects inaccurate subjects' use of an LCS which only peripherally included TAPVC.

TABLE 22

T-TESTS ON MEAN DIFFERENCES IN CORRECT HYPOTHESIS EVALUATION BETWEEN ACCURATE AND INACCURATE SUBJECTS

	N	X̄	sd	t	p (2-tail)
PDA					
Accurate	23	2.30	1.06	.79	.438
Inaccurate	12	1.92	1.50		
TAPVC					
Accurate	20	2.50	1.10	4.41	.0005
Inaccurate	15	.87	1.06		

The lack of differences on this measure for the typical case points to the possibility of differences in the structures of this case compared to the atypical case. It also warrants caution in generalizing such findings because they may be case specific.

Comparison of findings with previous research. A final point of interest regarding diagnostic accuracy was how the subjects in this study compared with the more experienced subjects in Feltovich's study. As with the earlier comparison on critical cue evaluation, this comparsion on diagnostic accuracy could only be done for the TAPVC case.

These results are as follows: Two of the four experts, or 50%, gave the correct diagnosis of TAPVC, while the other two experts gave PAPVC as their primary diagnosis. For the medical students and trainees, only two out of eight,

or 25%, gave a diagnosis of TAPVC. Primary diagnoses for the others included PAPVC, ASD, ECD, and diseases outside the LCS.

The comparable findings for the present study were 78% correct for experimental subjects and 35% correct for controls. The greater accuracy of subjects in the present study may be due to their smaller knowledge base and the relatively short time between learning the material and working on the CDPs. Also, perhaps more experienced clinicians tend to consider a wider range of cues, and therefore, consider a more complicated diagnostic picture.

Another relevant finding to compare between the two studies involved the mean differences in correct hypothesis evaluation between those who gave a correct versus incorrect diagnosis for the TAPVC case. These findings turned out to be quite similar for the two studies. Correct subjects in Feltovich's study evaluated TAPVC an average of 2.5 times for critical cues, compared to .62 times for incorrect subjects. In the present study, correct subjects also evaluated TAPVC an average of 2.5 times, compared to .87 times for incorrect subjects.

These findings suggest that for both studies, the extent to which a subject evaluated the correct hypothesis with respect to critical cues was a reliable indicator of diagnostic accuracy.

Cost of Workup

The cost of subjects' workups for the cases was computed by the CAMPS program. For each case, $25 was automatically charged for the patient visit. Additional costs were strictly due to laboratory tests ordered, because the other two sections of the workup that involved charges, treatment and consultation, were removed from the CDPs. The two most frequently ordered tests, the EKG and X-Ray, cost $35 and $50, respectively.

Although a large number of cardiology laboratory tests were available, such as catheterization, blood gas, and phonocardiogram, most subjects were not sufficiently familiar with them to be able to interpret the findings accurately. Moreover, the only laboratory tests discussed in subjects' knowledge bases were the EKG and X-ray, and the majority of subjects ordered one or both of these tests exclusively. For these reasons, the measure of cost of the workup was quite restricted in range.

Table 23 shows the mean cost of experimental and control subjects' workups for the three cases, as well as the results of t-tests computed on these mean differences.

Although for all three cases, the mean cost for experimental subjects was less than that for controls, none of the differences was significant. The mean group differences did show a trend for the prototypic case, however ($p = .079$). It might also be noted that for the prototypic and typical cases, the variance for the control subjects was significantly greater than that for experimental subjects.

The findings for the measure of cost of the workup do not unequivocally support the hypothesis of significantly lower cost for experimental subjects on the typical and atypial cases. However, experimental subjects' laboratory costs were slightly lower than controls across all three cases, suggesting the presence of a weak effect. Unfortunately, because of the restricted range of this measure, an accurate assessment of subjects' laboratory test ordering was not possible.

TABLE 23

T-TESTS ON MEAN DIFFERENCES IN COST OF WORKUP BETWEEN EXPERIMENTAL AND CONTROL SUBJECTS

	N	$\bar{X}$	sd	t	p (2-tail)
ASD					
Experimental	18	105.28	13.98	1.87	.079
Control	17	132.35	58.23		
PDA					
Experimental	18	104.17	21.64	1.06	.301
Control	17	118.82	53.02		
TAPVC					
Experimental	18	115.28	60.43	.70	.487
Control	17	130.29	65.87		

Summary of Results

The findings for the major hypotheses may be briefly summarized as follows:

1. The intervention produced strong effects regarding subjects' proficiency of cue acquisition. Experimental subjects acquired significantly fewer noncritical cues relative to critical cues than did controls, and this difference occurred across all three cases. These findings were somewhat stronger than predicted because group differences were found on the prototypic case, although they were not expected to occur there.

2. The effects of the intervention were less pronounced for the measure of early hypothesis generation, but they were in the predicted direction. Experimental subjects mentioned the correct hypothesis significantly sooner than control subjects for the atypical case. Although experimental subjects were also quicker than controls in their hypothesis generation for the other two cases, these differences were not statistically significant.

3. There were no significant group differences in the evaluation of LCS members with respect to critical cues, although there was a significant interaction between group and case. The interaction suggested that experimental subjects were more active than controls in their LCS evaluation for the typical case, but less active than control subjects for the atypical case. The hypothesis of more active critical cue evaluation by experimental subjects compared to controls on the typical and atypical cases was not supported by the results.

The findings for the secondary hypotheses are as follows:

1. There were no group differences in diagnostic accuracy for the prototypic case, but the two groups did differ significantly in their diagnostic accuracy for the typical and atypical cases. Control subjects were significantly more accurate than experimental subjects for the typical case, which was contrary to the hypothesis. On the other hand, experimental subjects were significantly more accurate for the atypical case, and this finding was hypothesized.

2. There were no significant group differences with regard to the cost of the workups, although there was a trend for the experimental subjects to incur less cost than controls. It appeared that this measure, based on the number of laboratory tests ordered, was severely restricted in range. Therefore, although

the hypothesis regarding the cost of the workup was not supported, the dependent measure appeared inadequate for an accurate assessment of this variable.

DISCUSSION

Primary Hypotheses

Proficiency of Critical Cue Acquisition

On all of the cases, experimental subjects acquired less noncontributory information than controls, while acquiring an equal amount of diagnostic information. This finding is not entirely surprising because the experimental intervention emphasized efficient problem-solving through the use of clusters of a few highly diagnostic cues.

Moreover, on the basis of relatively less information, experimental subjects were as accurate as controls in diagnosing the prototypic and atypical cases and even more accurate in diagnosing the atypical case. Thus, for these two cases, more efficient cue acquisition did not lessen diagnostic accuracy. The greater proficiency of critical cue acquisition demonstrated by experimental subjects is an important finding because it suggests that the intervention fostered a hypothesis-driven strategy in approaching the diagnostic task. This strategy involved a planned search for clusters of cues that could provide information with which to generate a good LCS for the case, and then distinguish among LCS members to diagnose the case.

In contrast, many control subjects seemed to be data-driven in their diagnostic reasoning; that is, they did not follow a particular plan but, rather, "were seemingly pushed from one hypothesis to another depending on the most

recent strong disease cue in the data" (Johnson, Duran, Hassebrock, Moller, Prietula, Feltovich, & Swanson, 1981, p. 253).

Also, because control subjects tended to acquire more noncontributory cues than experimental subjects, they ran the risk of using these cues to erroneously bolster their confidence in a diagnosis. The potentially deleterious effect of using noncontributory clinical information has been documented previously (Bergman and Beck, 1983).

It is also important to note here that research on clinical reasoning suggests that experienced clinicians tend to acquire less information than less experienced physicians in a diagnostic workup, but that the information they do acquire is highly diagnostic (McGuire, 1984; Barrows and Tamblyn, 1980; Kassirer & Gorry, 1978). Thus, the intervention seemed to produce an expert-like diagnostic approach in at least one aspect of clinical reasoning.

Proficiency of Early Hypothesis Generation

For the atypical case, experimental subjects generated the correct hypothesis on the basis of relatively less information than controls. For the prototypic and typical cases, experimental subjects showed a trend in the same direction. Again, these findings are consistent with the thrust of the intervention, because it stressed the rapidity of early hypothesis generation.

It is possible that control subjects were as capable as experimental subjects in generating the correct hypothesis early in their workups, but because control subjects were not told that this was the way in which experienced clinicians worked, they may have been less motivated to verbalize their hunches early in their workups. However, all subjects were frequently encouraged to report their hunches throughout the workups.

Regardless of whether control subjects were as capable as experimental subjects in generating early hypotheses, the major point here is that the intervention was moderately effective in fostering a form of reasoning demonstrated by experienced clinicians in studies of clinical problem-solving (Elstein et al, 1978; Barrows & Tamblyn, 1980; kassirer & Gorry, 1978). This type of expert reasoning involves the use of salient cues for rapid associative triggering of hypotheses in long term memory.

These results concur with Allal and Shulman's (1974) findings that early diagnostic problem formulations could be taught to medical students. However, Allal and Shulman used films of diagnostic encounters to teach problem formulations, coupled with process and outcome feedback from experienced physicians. Also, their intervention took place over three weeks. Thus, it is interesting to note that the present intervention produced a moderate effect given that it was of relatively short duration and that it did not provide any extra information to experimental subjects, but, rather, focused entirely on changing the way experimental subjects assimilated the information that was given to them.

The reason for experimental subjects' more proficient early hypothesis generation for the atypical case may be found in the structure of their knowledge base. It appears that cue-hypothesis associations for the atypical case were more quickly and effectively formed by experimental subjects because their knowledge base was more efficiently organized than that of control subjects.

In particular, flow diagram 1 demonstrated that symptom burden could be reliably assessed by combining cues related to poor growth, history of upper respiratory infections, and cyanosis. Because these cues were all more or less positive in the atypical case, experimental subjects may have generated a more

serious disease (for example, TAPVC, CTGV, PTA) sooner in their workups than controls, sheerly on the basis of increased symptom burden. Next, from the findings for the murmur and heart sounds, TAPVC should have been a strong candidate, because it was the only serious disease with an ASD type murmur and heart sounds. Finally, the X-Ray finding of an unusual vascular shadow would have served to confirm TAPVC for experimental subjects.

In contrast to using this diagnostic pathway, control subjects often did not synthesize cues from the history with which to generate a serious disease for consideration. For example, several control subjects stated that most of the diseases in their knowledge base presented with histories of upper respiratory infections, that the early cyanosis cue was not strong enough to consider a serious disease, and that the patient's poor growth could be due to causes other than a heart defect. These subjects were partially correct because, taken alone, the cues were not sufficiently salient to trigger the admixture lesion category. In combination, however, the cues formed a more serious picture.

The net result of this lack of an early problem formulation was twofold. First, when control subjects acquired the murmur and heart sounds on physical examination, they were typically led to consider the LCS of ASD, ECD, and PAPVC. Most did not include TAPVC because it was not generated earlier. Many of those subjects who did question whether the symptoms from the history were too serious for an uncomplicated ASD began to consider PAPVC at this point.

Second, when the X-Ray finding of an unusual vascular shadow was acquired, PAPVC was seen as confirmed by many control subjects. They may have also triggered TAPVC at this point but because they had not previously included TAPVC in their problem formulation, this disease was seen as a less likely candidate.

In sum, because control subjects were less likely to generate TAPVC early on in their workup, they were less likely to seriously consider it later, even in the presence of a fairly diagnostic cue. In addition, the lack of an early problem formulation which included TAPVC may have led to misdiagnosis of the case. This speculation will be explored further in the section on diagnostic accuracy.

Critical Cue Evaluation

Experimental subjects were no more active than controls in their evaluation of critical cues with respect to LCS members for the three cases.

This nonsignificant finding can best be interpreted in the following way. Both groups seemed to be relatively active in their use of LCS members. That is, given a fairly limited knowledge base, both groups were able to use most of the "good" hypotheses in their evaluation of critical cues. The knowledge base may have been small enough, and the LCSs sufficiently apparent, that subjects were at the ceiling on this measure, and no intervention could increase subjects' critical cue evaluation.

Some indirect support was found for this interpretation. Upon examination of control subjects' study materials, it was found that eight of the 17 control subjects constructed either a table or diagram detailing the relationships between the diseases in the knowledge base. Some of these tables were quite similar to the one used by experimental subjects, and several specifically grouped LCS members together.

A further observation was that those control subjects who constructed tables to help them "chunk" the information seemed to perform as well as experimental subjects on the CDPs, in terms of proficiency of critical cue acquisition, early hypothesis generation, and diagnostic accuracy.

The fact that almost half of the control subjects constructed tables of LCSs is not surprising; their knowledge base specified many of these relationships. For example, in the discussion of PTA, it was stated that, "In the absence of cyanosis, patients with PTA and congestive heart failure are clinically similar to those with VSD and PDA" (p. 5). Another example was the summary for TAPVC: "The clinical, EKG, and X-Ray findings of TAPVC without obstruction to pulmonary blood flow, resemble those of ASD because the effects upon the heart are similar" (p. 6). This kind of information enabled many of the control subjects to organize their knowledge base in a manner equally as efficient as that of the experimental subjects.

All references to disease similarities could have been eliminated from the control subjects' knowledge base to produce greater group differences, but this would have been somewhat artificial because the material was intended to represent a textbook presentation and these disease relationships were specified in Moller's pediatric cardiology textbook.

These findings indicate that some control subjects were able to use their knowledge base to form the appropriate LCSs without the benefit of any intervention. Moreover, the relatively small size of the knowledge base, and its lack of complexity, may have helped control subjects in the process of forming the LCSs.

Unfortunately, little light has been shed on the efficacy of the intervention itself. One can only speculate as to whether the intervention would have produced group differences if applied to a larger and more complicated knowledge base.

A final point regarding the critical cue evaluation variable is that its value as a measure of LCS use was rather limited. This is because subjects could score high on this measure yet misdiagnose the case.

The significant interaction found between the typical and atypical cases illustrates this point. For the typical case, experimental subjects were somewhat more active than controls in their critical cue evaluation, but experimental subjects were also more likely to misdiagnosis the case. Conversely, control subjects were somewhat more active than experimental subjects in their critical cue evaluation for the atypical case. However, for this case, the control subjects were also more likely to misdiagnose it. Perhaps more active critical cue evaluation occurred in a situation of increased diagnostic uncertainty.

Another anomalous finding for the critical cue evaluation measure was that it was significantly higher for the prototypic case compared to the typical and atypical cases. One might expect the opposite finding, because more active consideration of LCS members should have been prompted by more difficult cases.

Secondary Hypotheses

Diagnostic Accuracy

All but one subject correctly diagnosed the prototypic case, indicating that this case was relatively straightforward and did not demand subtle discriminations among ambiguous patient data. The proportions of subjects giving an incorrect diagnosis inreased to 34% for the typical case and 43% for the atypical case. These figures are consistent with the increasing difficulty of the cases.

Contrary to the hypothesis, control subjects were significantly more accurate than experimental subjects in diagnosing the typical case. A careful examination of experimental and control subjects' respective knowledge bases may provide clues to the source of this finding.

In the experimental subjects' flow diagrams, the major findings for PDA (the typical case) were a normal second heart sound, a continuous murmur, and absence of cyanosis. For the actual case of PDA, however, the murmur was not specifically described as continuous, and cyanosis due to congestive heart failure was present only on history, and not on physical examination. Thus, two of the major diagnostic cues used by experimental subjects to diagnose this case were obscured.

In addition, one of the most important cues for narrowing the LCS to PDA and PTA was the finding of a systolic ejection click. Although this finding was listed in experimental subjects' table of diseases, it was not part of their flow diagrams, and therefore was not emphasized.

Because of the seeming inconsistencies between their expectations and the actual findings for PDA on history and physical exam, many experimental subjects decided to use the laboratory test results to help them make a definitive diagnosis. However, the laboratory tests were not sufficiently diagnostic to confirm PDA because the EKG was positive for VSD, PDA, or PTA, and the X-Ray was positive for PDA or PTA. As a result, some experimental subjects became stuck at the laboratory section of the workup and did not know where to turn for more diagnostic information.

PDA was best diagnosed in the physical exam on the basis of the patient's heart sounds, with the following line of reasoning: A systolic ejection click on auscultation suggests a dilated aorta, which is positive for PDA and PTA. The fact that the auscultatory findings also show a loud pulmonary component of the second heart sound is strongly disconfirmatory for PTA, because PTA always has a single second sound. Finally, the X-Ray finding of aortic enlargement supports the auscultatory findings and confirms PDA.

In contrast to experimental subjects' abbreviated knowledge base for PDA, control subjects' knowledge base contained two clues that appeared to help them diagnose the case. The first was a rather strong statement regarding the X-Ray results for PDA: "PDA is the only cardiac defect with a left-to-right shunt with aortic enlargement. In the other left-to-right shunts, the aorta is normal or appears small. Therefore, if a distinctly enlarged aorta is present and a left-to-right shunt is suspected, PDA must be seriously considered" (p. 3). Given this datum, once subjects suspected aortic enlargement because of the systolic ejection click, they often triggered PDA as a good candidate and subsequently confirmed it with the X-Ray findings.

The second piece of information was considerably less diagnostic but it served to raise the suspicion of PDA early in the workup. This was the finding of rubella during the mother's first trimester of pregnancy. Some of the control subjects adopted a strategy of asking questions about rubella (as well as other findings, such as Down's syndrome) in the history in order to pursue early leads for hypotheses. The finding of rubella then served to trigger the possiblity of PDA early in the workup, and to help subjects decide in favor of PDA when they were considering competing alternatives. Although experimental subjects had the same rubella information, it was again de-emphasized by being placed in the table but not in the flow diagrams.

Two conclusions can be drawn from the findings on diagnostic accuracy for the PDA (typical) case. First, small changes of emphasis in a knowledge base may subsequently lead to large errors in diagnostic reasoning. Second, the use of an efficient diagnostic strategy that distills the major disease findings may be risky because certain important nuances in the data may be lost. Moreover, any diagnostic strategy that attempts to simplify the clinical information may

engender excessive dependence on it. The implications of these findings will be explored further in the section on theoretical issues.

A final point of interest regarding this case is that unlike other research findings on clinical reasoning (for example, Elstein, 1978), incorrect subjects were no less active than correct subjects in the extent to which they evaluated the target disease with respect to the critical cues.

The finding of greater diagnostic accuracy among experimental compared to control subjects for the atypical case was congruent with the hypothesis. It was particularly reassuring that this finding occurred on what was considered the most difficult case.

The reasons for the group differences in diagnostic accuracy can again be found in an examination of subjects' respective knowledge bases. This was done earlier for the findings related to early hypothesis generation. A brief recap of these findings is that experimental subjects more accurately assessed symptom burden early in the workup, thus triggering a consideration of a serious disease, such as TAPVC, which they later confirmed with the murmur, heart sounds, and X-Ray results.

In contrast, many control subjects did not synthesize the early cues for a serious disease and therefore did not actively consider TAPVC until later, at which point they had already chosen another likely disease, PAPVC, as their diagnosis.

Another phenomenon might have occurred to decrease control subjects' diagnostic accuracy, and this relates to how strongly TAPVC was connected to their LCS of atrial level shunts. Although some control subjects clearly included TAPVC in their LCS along with ASD, ECD, and PAPVC, as evidenced by the tables and diagrams they constructed, others had trouble even remembering

TAPVC, much less its associated findings. Some of these subjects would refer to TAPVC as "that last one on the list".

Another possible reason that more control subjects misdiagnosed this case was that their expectations for the X-Ray findings for TAPVC were too narrow. When presented with the finding of an unusual vascular shadow on X-Ray, some concluded that it probably was not TAPVC because there was no "snowman" heart, a type of cardiac silhouette. In this instance, subjects' expectations for a specific finding were too rigid to allow a consideration of an ambiguous finding that was nevertheless a type of cardiac silhouette.

Cost of Workup

The nonsignificant findings for cost of subjects' workups seemed to be due to a restriction in range of the number of laboratory tests ordered.

There were two reasons for this finding. First, subjects were generally unfamiliar with most of the cardiology tests available to them, except for the EKG and X-Ray, which were discussed in the knowledge base. Moreover, many subjects correctly assumed that the EKG and X-Ray would provide them with enough information for a definitive diagnosis. A few subjects did select and accurately interpret the blood gas test in order to assess whether the patient had a low blood oxygen content. Also, a few subjects seemed vaguely familiar with the type of results that the cardiac catheterization could provide. However, because this was an invasive procedure, particularly for a child, catheterization was rarely ordered.

The second reason for the restriction in range of laboratory tests ordered was that a majority of subjects were surprisingly cost conscious. This was reinforced when subjects were automatically presented with cost of the tests they ordered when they received the test results.

It appears that the present study design was inadequate to assess the effects of the intervention on the cost of subjects' workups. Given the probable reason for the nonsignificant results, little can be said regarding how the intervention might have influenced cost.

Additional Findings

Two general types of unexpected findings emerged from this study. Neither of the findings was hypothesized nor were they systematically measured. Therefore, the data to be presented are strictly impressionistic. some impressions are based on observations of the problem-solving sessions that were later coroborrated in discussions with subjects during the debriefings.

Individual Learning and Problem-Solving Styles

Regardless of their group membership, subjects demonstrated unique styles of organizing the knowledge base for the diagnostic task. One such style involved a heavy reliance on pictorial thought. That is, these subjects found the schematic diagrams of the heart defects crucial to their understanding of the knowledge base. Apparently such diagrams provided a great deal of valuable information that these subjects could quickly encode, such as the location of the shunt, the direction of blood flow through it, and the affected chambers of the heart. During the problem-solving sessions, some of these subjects would even draw pictures of the diseases in order to figure out whether the data they acquired was consistent with the pathophysiology of the disease under consideration.

Another learning style seemed to involve the use of the written disease information, particularly the introductory material, to better understand functional relationships among diseases. Often these subjects stated that they

needed to "see how it all fit together" in order to learn the material. The emphasis of this style seemed to be on information that illustrated functional relationships, such as the hemodynamics of left to right shunts and admixture lesions.

Experimental subjects who demonstrated either of these two styles tended to balk at the table and flow diagrams. Some even reorganized this information to better fit their learning style. The main problem appeared to be that the table and flow diagrams presented only disease-symptom associations without any explanations, and these subjects were unable to learn the associations effectively without an understanding of why the associations existed in the first place.

A third learning style involved an emphasis on associative thinking. This style was almost diametrically opposed to the others because it emphasized the learning of symptom-disease relationships without any knowledge of functional relationships. Control subjects who used this style did quite well on the diagnostic task because they tended to organize the information in a way similar to the table and flow diagrams used by experimental subjects. As might be expected, experimental subjects using this style remarked that the material was organized exactly as they would have organized it if they had been asked to do so. It is also interesting to note that three of the subjects who exemplified this learning style had experiences requiring a good deal of logic. One subject was a former mathematics major, one had a chemsitry background (this subject stated that her organization scheme was "the logical way to do it"), and one had extensive experience constructing and using algorithms and other types of decision trees.

The learning style most often used by subjects appeared to be the one involving the use of functional relationships. Roughly equal numbers of subjects used the associative and the pictorial styles.

Just as distinct learning styles emerged from observations of subjects' organization of the knowledge base, diagnostic styles were observed during the problem-solving sessions. Two general diagnostic styles were seen. The first might be called the intuitive approach because these subjects rapidly formed a fairly complete picture of the patient on the basis of relatively few cues. Subjects with the greatest amount of clinical experience tended to use this style, although it was also used by some inexperienced subjects.

In contrast, some subjects used what might be called a methodical style, that is, exhaustively collecting patient data, and then systematically sifting through it at the end of the workup in order to formulate the diagnosis.

Both diagnostic styles had advantages and disadvantages. The intuitive style was efficient and probably more akin to the way experienced clinicians diagnose diseases. However, the risk of not acquiring a critical cue was high for this style, and the failure to acquire such a cue sometimes led to an incorrect problem formulation. The methodical style was less efficient than the intuitive style but the chances of a correct problem formulation were greater because all of the critical cues were usually acquired. Thus, any errors associated with this style tended to center on cue interpretation, not cue acquisition.

It should be mentioned that the methodical style resembles the general style of clinical problem-solving taught in many introductory to clinical medicine courses. The rationale for this approach makes sense: In the absence of the knowledge of which cues provide the highest diagnostic payoff, it is best to teach a comprehensive, systematic cue acquisition strategy. With the accumulation of clinical experience, this exhaustive approach can be abbreviated and modified depending on the diagnostic problem.

Unique learning and problem-solving styles were readily observed in some subjects, but were not as easily observed in others, because these subjects seemed to use a combination of styles.

The major implication of these findings is that there may not be one particular instructional strategy for clinical problem-solving that works best for everyone. With regard to the present study, it appears that the intervention was highly useful for some subjects and counterproductive for others.

Assessing the Quality of Clinical Problem-Solving

It is well known that clinical problem-solving is complex and multidimensional, and that there are many techniques for its assessment. At best, one can hope to measure a fraction of the qualities that comprise good clinical problem-solving. In fact, some aspects of clinical problem-solving may be impossible to measure.

Two findings from the present study support this assertion. One finding was that certain intangible qualities emerged from the problem-solving sessions that seemed to represent good or poor clinical reasoning. These were not directly measured nor is it certain that they were measurable. For example, one such quality involved subjects' ability to synthesize various cues into an accurate picture of the patient. This was often done with minimal information, but was nevertheless accurate. However, some subjects were observed to synthesize findings accurately and yet arrive at an incorrect diagnosis. Either it must be acknowledged that good clinical reasoning form may not always be related to diagnostic accuracy, or that the form of clinical reasoning is irrelevant and the ultimate criterion of performance is diagnostic accuracy.

Another finding was that subjects took several different diagnostic pathways, many of which led to a correct diagnosis. Therefore, any consideration

of good clinical reasoning had to include a variety of different approaches that could lead to the same outcome.

Methodologic Issues

There are four major issues related to the measurement and interpretation of the results. These will be discussed next.

Limitations of the Measures

Critical cue acquisition, early hypothesis generation, and critical cue evaluation are all components of the medical problem-solving process that have been examined previously (Elstein et al, 1978; Barrows, 1975; Vu, 1979). Therefore, the measures chosen to assess diagnostic reasoning for this study were neither new nor were they particularly unusual.

In addition, these measures have been shown to discriminate between experienced and less experienced physicians, and to relate to such outcomes as diagnostic accuracy. However, the relationships reported for these measures are not always strong and unambiguous (McGuire, 1984), perhaps because they fail to capture all of the relevant dimensions of clinical problem-solving, as mentioned earlier. The measures of critical cue acquisition, early hypothesis generation, and critical cue evaluation only indirectly address a central issue in studies of clinical problem-solving expertise: The way in which the expert's knowledge is organized in long term memory and the production rules he or she has established for rapid retrieval and efficient use of this large knolwedge base.

Knowledge organization is difficult to measure directly and is most often inferred by the way the clinician evaluates competing sets of diseases with respect to critical cues. This method was adopted for the present study because

it was thought that an examination of subjects' LCS use would indicate their particular knowledge organization strategy. Although an examination of LCS use provides the best indirect measure of knowledge organization, a more direct measure would be to have subjects perform a sort task of the diseases that they would work with in a problem-solving simulation. In particular, it might be useful to have subjects sort diseases into similar categories before and after an intervention designed to facilitate an expert knowledge organization strategy. Pre-post changes in categorization could then be analyzed according to the degree to which subjects' knowledge organization schemes more closely resemble those of experts. Such methodology is routinely used in cognitive psychological studies of memory organization (Shavelson, 1973; Schoenfeld & Hermann, 1982).

It must be acknowledged, however, that the medical knowledge organization of experts involves a lattice-work of information that is flexible and dynamic. Experienced clinicians may categorize diseases a certain way for one type of clinical problem and quite differently for another type of problem. This type of knowledge organization is difficult to capture with a measure that relies on static, mutually exclusive categories.

Generalizability of Findings

The use of CDPs to assess clinical problem-solving has both advantages and disadvantages. The primary advantage of this method is that it allows a high degree of standardization in the presentation of patient findings. This was especially important for the present study, because a standard set of patient data was needed with which to assess the effects of the intervention. In addition, the CDPs provide a convenient method of examining subjects' cue acquisition.

The main disadvantage of the CDPs is that they have relatively low fidelity. This type of format does not include subtle visual cues from the patient and does not assess patient management skills. Moreover, the provision of data item categories tended to cue subjects on what information to seek. For these reasons, the external validity of the CDPs is questionable.

A second threat to external validity was the limited sample of cases used in the study. Because problem-solving has often been shown to be case specific, it is possible that the present results might only apply to diagnostic problems in pediatric cardiology. A more realistic conclusion might be that the cases in the present study could generalize to other medical problems that are comparable in the problem-solving process required (for example, sensing, defining, resolving), the clinical discipline involved, and the context of care (for example, chronic versus acute) (Bashook, 1976).

Validity of Verbal Reports

Another potential methodologic problem involved the use of process tracing, or thinking aloud protocols. Although this methodology produced interesting qualitative information on subjects' lines of reasoning, it may not have always assessed subjects' thoughts accurately. For example, subjects may have been unaware of, or unable to verbalize, part of their reasoning process because of the rapidity of their cue hypothesis associations. This phenomenon seemed to occur in subjects who said that a disease "popped into their heads", but when they were asked about it, could not pinpoint which cue triggered the disease. Often, the cues that triggered these diseases were more apparent to the experimenter than to the subject.

In addition, subjects may have altered the way in which they reported their thoughts if they were concerned about having their diagnostic thinking evaluated. This may have led to a reluctance to verbalize uncertain or incomplete hunches. Another potential for bias in verbal reports might have resulted from differences in subjects' level of comfort with discussing their thoughts. Some subjects were more verbal than others and discussed their reasoning more frequently and in more detail than subjects who were by nature less verbal.

All of these potential biases associated with veral reports warrant caution in generalizing from these findings to the actual diagnostic reasoning process. However, they were not expected to differentially affect the experimental or control groups because it was assumed that any differences in subjects' verbosity were randomly distributed between the two groups.

The possibility does exist, though slight, that the experimenter encouraged more thinking aloud in the experimental than control group without actually being aware of it. Because this possibility was acknowledged before the data collection began, every attempt was made to encourage subjects' verbalizations equally for both groups.

Relative Efficacy of the Intervention Components

The intervention had several components which, singly or in combination, may have produced the significant effects. The effect of the separate intervention components is impossible to disentangle, but future research comparing these specific components may shed light on this issue.

Impressionistic data from subjects' debriefings provided some clues to the relative efficacy of the different intervention components. Most experimental subjects said that the organization of the data into the table and flow diagrams

was more helpful to them than the lecture on problem-solving strategy and heuristics. Several subjects commented that the heuristics covered in the lecture seemed obvious to them. One exception was that the discussion on the rapidity of early hypothesis generation appeared to surprise some subjects. These subjects mentioned that they were unaware that hypotheses were triggered so quickly clinical workups.

Another interesting finding was that some subjects found the table and flow diagrams counterproductive to their own knowledge organization strategies, as discussed earlier.

Theoretical Issues

Implications of the Intervention

The intervention demonstrated a positive effect on some measures of problem-solving but appeared noneffective and possibly counterproductive for others. Therefore, an important question to explore is for what cases the intervention might be expected to facilitate problem-solving and for what cases it might have a small or negative effect.

A re-examination of the findings on diagnostic accuracy for the typical and atypical cases best illustrates this question. It was reasoned that experimental subjects were less accurate than controls in their diagnosis of the typical case because some of the salient cues were de-emphasized in the experimental flow diagrams. In cognitive psychological terms, diagnostic errors were due to a failure to use these salient cues to encode more features of the target disease, which would have allowed better discriminations between this disease and its competitors (Anderson, 1979). In this case, then, the experimental intervention

was counterproductive because it de-emphasized certain specific disease knowledge. One conclusion from this finding might be that a teaching intervention that is strictly focused on enhancing the formation of good disease groupings in memory addresses only part of the picture. For some cases, such as the typical case, it is perhaps more important to give appropriate weight to specific disease knowledge which is critical for a correct diagnosis. This point is particularly important to the present study, given that the medical students had a relatively sparse and imprecise knowledge base to begin with. Because of their limited knowledge bases, subjects had to rely on a small number of salient cues and consequently, the de-emphasis of only one or two of these cues could have led to a misdiagnosis.

With regard to the atypical case, experimental subjects were more accurate than controls in their diagnosis. Unlike for the typical case, diagnostic accuracy for the atypical case required the use of an LCS that crossed classical disease categories to include TAPVC. This is because many of the cues for the atypical case led subjects to consider milder diseases, so that if they did not include TAPVC in their LCS initially, they were unlikely to evaluate it actively in the presence of more serious cues occurring later in the workup.

The inclusion of TAPVC (an admixture lesion) in an LCS with ASD, ECD, and PAPVC (left-to-right shunts) represented an expert knowledge organization strategy. This strategy was emphasized in the intervention and appeared help experimental subjects with their problem-solving. On the other hand, this knowledge organization strategy may not have been immediately apparent to control subjects because their instructional materials (like the text from which they were taken) did not emphasize it. Thus, the atypical case provided a clear illustration of the importance of knowledge representation to clinical problem-

solving. As Feltovich notes, "Tight memory organization among competitor diseases, in a category or similar type of memory unit, supports diagnosis by providing interdisease activation; when one member is activated, other plausible candidates are likely to be considered" (p. 159).

In summary, these results suggest that the nature of the clinical problem had a substantial impact on the efficacy of the intervention. The findings warrant caution in applying the intervention in its present form to a variety of clinical problems. The intervention might be expected to oversimplify problems that require a knowledge base with precise, detailed disease knowledge for discriminating among diseases within LCSs that are readily apparent. The application of the intervention to this type of clinical problem might even be counterproductive for students who have a sparse and imprecise knowledge base. On the other hand, the intervention might be expected to be quite effective for atypical cases that require the use of LCSs that are not immediately obvious to the medical student.

Knowledge Representation

The primary question of this study was whether expert-like clinical reasoning skills could be taught to preclinical medical students. The answer to this question is both simple and complex. It is simple when clinical expertise is rather narrowly defined in terms of the amount of critical information acquired in a workup and the point at which the correct hypothesis is first generated. Theoretically, proficiency on these measures suggests both the use of a hypothesis driven data gathering strategy and the strengthening of salient cue hypothesis associations in long term memory. As discussed earlier, these measures were either strongly or moderately influenced by the intervention.

The answer to the question of whether expert-like clinical reasoning skills could be (and were) taught is more complicated when clinical expertise is defined in terms of knowledge representation and the use of strategies to rapidly access this information for problem-solving. This issue will be discussed next.

The purpose of the intervention was twofold. First, the table and flow diagrams were intended to help subjects organize the information in memory similar to the knowledge organization of more experienced physicians. This process involved enriching subjects' knowledge representations by creating disease categories that crossed over the classical categories described in introductory textbooks. Examples of classical categories are left to right shunts and admixture lesions. An example of a more expert-like category is that of atrial level shunts, or diseases with increased blood flow to the right side. The purpose of augmenting subjects' knowledge representations was to strengthen the associations they formed between logically competing diseases. Thus, when cues would trigger a more common disease in the LCS, for example, ASD, the other competing diseases would be activated as well. Once activated, these diseases could be systematically ruled in or out on the basis of additional information.

The effectiveness of expert-like knowledge representations was best demonstrated in cases where classical disease categories were overly restrictive. The atypical case (TAPVC) was the clearest example of this. Several of the cues for this case led subjects to trigger the left to right shunts of ASD, ECD, and PAPVC. Subjects who used only the category of left to right shunts ran the risk of not activating TAPVC, because this disease was an admixture lesion. In contrast, a relatively straightforward disease, such as ASD, did not require crossing over classical disease categories, and therefore, would not have been expected to elicit expert-novice differences in knowledge representation.

It must be stressed that the knowledge representation strategy used in the intervention was only an approximation of what would truly be considered diagnostic expertise. In reality, the knowledge representations of experts for the diseases used in the present study are far more dense, precise, and interconnected than those presented in the intervention. For example, one pediatric cardiologist working on the TAPVC case named nine different variations of TAPVC, each distinguished by slight anatomical differences (Feltovich, 1981).

The second purpose of the intervention was to provide an actual problem-solving strategy that involved the use of clusters of critical cues to group and differentiate diseases. Both the lecture on clinical problem-solving and the flow diagrams were designed to facilitate such a strategy. This strategy corresponds to the use of domain-specific procedural knowledge (Newell, 1969), "plans" (VanLehn & Brown, 1979), or "scripts" (Schank & Abelson, 1977). Another purpose of the problem-solving strategy was to limit the information available to clusters of the most diagnostic cues, thus reducing the burden on working memory that would be associated with using a large body of information for diagnostic decisions.

The knowledge representation and problem-solving strategies involved either reorganizing subjects' knowledge bases or emphasizing different parts of it. No specific knowledge was added so that the net content of experimental and control subjects' knowledge bases was comparable. As mentioned before, however, expertise involves more than the enrichment of classical disease categories or the use of well-developed procedural knowledge. It also involves the memorization of a great deal of extremely detailed information with which to make fine discriminations between diseases and disease variants. This aspect of expertise was not investigated because such knowledge accrues only through

extensive clinical experience, which allows a "fine tuning" of one's knowledge base.

Now that the general purposes of the intervention have been more fully described, the evidence for its efficacy can be examined. As noted earlier, the prototypic case did not discriminate between experimental and control subjects on any measures related to LCS use. This finding was hypothesized, given that an uncomplicated case was not expected to elicit differences between experimental and control sbjects' knowledge representation or problem-solving strategies.

For the typical case, no group differences were found in critical cue evaluation but control subjects were more accurate in their diagnoses than experimental subjects. The locus of this difference was speculated to be in the different emphasis that a critical cue, the systolic ejection click, received in the two knowledge bases. Control subjects were informed that a left to right shunt with aortic enlargement (as demonstrated by the systolic ejection click) was highly diagnostic of PDA. In contrast, experimental subjects' flow diagrams did not even include the systolic ejection click, and the table in which it was included did not emphasize its diagnosticity in this way. In terms of the preceding discussion, it appears that the control subjects inadvertantly received a powerful "script" for diagnosing PDA while experimental subjects did not.

The group differences in diagnostic accuracy that were found for the typical case seemed to be due more to experimental subjects' lack of a powerful script to use with ambiguous cues, rather than to 1) their failure to generate the appropriate LCS, or 2) their failure to trigger and actively evaluate PDA with respect to the critical cues of the case. This speculation is supported by the finding that experimental subjects were as active as controls in their evaluation of critical cues with respect to LCS members. Moreover, experimental and

control subjects did not differ in the extent to which they evaluated the correct disease with respect to the critical cues for the case. (It is also interesting to note here that there were no differences between diagnostically accurate and inaccurate subjects on the measure of correct hypothesis evaluation).

To summarize the findings for the typical case, the intervention may have had a negative impact on experimental subjects because, in the attempt to streamline their information, a problem-solving strategy was unintentionally omitted from their knowledge base. This strategy was included in control subjects' knowledge base and appeared to help them diagnose the case.

With regard to the atypical case, the findings are somewhat more encouraging, although not immediately apparent. First, there were no significant group differences in the extent of overall LCS evaluation or correct hypothesis evaluation with respect to the critical cues. On the surface, these findings suggest that control subjects were as active as experimental subjects in evaluating the appropriate LCS, as well as the most appropriate member of the LCS, for this case. This lack of group differences was due to the fact that many of the control subjects used an expert-like knowledge representation strategy, and thus washed out the effect of the intervention. As discussed earlier, almost half of the control subjects had actually constructed the LCS in their notes, specifically including TAPVC with the left to right shunts of ASD, ECD, and PAPVC. On the other hand, many other control subjects could barely recall TAPVC, much less actively evaluate it with respect to the critical cues. Therefore, some control subjects did not perform as well as their experimental counterparts, but this did not occur with enough frequency to produce significant group differences. Two additional findings lend support to this idea. First, control group subjects were significantly less accurate than experimental subjects

in their diagnosis of the TAPVC case. Second, a comparison of accurate and inaccurate subjects revealed that, on the average, accurate subjects were almost three times more active in the extent to which they evaluated the correct hypothesis with respect to the critical cues (see Table 21).

One interpretation of these findings might be that inaccurate subjects did not have TAPVC strongly associated with other LCS members, so that it was not triggered along with them. If TAPVC was triggered, it may not have been actively evaluated throughout the workup but perhaps mentioned in response to the X-Ray results. A second explanation, not mutually exclusive, is that inaccurate subjects did not have available (if they were control subjects) or did not use (if they were experimental subjects) the cue cluster for assessing symptom burden early in the workup. This "script" may have provided an early successful diagnostic pathway for subjects.

To summarize these findings, it appears that subjects who misdiagnosed the atypical case may not have represented the correct disease adequately in memory and may not have had available, or, if available, did not use, the procedural knowledge for activating and evaluating TAPVC early in the workup. Further, it is speculated that those most prone to such errors tended to be control subjects who failed to recognize and therefore learn the appropriate disease groupings for this case and, who could not benefit from the procedural knowledge given to experimental subjects. These findings suggest that although many control subjects did not need an intervention to help them group diseases well and to use good diagnostic strategy, others may well have benefited from such an intervention. If true, this speculation would suggest that the effect of the intervention was to help those subjects form LCSs and use good diagnostic strategy when they might not have been so inclined.

The complementary part of this discussion is that some control subjects demonstrated good clinical reasoning skills in the absence of any intervention. This finding argues for a consideration of the role of individual differences in problem-solving skills and how such differences may moderate the impact of a problem-solving intervention.

Fostering Problem-Solving Expertise

There are both similarities and differences between the present findings and those from problem-solving in other fields such as psychology (verbal learning), physics, and mathematics. The present study has demonstrated that, to some degree, knowledge representation strategies that facilitate problem-solving can be taught. This concurs with the findings of Wortman and Greenberg (1971), Shavelson (1972), and Ausubel (1960) that knowledge representation strategies specified by the experimenter can facilitate subjects' organization of information in long-term memory, help them retain and integrate the material, and facilitate problem-solving. Perhaps more important, the present results have shown that knowledge representation strategies taught to subjects can, in some cases and for some measures, result in problem-solving performance that closely resembles that of experts. Schoenfeld and Hermann (1982) showed similar results in their use of an instructional intervention designed to facilitate expert-like mathematics problem-solving in students.

There are also differences between research on problem-solving in medicine and problem-solving in such fields as physics and mathematics. The latter fields are circumscribed in comparison to medicine, and the rules for problem-solving are more straightforward and unambiguous. Moreover, the types of problem-solving strategies which are most successful in mathematics and physics are

somewhat different than those for medicine. For example, the key to successful mathematics problem-solving primarily involves the use of heuristics for simplifying complex equations (Schoenfeld, 1980). Successful physics problem-solving involves the ability to abstract underlying physics laws from problems (Chi, Feltovich, & Glaser, 1981). Clinical problem-solving requires a number of different skills, including the ability to group together logically competing diseases, accurately evaluate salient cues, and accurately interpret ambiguous clinical information, to name a few. Also, whereas mathematics and physics problems are fairly uniform, problems in clinical medicine make take an infinite variety of forms which require quite diverse problem-solving skills.

Little research is available on the teaching of expert-like clinical problem-solving strategies in medicine from an information processing paradigm. Most of the research is descriptive, focusing on differences in clinical problem-solving skills at different levels of experience (Bordage, 1983; Norman et al, 1979; Norman, 1983; Neufeld et al, 1981; Feltovich, 1981). The research that does exist on facilitating clinical problem-solving shows mixed results (Allal, 1974; Gordon, 1974). In light of the previous discussion on the diversity of clinical problems as well as the variety of skills required to solve these problems, mixed results are not surprising.

The present study provides some encouragement to the notion that problem-solving strategies can be taught in medicine, despite the complexity of the field. It also suggests two caveats. First, medical students cannot be expected to assimilate expert problem-solving strategies in tabula rasa fashion. Their unique prior experiences and problem-solving styles will modify the effect of the intervention, sometimes in a deleterious manner. Second, the efficacy of a clinical problem-solving intervention will depend on the nature of the problem

itself. It is reasonable to assume that no one type of intervention will be adequate for a variety of clinical problems.

Implications for Medical Education

McGuire (1984) rather brusquely summarizes much of the research in medical education this way: "...the process of clinical reasoning can be learned in a conscious, systematic way and...medical schools can facilitate and enhance that learning provided faculties are willing to abandon their compulsive and hypocritical advocacy of thorough, unguided data collection, as the first step in that process" (p. 4).

The results of the present study lend some support to this assertion. It seems that the organization of a knowledge base and the strategies for its use may be as important as the content of the knowledge base itself. Although medical students learn huge amounts of data, they are not formally taught what is important and what is not, nor what knowledge organization strategies are most efficient for later use in clinical problem-solving. The GPEP report (1985) descriptively labels this "dense pack" medical education. Perhaps medical education should shift its focus away from barraging students with endless quantities of details, and instead, call upon expert clinicians to teach their knowledge representation and problem-solving strategies to students. This way, students can benefit from the experiences of expert clinicians early on in their careers. At the very least, medical students could benefit from more extensive training in the following:

1) Assessing the relative salience of clinical information and using combinations of these salient data as triggers for competing sets of diseases.

2) Using initial hypotheses generated by early salient cues to focus and guide one's subsequent data collection.
3) Avoiding the collection of noncontributory clinical information, and avoiding the use of this information to bolster confidence in one's hypothesis.
4) Learning to organize disease knowledge according to similarity of underlying pathophysiology and hence, clinical presentation.
5) Learning to work with atypical cases that present with ambiguous clinical findings.

Although it is clear that much needs to be done to change medical education, several issues remain that warrant caution in planning for curricular changes. First, the most powerful clinical problem-solving strategies are those that are procedural (Newell, 1969), not general (for example, the hypothetico-deductive method). Unfortunately, procedural strategies are dependent on a specific knowledge base and therefore, cannot be readily generalized to other problem-solving domains. The teaching of procedural knowledge would be prohibitively cumbersome and time consuming unless some level of generality could be used, such as the organ system involved and the acuteness of the problem.

Second, the use of packaged diagnostic strategies, such as algorithms, might foster rigid adherence and excessive dependence, especially among those with the least amount of clinical experience. The result might be a loss of flexibility which is so necessary for solving more complex clinical problems involving subtle disease variations. This may have occurred for some subjects in the present study when they used the flow diagrams too concretely, and not merely as a guide for their workups.

Third, the question of when to teach clinical problem-solving remains unanswered. It is possible that clinical problem-solving concepts taught in the first year of medical school may ease the transition to the clinical years. However, it is not known at what point in medical school this type of training would have the most impact.

A fourth issue is that teaching some aspects of expert clinical problem-solvng may entail certain risks. For example, while this study indicated that early hypothesis generation could be facilitated in at least one case, the risk is that if the early hypothesis generated is incorrect, the student may prematurely focus and limit the workup, and, thus, miss the cues for the correct hypothesis entirely. Similarly, efficient cue acquisition can be taken to an extreme where too little information is acquired before making diagnostic conclusions (Voytovich, Rippey, & Suffredini, 1985).

A fifth issue regarding the teaching of clinical problem-solving is that individual differences in learning and problem-solving styles may modify the effect of instruction. It may not be feasible to tailor instructional materials to various learning styles, but perhaps students could be encouraged to better understand their own learning styles in order for them to reorganize the material in an optimal way for their particular style. This process might even be formalized by having students complete a test of learning style to provide them with specific feedback and guidance regarding the best way to approach instructional materials.

Directions for Future Research

The present study raised three major issues which warrant further exploration. The first issue is one of individual differences in learning styles. Because

individual differences exerted a powerful influence on the effect of the intervention, it is important to formally measure them in future research on teaching clinical problem-solving skills. Perhaps one of the several extant measures of cognitive style could be used to measure these individual differences. It would then be possible to examine the interaction between a problem-solving intervention and students' learning styles.

A second issue is one of generalizability. It is not known how effective the present intervention would be for different types of clinical problems. Further research is needed to determine the generalizability of the intervention.

A third issue involves the ecological validity of the problem-solving assessment format. The drawbacks of patient management problems, whether presented in computerized or paper and pencil format, are well known. It is therefore important to determine the degree to which a clinical problem-solving intervention can be expected to benefit students when they are faced with real clinical problems. Perhaps the effects of future clinical problem-solving interventions can be assessed with higher fidelity simulations.

REFERENCES

Allal, L.K. & Shulman, L. (1974). Training medical students to generate diagnostic problem formulations. Proceedings of the 13th Annual Conference on Research in Medical Education, Washington, D.C.

Ausubel, D.P. (1960). The use of advance organizers in the learning and retention of meaningful verbal material. Journal of Educational Psychology, 51, 267-272.

Anderson, J.R. (1978). Arguments concerning representations of mental imagery. Psychological Review, 8, 249-277.

Anderson, J.R., Kline, P.J., & Beasley, C.M. (1979). A general learning theory and its application to schema abstraction. In G.H. Bower (Ed.), The psychology of learning and motivation. New York: Academic Press.

Balla, J.J., Elstein, A.S., & Gates, P. (1983). Effects of prevalence and test diagnosticity upon clinical judgments of probability. Methods of Information in Science, 22, 25-28.

Barr, A. & Feigenbaum, E.A. (Eds.). (1981). Handbook of artificial intelligence. Los Altos, CA: Kaufman.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. Acta Psychologica, 44, 211-233.

Barro, A.R. (1973). Survey and evaluation of approaches to physician performance measurement. Journal of Medical Education, 48, 1047-1093.

Barrows, H.S. & Bennett, K. (1972). Experimental studies on the diagnostic (problem solving) skill of the neurologist, their implications for neurological training. Archives of Neurology, 26(3), 273-277.

Barrows, H.S. & Tamblyn, R.M. (1980). Problem based learning: An approach to medical education. New York: Springer-Verlag.

Bashook, P.G. (1976). A conceptual framework for measuring clinical problem-solving. Journal of Medical Education, 51, 109-114.

Bergman, D. & Beck, A. (1983). The influence of of clinical information variables on medical decision making. Paper presented at the Society for Medical Decision Making 6th Annual Meeting, Toronto, Canada.

Bergman, D. & Pantell, R. (1984). The art and science of medical decision making. Journal of Pediatrics, 104(5), 649-656

Bordage, G. (1983). The influence of case structure and knowledge structure on diagnostic reasoning in medicine. Proceedings of the 22nd Conference on Research in Medical Education, Washington, D.C.

Casscells, W., Schoenberge, A. & Graboys, T.B. (1978). Interpretation by physicians of clinical laboratory results. New England Journal of Medicine, 299, 999-1001.

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

Chi, M.T.H., Feltovich, P.J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. Cognitive Science, 5, 121-152.

Christensen-Szalanski, J.J., & Bushyhead, J.B. (1981). Physician's use of probabilistic information in a real clinical setting. Journal of Experimental Psychology: Human Perception and Performance, 7, 928-935.

Clancey, W.J. & Letsinger, R. (1983). NEOMYCIN: Reconfiguring a rule-based expert system for application to teaching. In W.J. Clancey & E. Shortliffe (Eds.), Readings in medical artificial intelligence: The first decade. Addison-Wesley.

Clancey, W.J. (1983). GUIDON. Journal of Computer-Based Instruction, 10, 1&2, 8-15.

Dawes, R.M. & Corrigan, B. (1974). Linear models in decision making. Psychological Bulletin, 81, 95-106.

deGroot, A.D. (1965). Thought and choice in chess. New York: Basic Books.

Detmer, D.E., Fryback, D.G., & Gassner, K. (1978). Heuristics and biases in medical decision making. Journal of Medical Education, 53, 682-683.

Duda, R.O., & Shortliffe, E.H. (1983). Expert systems research. Science, 220, 4594, 261-268.

Eddy, D.M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), Judgment under uncertainty: Heuristics and biases New York: Cambridge University Press.

Einhorn, H.J. (1974). Expert judgment: Some necessary conditions and an example. Journal of Applied Psychology, 59, 562-571.

Elstein, A.S. (1976). Psychological research and medical practice. Science, 194, 696-700.

Elstein, A.S. & Bordage, G. (1979). Psychology of clinical reasoning. In G. Stone, F. Cohen, & N. Adler (Eds.), Health psychology-a handbook. San Francisco, CA: Jossey-Bass.

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

Ericsson, K.A., & Simon, H.A. (1980). Verbal reports as data. Psychological Review, 87, 215.

Feltovich, P.J., Johnson, P.E., Moller, J.H., & Swanson, D.B. (1983). LCS: The role and development of medical knowledge in diagnostic expertise. In W.J. Clancey & E. Shortliffe (Eds.), Readings in medical artificial intelligence: The first decade. Addison-Wesley.

Feltovich, P.J. (1981). Knowledge based components of expertise in medical diagnosis. (Tech. Rep. No. PDS-2). Pittsburgh: University of Pittsburgh, Learning Research and Development Center.

Ferrell, W.R., & McGoey, P.J. (1980). A model of calibration for subjective probabilities. Organizational Behavior and Human Performance, 26, 32-53.

Fischhoff, B. (1975). Hindsight and foresight: The effect of outcome knowledge on judgment under uncertainty. Journal of Experimental Psychology: Human Perception, 1, 288.

Fleiss, J.L. (1973). Statistical methods for rates and proportions. New York: John Wiley and Sons.

Fryback, D.G. (1978). Baye's theorem and conditional non-independence of data in medical diagnosis. Computers in Biomedical Research, 11, 423-428.

Gilbert, J.P., McPeek, B., & Mosteller, F. (1977). Statistics and ethics in surgery and anesthesia. Science, 198, 684-689.

Goldberg, L.R. (1970). Man versus model of man: A rationale, plus some evidence for a method of improving on clinical inferences. Psychological Bulletin, 73, 422-432.

Goran, M.J., Williamson, J.W., & Gonnella, J.S. (1973). The validity of patient management problems. Journal of Medical Education, 48, 171-177.

Gordon, M. (1974). Heuristic training for diagnostic problem-solving. Proceedings of the 13th Annual Conference on Research in Medical Education, Washington, D.C.

Guenin, P. & Schwartz, M. (1982). Computer Assisted Medical Problem Solving System. Philadelphia: Dacis Software Corp.

Gustafsen, D.H., Kestly, J.J., Greist, J.H., & Jansen, N.M. (1971). Initial evaluation of a subjective bayesian diagnostic system. Health Service Research, 6, 204-213.

Hammond, K.R., Hursch, C.J., & Todd, F.J. (1964). Analyzing the components of clinical inference. Psychological Review, 71, 6, 438-456.

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

Hoffman, P.J. (1960). The paramorphic representation of clinical judgment. Psychological Bulletin, 69, 338-349.

Johnson, P.E., Duran, A.S., Hassebrock, J.M., Prietula, M., Feltovich, P.J., & Swanson, D.B. (1981). Expertise and error in diagnostic reasoning. Cognitive Science, 5, 235-283.

Kahneman, D., Slovic, P., & Tversky, A. (Eds). (1982). Judgment under uncertainty: Heuristics and biases. New York: Cambridge University Press.

Kassirer, J.P. (1983). Teaching clinical medicine by iterative hypothesis testing: Let's preach what we practice. New England Journal of Medicine, 309, 921-923.

Kassirer, J.P. & Gorry, G.A. (1978). Clinical problem solving: A behavioral analysis. Annals of Internal Medicine, 89, 245-255.

Kassirer, J.P., Kuipers, B.J., & Gorry, G.A. (1982). Toward a theory of clinical expertise. American Journal of Medicine, 73, 259.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. Journal of Experimental Psychology: Human Learning and Memory, 6, 107-118.

Langley, P. & Simon, H.A. (1980). The central role of learning in cognition. In J.R. Anderson (Ed.), Cognitive skills and their acquisition. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Lichtenstein, S., Fischhoff, B., & Phillips, D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), Judgment under uncertainty: Heuristics and biases. New York: Cambridge University Press.

Ludwig, D., & Heilbron, D. (1983). The design and testing of a new approach to computer-aided differential diagnosis. Methods of Information in Medicine, 22, 156-166.

Lyon, D. & Slovic, P. (1976). Dominance of accuracy estimation and neglect of base rates in probability estimation. Acta Psychologica, 40, 287-298.

Mandler, G. Organization and memory. In K.W. Spence & J.T. Spence (Eds.), The Psychology of learning and motivation: Advances in research and theory. Vol. 1, New York: Academic Press, 1967.

Marshall, J.R. (1983). How we measure problem-solving ability. Medical Education, 17, 319-324.

Martin, I.C. (1975). Empirical examination of the sequential management problem for measuring clinical competence. Proceedings of the 14th Annual Conference on Research in Medical Education, 11, 83-88.

McGuire, C.H. (1980). Assessment of problem-solving skills, 1. Medical Teacher, 2(2), 74-79.

McGuire, C.H., & Babbot (1977). Simulation technique in the measurement of problem-solving skills. Journal of Educational Measurement, 11, 1-10.

Miller, G.A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review, 63, 81-97.

Moller, J.H. (1978). Essentials of pediatric cardiology. Philadelphia: A.F. Davis.

Moore, R.A., Aitchison, J., Parker, L.S., & Taylor, T.R. (1974). Use of information in thyrotoxicosis treatment allocation. Methods of Information in Medicine, 13, 88-92.

Moss, A.J., Adams, F.H., & Emmanouilides, G.C. (1977). Heart disease in infants, children and adolescents. Baltimore, MD: Williams and Wilkins Company.

Neufeld, V.R., Norman, G.R., Feightner, J.W., & Barrows, H.S. (1981). Clinical problem-solving by medical students: A cross-sectional and longitudinal analysis. Medical Education, 15, 315-322.

Newble, D.I., Hoare, J., & Baxter, A. (1982). Patient management problems: Issues of validity. Medical Education, 16, 137-142.

Newell, A. (1969). Heuristic programming: Ill-structured problems. In J. Aronofsky (Ed.), Progress in operations research, vol. III. New York: Wiley.

Newell, A. & Simon, H.A. (1972). Human problem solving. Englewood Cliffs: Prentice-Hall.

Nisbett, R.E., & Wilson, T.D. (1977). Telling more than we can know: Verbal reports on mental processes. Psychological Review, 84, 231.

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

Norman, G. (1983). Features of expert memory and medicine. Proceedings of the 22nd Conference on Research in Medical Education, Washington, D.C.

Norusis, M.J. (1985). SPSSX advanced statistics guide. New York: McGraw-Hill.

Oskamp, S. (1965). Overconfidence in case-study judgments. Journal of Consulting Psychology, 29, 261-265.

Patel, V. (1983). Cognitive processes in clinical reasoning of medical students and physicians. Proceedings of the 22nd Conference on Research in Medical Education, Washington, D.C.

Patil, R.S., Szolovits, P. & Schwartz, W.B. (1983). Causal understanding of patient illness in medical diagnosis. In W.J. Clancey & E.H. Shortliffe (Eds.), Readings in medical artificial intelligence: The first decade. Addison-Wesley.

Pople, H.E. (1982). Heuristic methods for imposing structure on ill-defined problems: The structuring of medical diagnosis. In, P. Szolovits (Ed.), Artificial intelligence in medicine. Boulder, CO: Westview Press.

Schank, R. & Abelson, R. (1977). Scripts, plans, goals, and understanding. Hillsdale, NJ: Lawrence Erlbaum Associates.

Schoenfeld, A.H. (1980). Teaching problem-solving skills. American Mathematical Monthly, 87, 794-805.

Schoenfeld, A.H. & Hermann, D.J. (1982). Problem perception and knowledge structure in expert and novice mathematical problem solvers. Journal of Experimental Psychology: Learning, Memory, and Cognition, 8(5), 484-494.

Shakun, E.N., Taylor, W.C., & Osbaldeston, W. (1976). The relationship between computerized patient management problems and other pediatric certifying examinations. Proceedings of the 15th Annual Conference on Research in Medical Education, 167-171.

Shavelson, R.J. (1972). Some aspects of the correspondence between content structure and cognitive structure in physics instruction. Journal of Educational Psychology, 63, 225-234.

Shortliffe, E.H. (1976). Computer-Based Medical Consultations: MYCIN. New York: American Elsevier.

Shortliffe, E.H., Scott, A.C., Bischoff, M., Campbell, A.B., van Melle, W., & Jacobs, C. (1981). ONCOCIN: An expert system for oncology protocol management. In Proceedings of the Seventh International Joint Conference on Artificial Intelligence. Menlo Park, CA: AAAI.

Simon, H.A., & Chase, W.G. (1973). Skill in chess. American Scientist, 61, 393-403.

Slovic, p., Rorer, L.G., & Hoffman, P.J. (1971). Analyzing use of diagnostic signs. Investigative Radiology, 6, 18-26.

Szolovits, P., & Pauker, S.G. (1976). Research on a medical consultation system for taking the present illness. In Proceedings of the Third Illinois Conference on Medical Information Systems. Chicago: University of Illinois at Chicago Circle.

Szolovits, P. & Pauker, S.G. (1983). Categorical and probabilistic reasoning in medical diagnosis. In W.J. Clancey & E.H. Shortliffe (Eds.), Readings in medical artificial intelligence: The first decade. Addison-Wesley.

Taylor, P.J., Harasym, P.H., & Laurenson, R.D. (1978). Introducing first year medical students to early diagnostic hypotheses. Journal of Medical Education, 53, 402-409.

Thornbury, J.R., Fryback, D.G., & Edwards, W. (1975). Likelihood ratios as a measure of the diagnostic usefulness of excretory urogram information. Radiology, 114, 561-565.

VanLehn, K., & Brown, J.S. (1979). Planning nets: A representation for formalizing analogies and semantic models of procedural skills. In R.E. Snow, P.A. Federico, & W.E. Montague (Eds.), Aptitude, learning and instruction: Cognitive process analyses. Hillsdale, NJ: Lawrence Erlbaum Associates.

Voytovich, A.E., Rippey, R.M., & Suffredini, A. (1985). Premature conclusions in diagnostic reasoning. Journal of Medical Education, 60, 302-307.

Vu, N.V. (1979). Medical problem-solving assessment: A review of methods and instruments. Evaluation and the Health Professions, 2, 281-307.

Wallsten, T.S. (1978). Three biases in the cognitive processing of diagnostic information. Unpublished manuscript, Psychometric Laboratory, University of North Carolina, Chapel Hill.

Wallsten, T.S. (1981). Physician and medical student bias in evaluating diagnostic information. Medical Decision Making, 1, 145-164.

Ways, P.O., Loftus, G., & Jones, M. (1973). Focal problem teaching in medical education. Journal of Medical Education, 48, 565-570.

Weiss, S., Kulikowski, C., & Safir, A. (1978). Glaucoma consultation by computer. Computers in Biology and Medicine, 8, 1, 25.

Wood, G. (1983). Cognitive Psychology: A Sills Approach. Belmont, CA: Brooks/Cole Publishing Co.

Wortman, P.M. (1972). Medical diagnosis: An information processing approach. Computers and Biomedical Research, 5, 315-328.

Wortman, P.M. & Greenberg, L.D. (1971). Coding, recoding, and decoding of hierarchical information in long-term memory. Journal of Verbal Learning and Verbal Behavior, 10, 234-243.

Wortman, P.M. (1966). Representation and strategy in diagnostic problem solving. Human Factors, 8, 48-53.

Wright, H.J., Stanley, I.M., & Webster J. (1983). The assessment of cognitive abilities in clinical medicine. Medical Education, 17, 31-38.

APPENDIX A

Outline of Lecture on Clinical Problem-Solving for
Experimental Group Subjects

I. Clinical Reasoning.

Research on clinical reasoning has shown that experienced physicians rapidly generate initial diagnostic hunches early in their workups. In fact, the first hypothesis is typically generated within the first 5-10 seconds of the workup. The first correct diagnostic hypothesis is often generated within 300 seconds (5 minutes).

This rapid generation of hypotheses points to the physician's development and use of strong associations between patient cues and disease prototypes in his or her memory. Essentially, this is a process of pattern matching between cues and diseases.

In addition to rapid hypothesis generation, the experienced physician often works with sets of 3 to 5 hypotheses at a given time. These diseases are grouped together because they present with similar clinical findings. The use of 3 to 5 hypotheses also prevents too narrow a focus on 1 hypothesis early in the workup.

II. Data Driven Hypothesis Generation and Hypothesis Driven Data Gathering.

Experienced physicians work in two complementary modes during clinical problem-solving. The first is called data driven hypothesis generation and simply involves generating hunches based on the findings from the patient. Once an initial set of hunches (or initial problem formulation) is generated, the physician can then reason forward to the kinds of findings he or she would expect if these hunches were correct. The physician then proceeds to the step of gathering data to support or refute the initial set of hunches. This is the second step, called hypothesis driven data gathering. Finally, the initial hunches are revised, discarded, etc., and the process continues.

While these two types of reasoning may seem simple and even obvious, they are important in clinical reasoning because they provide a strategy for organizing an overwhelming amount of clinical information.

It is important to remember that these two problem-solving modes are very fluid, not rigid, and that normally, the physician moves back and forth between the two several times during the workup.

III. Clinical Problem-Solving Heuristics

There are a few problem-solving heuristics, or rules of thumb, that may help in the clinical problem-solving process.

The first heuristic was mentioned earlier, and involves the use of 3 to 5 hypotheses at a given time. This prevents premature closure on an erroneous diagnosis.

Two other heuristics are probably well-known to you as part of the differential diagnostic process. The first is the discrimination heuristic. This heuristic involves searching for clinical information that will rule out one or more diseases in the set actively under consideration.

The second heuristic is the confirmation heuristic, and simply involves the use of clinical information to confirm one of the diseases in the set under consideration as the most likely candidate for a primary diagnosis.

Obviously, the most useful clinical information is that which simultaneously rules out diseases in the differential while ruling in others.

A second point about these heuristics is that they can be applied to whole sets of diseases as well as single diseases. Thus, some patient cues might be used to rule in or out an entire class of diseases.

A final, very general heuristic is to get the whole picture of the patient, rather than relying on one or two pieces of clinical information you think are most diagnostic. Rarely can you make a definitive diagnosis on the basis of one or two findings, and more important, these findings are often unreliable. Thus, don't perseverate on a few questionable findings. Consider the weight of all of the findings when making your diagnosis.

IV. Variability of Findings.

As mentioned above, clinical findings are notoriously unreliable. Often, they demonstrate a much greater degree of variability than medical texts would have you believe. While it is only through years of experience that you can "fine tune" the range of expected findings for various diseases, for now, it is important that you appreciate the frequency with which ambiguous findings occur in clinical medicine. Therefore, It is important not to discard key findings if they don't perfectly match your expectations for them.

Instead, there are two strategies to adopt when confronted with "grey area" findings. The first is to determine whether this ambiguous finding might better match a variant of the disease you have under consideration. If the finding doesn't match a less common variant of the disease under consideration, then you may need to discard your group of diseases under consideration and adopt another one that matches the finding better. This latter strategy is a difficult judgment call, but the important point is not to get too locked into one set of diseases when the findings don't seem to fit too well (even if the set of diseases seemed appropriate initially). You need to be flexible enough to "jump" disease categories when the findings become increasingly discrepant with your expectations for your initial hypotheses.

APPENDIX B

INSTRUCTIONS TO SUBJECTS

This is a study of diagnostic thinking. You will be presented with computerized exercises in which you are to reach a diagnostic conclusion given a set of data from a patient case.

The data you will be given are based upon actual cases that were evaluated for congenital heart disease. These cases were randomly selected from a large number of medical records. The cases selected may differ in how common they are and how complex they are. Further, it is possible that a normal case (i.e., no heart defect) was randomly selected from the records, or that 2 cases with the same diagnosis were selected. Hence, you should try to diagnose each case independently of the others.

For each exercise, you will be presented with a computerized "patient file" consisting of patient data items. The data are presented in the order: Introduction to case, history, physical exam, and laboratory tests. Within these major categories, data are segmented into small numbered groups. In order to select an item of interest, type in the number corresponding to that item and the information will appear on the video screen.

When you begin each case, I will read the introduction out loud and then you should select additional data items from the major categories that you think are of diagnostic relevance. You should select data items in the following order: Start with the history first, then proceed to the physical exam, and finish with the laboratory data. After you have completed your ordered search, you may follow any order you wish to review previously selected items or to select additional data items from a previous category that you might have overlooked.

Many of the patient data items that you will want to select are under a variety of different sub-categories within the major options. Although you may expect certain groups of findings to be together in one data item, they may be under separate items. Therefore, if the information you expect to find on a data item is not there, do not immediately assume that it is normal. Instead, look under different categories for that item. I will tell you if the item is actually not available, and if so, then you may consider it a normal finding.

While you should be thorough in your selection of patient information, you should also try to be reasonably efficient. You should approach these cases as you might in an actual clinical setting, with a consideration of "real world" constraints, such as the cost of the workup and the time involved.

Prior to your selection of each data item, I would like you to tell me why you are selecting that item and what you expect to find. Once you select that item, I will read it aloud and then you should think out loud about its significance toward formulating a diagnosis for the patient. When you have finished thinking about a data item, go on and select another.

Please try to be as thorough as possible in reporting your thoughts as they arise, even if they seem unimportant to you. In particular, try to make clear when you first think of something, for example, a possible diagnosis, whether the data are consistent or inconsistent with "hunches" you have, and when you eliminate a diagnosis you had been considering.

At three points during each exercise, after history, physical exam, and laboratory tests, I will say "please tell me about hunches." At these points, I would like you to just tell me what diagnoses (if any) you are actively considering for the patient at the time I interrupt. The purpose of my probe is simply to get an explicit listing of the hypotheses you are considering. Report your hypotheses in the manner that best represents the way you are thinking about them. If you have no hypotheses or "hunches", when I interrupt, say so and go on. Throughout the exercise, whenever I judge that an unusual amount of time has passed without your

saying anything, I will say "please talk more." This is just to encourage you to report your thoughts.

At the end of each exercise, I will ask you to give a primary diagnosis. This is the diagnosis you think is the best description of the patient's condition. Also, I would like you to tell me how likely you think that the diagnosis you give is in fact the actual diagnosis. For this estimate, use a five point scale, with 1 referring to "a little likely", and 5 referring to "highly likely".

I will also ask you to give secondary diagnoses. These are diagnoses you feel might apply to the patient, but about which you are not as confident as you are about the primary diagnosis. You may give as many as two secondary diagnoses; you may also give one or none. If you give one or two secondary diagnoses, please rate their likelihood on the same five point scale that you use for the primary diagnosis.

This is a research project and not a test. Your participation will be confidential as described in the consent form; hence, I hope you will be relaxed in doing the exercises.

Do you have any questions?

APPENDIX C

INTRODUCTION

THE NORMAL CARDIOVASCULAR SYSTEM

Figure 1 shows the normal heart and other major components of the cardiovascular system. Starting on the right side of the heart, the right ventricle (RV) of the heart pumps blood across the pulmonary valve (PV), through the pulmonary artery (PA), and into the lungs where the blood receives oxygen. Blood then returns to the heart via the pulmonary veins (PVn) into the left atrium (LA). From the left atrium, oxygenated blood proceeds across the mitral valve (MV) into the left ventricle (LV), where it is pumped across the aortic valve, through the aorta (Ao), and to the body. in the body, oxygen is extracted from the blood which then flows back to the right atrium (RA) of the heart via the vena cavae (VC). Deoxygenated blood from the right atrium flows across the tricuspid valve (TV) into the right ventricle and the cycle repeats. The "upper" chambers of the heart, the atria, are normally separated by the atrial septum, while the "lower" chambers, the ventricles, are normally separated by the ventricular septum.

Role of the ductus arteriosus. The ductus arteriosus, a large channel found normally in all mammalian fetuses, develops from the distal portion of the left sixth aortic arch and connects the main pulmonary trunk (which arises from the right ventricle) with the descending aorta about 5 to 10 mm distal to the origin of the left subclavian artery in a full term infant.

The purpose of the ductus arteriosus is to permit blood to flow to the umbilical placental circulation for gas exchange, rather than to the pulmonary circulation, the normal site of gas exchange in adults. A large pulmonary blood flow during fetal life would represent wasted circulation and the ductus arteriosus therefore reduces the total workload of the fetal ventricles.

The primary change in circulation after birth is a shift of the blood flow for gas exchange from the placenta to the lungs. This is accomplished by the closure of the ductus arteriosus which results from expansion of the lungs and the ensuing increase in arterial 02 saturation. Functional closure of the ductus arteriosus occurs 10-15 hours after birth, and anatomical closure is accomplished by 2-3 weeks of age.

The closure of the ductus arteriosus is important in understanding some congenital heart diseases. For example, patent ductus arteriosus represents a failure of the ductus arteriosus to close after birth, causing a shunting of blood from the arterial to the pulmonary circulations.

HEART SOUNDS

Auscultation

First heart sound. The first heart sound represents closure of the mitral and tricuspid valves and occurs as the ventricular pressure exceeds the atrial pressure at the onset of systole. In children, the first heart sound usually appears single. The first heart sound is accentuated in conditions with increased pulmonary blood flow.

Second heart sound. The second heart sound is of great diagnostic significance in children with congenital cardiac disease.

Splitting of second heart sound. The normal second sound has two components

representing the asynchronous closure of the aortic and pulmonary valves. These sounds signal the completion of ventricular ejection. Aortic valve closure normally precedes closure of the pulmonary valve because right ventricular ejection is longer. The presence on auscultation of the two components, aortic (A2) and pulmonic (P2), is called splitting of the second heart sound.

The time interval between the components varies with respiration. Normally, on inspiration the degree of splitting inreases, while on expiration it shortens. This variation is related to the greater volume of blood that returns to the right side of the heart during inspiration. Since the ejection of this augmented volume of blood requires a longer time, the second heart sound becomes more widely split on inspiration.

Conditions prolonging right ventricular ejection lead to wide splitting of the second heart sound because P2 is delayed further. In addition, the wide splitting becomes fixed, that is, it does not vary in length with respiration. This phenomenon of wide, fixed splitting is present in atrial septal defect because the right ventricle ejects an increased volume of blood.

Intensity of P2 The intensity of the pulmonary component (P2) of the second heart sound is also important. The pulmonic component of the second sound is accentuated whenever the pulmonary arterial pressure is elevated, as in conditions of increased pulmonary arterial blood flow. In general, as the level of pulmonary arterial pressure increases, the pulmonic component of the second sound becomes louder.

Single second heart sound. The finding of a single second heart sound usually indicates that one of the semilunar valves is atretic or severely stenotic because the involved valve does not contribute its component to the second sound. The second heart sound is also single in patients with persistent truncus arteriosus because there is only a single semilunar valve.

Systolic ejection clicks Systolic ejection clicks occur when the semilunar valves open and, therefore, mark the transition from the isovolumetric contraction period to the onset of ventricular ejection. Ordinarily this event is not heard, but in specific cardiac conditions a systolic ejection click may be present. Systolic ejection clicks indicate the presence of a dilated great vessel.

Murmurs

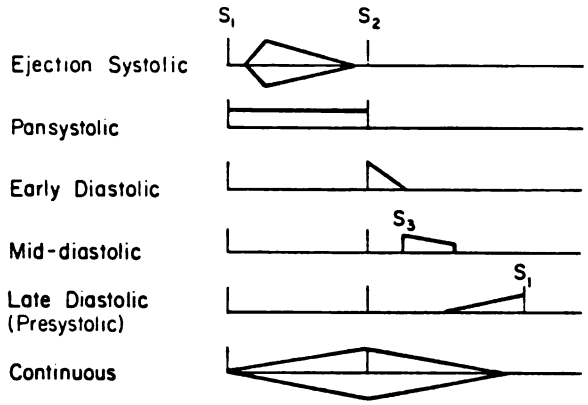
Cardiac murmurs are generated by increased turbulence in the normal pattern of blood flow through the heart. Turbulence results from narrowing of the pathway of blood flow, abnormal communications, or increased blood flow.

Systolic murmurs There are two types of systolic murmurs: pansystolic and ejection systolic. Pansystolic murmurs start with the first heart sound and continue to the second heart sound. This murmur occurs with ventricular septal defect because a pressure difference exists between the left and right ventricles throughout systole.

Ejection systolic murmurs result from turbulent forward blood flow across either the aortic or the pulmonary valve. Ejection murmurs occur with atrial septal defect.

Diastolic murmurs Diastolic murmurs can be classified according to their timing in the cardiac cycle as early, mid, late, or continuous. Early diastolic murmurs occur

immediately following the second heart sound and indicate regurgitation across a semilunar valve (aortic or pulmonary insufficiency). Mid-diastolic murmurs occur at the transition of rapid and slow filling pressure and result from increased volume of forward blood flow across a normal atrioventricular valve. In children, they most commonly occur with increased pulmonary blood flow and therefore increased blood flow into the ventricles. They are sometimes called an inflow murmur. Late diastolic murmurs represent organic obstruction of an atrioventricular valve. These murmurs are crescendo and low pitched. A continuous murmur indicates turbulence throughout the cardiac cycle. Usually this occurs when a communication exists between the aorta and the pulmonary artery or other portions of the venous side of the heart or circulation. Patent ductus arteriosus is the classic example. An illustration of the various murmurs is provided below.



Classification of murmurs, showing location within cardiac cycle and usual contour. S_1 = first heart sound. S_2 = second heart sound. S_3 = third heart sound.

Location of a murmur The location of the maximal intensity of murmurs on the thorax provides information about the anatomic origin of the murmur. Auscultatory areas on the thorax have been described as:

aortic area--along the mid-left sternal border to beneath the right clavicle.
pulmonary area--along the upper left sternal border and beneath the left clavicle
tricuspid area--along the lower left sternal border
mitral area--the cardiac apex.

Loudness of a murmur The loudness of a cardiac murmur is graded on a scale in which grade VI represents the loudest murmur. Conventionally, loudness is indicated by a fraction in which the numerator indicates the loudness of the patient's murmur and the denominator indicates the maximum grade possible (VI); therefore, grade I/VI would be very soft and grade VI/VI would be very loud. A grade IV/VI murmur is associated with a thrill. Thrills are coarse, low frequency vibrations which are palpable. They occur with loud murmurs and are located in the same areas as the maximal intensity of the murmur.

Functional murmurs Distinction between a functional (innocent) murmur and a

significant murmur can be a difficult problem in some children. Although we will describe the characteristics of the commonly heard functional murmurs, only by experience and careful auscultation can one become proficient in distinguishing the functional from the significant murmur.

Functional mumurs have five features that help to distinguish them from significant murmurs:

1. The heart sounds are normal
2. The heart size is normal
3. There are no significant cardiac symptoms
4. The murmurs are grade III/VI or less
5. No thrill is present

There are five types of functional murmurs:

1. Twangy string murmur. This is a low pitched, soft (grade I-III/VI) midsystolic murmur heard along the lower left sternal border. It derives its name from its vibratory character. Because of its location on the thorax it may be misinterpreted as a ventricular septal defect. It can be distinguished because it begins AFTER, not with, the first heart sound as in ventricular septal defect.

2. Pulmonary flow murmur. This soft (grade I-III/VI) low pitched systolic ejection murmur is heard in the pulmonary area. The murmur itself may be indistinguishable from atrial septal defect. With this functional murmur, however, the characteristics of the second heart sound are normal; whereas in atrial septal defect the components of the second heart sound show wide, fixed splitting.

3. Venous hum. This murmur might be confused with a patent ductus arteriosus because it is continuous. It is, however, heard best in the right infraclavicular area. Venous hum originates from turbulent flow in the jugular venous system. It has several characteristics distinguishing it from patent ductus arteriosus: it is louder in diastole, is best heard with the patient sitting, diminishes when the patient reclines, and changes in intensity with movements of the head or pressure over the jugular vein.

4. Bruits in the neck. In nearly every child, soft systolic arterial bruits may be heard over the carotid artery and are believed to originate at the bifurcation of the carotid arteries. The bruit should not be confused with the transmission of cardiac murmurs to the neck, as in aortic stenosis. Aortic stenosis is associated with a suprasternal notch thrill.

5. Cardiopulmonary murmur. This sound originates from compression of the lung between the heart and the anterior chest wall. This murmur or sound occurs during systole, is loudest in mid-inspiration, and sounds close to the ear.

In most children with a functional heart murmur, neither an X-ray nor EKG is indicated, as the diagnosis can be made with certainty from the physical examination. In a few patients additional studies may be necessary to distinguish significant murmurs from functional murmurs. Most functional murmurs disappear in adolescence.

ELECTROCARDIOGRAPHY

EKG plays an integral part in evaluation of a child with cardiac disease. It is

most useful in reaching a diagnosis when combined with patient data obtained from the history, physical exam, and X-ray.

QRS complex The QRS complex represents ventricular depolarization. It should be analyzed for its axis and amplitude.

QRS axis. The QRS axis represents the net direction of ventricular depolarization. By three months of age, the QRS axis has a normal range of 0 to +120 degrees.

Right axis deviation is diagnosed when the calculated value for the QRS axis is greater than the upper range of normal, which for older children is more than +120 degrees. Right axis deviation is almost always associated with right ventricular hypertrophy.

Left axis deviation is indicated when the calculated QRS axis is less than the smaller value of the normal range. Left axis deviation is associated with myocardial disease or ventricular conduction abnormalities, such as occur in endocardial cushion defect, but it is rarely associated with left ventricular hypertrophy.

QRS amplitude. QRS amplitude is used to determine ventricular hypertrophy. The term ventricular hypertrophy is partly a misnomer, as this term is applied both to the EKG patterns associated with ventricular chamber enlargement, as well as to an abnormal thickening of the ventricular walls. Hypertrophy is the response to pressure loads upon the ventricle, whereas enlargement reflects augmented ventricular volume.

Interpretation of an EKG for ventricular hypertrophy must be made in relation to the amplitude of the R and S waves in leads V1 and V6.

In right ventricular hypertrophy, the major QRS forces are directed anteriorly and rightward, usually leading to right axis deviation, a taller than normal R wave in lead V1, and a deeper than normal S wave in lead V6. QRS patterns reflecting increase in right ventricular muscle mass ("hypertrophy") usually show an R wave in lead V1, whereas patterns showing right ventricular enlargement usually show an rsR' pattern in lead V1. This distinction is not absolute and variations occur.

In left ventricular hypertrophy, the major QRS forces are directed leftward and sometimes posteriorly. It can be diagnosed when the both the R wave in lead V6 and the S wave in lead V1 are greater than 25 mm. Distinction between left ventricular hypertrophy and left ventricular enlargement is also difficult.

Biventricular hypertrophy is diagnosed by criteria for both right and left ventricular hypertrophy.

PATHOPHYSIOLOGY OF CONGENITAL HEART DISEASES

Two general types of cardiac defects will be covered, left-to-right shunts and admixture lesions.

Left-to-right shunts are defects that cause increased blood flow from the arterial to the pulmonary circulation. The actual defects consist of holes in the atrial or ventricular septum or abnormally connected vessels. Left-to-right shunts lead to an excess blood volume in the right side of the heart and the pulmonary circulation. The specific cardiac chambers affected by the volume overloading differ depending on the location of the defect, but the typical response to volume overloading is enlargement of the affected chamber or vessel.

Left-to-right shunts usually are acyanotic, that is, there is enough oxygenated blood flowing to the systemic circulation so that the coloration of the skin remains normal.

Specific left-to-right shunts and their hemodynamics are discussed in more detail below.

In admixture lesions, usually a single cardiac chamber receives the total systemic and pulmonary venous returns. As with left-to-right shunts, admixture lesions are characterized by increased pulmonary blood flow. Admixture lesions lead to the mixing of oxygenated and deoxygenated blood. Therefore, blood flowing through the systemic circulation contains less oxygen than usual, leading to cyanosis, a bluish or purplish coloration to the skin or fingernails.

Specific admixture lesions and their hemodynamics are discussed in more detail below.

1. Left-to-right shunts. Five defects account for most left-to-right shunts:

Atrial septal defect (ASD)
Endocardial cushion defect (ECD)
Ventricular septal defect (VSD)
Patent ductus arteriosus (PDA)
Partial anomalous pulmonary venous connection (PAPVC)

ASD and ECD (Figs. 2 and 3) consist of holes in the atrial septum. ASD is a hole in the upper portion of the atrial septum, the ostium secundum. ECD is a hole in the lower portion of the atrial septum, the ostium primum.

In PAPVC (Fig. 6), a subset of the pulmonary veins connect abnormally to the right atrium, with the remainder connecting, as they should, to the left atrium. PAPVC is often accompanied by a hole in the atrial septum as well.

VSD (Fig 4) consists of a hole in the ventricular septum, the size of which may vary considerably.

PDA is a communication between the aorta and pulmonary artery. It represents the persistence of fetal communication between the aorta and the pulmonary trunk.

Hemodynamics of left-to-right shunts. While all of the defects mentioned above cause the blood to flow from the left to the right side of the heart, the actual cause of the left-to-right shunting depends on the location of the defect.

Shunts that occur at the atrial level (e.g., ASD) are usually large, so there is no pressure gradient across the shunt. Thus, pressure differences do not determine the direction of blood flow. Instead, the direction of flow is determined by the relative compliances of the atria and ventricles. Since both the left atrium and left ventricle are less compliant than the right atrium and ventricle, the blood flows from the left to the right side of the heart.

The direction and magnitude of blood flow through shunts at the ventricular or great vessel level (e.g., VSD, PDA) are usually determined by the pressure gradient across the shunt during systole. In most cases, the pressures on the right side of the heart and the pulmonary arterial system are less than on the left side of the heart, and a left-to-right shunt also occurs. If the VSD is very large, then the flow is determined by the level of pulmonary and systemic vascular resistances rather than the pressure gradient across the shunt.

2. Admixture lesions. Three admixture lesions will be discussed:
Complete transposition of the great vessels (CTGV)
Persistent truncus arteriosus (PTA)
Total anomalous pulmonary venous connection (TAPVC)

In CTGV (Fig. 7), the aorta arises from the right ventricle and the pulmonary

artery from the left ventricle. Thus, two parallel and separate circulatory systems exist, one pulmonary and one systemic. A communication must exist between the left and right sides of the heart to allow some mixing of the pulmonary and systemic venous returns. The communication may include one of the following: ASD, VSD, or PDA.

In PTA (Fig. 8), a single arterial blood vessel leaves the heart and gives rise to both the pulmonary and systemic circulations. This malformation is always associated with a large VSD, through which both ventricles empty into the truncus arteriosus.

In TAPVC (Fig. 9), all 4 pulmonary veins connect to the right atrium (RA) of the heart rather than to the left atrium (LA), their normal site of connection.

It should be noted that while PAPVC was classified as a left-to-right shunt, in some cases it can also present as an admixture lesion, similar to TAPVC.

Admixture lesions commonly present with a clinical triad of cyanosis, congestive heart failure, and increased pulmonary arterial markings on X-ray.

Hemodynamics of admixture lesions. The hemodynamics of admixture lesions resemble those of the left-to-right shunts occurring at the same level. For example, relative resistances to systemic and pulmonary flow control the distribution of blood in patients with PTA in a way similar to the case in a large VSD. The direction and magnitude of blood flow in TAPVC is governed as an ASD by the relative ventricular compliances.

Cyanosis Cyanosis is an important diagnostic finding for admixture lesions. However, it may be difficult to determine or may be caused by other factors than the admixture lesion itself. Therefore, A brief overview of cyanosis in general and in admixture lesions is presented below.

Cyanosis is a bluish or purplish color of the skin caused by reduced hemoglobin in the capillary beds. The degree of cyanosis reflects the magnitude of unsaturated blood. Mild degrees of arterial desaturation may be present and cyanosis may not be noted clinically. There are two general types of cyanosis, peripheral or central.

Peripheral cyanosis Peripheral cyanosis is associated with normal cardiac and pulmonary function; it is related to sluggish blood flow through the capillaries so that continued oxygen extraction leads to increased amounts of desaturated blood in the capillary beds. Typically the extremities are involved but the trunk and mucus membranes are not. Exposure to cold is the most frequent cause of peripheral cyanosis, leading to blue hands and feet in neonates and a bluish color around the mouth in older children.

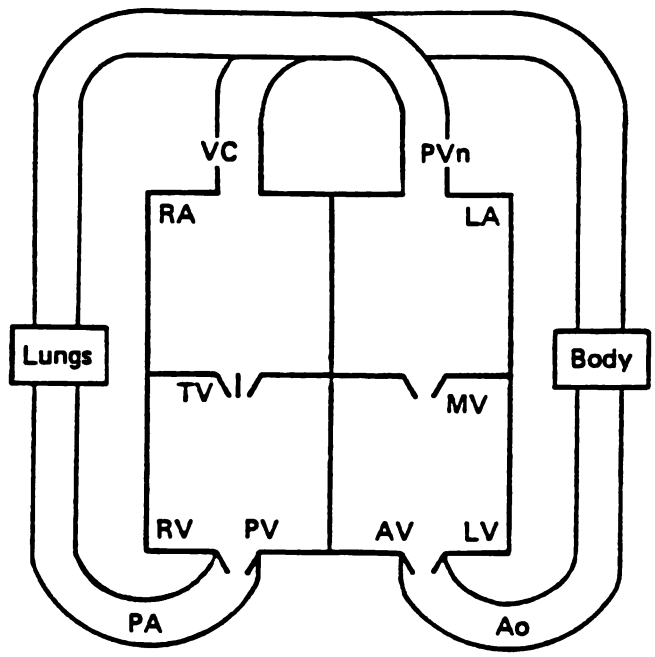
Central cyanosis Central cyanosis is related to an abnormality of the lungs or heart that interferes with oxygen transport from the atmosphere to the pulmonary capillary. It invovles the trunk and mucus membranes as well as the extremities. There are two mechanisms of central cyanosis:

1. Structural abnormalities which cause the mixing of systemic and pulmonary venous returns before being ejected. Admixture lesions fall into this category.

2. Pulmonary edema. With increased pulmonary capillary pressure, fluid crosses the capillary wall into the alveolus. Fluid accumulation interferes with oxygen transport from the alveolus to the capillary so that the hemoglobin leaving the

the capillaries remains desaturated. Cyanosis from pulmonary edema may be strikingly improved by oxygen administration, whereas oxygen will not reduce peripheral cyanosis or cyanosis caused by structural abnormalities.

Cyanosis in admixture lesions In admixture lesions, The degree of cyanosis is inversely related to the volume of pulmonary blood flow. In patients with large pulmonary blood flow the degree of cyanosis is slight, since large amounts of fully saturated blood return from the lungs and mix with a relatively smaller volume of systemic venous return. Should the patient develop pulmonary vascular disease or another factor that limits pulmonary blood flow, the amount of fully oxygenated blood returning from the lungs and mixing with the systemic venous return is reduced, so the patient becomes more cyanotic.



LEGEND

Ao = Aorta	PV = Pulmonary Valve
AV = Aortic Valve	PVn = Pulmonary Veins
LA = Left Atrium	RA = Right Atrium
LV = Left Ventricle	RV = Right Ventricle
MV = Mitral Valve	TV = Tricuspid Valve
PA = Pulmonary Artery	VC = Vena Cavae

Figure 1. The normal heart and cardiovascular system.

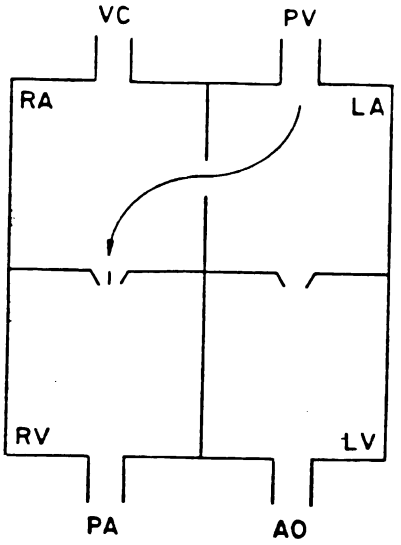


Fig. 2 Central circulation in atrial septal defect.

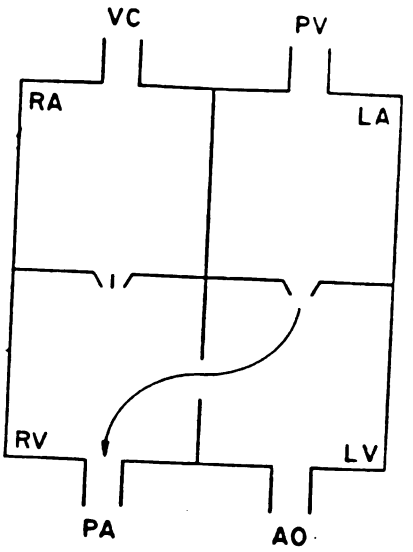


Fig. 4 Central circulation in isolated ventricular septal defect.

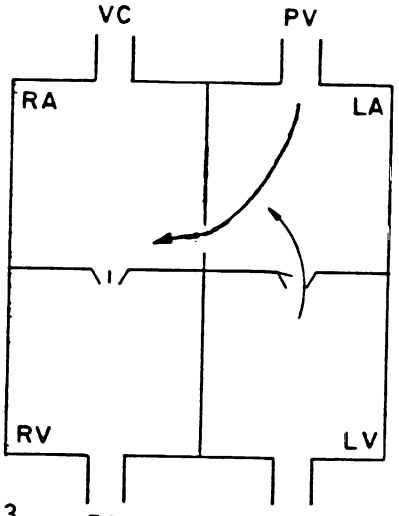


Fig. 3 Central circulation in endocardial cushion defect.

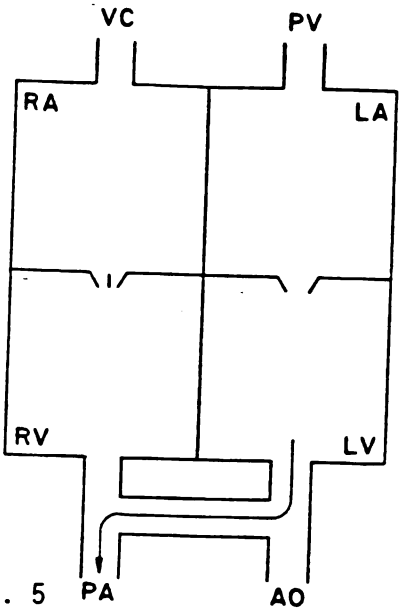
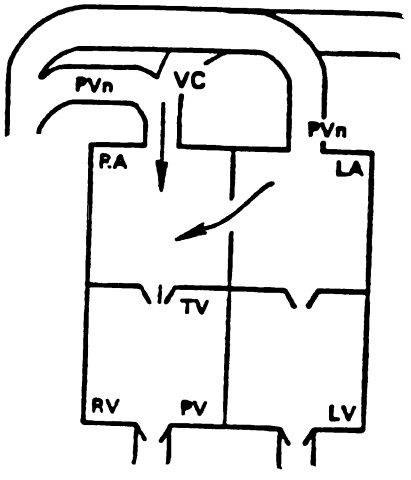


Fig. 5 Central circulation in patent ductus arteriosus.



PAPVC

Fig. 6

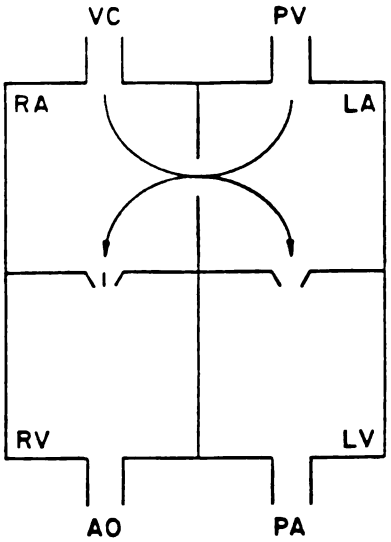


Fig. 7

Central circulation in complete transposition of the great vessels.

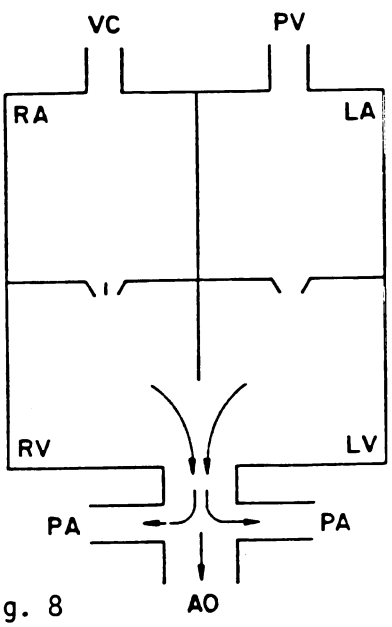


Fig. 8

Central circulation in persistent truncus arteriosus.

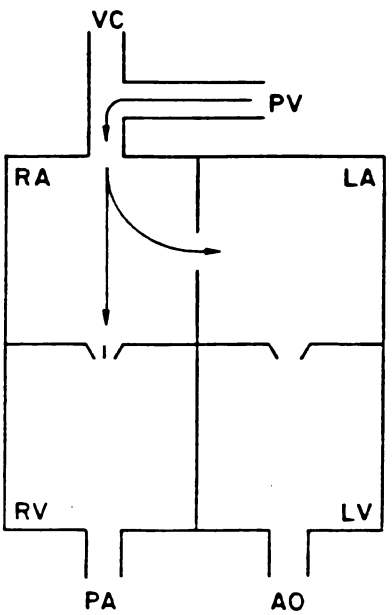


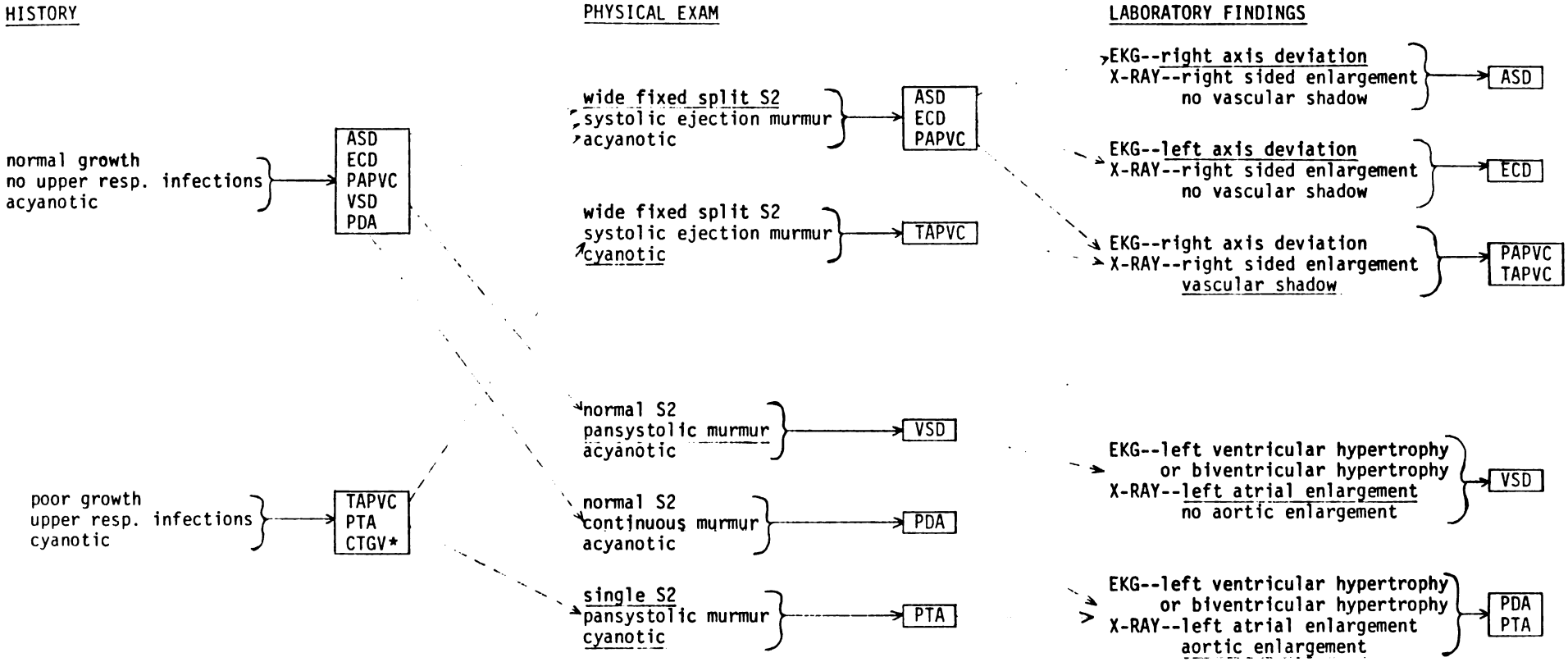
Fig. 9

Central circulation in total anomalous pulmonary venous connection.

APPENDIX D

HISTORY	PHYSICAL EXAM	EKG	X-RAY
ASD incidence: 5-10% more common in girls usually asymptomatic normal growth no congestive heart failure (e.g., shortness of breath, increased respiratory rate) acyanotic	loud S1 wide fixed split S2 systolic ejection murmur (2-3/6), upper left sternal border middiastolic tricuspid flow murmur, lower left sternal border acyanotic	right axis deviation of +90 to +180 degrees right atrial enlargement right ventricular hypertrophy rSR' pattern in lead V1	right sided enlargement (e.g., right atrium, right ventricle) increased pulmonary vascular markings
ECD incidence: 1-2% equally common in boys and girls often occurs with Down's syndrome otherwise, similar to ASD	same as ASD	left axis deviation of 0 to -150 degrees otherwise, same as ASD	same as ASD
PAPVC incidence: less than 1% similar to ASD	same as ASD	same as ASD	crescent shaped vascular shadow ("scimitar syndrome") otherwise, same as ASD
TAPVC incidence: 1% history of upper respiratory infections poor growth congestive heart failure cyanosis may or may not be present	mild to moderate cyanosis (may be difficult to detect clinically) otherwise, same as ASD	same as ASD	"snowman" shaped heart silhouette otherwise, same as ASD
VSD incidence: 25% usually asymptomatic normal growth no congestive heart failure (with larger defect, poor growth, congestive heart failure) acyanotic	pulmonary component of S2 accentuated loud, harsh pansystolic murmur (3-4/6) lower left sternal border soft middiastolic mitral flow murmur at apex acyanotic	left ventricular or biventricular hypertrophy normal QRS axis	left atrial enlargement aorta is normal size increased pulmonary vascular markings
PDA incidence: 10% more common in girls, premature infants, and children whose mothers had rubella during 1st trimester otherwise, similar to VSD	pulmonary component of S2 may be accentuated continuous murmur (1-4/6), left infraclavicular area (continuous murmur not always heard in 1st 6 months of life) soft middiastolic mitral flow murmur wide pulse pressure systolic ejection click from dilated aorta	same as VSD	left atrial enlargement aortic enlargement
PTA incidence: less than 1% history of upper respiratory infections poor growth congestive heart failure cyanosis in neonatal period	Single S2 (representing truncus) loud systolic murmur, left sternal border systolic ejection click wide pulse pressure mild cyanosis	biventricular hypertrophy similar to VSD	left atrial enlargement prominent ascending aorta right aortic arch seen in 1/4 of cases
CTGV incidence: 5% more common in boys intense cyanosis neonatally congestive heart failure (death usually occurs by 6 months)	often no murmur or VSD type murmur single S2 intense cyanosis	right ventricular or biventricular hypertrophy right axis deviation, +90 to +180	left atrial enlargement egg shaped heart

DIAGRAM 1



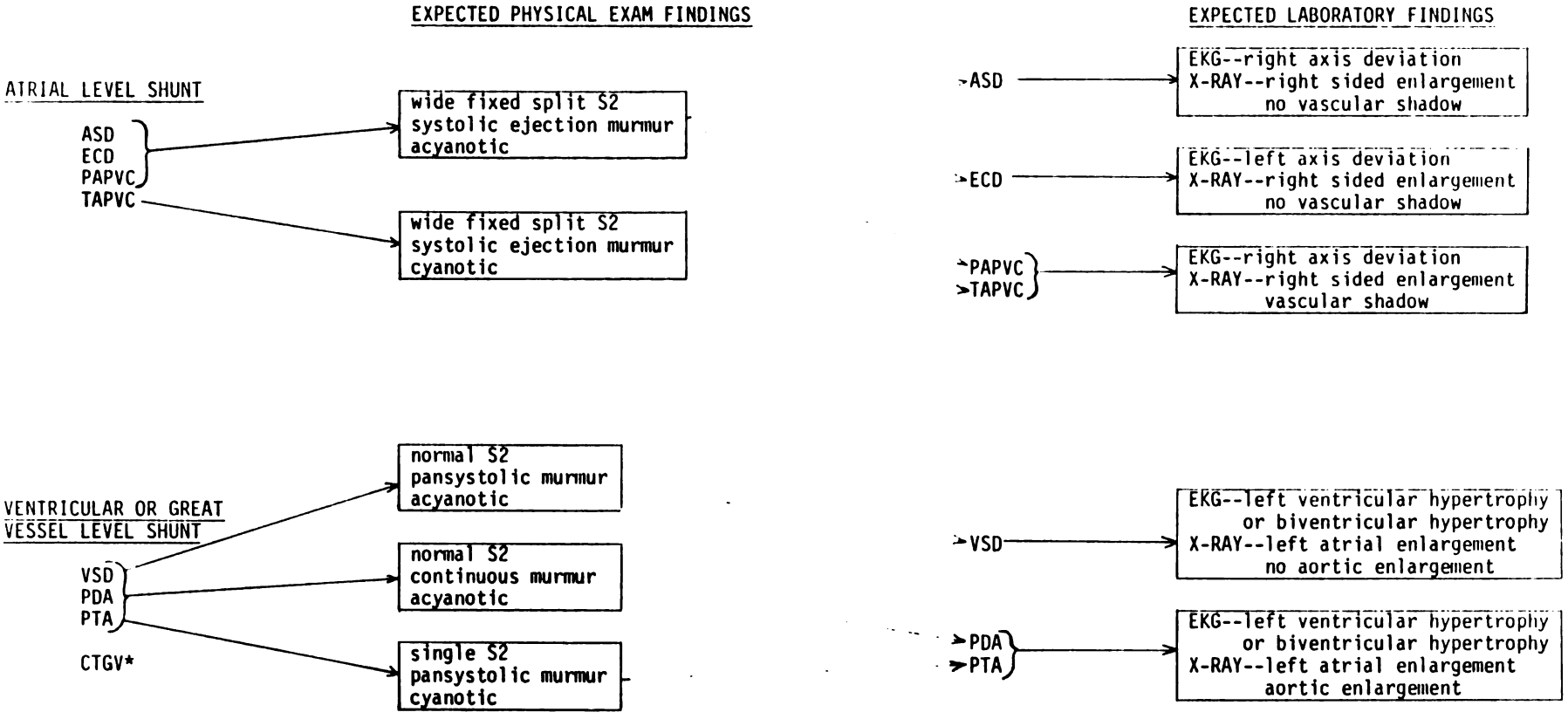
Note: CTGV can usually be diagnosed by findings of intense cyanosis in the newborn period. Most of these patients die by 6 months of age.

Exceptions in History:
For VSD and PDA, the severity of the symptoms may vary depending on the size of the shunt.

Exceptions in Physical:
1. The continuous murmur of PDA is not always present in the 1st 6 months of life; it may sound like a VSD murmur then.
2. Cyanosis may not always be clinically evident.

Note for Laboratory Findings:
Increased pulmonary vascular markings are often found on X-RAY with most of these defects.

DIAGRAM 2



* Note: CTGV can usually be diagnosed by findings of intense cyanosis in the newborn period. Most of these patients die by 6 months of age.

Exceptions in Physical:

1. The continuous murmur of PDA is not always present in the 1st 6 months of life; it may sound like a VSD murmur then.
2. Cyanosis may not always be clinically evident.

Note for Laboratory Findings:
Increased pulmonary vascular markings are often found on X-RAY with most of these defects.

APPENDIX E

CONGENITAL CARDIAC DEFECTS

ATRIAL SEPTAL DEFECT

History ASD is one of the more common congenital cardiac defects, accounting for 5% to 10 % of all congenital heart disease. ASD occurs 2-3 times more frequently in girls. Children with ASD rarely develop congestive heart failure because the volume overload of the right ventricle is well tolerated. Thus, these children are usually asymptomatic. Occasionally there is a history of frequent respiratory infections. Typically, ASD is first recognized as late as the preschool physical exam because the murmur is soft and may be mistaken for a functional murmur or obscured during examination of an active or fearful infant.

Physical Exam Enlargement of the right ventricle may cause a precordial bulge. There is an accentuated first heart sound in the tricuspid area, and a pulmonary systolic ejection murmur (of increased pulmonary valve flow), varying from grade I-III/VI ("scratchy" in sound), and rarely associated with a thrill. A mid-diastolic murmur is present along the lower left sternal border (i.e., a diastolic tricuspid flow murmur). There is a wide, fixed splitting of the second heart sound.

EKG The right atrium and right ventricle are enlarged and the EKG reflects this. Right axis deviation of +90 to +180 degrees is usually present, and right atrial enlargement is found in many patients. right ventricular hypertrophy is also found and is reflected by an rSR' pattern in lead V1 of the EKG. It is difficult to diagnose ASD in the absence of this EKG finding.

X-ray The X-ray shows increased pulmonary vasculature and enlargement of the right side of the heart. The left atrium is not enlarged since it is decompressed by the atrial communication. A large pulmonary artery and normal aorta are typically seen.

Summary In ASD, fixed splitting of the second heart sound indicates the presence of an atrial communication. Findings of a pulmonary systolic ejection murmur, mid-diastolic murmur, rSR' on EKG, cardiomegaly and increased pulmonary blood flow all reflect the increased volume of flow through the right side of the heart. However, pressure in the pulmonary artery is normal. ASD rarely results in congestive heart failure in the pediatric age range.

ENDOCARDIAL CUSHION DEFECT

History ECD accounts for 2% of all congenital heart disease. The histories of patients with ECD vary considerably. Those with an ostium primum defect are typically asymptomatic. When present, symptoms are usually related to congestive heart failure, poor growth and frequent respiratory infections. Frequently the murmur is heard early in life, even if the patient is asymptomatic. Down's syndrome is frequently found in association with ECD (1/3 of cases occur in Down's syndrome).

Physical Exam The general appearance of the child is usually normal, but infants with congestive heart failure may be scrawny and show labored breathing on exertion. In patients with cardiac enlargement, the precordial bulge and cardiac apex are displaced toward the left and inferiorily. There is a left-to-right shunt at atrial level. In patients with an ostium primum defect, three findings are present:

1. Pulmonary systolic ejection murmur.
2. Tricuspid, diastolic murmur.
3. Wide, fixed splitting of the second heart sound.

EKG There are five commonly observed features:

1. Left axis deviation ranging from 0 to -150 degrees (sometimes called a "Northeast" axis); greater degrees of left axis deviation occur in patients with increasing degrees of right ventricular hypertrophy, secondary to elevated pulmonary arterial pressure.
2. Right atrial enlargement.
3. Right ventricular hypertrophy.
4. Cleft mitral valve which leads to the regurgitation of blood across the mitral valve. Mitral regurgitation is not significant from a hemodynamic point of view because blood regurgitated into the left atrium is immediately shunted to the right atrium; therefore, the left atrium remains normal sized.
5. rSR' pattern in lead V1.

X-ray X-ray findings are similar to those of ASD; there is increased pulmonary vasculature, and enlargement of the right side of the heart (right atrium and ventricle).

Summary The clinical and lab findings reflect a left-to-right shunt at the atrial level. EKG features are most diagnostic for ECD, showing left axis deviation, and right atrial and ventricular hypertrophy. The X-ray reveals enlargement of the right side of the heart.

VENTRICULAR SEPTAL DEFECT

History VSD is the most common defect, accounting for nearly 25% of all congenital heart disease. There is marked variation in the size of the defect and and the associated symptoms. Most patients are asymptomatic. The defect is usually detected by the discovery of a murmur either prior to discharge from the newborn nursery or, more commonly, at the first postnatal visit to the MD. A few patients develop congestive heart failure. Most patients show normal growth and development.

Physical Exam Cardiomegaly may or may not be present. A loud, harsh (grade III-IV/VI) systolic murmur is present along the lower left sternal border and may be associated with a thrill. In most patients, the murmur is pansystolic. The pulmonary component of the second heart sound is accentuated and there is a soft apical diastolic murmur.

EKG There is a pattern of left atrial enlargement with left ventricular or biventricular hypertrophy, indicating increased volume of blood flow to the left ventricle and elevation in pressure of the right ventricle.

X-ray The appearance of the heart varies according to the magnitude of the shunt and the level of pulmonary arterial pressure. The X-ray may be normal or may show left ventricular or biventricular hypertrophy, left atrial enlargement, a large pulmonary artery, and increased pulmonary vasculature. There is no characteristic contour of the heart in VSD.

Summary In VSD, the magnitude of the shunt depends on the size of the defect and the

relative levels of pulmonary and systemic vascular resistances. The primary finding of VSD is a pansystolic murmur along the left sternal border. Secondary features include left atrial enlargement and left ventricular or biventricular hypertrophy on EKG from excess blood flow to left ventricle and elevated right ventricular pressure. The X-ray reveals left or biventricular hypertrophy and left atrial enlargement.

PATENT DUCTUS ARTERIOSUS

History PDA accounts for 10% of all congenital heart disease, excluding premature infants. It occurs more frequently in females and prematurely born infants. It is the most commonly observed defect in children whose mothers had rubella during the first trimester of pregnancy. Many patients are asymptomatic and the ductus may be identified only by the presence of a murmur. On the other hand, congestive heart failure can develop early in infancy because of volume overload of the left ventricle, although this typically does not occur for at least 3 months. symptomatic children may also present a history of frequent respiratory infections and easy fatigability.

Physical Exam The classical finding for PDA is a continuous or machinery type systolic murmur best heard over the upper left chest under the clavicle. It may be associated with a thrill. However, a continuous murmur may not always be present in the first 6 months of life. An aortic systolic ejection click is frequently heard because the aorta is dilated. The pulmonary component of the second heart sound is accentuated and sometimes a soft apical diastolic murmur is heard. A wide pulse pressure is often present.

EKG EKG patterns are similar to VSD since the potential hemodynamic burdens are volume overload of the left ventricle and pressure overload of the right ventricle. In many patients with PDA, the major hemodynamic burden is volume overload of the left atrium and left ventricle, the EKG revealing left ventricular hypertrophy and perhaps left atrial enlargement. In infants and children with increased pulmonary pressure, right ventricular hypertrophy coexists with a pattern of left ventricular hypertrophy (i.e., biventricular hypertrophy).

X-ray X-ray findings of PDA typically exhibit increased pulmonary vasculature, and left atrial and left ventricular enlargement. Usually both the aorta and the pulmonary trunk are enlarged, although in infants the aortic knob may be obscured by the thymus. PDA is the only major cardiac defect with a left-to-right shunt with aortic enlargement. In the other left-to-right shunts, the aorta is normal or appears small. Therefore, if a distinctly enlarged aorta is present and a left-to-right shunt is suspected, PDA must be seriously considered.

Summary The primary features of PDA include a continuous murmur (thought not always present in the first 6 months), and findings associated with a wide pulse pressure. In general, PDA results in an excessive volume of blood flow to the left ventricle combined with a pressure overload on the right ventricle. However, the direction and magnitude of flow through the ductus depend on the size of the ductus and the relative systemic and pulmonary vascular resistances.

PARTIAL ANOMALOUS PULMONARY VENOUS CONNECTION

History PAPVC accounts for less than 1% of all congenital heart disease. symptoms are uncommon during childhood but there may be some dyspnea on exertion. Cyanosis is unusual during childhood even though a small right-to-left shunt may exist. Upper respiratory infections are often seen.

Physical Exam In the presence of an associated ASD, the physical findings are similar to those noted in uncomplicated ASD. A precordial bulge is common from right ventricle enlargement, and the second heart sound shows wide, fixed splitting. There is an accentuated first heart sound and a pulmonary systolic ejection murmur (grade I-III/VI) is usually present. A diastolic tricuspid flow murmur may also be present.

EKG The EKG findings are comparable to those seen in uncomplicated ASD. Right axis deviation is often seen along with right atrial enlargement and right ventricular hypertrophy. An rSR' pattern is most commonly seen in lead V1, although the EKG is occasionally normal.

X-ray There is increased pulmonary vasculature and enlargement of the right side of the heart. The left atrium and aorta are normal size. There may be distinctive x-ray features depending on the site of the anomalous connection of the pulmonary veins. Patients with anomalous connection of the right pulmonary veins to the inferior vena cava have a crescent-like shadow in the right lower lung field, called scimitar syndrome. when the left innominate vein is the site of the connection of the left pulmonary vein, the X-rays reveal a prominent supracardiac shadow composed of the vertical vein on the left, the innominate vein above, and the superior vena cava at the right. These structures are the same ones that account for the characteristic "snowman" appearance of TAPVC, but the enlargement is not as prominent.

Summary PAPVC presents a clinical picture of enlargement of the right-sided cardiac chambers due to increased pulmonary blood flow. In the presence of an associated ASD, many of the findings of PAPVC are similar to those of uncomplicated ASD. PAPVC often has the unique feature of the presence of a crescent-like shadow in the right lower lung field on X-ray (scimitar syndrome).

COMPLETE TRANSPOSITION OF THE GREAT VESSELS

History CTGV accounts for 5% of all congenital heart disease and occurs more frequently in male infants. Cyanosis is evident shortly after birth, and dyspnea and other signs of heart failure are uniformly seen in the first month of life. in the absence of operative relief, death occurs in almost every patient by 6 months of age.

Physical Exam cyanosis is usually intense and heart failure is typically seen. Physical findings vary, depending upon the defect associated with complete transposition of the great vessels. With an intact ventricular septum, no murmur or a soft murmur is heard. With an associated VSD, a louder murmur is present. The type or presence of a murmur is not helpful in diagnosing transposition of great vessels, although it may indicate the type of associated defect (ASD, VSD, or PDA).

EKG Since the aorta arises from the right ventricle, pressure in the right ventricle

is elevated to systemic levels and is associated with a thick-walled right ventricle. The EKG reflects this by a pattern of right axis deviation and right ventricular hypertrophy. Right atrial enlargement may also be seen. Patients with a large volume of pulmonary blood flow may also exhibit left ventricular hypertrophy because of the volume load on the left ventricle.

X-ray cardiomegaly is almost always present. The cardiac silhouette has a characteristic egg-shaped appearance. Left atrial enlargement is present in the unoperated patient.

Summary complete transposition of the great arteries is a common cardiac anomaly and results in neonatal cyanosis and heart failure. Diagnosis is usually indicated by a combination of rather intense cyanosis in the neonatal period and X-ray findings of increased pulmonary vasculature, cardiomegaly, and a characteristic cardiac contour.

PERSISTENT TRUNCUS ARTERIOSUS

History PTA accounts for less than 1% of all congenital heart disease. The hemodynamics of PTA are similar to those of VSD and PDA because the respective volumes of systemic and pulmonary blood flow depend upon the relative resistance to flow into the systemic circulation and into the pulmonary circulation. Increased pulmonary blood flow leads to three effects:

1. The degree of cyanosis lessens as pulmonary blood flow increases.
2. Congestive heart failure usually develops after several weeks of age because of left ventricular volume overload.
3. The pulse pressure is widened because during diastole the blood leaves the truncus arteriosus to enter the pulmonary arteries.

Symptoms vary with the volume of pulmonary blood flow. Neonatally, cyanosis is a major symptom, but lessens as pulmonary blood flow increases. In the absence of cyanosis, patients with PTA and congestive heart failure are clinically similar to those with VSD and PDA. Dyspnea on exertion, easy fatigability, and frequent respiratory infections may be common symptoms.

Physical Exam cyanosis may or may not be clinically evident. A wide pulse pressure may be present if there is increased pulmonary blood flow. Cardiomegaly and a precordial bulge are commonly seen. The major auscultatory finding is a loud systolic murmur along the left sternal border. An apical diastolic rumble is present in most patients. There are 3 distinct auscultatory findings:

1. The second heart sound is single, since there is only a single semilunar valve.
2. A high pitched, early decrescendo diastolic murmur may be present if truncal valve insufficiency coexists.
3. An apical systolic ejection click is usually heard and indicates the presence of a dilated great vessel, which in this case is truncus arteriosus.

EKG There is a normal QRS axis and biventricular hypertrophy.

X-ray Increased pulmonary vasculature is seen. Usually a prominent "ascending aorta" is found, representing the truncus arteriosus. A pulmonary artery segment

may also be present. Most patients show cardiomegaly proportionate to the volume of pulmonary blood flow and amount of truncal insufficiency. Left atrial enlargement is found with increased pulmonary blood flow. A right aortic arch is found in 25% of the patients. This finding combined with increased pulmonary vascular markings and the presence of cyanosis is virtually diagnostic of truncus arteriosus.

Summary PTA is an infrequently occurring cardiac anomaly resulting in excess flow to the left ventricle and excess pressure on the right ventricle. PTA can be suspected in a cyanotic patient who has a loud systolic murmur along the left sternal border and 2 characteristic features: a single second heart sound and an early systolic ejection click.

TOTAL ANOMALOUS PULMONARY VENOUS CONNECTION

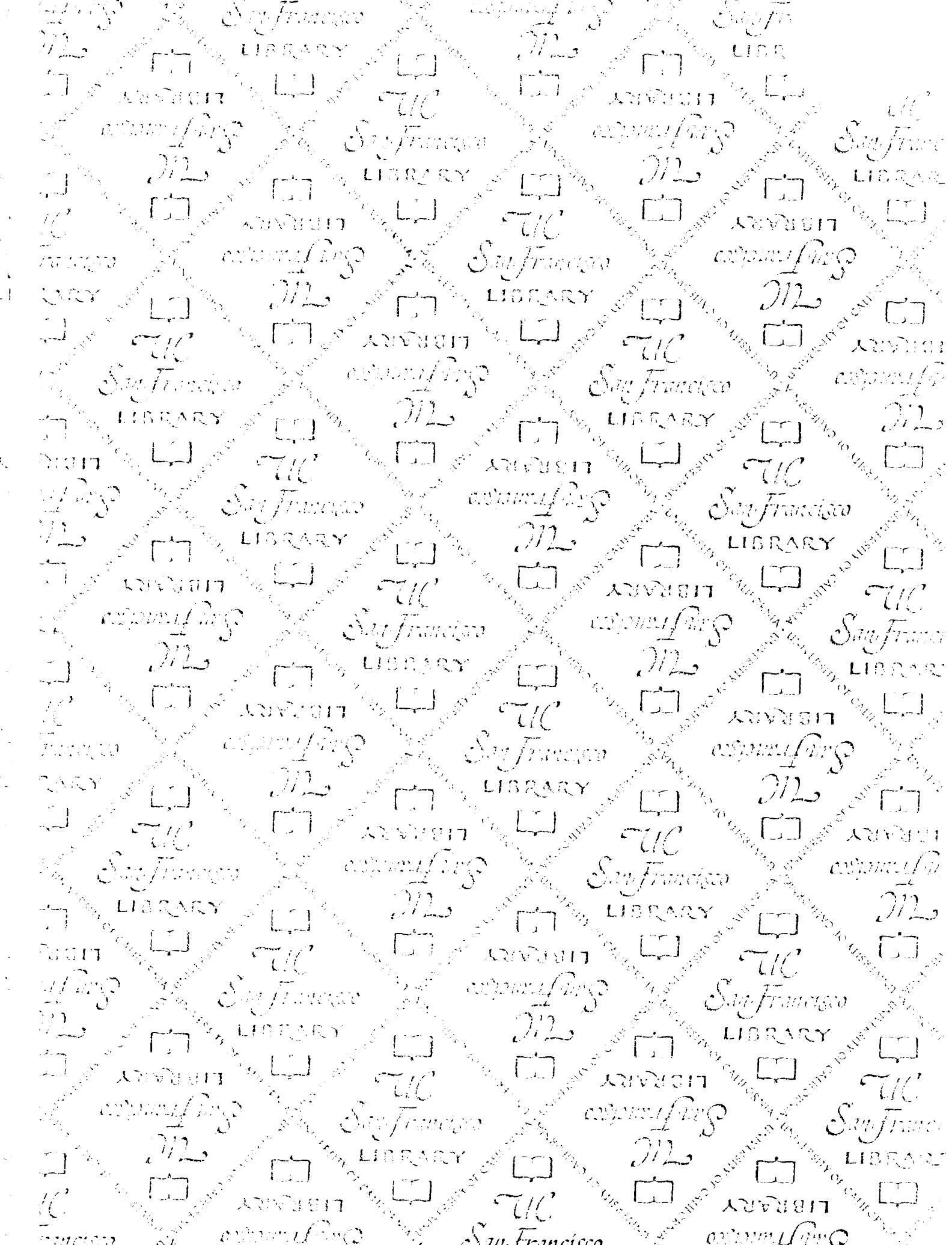
History TAPVC accounts for 1% of all congenital heart disease. The age of onset and clinical manifestations vary considerably. Most patients develop congestive heart failure in infancy, grow slowly, and have frequent respiratory infections, but a few may be asymptomatic into later childhood.

Physical Exam The degree of cyanosis varies because of differences in the amount of pulmonary blood flow. Most children appear acyanotic or show only slight cyanosis, although systemic arterial desaturation is always present. The physical findings for TAPVC are similar to those of isolated ASD. Cardiomegaly and a precordial bulge are commonly seen. A grade II-III/VI pulmonary systolic ejection murmur is present along the upper left sternal border. Wide, fixed splitting of the second heart sound is present, and the pulmonary component may be accentuated, reflecting pulmonary hypertension. A diastolic murmur is present along the lower left sternal border, and is associated with greatly increased pulmonary blood flow.

EKG The EKG reveals enlargement of the right-sided cardiac chambers by a pattern of right axis deviation, right atrial enlargement, and right ventricular hypertrophy. An rSR' pattern in lead V1 reflects right ventricular hypertrophy.

X-ray X-ray findings resemble those of isolated ASD, showing increased pulmonary vasculature and enlargement of right side of the heart. The left atrium is not enlarged, in contrast to most admixture lesions. There is a large pulmonary artery and a normal aorta. Except for TAPVC of the left superior vena cava, the X-ray contour is not characteristic. In this form, the cardiac silhouette has been described as a figure eight or "snowman heart".

Summary The clinical, EKG, and X-ray findings of TAPVC without obstruction to pulmonary blood flow, resemble those of ASD because the effects upon the heart are similar. Cyanosis is a distinguishing feature of TAPVC, although it may be minimal or not clinically evident. Unlike the case in uncomplicated ASD, congestive heart failure and elevated pulmonary aterial pressure may be found with TAPVC.


UC
San Francisco
LIBRARY

FOR REFERENCE

NOT TO BE TAKEN FROM THE ROOM