COGNITIVE CORRELATES OF CLINICAL PERFORMANCE

BY

JERRY WEISS

A THESIS PRESENTED TO THE FACULTY
OF GRADUATE STUDIES AND RESEARCH,
McGILL UNIVERSITY, MONTREAL, IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF ARTS.

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY
AND COUNSELLING
JULY, 1982

ABSTRACT

This investigation explored the nature of the relationship between successful performance on computerized multiple choice exams and clinical performance in third year medical students. Included was an examination of the process of medical problem solving and its characteristic properties. Specifically, scores obtained on 18 computer assisted multiple choice questions were correlated with clinical evaluations that were secured through a 10 category checklist.

The major finding of the study was that there were some significant correlations between certain categories of the clinical checklist and various measures obtained via the computer. There was some evidence of differing strategies among subjects. In the light of other research it is argued that these differences may reflect a varying knowledge base.

The display of similar behaviors on successful performance in both settings led to the conclusion that computerized multiple choice exams do have a predictive capability with reference to clinical performance. The lack of significant correlations among all variables led to the conclusion that a mixed array of criteria should be used in order to evaluate clinical performance.

It was also felt that as a result of the existing evidence the process of medical education would possibly be enhanced by the use of a problem solving mode of instruction.

RESUME

Cette étude porte sur le rapport éventuel entre de bons résultats aux examens informatisés à réponses multiples et de bon résultats cliniques obtenus par des étudiants en troisième année de médicine. On a également étudié la résolution d'un problème de médicine et ses propriétés. Plus précisément, on a comparé les résultats obtenus à 18 questions à réponses multiples aux évaluations cliniques réalisées au moyen d'une liste de contrôle à dix catégories.

La principale conclusion de cette étude est qu'il existe des rapports étroits entre plusieurs des catégories de la liste de contrôle clinique et les diverses mesures obtenues par ordinateur. On a également constaté l'existence de stratégies variables entre les sujets. Si l'on en croit d'autres études menées à ce sujet, on peut affirmer que ces différences témoignent de divergences au niveau de la base des connaissances.

La constatation de comportements analogues relativement à de bons résultats dans les deux milieux nous a amené à conclure que les examens informatisés à réponses multiples permettent de prévoir les résultats cliniques des étudiants. L'absence de corrélation notable entre les variables donne à penser que pour évaluer les résultats cliniques des étudiants, il faut se baser sur un éventail de critères très diversifié.

On a également conclu que devant les preuves existantes, l'enseignement de la médecine pourrait bénéficier de méthodes basées sur la résolution de problèmes.

ACKNOWLEDGEMENTS

The author wishes to express his thanks to the medical students who so kindly and generously volunteered their time for this study. Their unselfish cooperation was the most important factor in the project reaching fruition.

I would like to acknowledge the important and invaluable contribution of Dr. W. Dale Dauphinee, Associate Dean for Medicine Education, who greatly aided in the procurement of participants. In addition, his ideas and suggestions were extremely useful particularly when the project was in its embryonic stages.

I am greatly appreciative of the role played by my supervisor, Dr. Guy Groen whose steady guidance and patience aided in the elimination of many of the obstacles present when the project was exhibiting growing pains. He was also responsible for my acquisition of the analytical and organizational skills necessary for the completion of this study. Lastly, his optimistic attitude served to supersede several crisis laden periods.

I would like to thank Dr. Patricia Cranton of the Centre for Teaching and Learning Services, McGill University, who acted as a consultant on the study. Her critical analyses of the technical aspects involved in the construction of the manuscript as well as her moral support during all phases of the study greatly contributed to its successful completion.

I would like to acknowledge the input of Dr. Vimla Lodhia-Patel, Research Associate, Centre for Medical Education, McGill University. Her expertise in the field of medical education itself, most important of which were her guidelines concerning clinical evaluation which facilitated the study's attainment of a definitive structure. Furthermore, her advice on problems of both a minor and major nature boded well at fostering the continuity of the study.

I am especially grateful for the help accorded me by Mary McQueen whose role as a computer consultant proved to be invaluable. In addition, she gave freely of her time, lending support throughout the entire process. Without her help the study would not have been completed.

Finally, special thanks to Avril Bray and Kathryn Boyd, administrative assistants at the Centre for Medical Education, McGill University. Ms. Bray was especially helpful in dealing with many of the problems encountered during the study's infancy. The work of Ms. Boyd ensured that the scheduling of participants occurred in a smooth and efficient manner.

TABLE OF CONTENTS

Page

ABSTRACT i

RESUME ii

ACKNOWLEDGEMENTS iv

LIST OF TABLES vii

CHAPTER

I. INTRODUCTION 1

II. REVIEW OF THE LITERATURE 4
- Problems Involving the Use of Rating Scales in Medical Education
- Measurement of Interviewing Skills
- The Assessment of Clinical Skills
- A Final Note on Interviewing Skills
- Medical Problem Solving
  - Protocol Analysis
  - Possible Limitations of Information Processing Theory
  - Specific Applications of Medical Problem Solving
- Present Research
- Conclusion

III. METHODOLOGY UTILIZED 43
- Sample
- Materials
- Equipment
- Procedure
  - Clinical Component
  - Computer Based Multiple Choice Question Experiment

IV. RESULTS OF THE STUDY 49
- Descriptive Statistics for the Two Paradigms
- Relationships Between Measures on Computer Exam and Clinical Checklist
- Analysis of Data Obtained from Computerized Multiple Choice Exam Paradigm

V. DISCUSSION 82

REFERENCES 86

APPENDIX 89

LIST OF TABLES

1. Means and Standard Deviations of Measures Obtained on Computerized Multiple Choice Questions 50

2. Analysis of Reversals Obtained on Computer Measures 51

3. Summary of Data Across Problems on Computerized Paradigm 52

4. Intercorrelation Matrix of Problem Data on Computerized Multiple Choice Question Paradigm ... 53

5. Item Analysis of Group's Responses on Computerized Multiple Choice Question Model 54

6. Means and Standard Deviations of Measures on Clinical Evaluation Checklist 55

7. Intercorrelations of Measures Obtained on Clinical Evaluation Checklist 56

8. Intercorrelations of Measures on Computerized Multiple Choice Paradigm and Clinical Evaluation Checklist 58

9. Relationship Between Subjects' Performance on Computerized Multiple Choice Exam and Clinical Competence 60

10. Relationship Between Subjects' Performance on Computerized Multiple Choice Exam and Clinical Competence 61

11. Relationship Between Subjects' Performance on Computerized Multiple Choice Exam and Clinical Competence 62

12. Relationship Between Subjects' Performance on Computerized Multiple Choice Exam and Clinical Competence 63

13. Relationship Between Subjects' Performance on Computerized Multiple Choice Exam and Clinical Competence 64

14. Relationship Between Subjects' Performance on Computerized Multiple Choice Exam and Clinical Competence 65

LIST OF TABLES
(Continued)

15. Relationship Between Subjects' Performance on Computerized Multiple Choice Exam and Clinical Competence 66

16. Relationship Between Subjects' Performance on Computerized Multiple Choice Exam and Clinical Competence 67

17. Intercorrelation of Computer Measures for Third Year Medical Students 69

18. Classification of Students According to Speed and Degree of Backtracking 71

19. T-Tests Comparing Mean Times for Problems Answered Correctly and Incorrectly on Computerized Multiple Choice Question Model 72

20. The Relationship Between Time Spent on All Problems and Time Spent on Problems Answered Correctly on Computerized Multiple Choice Exam ... 73

21. The Relationship Between Time Spent on Problems Answered Correctly and Incorrectly on Computerized Multiple Choice Exam 74

22. The Relationship Between Time Spent on all Problems and Time Spent on Problems Answered Incorrectly on Computerized Multiple Choice Exam . 76

23. The Relationship Between Backtracking and Overall Reversals on Computerized Multiple Choice Exam ... 77

24. The Relationship Between Backtracking and Reversals on Questions Answered Correctly on Computerized Multiple Choice Exam 77

25. The Relationship Between Backtracking and Reversals on Questions Answered Incorrectly on the Computerized Multiple Choice Exam 78

26. Evaluation of Group's Responses on Computerized Multiple Choice Question Model 80

CHAPTER I

INTRODUCTION

Researchers in medical education have encountered difficulties in the evaluation of medical students. There have been problems in constructing rating instruments that provide formal, discrete and unambiguous data. Medicine as a discipline has a structure that is dichotomous; there is a potential for a significant variance between success in artificial environments such as exams, and expertise exhibited on the job. It will be shown that attainment of superior marks is not necessarily a good predictor of consistently good performance in the actual doctor-patient encounter.

The primary goal in medical education is to delineate whether or not the student is or will be clinically competent. Historically, one contentious issue has been the locus of debate: the view of some has been that the approaches utilized to rank medical students, interns and residents have been both haphazard and subjective. This literature review centers on some of the problems faced by researchers who attempted to formulate rating procedures that would serve to circumscribe all the characteristics pertaining to total physician performance in both hypothetical and real-life settings. Accordingly, the literature review in this thesis will consist of two parts. The first section will explain and expand upon the use of performance checklists for rating purposes. Evidence will

also be presented that will reveal the serious problem of the inability to obtain interrater reliability as well as the issue of questionable item validity.

As a function of the above mentioned problems, the second section of the review will serve to explain why a different methodological approach vis-à-vis physician evaluation has emerged. The new form of analysis is cognitive in nature and the conclusion is that heuristically, the physician is a problem solver. There will be an extensive explanation of medical problem solving, what the latter entails, and how it measures the skills and abilities inherent in the medical context. The review will conclude with an explanation of the goals of the present inquiry, one of which is to discover whether or not a correlation exists between the two different analytical frameworks, i.e. checklisting and medical problem solving for the purposes of identifying and rating clinical competence.

Chapter III discusses the methodology utilized with descriptions of the sample, design and specific procedures necessary to administer the project.

The results of the study are presented in Chapter IV; included is a descriptive text accompanied by a number of tables tables exhibiting all the behavioral characteristics that were measured.

Chapter V consists of a discussion of the presented results examining and analyzing the data in a critical mode for the purpose of evaluating its alleged significance. In

addition, problems normally encountered in projects such as the present one that pertain to the framework of medical education and its limitations are examined. Final comments concern possible implications of the present study as well as suggestions of potential investigatory loci of future research.

CHAPTER II

REVIEW OF THE LITERATURE

The purpose of this study is to investigate whether or not a relationship exists between success on computerized multiple choice exams and performance in a clinical milieu by medical students. It will be pointed out that historically, the construction of distinct quantitative assessment criteria for the purpose of clinical evaluation has proven to be a difficult process. Researchers have also found it difficult to delineate a consistent confluence in performance among differing medical contexts. The literature review will conclude with a description of medical problem solving with the viewpoint being posited that some relatively new aspects of research in medical problem solving can better address current problematical issues in medical education.

Problems Involving the Use of Rating Scales in Medical Education

Studies rating clinical competence have mostly involved the use of checklists. At times reliability and validity have been difficult to obtain. It has been discovered that technical skills are easier to evaluate then interpersonal interactions which subsume subjective criteria. Harden (1979) comments on the need for improvement in both the reliability and the validity of the clinical examination. The author cites studies in which marks awarded by one

examiner differed by as much as 25% from those awarded by another examiner for the same performance and it was also discovered that inconsistency was evident in the same examiner. A study conducted by Fleming, Manderson, Matthews, Sanderson, and Stokes (1974) was mentioned for the purpose of illustrating the discovery of the existence of examiners who tended to favor high or low marks. It was revealed that one examiner's influence on his colleagues resulted in the lowering substantially below the expected level the pass rate for the candidates that he examined. With reference to the issue of validity the comment was made that frequently during clinical examinations what was measured was not what should be measured. The operational definition of clinical competence utilized by the author was the one formulated by Hubbard, Levit, Schumacher, and Schnabel (1965) who defined it as the skill in obtaining pertinent information from a patient, the ability to detect and interpret symptoms and abnormal signs, acumen in arriving at a reasonable diagnosis, and judgment in the management of patients. Harden (1979) also notes that the doctor's ability to frame a diagnosis depends on his skill in collecting the appropriate information and recognizing patterns such as abnormal gait in certain patients due to underlying pathology. It is felt by the author that a clinical examination should entail a student's application of knowledge in relation to a patient and with his clinical skills and attitudes as opposed to solely focusing on the extent of his factual knowledge. A cogent and important

finding was noted: students found to have the best record in undergraduate examinations frequently did not make the best house physicians or surgeons while students exhibiting an average performance on examinations were found to be excellent house staff as well as enjoying a very successful career in medicine. However, the experimenters did not reveal upon what standards the latter formulated conclusions were based. Other comments made in the critique include the belief that several observers should be used to assess clinical competence in order to cancel possible individual observer variation. Board examinations (e.g. National Board of Medical Examiners) are criticized by the author citing Newble and Elmslie (1978) who point out that the boards, in the interest of reliability, exclude the clinical component. Board examinations therefore are limited for use as an instrument to assess clinical competence.

The efficacy of performance checklists for the delineation of technical competence should not be minimized. McCaffrey (1978) developed a checklist that rated routine neurological vital signs. In order to clearly define and identify symptoms, a nominal (i.e. yes-no) format was utilized. It is felt by the author that checklists are advantageous in that the learner is provided with specific step-by-step instructions regarding the performance of a basic neurological exam and that the instructor is thereby provided with a reliable valid measurement tool. The checklist in question was constructed in order to assess a nurse practitioner's clinical ability in neurology. Parts

of it focus on the nurse's ability to give the patient correct instructions because it was noted that to accurately rate a neurological baseline a nurse must give appropriate instructions as well as make observations. The methodology involved the instructor checking off whether the behaviors were performed correctly or incorrectly (i.e. yes or no on the checklist) while the subject (nurse) conducted the examination. Immediately following the examination, the nurse was required to chart the findings on a worksheet. It was revealed that prior to the use of the performance checklist aproximately one out of ten nurses successfully performed the neurological exam. After implementation of the checklist unsuccessful clinical performance ws found to be rare. It was further posited that checklists preclude the use of unnecessary redundant verbal teaching. Finally, the notion was put forward that checklists provide immediate feedback on performance, and are a self-paced learning method.

Reliability-validity issues were addressed by O'Donohue Jr., and Wergin (1978) in a study that consisted of medical students being evaluated during a clinical clerkship in internal medicine. Specifically, 175 third year medical students were evaluated during a three month clinical clerkship in medicine. They were evaluated by full-time attending physicians and medical residents. Every student was evaluated by at least five instructors, each of whom submitted an independent assessment form. At the end of the three month period all students were given a written and an

oral exam. The latter contained mostly multiple choice items. The experimenters were concerned with investigating the alleged reliability of independent faculty ratings of student clinical competence. A sub-goal was to determine the reliability of independent faculty judgments of student competence in an oral exam. A further aim was to delimit the nature and degree of relationship among three measures of student clinical competence: written test, oral exam and clinical performance ratings. Results indicated a good correlation (r=.70) between the ratings assigned by the attending physicians and the residents. It was also discovered that resident physicians tended to give students slightly higher scores than did attending physicians. Reliability was found to exist between two independent judges (r=.75) via-à-vis the oral examination. It was found however that intraclass correlations among attending physicians and among residents were low (.32 and .38 respectively) which led the experimenters to conclude that a greater variation of ratings existed within students when evaluated separately and at different times by the attending physicians or residents. In essence, individual differences overshadowed possible rating disparities between the attending physicians and residents. Finally, it was discerned by virtue of the intercorrelations that the relationships among the three sources of student evaluation were small which would seem to indicate a definite dichotomy between successful performance in an artificial environment (i.e. and exam) and performance in a real-life setting. The

latter conclusion was confirmed by the fact that neither the written nor the oral exam alone correlated well with clinical performance.

Measurement of Interviewing Skills

In order to gather data (elicit symptoms) from a patient a doctor must possess a certain degree of interviewing skill. However, the latter component of physician performance has been found to be difficult to evaluate. Again, the problem found to be present was the difficulty in obtaining significant interrater reliability. In essence, the objective measurement of interviewing skills has not been easily attainable.

Helfer and Hess (1970) comment on the difficulty that medical educators experience when trying to objectively measure interpersonal communication skills of medical students. It is noted that traditionally ratings have been found to be extremely subjective and done by a single observer. It is also stated that more commonly, interpersonal skills are not quantitatively assessed by medical educators. It is claimed that students require feedback and constructive criticism. It is also posited that standardization of the interviewing technique was difficult due to the at times wide variation between patients; for example differing problems and backgrounds, lack of motivation and wide-ranging verbal skills from poor

to excellent. The experimenters felt that the problem of standardization could be solved by training simulators to portray patients (in this case mothers). Two measuring instruments to rate the interviewing technique of six randomly selected senior medical students were utilized: an interactional analysis form which listed 11 behavioral categories presumably exhibited by the physicians in varying degrees during the interview with the mother and a second instrument which dealt with factual information of both and organic and psychological nature that the physician recorded. Five sophomore medical students were utilized as raters. They were provided with operational definitions of the 11 categories on the interactional analysis form and all raters were trained together. With regard to interrater reliability, it was found that the latter coefficient rarely fell below .90 on the 11 categories in the interaction analysis form or below .75 when several different interviews were rated on any specific category contained in the form. Not surprisingly, it was discovered by the experimenters that much less time was required for the raters to reach similar levels of agreement for the factual checklist. This underscores the previously mentioned presence of a subjective variable when rating total physician performance.

In a similar vein, Barbee, Feldman and Chosy (1967) found it difficult to obtain interrater reliability when measuring interviewing skills of medical students. Like many others it was their contention that the evaluation of medical education lacks objectivity and precision especially

during the clinical years. They were aware of the problem of obtaining interrater agreement as well as the difficulty of formulating quantitative criteria to assess performance. The subjects consisted of ten sophomore medical students who were enrolled in an introductory course in clinical medicine. Audiotape recordings were made of the present illness history taken by the students during their first patient interviews and on each of the succeeding six interviews during the semester. Forty tapes, i.e. the first two and the final two interviews by each of the ten students were evaluated. The evaluation form itself was adapted from research done by Hinz (1966) and Matarazzo, Phillips, Wiens and Saslow (1965).

The raters consisted of four faculty members of the Department of Medicine of which three participated in three two hour training sessions in which the use of the form was explained and the practice tapes were evaluated. The fourth faculty member received no training and served as a control. Each rater independently evaluated all 40 of the randomly numbered tapes. After completing the tape evaluation each rater was required to assign a number weight of 1-7 to each of the 12 items on the present illness history form indicating what the rater felt each item represented in terms of relative importance in the overall evaluation of the present illness history. The original score given by the rater to each item was multiplied by this factor to arrive at a weighted score for each of the tapes. The latter were then compared to the unweighted present-illness

history scores. Six months later (to test reliability) each rater rerated eight of the 40 tapes. The two sets of ratings were compared. The results showed that following the six hour training session significant interrater agreement was achieved by the three trained raters. However, little agreement was found to be present between the untrained rater and any of the other three raters. When the scores were recalculated according to the weight or importance attached to each of the 12 items by the individual raters, the previous level of agreement was unchanged and it was also found that the untrained rater showed no significant agreement with the other raters. The correlation was .96 between the weighted and unweighted ratings.

Significantly, it was also determined in the experiment that less satisfactory levels of agreement were elicited on the interview technique section of the form. The summary question was found to have the best correlation (r=.47). For the eight tapes that were rerated six months later, two of the three trained raters were able to duplicate their previous ratings on the history section of the form with high correlations of .90 and .82 respectively. The third trained rater and the control were unable to repeat their previous performance, the correlations being .12 and -.13 respectively. However, the test-retest reliability on the interview technique section was found to be extremely poor. None of the raters made second evaluations which significantly correlated with their initial ratings. It was

further noted that the importance of using trained raters was underscored by the consistent lack of agreement characteristic of the untrained rater. The results also revealed a need for periodic retraining as evidenced by the fact that one of the trained raters could not duplicate his original ratings when rerating the 8 tapes 6 months later. It was further noted that a more permanent interrater agreement could have been achieved had practice tapes been utilized intermittently throughout the study. It was also concluded that since agreement between raters proved to be higher on the present illness history section of the form than on the interview technique part then a safe inference could be made that the latter result was a function of the more subjective and individualized nature of the interview technique.

However, the above paradigm contained several flaws. The use of audio rather than videotape is somewhat limiting and may not consititute a full revelation of the differing vicissitudes inherent during the doctor-patient encounter. The use of videotape could be helpful as facial expressions, cues, or other physical demeanours might be identified and rated. Finally, the discovery that the interview section of the form contained little interrater reliability would seem to identify a need to better quantify interviewing skills into more accurate, discrete components.

The Assessment of Clinical Skills

Hinz (1966) experienced considerable problems when trying to construct an objective instrument that could be utilized to clearly assess clinical skills. The study was not successful as evidenced by the failure to obtain interrater reliability. The experimenter believed that a conundrum exists in the field of medicine in that although medical knowledge is measured by objective examination, he did not believe that suitable objective measurements of student performance in the area of clinical skills existed. It is further observed that performance is seldom witnessed by faculty and is mostly judged indirectly via case reports or the presentation and discussion of cases by the student on ward rounds. Hinz (1966) states his belief that the latter methods are subjective and that there is no uniform method of determining excellence in the case method based upon performance. His study focuses upon the development of a method of direct observation of students used as both a teaching and evaluating device. The objectives were to determine if teaching is improved by having an instructor observe at the bedside during a student's case workup, to develop objective scoring scales to judge performance in the case method, and to learn if direct observation identifies aspects of student performance that normally are not apparent.

The method involved the use of a rating scale by trained physician observers working with third year medical students on an introductory medical clerkship. The rating scale was constructed in a manner such that items were categorized according to various portions of the patient examination focusing attention on the content of the present illness and the technique of eliciting the history. To keep the number of items limited it was decided to sacrifice specificity for general applicability. There were 50 items on the list. They were scored on a 4-point scale according to whether they were performed completely or incompletely. The scale provided for considerable space for notations during observation at the bedside to aid in completeness of scoring and to facilitate later discussion and feedback with the student. A separate checklist for all individual parts of the physical exam was also used. As predicted, there existed the problems of lack of interrater reliability and rater consistency despite the fact that they were given extensive training which involved the use of both video and audiotape films, all of which utilized the rating scale. The raters consisted of four senior residents who performed repeated ratings on five or six students.

The procedure involved each student performing a complete workup of a newly admitted hospital patient. The rater sat at the bedside during the entire history and physical exam and recorded the observations. At the end of the workup the students summarized their findings and presented them to the rater. The subjects were observed on

the first patient they examined and also after two and four months of clerkship experience.

It would appear that several flaws were present in the experiment. Many of the criteria were imprecise and lacking in definitive quantification procedures. For example, it was revealed that the degree of interrater variation was not calculated statistically but the experimenter stated that it was recognized to be great. The latter is an indicator of faulty methodological standards. Another design imperfection was the lack of precise numerical scoring as individual items were not weighted even though the experimenters recognized that they varied in importance. The forms were scored according to the total number of items performed completely or nearly completely compared with those that were imcomplete or omitted. Items omitted were noted in order to identify consistent patterns of error. The fact that individual items could not be weighted would seem to denote that the scale lacked reliability. However, it was discovered that direct observation was useful in that it revealed aspects of student performance that were not readily apparent on ward rounds; an illustration was that some subjects were found to perform well on direct observation but presented cases badly on ward rounds and vice versa.

Hinz (1966) expanded the project by conducting relability studies of direct observation in which senior medical students were the raters. However, it is unclear how reliability can be established if the knowledge base of

the raters (i.e. senior medical students) is less than optimal. The method was rather flawed in that it consisted of five three hour group sessions in which the raters independently scored films or audiotapes of actual interviews. Next followed item by item discussion of the ratings. At the end of the training period, item agreement between individual pairs of observers averaged 75 and 68% respectively in two groups of four raters each. Conversely, in the actual test the four observers worked in pairs, each pair rating 10 different students at the bedside and making an audiotape recording of the interview. Each observer pair next rated the tapes of the 10 interviews made by the other rater pair. In essence, each of the 20 interviews was rated by each of the four observers.

The results showed that item agreement between pairs of observers remained at 75% during the experimental period and did not improve during the period of data collection. The latter finding was an indication that the raters had reached maximum levels of agreement only during the training period. No single item was deemed to be responsible for the disagreements. The experimenter conceded that the broad distribution of disagreement was an indicator that sources of inconsistency were inherent in the method used rather than isolated to a few items or definitions. As a result, an adequate test of reliability could not be devised by the experimenter. It is further stated by the experimenter that in order to achieve a significant test of relability multiple observations of a large number of individual

students by multiple observers would be needed.

A Final Note on Interviewing Skills

There have been further attempts at creating a measuring instrument that could discretely assess certain aspects of clinical performance (e.g. interviewing skills). The thesis posited is that successful physician performance is not solely dependent on technical knowledge but also on humanistic expertise. The argument can be put forward that the latter qualities cannot be objectively measured but Stillman (1980) disagrees. As will be discussed later within this review, it is imperative that the physician when in a problem solving mode possess the ability to delineate the appropriate cues (symptoms) from a patient in order to construct a diagnosis. Accordingly, a poor interviewer would not be able to tease out or intuitively discern the pathology ostensibly present in a patient. Stillman, Sabers and Redfield (1976) pioneered the use of the Arizona Clinical Interview Rating (ACIR) Scale as a device for grading interviewing skills. The scale was developed after listening to and observing interviews conducted by experienced clinicians who were considered to be excellent interviewers. It was attempted by the constructors of the scale to identify characteristics of their techniques that were unrelated to the information that they collected. Fourteen criteria that were characteristic of exceptional

technique and applicable to any type of interview were formulated by the experimenters. Each category was evaluated using a five-point descending scale which described performance that ranged from excellent through average to poor.

Stillman (1980) did not discuss any problem in obtaining inter or intra-rater reliability. Correlations of .85 and .90 are recommended in order to safely conclude that both inter and intra-rater reliability are present.

It is somewhat curious that implementation of the ACIR (unlike most of the previously cited studies in the area) was not plagued by relability factors. Perhaps the latter was a function of better rater training or higher educational levels on the part of all concerned. However, the ACIR seems to possess some limitations: a five point scale is utilized although only three points are defined; also there are no equal intervals between scale points. Other problems include the fact that the scale is not behavioral, it does not allow the patient to complete a train of thought, and also some items seem to measure more than one behavior. Differing items are related to each other and overlap in a variety of ways. In effect, the scale does not appear to be additive.

As evidenced by the problems cited, it appears to be somewhat difficult to accurately gauge clinical competence of both a technical and more so of an interpersonal nature. Depending on the subspecialty focused upon, the use of checklists containing discretely defined nominal criteria

have proven to be useful in some instances. However, accurate evaluation of total physician performance which would include all aspects of doctor-patient interaction has remained an elusive goal.

Medical Problem Solving

A new epistemological paradigm pertaining to medical education has emerged. The new and innovative frame of reference being posited is that the physician is essentially a problem solver. It has been attempted to relate and apply the Newell and Simon (1972) information-processing theory of human problem solving to the physician. Recent research which focused upon how doctors typically diagnose (i.e. solve problems) patients has identified various stable and invariant characteristics among all subjects in the differing paradigms. One of the principal tenets of the Newell and Simon model is that the individual operations performed by the problem solver do not vary over disciplines; the notion being that a musician composing a fugue or a surgeon planning an intricate operation utilize similar cognitive stfategies. In other words, conceptual knowledge facilitates success in an endeavour, however the methods utilized to reach a goal seem to be identical regardless ofthe information possessed by the individual.

Newell (1977) states that problem solving occurs in a problem space. This "area" consists of states of knowledge (nodes) pertaining to the problem. Newell goes on to say that the initial situation and the desired situation co-exist within the problem space. The latter also contains operators and analysts. Newell and his colleague Simon say that complex thinking processes that occur when attempting to solve a problem take place in what is termed a production system, the principal components consisting of a condition and an action. Apparently, at each node a decision is reached by an analyst. An example might be a decision to cross the street after a traffic light turned green. If a specified condition of the production system at each state of knowledge (node) is met, then the action part of the production system is evoked. Many productions may occur at a node. According to the authors it is not inconceivable that a node may have to satisfy many conditions. Each node contains an operator; the latter can produce new nodes (i.e. states of knowledge) if conditions are met. In essence, there seems to be a definitive hierarchy.

Simon (1978) gives a more complete breakdown of the progression followed in the information-processing theory. It is stated that problem solving involves an interaction between an information processing system, the problem solver and the task environment. The latter is the task as described by the experimenter. As noted earlier, the problem solver represents the situation in terms of a problem space. Simon does not believe that the information

processing system varies over task and problem solver. It is his contention that the structure of the problem space determines the possible strategies that can be utilized for problem solving. The information processing system operates serially; there is apparently a chunking of four to seven pieces of information in limited short-term memory. The latter conforms to the Miller (1956) thesis of the "magical number 7 plus or minus two" where it is stated that apparently short-term memory can only effectively encode between five and nine bits of information. The ability to solve problems also involves the efficacious utilization of information stored in long-term memory which has as one of its properties a seemingly unlimited capacity.

Simon (1978) further states that search for information is done in a sequential manner. The relative ease in solving a problem is dependent upon how successful the subject has been in representing critical features of the task environment in the problem space. It is also felt by the author that trial and error search is involved but only to a limited extent. There is some discussion of means-end analysis which is defined as a strategy in which differences between current and desired situations are discovered and next an operator relevant to each difference is applied to reduce the disparity. With reference to the physician, it will be clarified in the latter part of the review that more often than not, doctors search for positive cues to confirm an hypothesis. To a much lesser extent they subtract negative from positive cues to arrive at a conclusion.

Simon (1978) theorizes that the nature of the task environment is critical in order to solve the problem. Some of the relevant variables are whether or not the instructions are clear, and if there might exist multiple solutions. It would appear then that for a doctor, if the patient exhibits definitive symptoms, the task environment would be unambiguous. The goal would be to ascertain the nature of the problem and institute remedial action. It is further noted that one of the disadvantages found to be present in ill-structured problems is that the latter hampers recall of possibly helpful information from long-term memory.

Protocol Analysis

One of the principal characteristics of information-processing theory is the use of verbal protocols to analyse thinking. Verbal protocols involve the subjects stating out loud their thought processes as they go about solving a problem. Essentially the subject is thinking out loud. The use of verbal protocols does have limitations. The principal contentious issue is whether or not thinking aloud constitutes an accurate depiction of normally silent internal cognitions.

Simon (1978) does not believe that the thinking aloud methodology causes gross changes in problem solving behaviour, although he does note that the thinking aloud

problem solver is slower, more planful and somewhat more deliberate in demeanour than when silent. A study performed by Ericsson (1975) is cited in which it was found that subjects did not vocalize goals that could be realized immediately as often as longer range goals that were reachable though intermediate sub-goals. Vocalization was found to decrease as subjects became more proficient in the task; at this point responses seemed to be automatic. It was the conclusion of Ericsson (1975) that the subject's goal statements were predictive of subsequent moves.

Nevertheless, it is virtually impossible to quantify thinking processes or to really know whether or not the subject is revealing everything critical concerning his internal cognitive mechanisms. Researchers must be cognizant of the fact that there do not exist any completely reliable methods of making accurate measurements on the so-called internal space notwithstanding the complicated procedure of protocol analysis formulated by Newell (1977). The latter method appears to contain some ambiguity and the possibility does exist that information not actually present could inadvertently be created. Another limitation present when utilizing verbal protocols is that it remains difficult to discover what cognitive strategies preceded the verbalization; hence it would be somewhat onerous to accurately gauge the progression of problem-solving ability. Lindsay and Norman (1972) state that possibly only some of the subject's internal cognitions are revealed during the protocol and that intermediate states are missed.

Essentially, problem-solving ability is inferred because at times verbal protocols do not make sense. It is, however, quite conceivable that verbalizations might represent the end product of a previous cognitive function that was not revealed by the subject.

Possible Limitations of Information Processing Theory

Greeno (1978) criticizes the information-processing model of problem-solving postulated by Newell and Simon (1972). It is stated that the theory contains strong concepts for use in analysing specific tasks, but has not developed a coherent body of theory composed of general psychological principles. Ostensibly the latter method would better explicate performance in broad classes of problems. Greeno defines problem solving as the process of identifying relations among components and fitting the relations together into a pattern. Similarly, it will be illustrated that a doctor identifies relationships between cues (symptoms) and fits them together to form an hypothesis. Greeno does concede that problem-solving skill interacts with the individual's level of general conceptual knowledge.

One possible imperfection in the hypothesis of an alleged relationship between information-processing theory and medical problem solving is that the consequences of not being able to solve artifically constructed puzzle-type

problems such as the Tower of Hanoi or Donald and Gerald=Robert in no way approach the severity of a doctor's formulation of an inaccurate diagnosis. Thus the subject's motivation to solve a problem might vary over disciplines or categorical modes.

Memory does play a role in generating solutions to problems. Greeno (1978) cites a study performed by De Groot (1966) in which it was discovered that chess masters exhibited better recall of the position of most pieces on the board than did average players. This was interpreted by Simon and Gilmartin (1973) as indicating that chess masters have stored in memory a large number of patterns of a few pieces which they can recognize as units. Similarly, it will be illustrated later in the review that a good and experienced physician stores in memory a large number of combinations (cues) of a few competing hypotheses which are then constructed into units. The physician eventually discards the irrelevant hyoptheses and settles on one.

Specific Applications of Medical Problem Solving

Now that the theoretical underpinnings as well as the inherent problems of information processing theory have been discussed and evaluated, a selected sample of experimental manipulations which focused on the physician-problem solver will be presented. The goal will be to highlight the salient characteristics manifested by the doctor while in a

problem solving work situation mode.

To expand on the previously cited De Groot (1966) chess findings, Norman, Jacoby, Feightner and Campbell (1979) were of the opinion that the thinking processes involved in chess are somewhat similar to those of clinical reasoning (i.e. medical problem solving). De Groot (1966) discovered that chess players (like doctors) generated a number of competing hypotheses about possible moves. It was uncovered that the number of hypotheses, depth of strategy and early hypotheses did not vary between chess masters and average players. The latter findings constitute confirmation of the Newell and Simon (1972) "univariate" tenet of problem solving. However, De Groot (1966) did discover that the specific nature of the moves between average players and masters did in fact vary. For example, when viewing a typical mid-game position chess masters remembered more than average players about position location of each piece. The two groups did not differ when exposed to pieces on a board distributed at random. Based upon the chess studies it was the prediction of Norman, Jacoby, Feightner and Campbell (1979) that an experienced physician when presented a case history of a typical case should recall more details of the latter than would a novice. It is further stated that if presented with a random array of signs and symptoms (i.e. an atypical case), there should be no significant differences in recall between expert and novice groups. The subject population consisted of four groups with each containing five participants. Group one consisted of second year medical

undergraduates, group two contained first year residents in family medicine, group three had third year residents in family medicine, and group four had the practising family physicians. After being presented with atypical and typical written case protocols in which each subject was required to verbalize whatever they remembered about each case, the results confirmed the experimenters' prognostication. There was no difference in recall between typical and atypical cases for medical students and first year residents. Third year residents had slightly better recall of typical cases but it was found that physicians showed significantly better recall of typical cases. Higher educational level was associated with a greater ability to recall information. There was a significant type by level interaction indicating that increased recall of typical cases is associated with increased experience and education. Since recall of atypical cases did not vary significantly over groups, memory ability for problem solving may not be a significant factor for atypical cases. With reference to better recall of typical cases, presumably the experienced physician knows what symptoms to look for, i.e. to what to attend to and what to encode. However, the processing similarities of atypical cases reveal quite explicitly that the methods of problem solving do not widely differ across individuals.

Norman and Feightner (1981) compared the performance of medical students on simulated patients and on patient management problems (P.M.P.s). The latter is operationally defined as a written problem that describes characteristics

and symptoms of a clinical case. Information is elicited by the erasing or rubbing out of a response box. A simulated patient encounter is one in which actors are trained to masquerade as patients complete with the alleged presence of appropriate symptoms. One important discovery via-à-vis medical problem solving was that the clinical measures adopted by each student were specifically a function of the nature of the problem. The number of critical options selected by each participant was deemed to be strongly influenced by the specific features of the problem. The experiment revealed that the physician or medical student's actions were content specific with reference to the problem at hand. It can also be noted that the relatively low level of individual differences measured in the study lends credence to the Newell and Simon (1972) theory that the methods utilized to solve a problem are invariant over individuals. However, superior conceptual knowledge facilitates efficiency in problem solving.

The following set of experiments extracted and identified similar characteristics and properties of the physician-problem solver.

Barrows, Norman, Neufield and Feightner (1981) examined the clinical reasoning processes of 18 family physicians and 19 general internists via the use of the previously defined simulated patient method. Each subject was exposed to four different cases and was required to formulate a definitive diagnosis. The results revealed that all subjects generated early, multiple and at time competing hypotheses. The first

hypothesis was advanced on an average of 28 seconds after knowledge of the first complaint. The correct hypothesis occurred one to seven minutes into the encounter. There was a mean of 5.5 hypotheses per patient. Questions asked the patient by the doctor were discovered to be for the purpose of testing hypotheses. It would appear that physicians are not of the "null hypothesis" school. They do not begin an analysis with the assumption that no relationship exists between variables. Conversely, they ask questions in order to confirm existing hypotheses, they look for positive cues and they do not actively seek information that would serve to eliminate hypotheses. It was found that fifty-two out of fifty-three physicians who had obtained the correct hypothesis had thought of it during the patient encounter. Thoroughness was not found to be a function of accuracy as the experimenters discovered that neither the total time of the encounter nor the amount of data gathered was significant in predicting the correct diagnostic outcome. The data in this experiment are virtually identical to previous and subsequent studies in this review, especially the finding that physicians formulate early and at times competing multiple hypotheses, the number of which range from approximately four to seven.

It may be somewhat presumptuous to state that the above data serve as confirmation of the Newell and Simon (1972) model. However, certain elements might in fact constitute positive evidence for their approach. The paucity of individual differences lend weight to the invariance theorem

of problem solving defined earlier in the review. The limited number of formulated hypotheses (4-7) are an indication of the limited capacity of short-term memory as well as proof of the Miller (1956) pioneer finding that culminated in the "Magical number 7 plus or minus two" treatise. The Barrows, Norman, Neufeld and Feightner (1981) study as well as future studies to be described later in the review reveal that multiple early hypothesis formulation is a central element of the physician-problem solver. This multiple hypothetico- deductive model has been found to be stable across physicians and students.

In an earlier study that focused upon medical problem solving Neufeld, Norman, Feightner and Barrows (1975) were primarily interested in two basic issues. Specifically, the goal was to determine whether or not a relationship exists between educational level and certain clinical problem solving strategies. A sub-goal was to delineate how the strategies of medical students compared with those of experienced physicians. Results indicated that greater experience and higher educational levels are indicators of more expertise in medical problem solving, however in a general contextual framework the process is similar at all levels. Both groups displayed early diagnostic hypotheses, the latter usually occurring within the first minute of the encounter. The specificity of the hypotheses correlated positively with educational level but their number or time of onset did not: One half of the questions asked by the students was determined to be specifically

testing hypotheses, one third of the questions was classified as non-routine. Results indicated that as a function of increasing education, students were more thorough and efficient in their data gathering although the experimenters did not feel that the relationship was strong. It was further concluded that among both physicians and students, the earliness of hypothesis generation and the number of hypotheses did not reveal the existence of a relationship with the outcome measures; however a low association was found to be present between thoroughness and early hypothesis formation. A further identifiable result was that successful problem solving in students was characterized by the generation of rather specific hypotheses and a highly problem oriented search strategy. The latter subjects utilized specific questions for the purpose of eliciting significant findings as opposed to the use of extensive routine questions. The experimenters ascertained that the genesis of the hypotheses appeared to correlate with previous patient experience; the latter finding appears to constitute a reinforcement for the notion of problem based learning as a device to acquire the method of problem solving and to increase the number of patient cases to which a student might refer to generate hypotheses. A final result of the study was that there was a positive correlation between the ability to diagnose and higher educational level.

In another exploration of clinical reasoning Neufeld, Norman, Feightner and Barrows (1981) compared the problem solving abilities of physicians, pre-clerkship students and recent post graduates. The simulated patient method was used. Each patient-participant interaction was videotaped. As in previous experiments concerning this topic by this group of researchers subjects viewed a videotape of their individual encounter with the simulated patient while at the same time verbally recalling the thinking processes. Non-directive open-ended questions were utilized for the purpose of stimulating memory. The results could be viewed as a confirmation that the information processing theory paradigm is applicable to medical problem solving. All groups were discovered to have constructed early hypotheses, approximately six in number which serves to illustrate the limited capacity of short term memory (S.T.M.). Hypotheses were generated roughly 30 seconds into the doctor-patient encounter. It was determined that all groups uncovered the correct hypothesis (diagnosis) after a time period ranging from seven to eleven minutes. An important finding was that the content and specificity of the hypotheses were a function of educational level; however the accuracy of diagnostic outcome was strongly correlated with increasing education.

Notwithstanding the role that is played by educational level and experience in the ability to compose a correct diagnosis, the above experiment is a clear illustration that medical students and physicians approach problem solving in

a similar manner. The striking and most important finding with reference to information processing theory was that all groups exhibited parallel processing, i.e. the formulation of early, multiple and at times competing hypotheses. The only differences found to exist between physician and student were that the former group performed more physical maneuvers during the workup and elicited more critical findings. These differences were attributed to educational level and not processing variations.

The results should not be accepted without reservation. Certain contentious issues can be raised. Stimulated recall via the employment of non-directive open-ended questions is not a reliable procedure. It is too retrospective and it is not inconceivable that the subject would not be able to remember all of the thinking processes used during the patient encounter. Another possible dispute concerns whether or not the subject is telling the truth. In order to mask an apparent inability at solving a problem, the subject may exaggerate thinking processes to the experimenter. Another possible limitation was that the experimenters did not operationally define non-directive open-ended questions.

The most definitive and exhaustive investigation of medical problem solving was the Medical Inquiry Project at Michigan State University which was administered by Elstein, Shulman and Sprafka (1978).

Since the findings are strikingly similar to those described earlier, only salient highlights of the study will be described. The experimenters placed medical problem solving in the context of the Newell and Simon (1972) theory of information processing. The problem solving behavior of experienced physicians and medical students at varying levels of their training was studied in much the same manner as in the previously described studies. The designs included P.M.P.s, the simulated patient method and fixed order problems.

Elstein, Shulman and Sprafka (1978) postulated the existence of a four-stage model of medical problem solving: cue acquisition (i.e. symptom compilation), hypothesis generation, cue interpretation and hypothesis evaluation. The most consistent finding across all subjects and designs was the generation of early hypotheses. The latter, according to information processing theory, serve as organizers of the data in short term memory. The results also indicated a high degree of content specificity in the process of solving medical problems. Indeed, one of the barriers that prevent formulation of discrete rules or characteristics of medical problem solving is the difficulty of predicting physician performance based upon one problem on problems in different domains. Accordingly, the experimenters propose that content specificity is an implication that excellence in medical problem solving depends upon the type of problem as well as the thinking strategies employed.

Other results of Elstein, Shulman and Sprafka (1978) include the fact that there was an inability to discern differences between criterial (expert) and non-criterial (non-expert) physicians. At times, non-criterial physicians posed more questions prior to generating a first hypothesis; however, they did collect more data and interpreted the latter more accurately.

However, none of the findings concerning alleged differences between criterial and non-criterial doctors were statistically significant. Furthermore, the experimenters did not extricate a significant correlation between thoroughness of data collection and accuracy in diagnostic formulation. All subjects were found to have adopted a hypothetico-deductive approach, i.e. early hypothesis formulation and a search for positive cues for confirmation purposes. To a lesser extent negative cues were subtracted from positive cues and subjects almost never searched solely for negative cues.

There appears to be an interesting conundrum present in the Elstein, Shulman and Sprafka (1978) studies: the generation of early hypotheses regardless of the size of the data base and the revelation of very little intra-individual consistency across problems. The two pieces of data appear to be a contradiction of one another. Individual differences therefore appear to be in evidence. Since both subject (participant) and problem are unique, researchers are precluded from stating that problem solving behavior across individuals is identical. The fact that medical

reasoning appears to be case specific has led Elstein, Shulman and Sprafka (1978) to conclude that the psychological problem space reflects the characteristics of the task environment rather than personality variables. Nevertheless, the consistently stable finding of early hypothesis generation across all subjects and at all levels does indicate that there are certain unitary aspects of problem solving. The fact that all subjects in this (and previous) experiments simultaneously considered between four and seven hypotheses (i.e. parallel processing) does lend some weight to the Newell and Simon thesis.

According to McGaghie (1980), the Elstein, Shulman and Sprafka (1978) project contains several methodological flaws. The manner in which 15 measures of diagnostic reasoning were compiled to represent medical problem solving is criticized. Elstein, Shulman and Sprafka (1978) stated that the measures were built from the three base variables of information search units, cues and hypotheses. It is the contention of McGaghie (1980) that a numerical reliability coefficient for each dependent variable should have been reported. Elstein, Shulman and Sprafka (1978) stated that the goal was not to develop scoring keys but to evaluate clinical competence. However, McGaghie (1980) remains convinced that the validity of the data suffers becuse the variables said to represent clinical competence cannot be reliably measured. It is further stated in the latter critique that the sample size utilized across problems was too small for the use of factor analysis. The reply of

Elstein was that a prohibitive sample of 200 would have been required. It is McGaghie's conclusion that close inspection of the data does not confirm the model's postulation of the three units of analysis contructed from the 15 variables. Accordingly, McGaghie does not feel that the Elstein, Shulman and Sprafka (1978) model of cue acquisition, hypothesis generation, cue interpretation and hypothesis evaluation is valid since the model lacks internal consistency. All concerned agree however that future studies should focus on a wider range of medical cases.

Present Research

The present thesis will examine the previously discussed issues of early hypothesis generation, performance on computerized multiple choice exams versus clinical performance and will attempt to close the previously described gaps that currently encumber medical education. It is an outgrowth of work done by Groen, Dauphinee and McQueen (1981) in which the possible relationship between speed and success on multiple choice examinations was examined. The subjects were 17 fourth year medical students from McGill University. An APPLE microcomputer was used for the study in which each subject was required to answer 20 multiple choice questions. The first two questions were utilized for practice purposes. Ten questions were from Medicine, six from Surgery, two from Obstetrics and two from

other subjects. All the problems were of a form which included, in the following order, a description of the patient and that patient's symptoms, a question requiring that a diagnosis be constructed based upon the preceding information and a list of six alternative diagnoses. Each question was presented in segmented form on a video display screen; i.e. each symptom or patient characteristic was shown individually. The subject was required to press F (for forward) on a keyboard to be shown succeeding segments. If a subject wished to review previously shown segments B (for backtrack) was to be pressed. Backtracking also conformed to the segmented procedure. The final segment of the problem always consisted of the question and the six given diagnoses. The subject was required to press (on the keyboard) a number (1-6) corresponding to what was felt to be the correct answer. The dependent variables were solution times and the amount of backtracking done. Results indicated that slow subjects who did a lot of backtracking during the experimental procedure tended to do poorly in the actual final exam in Medicine. It was also noted that fast performance in general during the experiment was related to good performance in general on the examination. For the actual experiment mean time per subject per segment was measured and it was determined that subjects spent more time on problems answered incorrectly. The latter was also found to be true when measuring overall mean time per problem. T-tests comparing mean times for problems answered correctly and incorrectly were significant ($t=3.91$, $p<.005$).

Recently, McQueen (1981) utilized 13 residents as subjects for the experimental procedure. The problem solving characteristics of residents appeared to be identical to those of fourth year medical students. Residents were found to have spent more time overall on problems answered incorrectly and more time per segment on the problem that proved to be insoluble to them. A one-tailed t-test comparing mean times answered correctly and incorrectly was significant ($t=2.93$, $p<.01$). Not surprisingly, and probably due to their more extensive knowledge base, the residents were found to be more accurate than the medical students.

Upon comparison of the backtracking done by the residents and medical students, some interesting differences were found by McQueen (1981). The fourth year medical students backtracked on between 0 and 11 problems, and fell into two distinct groups, backtracking on a few or on many problems. Three medical students did not backtrack at all while all residents backtracked on at least one problem. Residents did not backtrack on more than 10 problems. The variance of backtracking was less than that of the medical students which is an indication that the former appear to be a more homogeneous group than the latter.

Conclusion

The literature review has attempted to examine the possible link between the Newell and Simon (1972) information processing theory of problem solving and medical problem solving. In one aspect (early multiple hypothesis formation and limited capacity S.T.M.) the evidence is incontrovertible. On the other hand, the relationship can be perceived as being somewhat tenuous as evidenced by the finding that medical problem solving behaviour is case specific. There exists both variance and invariance. The consistent finding of early hypothesis generation among medical people at all levels does indicate that globally the process of medical problem solving is done in a similar manner. There do seem to exist preconceived notions about the nature and aetiology of the problem at hand. However it was shown that successful medical problem solving is a function of increasing educational level. The discovery of early hypothesis generation should not be regarded as a dramatic breakthrough. It is imperative that physicians classify patients along a discrete continuum at an early stage in order to ascertain what or if remedial action should be pursued. The most cogent finding among all research paradigms was the Elstein, Shulman and Sprafka (1978) discovery that criterial and non-criterial doctors could not be differentiated with statistical significance. The present thesis, an offshoot of the Groen, Dauphinee and McQueen (1981) study, attempted to explore whether or not a

relationship exists between level of achievement attained in hypothetical situations such as computer based multiple choice exams and performance in actual clinical situations.

CHAPTER III

METHOD

The study was conducted in two distinct frameworks: (1) the assessment of clinical performance via a 10 category checklist, and (2) the measurement of performance on multiple choice questions via a computer based paradigm.

Sample

Twenty-six third year medical students participated in the study. There were 11 females and 15 males. The members of the total group's ages ranged from 21 to 32 with a mean age of 24.5 years. The females alone ranged in age from 21 to 31, the average age being 25. Males ranged in age from 21 to 32, the average age was 24.2.

Materials

The clinical component utilized a ten category rating checklist; the first nine categories consisted of a four point ascending scale while the last category (attitudinal characteristics) was om a three point ascending scale.

The microcomputer based multiple choice questions were gleaned from McGill University final exams in the Faculty of Medicine of recent years. They were of the "one best

response" type. In addition, each item contained five distractors. There were 20 questions in all, two of which were for practice purposes. The questions were selected from several different sections of the exams (10 from Medicine, six from Surgery, two from Obstetrics and two from other sections). The questions were chosen to represent a cross-section of medical problems, necessitating diagnostic ability and clinical experience, as well as recall of factual knowledge.

As in previous studies, the problems were of a form which included, in the following order, (1) a description of the patient and that patient's symptoms, (2) a question requiring that a diagnosis be made based on the preceding information, and (3) a list of six alternative diagnoses. The description of the patient included presenting symptoms and sometimes results of lab tests or physical examinations. All the questions were of a similar form, each asked for the most appropriate explanation or the most likely diagnosis. The ordinal position of the correct response among the six choices was randomly assigned.

The questions were prepared in a 38 column format for presentation. Each question was separated into segments. The final segment of a problem always consisted of the question and the six given diagnoses. The preceding information about the patient was divided into a number of segments, one or two sentences in length, on the basis of the way such information would group together if obtained in a natural setting. The questions and the segments were

selected by an experienced physician.

Of the 20 multiple choice questions, the two practice items were selected at random and presented in fixed order. The remaining 18 problems were the designated experimental items. They were presented in four counterbalanced random orders, the sequential pattern utilized was 4-1-3-2.

Equipment

The multiple choice question paradigm was run under the control of an APPLE computer, modified to measure reaction time through add-on equipment developed by Digitry, Inc. The experimental procedures were implemented using the programming language PASCAL. The experimental stimuli were presented on the screen of a small television monitor. The subjects responded by using a special keyboard connected to the APPLE computer.

Procedure

Clinical Component

All aspects pertaining to the various processes of the subjects' clinical performance were evaluated via the checklist. Each individual subject was not rated by an

equal amount of instructors. The number of raters ranged from one to four per subject. All aspects related to the course of working up the patient and recommending remedial action were graded via the checklist (see appendix).

Computer Based Multiple Choice Question Experiment

The subject was seated in front of a television screen and a keyboard. The next step consisted of the multiple choice question being presented on the screen, one segment at a time. The subject was able to control the presentation of segments and answers via the keyboard. The latter consisted of eight keys; the first six, numbered one to six, were for the purpose of answering the multiple choice question, while the B key -- the backward key, and the F key -- the forward key, controlled segment presentation.

Prior to each question, a message appeared asking the subject to wait for the ready signal. Shortly thereafter, a ready message was displayed: the problem number was given, and the subject was asked to press the F key to initiate the question. Initial pressing of the F key resulted in the presentation of the first segment of the problem. Next, the subject was permitted to see the subsequent segment or the preceding segment (provided they existed) by pressing the F or the B key respectively. On the first segment, pressing the B key would have no effect and, on the final segment, pressing the F key would have no effect. On the last

segment, the subject may choose to press the numbered key corresponding to the diagnosis selected. Answering the question terminated the presentation of the problem.

Time was recorded whenever a new segment was displayed. It stopped as soon as a key was pressed. The time and the key were then automatically recorded, and the timer was reset to zero. Timing began again only when a new segment was displayed.

Subsequent to a subject answering a question, no feedback was given. The message to wait for the ready signal appeared immediately, then after a short pause, the ready signal was displayed. Presentation of the next problem began when the F key was pressed, thereby restarting the timer.

Subjects were told to always use the one finger on the preferential hand when responding. Subjects were further instructed to remove the finger immediately following the depressing of a key. The computer emitted a warning beep if a subject rested a finger against a key for too long a period of time.

When doing the first two practice problems the subjects were permitted to ask any questions that concerned the experimental procedure during this time or during a pause which followed. However, during the actual experiment subjects were not allowed to ask questions, with the exception of between the presentation of problems, due to the sensitive time measurement in progress. Subjects were instructed to respond as quickly and as accurately as

possible.

CHAPTER IV

RESULTS

The outcome of the research will be presented in three parts. Tables 1 through 7 contain the descriptive statistics as well as other descriptive data pertaining to all the measures obtained. Tables 8 through 16 present the results pertaining to the goal of the project--the investigation of whether or not a relationship existed between performance on computer assisted multiple choice exams and clinical performance in third year medical students. Tables 17 to 26 serve as an examination of the data obtained on the computerized paradigm alone.

Descriptive Statistics for the Two Paradigms

Tables 1 to 7 provide descriptive data obtained on the computerized multiple choice exam and the clinical evaluation checklist.

Table 1 separately lists the means and standard deviations for all computer measures including reversals (i.e. the number of times per problem a subject went back to examine previously given information).

TABLE 1

Means and Standard Deviations of Measures
Obtained on Computerized Multiple Choice Questions

	Mean	Standard Deviation
Experimental Score	11.15	2.01
Number of Problems Backtracked Upon	8.23	4.03
Overall Mean Time Per Problem (Seconds)	69.87	16.42
Mean Time on Problems Answered Correctly	61.80	17.10
Mean Time on Problems Answered Incorrectly	83.60	21.26
Reversals Over 18 questions (percentage)	.58	.32
Reversals on questions answered correctly (percentage)	.48	.32
Reversals on questions answered incorrectly (percentage)	.72	.39

Table 2 illustrates the number of reversals per subject over all questions, and over questions answered correctly and incorrectly.

On total reversals over all questions the numbers ranged from a high of 21 to a low of one. The mean number of reversals was 10.42.

For reversals on questions answered correctly the range was from 14 to 1. The mean was 5.42.

On reversals for questions answered incorrectly the range was from 15 to 0. The mean was 5.

The above results may be indicative of a stable pattern of medium to high backtracking.

TABLE 2

Analysis of Reversals Obtained on Computer Measures

Subject	Total Reversals Over 18 Questions		Reversals over Questions Answered Correctly		Reversals over Questions Answered Incorrectly	
1	20/18	1.11	14/12	1.17	6/ 6	1.00
2	9/18	.50	3/10	.30	6/ 8	.75
3	18/18	1.00	13/14	.93	5/ 4	1.25
4	5/18	.28	3/12	.25	2/ 6	.33
5	14/18	.78	11/14	.79	3/ 4	.75
6	17/18	.94	8/11	.73	9/ 7	1.29
7	9/18	.50	3/12	.25	6/ 6	1.00
8	2/18	.11	2/11	.18	0/ 7	0.00
9	11/18	.61	7/13	.54	4/ 5	.80
10	13/18	.72	5/ 8	.63	8/10	.80
11	4/18	.22	2/11	.18	2/ 7	.29
12	13/18	.72	6/12	.50	7/ 6	1.17
13	21/18	1.17	6/ 7	.86	15/11	1.36
14	6/18	.33	2/12	.17	4/ 6	.67
15	9/18	.50	4/ 8	.50	5/10	.50
16	5/18	.28	1/12	.08	4/ 6	.67
17	9/18	.50	5/ 9	.56	4/ 9	.44
18	12/18	.67	2/ 9	.22	10/ 9	1.11
19	8/18	.44	5/12	.42	3/ 6	.50
20	2/18	.11	2/13	.15	0/ 5	0.00
21	7/18	.39	2/ 9	.22	5/ 9	.56
22	12/18	.67	5/10	.50	7/ 8	.88
23	1/18	.06	1/13	.08	0/ 5	0.00
24	19/18	1.06	14/13	1.08	5/ 5	1.00
25	17/18	.94	13/14	.93	4/ 4	1.00
26	8/18	.44	2/ 9	.22	6/ 9	.67
	mean	10.42	mean	5.42	mean	5

Table 3 focuses on the characteristics of the multiple choice questions themselves and lists the means and standard deviations.

TABLE 3

Summary of Data Across Problems on Computerized Paradigm

Variable	Mean	Standard Deviation
Segments	5.28	1.96
Time Per Segment	13.41	3.32
Time Across Problems	69.87	29.67

Table 4 basically focuses on the varying number of segments per multiple choice question. The intercorrelated variables consisted of the number of segments, time per segment and time per problem. As would be expected, there was a significant correlation $r=.81$, $p<.001$ between the number of segments per problem and the time spent per problem. It was also expected that the time spent per segment would correlate significantly with the time spent per problem. The latter proved to be the case: $r=.46$, $p<.05$.

TABLE 4

Intercorrelation Matrix of Problem Data
on Computerized Multiple Choice Question Paradigm

	Number of Segments	Time per Segment	Time per Problem
Number of Segments	---	-.13	.81**
Time per Segment	---	---	.46*
Time per Problem	---	---	---

* Significant at $p<.05$
** Significant at $p<.001$

Table 5 lists various characteristics of subject performance for each of the 18 computer-based multiple choice questions.

Question number 10 proved to be the easiest as evidenced by the fact that 24 out of the 26 subjects answered it correctly. Question number 17 was the most difficult, only nine subjects were correct in their responses. The table also reveals a lower mean time per subject per segment (12.39 seconds) on questions answered correctly than on incorrect questions (14.84). The latter result was expected due to the subsequently reported finding of a significant difference between overall mean time on correct and incorrect questions. Subjects spent more time on problems that were later found to be answered incorrectly than on questions where they proved to be successful.

TABLE 5

Item Analysis of Group's Responses on
Computerized Multiple Choice Question Model

Problem	Number of Segments	Number of Subjects Answering Correctly	Mean Time (sec) per Subject	Mean Time per Subject per Segment: All Problems	Problems Answered Correctly	Problems Answered Incorrectly
1	6	14	98.59	16.43	14.73	18.42
2	5	22	45.87	9.17	9.16	9.26
3	10	18	113.49	11.35	10.74	12.71
4	7	13	85.21	12.17	11.82	12.53
5	4	16	57.46	14.36	11.54	18.88
6	7	16	114.00	16.29	16.10	16.60
7	7	11	100.83	15.55	13.10	17.35
8	7	12	97.86	13.98	12.03	15.65
9	5	23	40.54	8.11	7.91	9.66
10	5	24	85.13	17.03	17.60	10.18
11	4	18	47.72	11.93	11.60	12.68
12	2	20	33.30	16.65	14.85	22.67
13	6	10	57.54	9.59	10.50	9.03
14	6	15	54.93	9.15	8.51	10.03
15	3	18	47.77	15.92	14.16	19.90
16	3	12	38.01	12.67	9.90	15.05
17	5	9	98.78	19.76	19.65	19.81
18	3	20	32.63	10.88	9.13	16.72

Mean Number of Segments	Overall Mean Time/Subject/Segment: All Problems	Correct	Incorrect
5.3	13.39	12.39	14.84

Table 6 lists the means and standard deviations of scores obtained on the clinical evaluation checklist.

TABLE 6

Means and Standard Deviations of Measures Obtained on Clinical Evaluation Checklist

	Mean	Standard Deviation
Global Checklist Score	2.80	.39
Investigation	2.69	.67
Differential Diagnosis/Problem List	2.84	.50
History	2.77	.47
Physical Exam	2.73	.45
Case Reports	2.82	.44
Knowledge	2.82	.59
Oral Presentations	2.70	.43
Communication Skills	2.99	.47
Self Education	2.83	.52
Attitude	2.77	.29

Table 7 lists the intercorrelations obtained among items on the clinical evaluation checklist. The great majority of significant inter-item correlations are an indication of the homogeneity of the evaluative criteria and the construct validity of the checklist; it also may have been an indicator of the existence of a halo effect.

TABLE 7

Intercorrelations of Measures Obtained on Clinical Evaluation Checklist

	Mean Score Overall	Investigation	Differential Diagnosis/ Problem List	History	Physical Exam	Case Reports	Knowledge	Oral Presentations	Communication Skills	Self Education	Attitude
Mean Score Overall	-	.68** p=.001	.70** p=.001	.76** p=.001	.86** p=.001	.71** p=.001	.80** p=.001	.82** p=.001	.72** p=.001	.78** p=.001	.81** p=.001
Investigation	-	-	.36* p=.037	.39* p=.024	.46* p=.009	.61** p=.001	.44* p=.012	.55* p=.002	.34* p=.045	.40* p=.021	.51* p=.004
Differential Diagnosis Problem List	-	-	-	.66** p=.001	.58** p=.001	.39* p=.024	.57** p=.001	.51* p=.004	.57** p=.001	.48* p=.006	.62** p=.001
History	-	-	-	-	.70** p=.001	.49* p=.005	.72** p=.001	.63** p=.001	.47* p=.008	.56** p=.001	.65** p=.001
Physical Exam	-	-	-	-	-	.62** p=.001	.77** p=.001	.59** p=.001	.64** p=.001	.65** p=.001	.66** p=.001
Case Reports	-	-	-	-	-	-	.41* p=.019	.52* p=.003	.35* p=.038	.45* p=.011	.62** p=.001
Knowledge	-	-	-	-	-	-	-	.68** p=.001	.48* p=.006	.64** p=.001	.67** p=.001
Oral Presentations	-	-	-	-	-	-	-	-	.64** p=.001	.72** p=.001	.67** p=.001
Communication Skills	-	-	-	-	-	-	-	-	-	.70** p=.001	.68** p=.001
Self Education	-	-	-	-	-	-	-	-	-	-	.68** p=.001
Attitude	-	-	-	-	-	-	-	-	-	-	-

* Significant at p<.05
** Significant at p<.001

Relationships Between Measures on Computer Exam and Clinical Checklist

As illustrated in Table 8, no consistent definitive relationship was found to exist between performance on computerized multiple choice exams and clinical competence. However, there were significant correlations between various components of the two task environments. The total multiple choice exam score correlated significantly with the clinical investigation category (r=.42, p=.017). These latter two designations were not found to correlate significantly with any other element of the respective opposing task environments. It was discovered that similar levels of correlations were found to be present between the number of multiple choice questions backtracked upon and the clinical category of differential diagnosis/problem list (r=.34, p=.046); and number of backtracks over problems with physical exam (r=.35, p=.04) and finally number of backtracks over problems and the clinical category titled knowledge (r=.33, p=.052). The process of backtracking was not discovered to have exibited a significant relationship with any of the remaining seven clinical categories or with the mean score on the clinical checklist.

TABLE 8

Intercorrelations of Measures on Computerized Multiple Choice Paradigm and Clinical Evaluation Checklist

	Mean Score On Checklist	Investigation	Differential Diagnosis/ Problem List	History	Physical Exam	Case Reports	Knowledge	Oral Presentations	Communication Skills	Self Education	Attitude
Total Exam Score	.07	.42* p=.017	-.08	-.09	.05	.16	.14	.11	.04	-.05	.16
Backtracks	.16	-.04	.34* p=.046	.29	.35* p=.04	.05	.33* p=.052	.00	-.05	-.06	.23
Overall Mean Time Per Problem	-.27	-.10	-.21	-.21	-.06	-.07	-.17	-.27	-.30	-.35 p=.041	-.21
Mean Time Spent On Problems Answered Correctly	-.32* p=.054	-.09	-.34* p=.045	-.28	-.13	-.14	-.20	-.30	-.30	-.41* p=.010	-.30
Mean Time Spent On Problems Answered Incorrectly	-.14	.04	-.04	-.13	.02	.09	-.12	-.16	-.23	-.21	-.00

* Significant at $p<.05$

The overall mean time spent on the multiple choice questions significantly correlated with only one clinical category: self-education (r=-.35, p=.041). The negative correlation is indicative of an inverse relationship. It was also noted that self-education did not exhibit significant correlations with any other part of the computer paradigm except for mean time spent on problems answered correctly.

The similarity of performance on the two task environments was confirmed by the fact that significant negative correlations were uncovered between the computer measures of mean time spent on problems answered correctly and the mean score on the clinical checklist (r=-.32, p=.054), and mean time on problems answered correctly with differential diagnosis/problem list (r=-.34, p=.045) and finally mean time on problems answered correctly with self-education (r=-.41, p=.018).

In essence, the above would seem to indicate that low mean times on the computer-based multiple choice questions answered correctly would tend to serve as a precursor for high clinical scores on some categories.

Tables 9 to 16 serve as illustrations as to where subjects clustered in relationship to their scores on the computerized multiple choice exam and their clinical ratings. Cut-off points were arbitrarily determined in order to investigate the possible clustering of relationships between scores.

TABLE 9

Relationship Between Subjects' Performance on Computerized Multiple Choice Exam and Clinical Competence

Total Score on Computer Measures	Clinical Rating on Investigation High (3.5-4.0)	Medium (2.5-3.4)	Low (1-2.4)
High (10-14)	3	11	5
Medium (7-9)	1	3	3
Low (1-6)	0	0	0

Note: The numbers at each level represent the number of subjects falling into a particular category.

Table 9 indicates a lack of consistent uniformity in performance on the two task environments. Eleven subjects who had a high total score on the multiple choice questions (10-14) were evaluated as being average (i.e. medium) (2.5-3.4) when clinically rated on the investigation category. Only three subjects were judged superior for both settings, (10-14 on exam, 3.5-4.0 clinically). Three subjects fell into the medium category (7-9 on exam, 2.5-3.4 clinically). Five subjects received a low (1-2.4) clinical score but fell into the high (10-14) category on the exam. An additional three subjects graded low (1-2.4) clinically but scored in the medium range (7-9) on the exam. One subject received a high clinical score (3.5-4.0) but achieved a medium score (7-9) on the exam.

TABLE 10

Relationship Between Subjects' Performance on Computerized Multiple Choice Exam and Clinical Competence

Number of Problems Backtracked Upon	Clinical Rating on Differential Diagnosis/Problem List		
	High (3.5-4.0)	Medium (2.5-3.4)	Low (1-2.4)
High (10-15)	1	7	0
Medium (5-9)	3	7	4
Low (1-4)	0	4	0

Note: The numbers at each level represent the number of subjects falling into a particular category.

Table 10 charts the number of backtracks high (10-15), medium (5-9) and low (1-4) in comparison to the clinical evaluation (high, 3.5-4.0; medium, 2.5-3.4 and low, 1-2.4) on differential diagnosis/problem list. Seven high backtrackers were in the medium clinical range as were 7 medium backtrackers. Three medium backtrackers scored high clinically while 4 medium backtrackers achieved low clinical scores. Four low backtrackers achieved medium clinical scores while 1 subject received high scores in both settings.

TABLE 11

Relationship Between Subjects' Performance on Computerized Multiple Choice Exam and Clinical Competence

Number of Problems Backtracked Upon	Clinical Rating on Physical Exam		
	High (3.5-4.0)	Medium (2.5-3.4)	Low (1-2.4)
High (10-15)	1	7	0
Medium (5-9)	0	10	4
Low (1-4)	0	3	1

Note: The numbers at each level represent the number of subjects falling into a particular category.

Table 11 compares backtracking to the clinical rating on the physical exam category. The parameters vis-a-vis high, medium and low are identical to those of Table 10. One subject fell into the high level for both categories. Seven subjects were high on backtracks and medium clinically, while 10 subjects scored in the medium range for both categories. Four subjects were found to be medium backtrackers but were given low grades clinically. Three subjects were low backtrackers but achieved medium clinical performance while one subject was rated low in both environments.

Table 12 illustrates possible parallels between backtracking and clinical ratings on the knowledge category. The parameters (high, medium, low) are the same as those of Table 11.

TABLE 12

Relationship Between Subjects' Performance on Computerized Multiple Choice Exam and Clinical Competence

Number of Problems Backtracked Upon	Clinical Rating on Knowledge		
	High (3.5-4.0)	Medium (2.5-3.4)	Low (1-2.4)
High (10-15)	3	4	1
Medium (5-9)	2	7	5
Low (1-4)	1	1	2

Note: The numbers at each level represent the number of subjects falling into a particular category.

Three subjects achieved the high classification in both settings. Four high backtrackers received medium clinical grades while one high backtracker was judged to be low clinically. Two medium backtrackers received high clinical ratings while seven subjects scored in the medium range for both task environments. Five medium backtrackers were graded low clinically while one low backtracker received high clinical ratings and another low backtracker was judged to be medium clinically. Two low backtrackers received low clinical grades.

Table 13 focused on the possible relationship between the overall mean time (sec) spent on the computerized multiple choice questions (high, 80-115, medium, 55-79, and low, 40-54) and the clinical rating on self-education. The latter's parameters are identical to those of previous

tables.

TABLE 13

Relationship Between Subjects' Performance on Computerized Multiple Choice Exam and Clinical Competence

Overall Mean Time on Computer Measure (sec)	Clinical Rating on Self Education		
	High (3.5-4.0)	Medium (2.5-3.4)	Low (1-2.4)
High (80-115)	1	4	2
Medium (55-79)	3	12	1
Low (40-54)	1	2	0

Note: The numbers at each level represent the number of subjects falling into a particular category.

One subject was found to have a high mean time and a high clinical evaluation on self-education. Four subjects with high mean times received medium clinical ratings while two subjects with high mean times were rated low clinically. Three subjects in the medium mean time range received high clinical ratings but 12 subjects were found to be in the medium range for both tasks. One subject had a medium overall mean time but was given a low clinical rating. One subject had a low overall mean time and a high clinical rating. Finally, two subjects with low mean times on the multiple choice questions were given medium clinical scores.

Table 14 illustrates comparisons between the mean time (sec) spent on multiple choice questions answered correctly (high, 80-105, medium, 55-79, and low, 30-54) with the mean score attained on the clinical evaluation checklist. The clinical parameters are the same as previously described. Two subjects with high mean times had a medium mean clinical rating. One subject with a high mean time achieved a low mean clinical rating. Eleven subjects fell into the medium mean range for both settings. Two subjects with medium mean times had a low mean clinical rating. Eight subjects with low mean times had medium mean clinical evaluations while two subjects had low mean scores for both settings.

TABLE 14

Relationship Between Subjects' Performance on Computerized Multiple Choice Exam and Clinical Competence

Mean Time Spent on Problems Answered Correctly (sec)	Mean Score on Clinical Evaluation Checklist		
	High (3.5-4.0)	Medium (2.5-3.4)	Low (1.2.4
High (80-105)	0	2	1
Medium (55-79)	0	11	2
Low (30-54)	0	8	2

Note: The numbers at each level represent the number of subjects falling into a particular category.

Table 15 compared the mean time (sec) spent on questions answered correctly with the clinical evaluation on differential diagnosis/problem list. The standards set for high, medium and low are the same as those in Table 14.

TABLE 15

Relationship Between Subjects' Performance on Computerized Multiple Choice Exam and Clinical Competence

Mean Time (sec) Spent on Problems Answered Correctly	Clinical Rating on Differential Diagnosis Problem List		
	High (3.5-4.0)	Medium (2.5-3.4)	Low (1-2.4)
High (80-105)	0	2	1
Medium (55-79)	0	12	1
Low (30-54)	4	4	2

Note: The numbers at each level represent the number of subjects falling into a particular category.

Two subjects with high mean times were rated in the medium range for their clinical performance. One subject with a high mean time received a low clinical rating. Twelve subjects were in the medium range for both environments. One subject had a medium mean time and a low clinical rating. Four subjects with low mean times had high clinical ratings and four subjects with low mean times were judged to be medium clinically. Two subjects were designated as low in both settings.

Table 16 reported the possible similarities between the mean time (sec) spent on problems answered correctly and the clinical rating on self-education. The parameters vis-a-vis high, medium and low are identical to those in Table 15.

TABLE 16

Relationship Between Subjects' Performance on Computerized Multiple Choice Exam and Clinical Competence

	Clinical Rating on Self Education		
Mean Time (sec) Spent on Problems Answered Correctly	High (3.5-4.0)	Medium (2.5-3.4)	Low (1-2.4)
High (80-105)	0	2	1
Medium (55-79)	2	10	1
Low (30-54)	3	6	1

Note: The numbers at each level represent the number of subjects falling into a particular category.

Two subjects with high mean times were judged to be at a medium level clinically while one subject showing a high mean time was accorded a low clinical rating. Two subjects with medium mean times were given high clinical ratings but 10 subjects proved to be in the medium range for both settings. One subject with a medium mean time was given a low clinical evaluation. Three subjects with low mean times achieved high clinical ratings while six subjects with low mean times achieved a medium clinical rating for

self-education. One subject was in the low classification for both criteria.

Tables 8 through 16 indicate the existence of processing similarities between performance on an exam and clinical performance. There were exceptions, however there was a consistent pattern of high-medium and medium-medium pairings. The results constitute an indicator of the students' use of their knowledge to perform their required interaction with a patient. The medical students seem to be utilizing similar strategies and goal directed behavior in both environmental situations.

There existed a lesser amount but not insignificant pattern of low-medium classifications. The medium category may in fact constitute a cut-off point for superior performance as evidenced by the majority of medium-medium, medium-high and medium-low patterns. There were few high-high or low-low pairings.

Analysis of Data Obtained from Computerized Multiple Choice Exam Paradigm

Table 17 shows the significant intercorrelations between measures obtained on the computerized multiple choice question paradigm itself. The experimental score was not found to correlate with any other measure. However, all other measures correlated significantly with at least one other category.

TABLE 17

Intercorrelation of Computer Measures for Third Year Medical Students
Mean Time per Problem

Correlation Coefficients	Experimental Score	Number of Problems Backtracked Upon	All Problems	Problems Answered Correctly	Problems Answered Incorrectly	Total Reversals Over 18 Questions	Reversals Over Questions Answered Correctly	Reversals Over Questions Answered Incorrectly
Experimental Score	-	.07	.09	.21	.25	-.02	.15	-.08
Number of Problems Backtracked Upon	-	-	.18	.05	.348* p=.041	.960** p=.001	.885** p=.001	.865** p=.001
Overall Mean Time Per Problem (All Problems)	-	-	-	.919** p=.001	.844** p=.001	.18	.13	.18
X Time On Problems Answered Correctly	-	-	-	-	.636** p=.001	.04	.11	-.01
X Time On Problems Answered Incorrectly	-	-	-	-	-	.324* p=.053	.24	.360* p=.036
Total Reversals Over 18 Questions	-	-	-	-	-	-	.916** p=.001	.878** p=.001
Reversals Over Questions Answered Correctly	-	-	-	-	-	-	-	.636** p=.001
Reversals Over Questions Answered Incorrectly	-	-	-	-	-	-	-	-

* Significant at p<.05
** Significant p<.001

One correlation of interest would be the one between backtracking and time spent on problems answered incorrectly (r=.348, p<.05). It appears to be clear that lack of knowledge would require the subject to recheck previously given information. The latter is borne out by the significant correlation with mean time on problems answered incorrectly.

The number of problems backtracked upon was also significantly correlated with total reversals (i.e. the number of times per problem subjects sought to retrieve information over 18 questions) (r=.960, p<.001), with reversals over questions answered correctly, (r=.885, p<.001) and with reversals over questions answered incorrectly, (r=.865, p<.001).

Results indicate that subjects with superior knowledge would have no need to backtrack; this is confirmed by the lack of a significant correlation between the number of problems backtracked upon and mean times on problems answered correctly.

The significant correlations found to exist between mean time spent on problems answered correctly and mean time spent on problems answered incorrectly (r=.636 p<.001) were expected and assume minimal importance. The significant result constitutes an artifact because essentially, similar processes were being measured.

Despite the lack of a significant correlation between the number of problems backtracked upon and mean time spent on problems answered correctly, the medical students

exhibited definitive characteristics pertaining to the interaction between the two variables.

Table 18 reveals that over 50% of all subjects with high mean times on problems answered correctly were high backtrackers. In all cases, cut-off points are 50 seconds for mean time correct and 6 problems for degree of backtracking. Specifically, 14 subjects wre found to be high in both categories while one subject was deemed to be low in the two settings. Seven subjects with high mean times were low backtrackers while four subjects with low mean times were high backtrackers.

TABLE 18

Classification of Students According to Speed and Degree of Backtracking*

Frequencies		Mean Time Correct	
		Low	High
Number of Problems Backtracked	Low	1	7
	High	4	14

* In all cases, cut-off points are 50 seconds for mean time correct and 6 problems for degree of backtracking.

In essence, high mean times and high backtracking were conspicuous features of the group's performance. As illustrated in Table 18, 81% of all students had high mean

times while 69% of all students proved to be high backtrackers.

It was determined however that students spent significantly less time on problems answered correctly than on problems answered incorrectly. T-tests comparing the latter two variables were significant over subjects (t=-6.61, p<.05) and over problems (t=-2.78, p<.05) (Table 19).

TABLE 19

T Tests Comparing Mean Times For Problems Answered Correctly and Incorrectly on on Computerized Multiple Choice Question Model

	Group	
	Over Subjects	Over Problems
T value	-6.61*	-2.78*

* p<.05

Tables 20 through 22 divide the element of time into high, medium and low categories and show the interrelationships among levels.

Table 20 compares the overall mean time spent on all problems with mean time spent on problems answered correctly. Three subjects had high mean times for both categories while four subjects with high overall mean times

exhibited medium mean times on correct questions. One subject with a low mean time on correct questions had a high overall mean time. Nine subjects showed medium mean times for both categories while seven subjects with low mean times on correct problems had a medium overall mean time. Two subjects were attributed a low designation for both categories.

TABLE 20

The Relationship Between Time Spent on All Problems and Time Spent on Problems Answered Correctly on Computerized Multiple Choice Exam

Overall Mean Time (sec) per Problem	Mean Time on Problems Answered Correctly (sec)		
	High	Medium	Low
	(80-105)	(55-79)	(30-54)
High (80-115)	3	4	1
Medium (55-79)	0	9	7
Low (40-54)	0	0	2

Note: The numbers at each level correspond to the number of subjects falling into a particular category.

Table 21 compares the mean time spent on problems answered correctly with mean time spent on incorrect problems. Three subjects scored high mean times for both categories. Six subjects who fell into the medium range on questions answered correctly had high mean times on questions answered incorrectly. Seven subjects proved to be

in the medium mean time range for both classifications. Three subjects who were low on correct were high on incorrect while seven subjects with low mean times on correct questions achieved medium times on incorrect questions.

TABLE 21

The Relationship Between Time Spent on Problems Answered Correctly and Incorrectly on Computerized Multiple Choice Exam

Mean Time (sec) on Problems Answered Correctly	Mean Time (sec) on Problems Answered Incorrectly		
	High (80-140)	Medium (55-79)	Low (40-54)
High (80-105)	3	0	0
Medium (55-79)	6	7	0
Low (30-54)	3	7	0

Note: The numbers at each level correspond to the number of subjects falling into a particular category.

Table 22 charts the relationship between overall mean time per problem and mean time spent on problems answered incorrectly.

TABLE 22

The Relationship Between Time Spent on All Problems and Time Spent on Problems Answered Incorrectly on Computerized Multiple Choice Exam

Overall Mean Time (sec) per Problem	Mean Time (sec) on Problems Anwered Incorrectly		
	High (80-140)	Medium (55-79)	Low (40-54)
High (80-115)	7	0	0
Medium (55-79)	5	11	0
Low (40-54)	0	3	0

Note: The numbers at each level correspond to the number of subjects falling into a particular category.

Seven subjects exhibited high mean times for both taxonomies. Five subjects with medium overall mean times showed high mean times on problems answered incorrectly. Eleven subjects had medium mean times for both classifications while 3 subjects with low overall mean times spent a medium amount of time on problems answered incorrectly.

Tables 20 through 22 exhibit a large amount of individual variance vis-a-vis time spent on problems however globally, as confirmed by the t-tests, there was a significant difference between the amount of time spent on questions answered correctly versus questions responded to incorrectly. More time was spent on problems answered incorrectly.

Tables 23 through 25 illustrate the relationship between the number of problems in which a subject backtracked and the number of reversals (i.e. how many times per problem a subject went back to examine previously given information). Table 23 compares the number of problems backtracked upon with the total number of reversals over the 18 multiple choice questions. Seven high backtrackers had a high amount of reversals while one high backtracker exhibited a medium number of reversals. Ten subjects proved to be in the medium range for both categories but four medium backtrackers exhibited a low amount of reversals. Four subjects were rated low for both classifications.

TABLE 23

The Relationship Between Backtracking and Overall Reversals on the Computerized Multiple Choice Exam

Number of Problems Backtracked Upon	Total Number of Reversals Over 18 Questions		
	High	Medium	Low
	(15-21)	(8-14)	(1-7)
High (10-15)	7	1	0
Medium (5-9)	0	10	4
Low (1-4)	0	0	4

Note: The numbers at each level represent the number of subjects falling into a particular category.

Table 24 compares backtracking with reversals over questions answered correctly. Five subjects were judged high for both categories while three high backtrackers showed a medium amount of reversals. Five subjects fell into the medium range for both categories however nine medium backtrackers showed a low number of reversals. Four subjects were low for both criteria.

TABLE 24

The Relationship Between Backtracking and Reversals on Questions Answered Correctly on the Computerized Multiple Choice Exam

Number of Problems Backtracked Upon	Reversals Over Questions Answered Correctly		
	High (10-14)	Medium (5-9)	Low (1-4)
High (10-15)	5	3	0
Medium (5-9)	0	5	9
Low (1-4)	0	0	4

Note: The numbers at each level represent the number of subjects falling into a particular category.

Table 25 delineates the relationship between the number of problems backtracked upon with the number of reversals over questions answered incorrectly. One subject was designated into the high category for both criteria while three high backtrackers exhibited a medium number of

reversals. Four high backtrackers had a low number of reversals. One medium backtracker had a high number of reversals but five medium backtrackers scored in the medium range on reversals. Eight medium backtrackers were low on reversals and four subjects were low in both backtracking and reversals.

TABLE 25

The Relationship Between Backtracking and Reversals on Questions Answered Incorrectly on the Computerized Multiple Choice Exam

Number of Problems Backtracked Upon	Reversals Over Questions Answered Incorrectly		
	High (12-17)	Medium (6-11)	Low (0-5)
High (10-15)	1	3	4
Medium (5-9)	1	5	8
Low (1-4)	0	0	4

Note: The numbers at each level correspond to the number of subjects falling into a particular category.

Tables 23 through 25 exhibit considerable individual variance however the number of reversals appear to be a function of backtracking. The correlations confirm a specific link between the two variables. The medium range was the cut-off point for the establishment of pairings. Most of the latter were of the medium-low, medium-high, medium-medium type. There were some high-high pairings

however significantly there were very few high-low or low-low pairings. Subjects exhibited consistent strategies for the two variables. Scores did not tend to cluster on the extreme (low) end of the spectrum. Results indicate that the number of problems backtracked upon are closely related to the number of reversals. It should be noted however that the findings pertaining to the data in Tables 23, 24 and 25 were expected because the processes of backtracking and reversing are not independent.

Table 26 lists each individual subject's performance on all criteria pertaining to the computerized multiple choice questions. The experimental score on the 18 problems ranged from a high of 14 to a low of 7, the mean score was 11.15 and 12 was the mode. The overall time per problem ranged from a high of 114.42 to a low of 41.01 seconds with a mean of 69.87. The time spent on problems answered correctly ranged from a high of 103.78 seconds to a low of 32.59; the mean was 61.80. The time spent on problems answered incorrectly ranged from a high of 136.59 seconds to a low of 58.42; the mean was 83.60. The number of problems backtracked upon ranged from a high of 15 to a low of one; the mean was 8.23 and the mode was seven.

TABLE 26

Evaluation of Group's Responses on Computerized Multiple Choice Question Model

Subject	Experimental Score on 18 problems	Mean Time per Problem (sec) All Problems	Problems Answered Correctly	Problems Answered Incorrectly	Problems Backtracked Upon
1	12	82.73	76.5	95.1	13
2	10	82.30	61.9	107.83	7
3	14	80.47	68.83	121.20	15
4	12	73.55	70.91	78.84	5
5	14	77.88	78.41	76.01	12
6	11	86.20	74.99	103.80	11
7	12	62.63	53.33	81.22	8
8	11	58.72	55.54	63.71	2
9	13	41.01	32.59	62.92	9
10	8	55.40	51.43	58.58	9
11	11	58.30	53.63	65.62	4
12	12	62.15	60.37	65.71	11
13	7	73.84	51.68	87.94	15
14	12	98.55	100.17	95.29	6
15	8	53.89	48.22	58.42	9
16	12	58.14	47.62	79.18	5
17	9	61.41	61.89	60.93	7
18	9	62.21	35.55	88.87	9
19	12	51.96	43.19	68.24	7
20	13	62.24	58.36	72.30	2
21	9	114.42	103.78	125.05	5
22	10	57.04	51.24	68.05	7
23	13	73.80	67.01	91.43	1
24	13	63.99	56.88	82.44	13
25	14	94.66	82.68	136.59	15
26	9	69.24	60.22	78.25	7

Mean Score	Overall Mean Time Per Problem	Problems Answered Correctly	Problems Answered Incorrectly	Mean Number of Backtrack
11.15	69.87	61.80	83.60	8.23

CHAPTER V

DISCUSSION

It is uncertain from these results whether or not success on multiple choice exams would serve as a predictor of successful clinical performance by medical students (or for that matter vice versa). The significant correlations constitute an indicator of the validity of the examination. However a serious limitation to complete predictive validity was the fact that five clinical categories did not correlate significantly with any of the measures on the computerized model. Since there was a high degree of significant correlation among the clinical criteria (i.e. evidence of homogeneity and a halo effect), it remains problematical as to reasons for the lack of total significant correlations between measures taken in the two environments. The most puzzling obstacle was the lack of a significant correlation between the total score obtained on the multiple choice questions and the mean score obtained on the clinical checklist. This evident dichotomy precludes the formulation of the existence of a total relationship between the two task environments. Essentially, there exists a considerable degree of performance variance between hypothetical situations such as examinations and actual interactions in the medical context.

Nevertheless, the significant correlations function as evidence that components of the two sub-paradigms appear to be measuring aspects of similar processes. The correlation

between total exam score and the clinical rating on investigation would serve as an example. These latter two criteria involve the procedures of analysing existing data for the purpose of the construction of a solution (diagnosis) to the given problem.

The significant correlations found to be prevalent between the number of problems backtracked upon and the clinical grades obtained on differential diagnosis/problem list, physical exam and knowledge provide grist for theorists' postulating a link between the information processing model of human problem solving and medical problem solving. These measures appear to be compatible to the previously described findings of early hypothesis generation and the consideration of a limited number of hypotheses (between five and nine). The latter conforms to the limited capacity channel characteristics of short-term memory. The reason for the lack of a significant correlation between backtracking and the remaining clinical categories is unknown and would serve as a question for further investigation. The present findings have some consistency with the Elstein, Shulman and Sprafka (1978) finding of successful medical problem solving being a function of educational level. The relatively high amount of backtracking and high mean time spent on problems for the total group symbolizes the novice-type characteristics of the group who were in fact third year medical students undergoing their first experience in clinical medicine. They have not as yet acquired the efficient processing

strategies that have proven to be endemic to experienced physicians.

The above is borne out by the discovery of a negative correlation between the mean time spent on problems answered correctly and mean score obtained on the checklist and between mean time on correct problems with the clinical categories of differential diagnosis/problem list and self-education. Accordingly, more time spent on problems is indicative of a lack of knowledge, hence the lower scores on the three checklist categories. The anomaly appears to be the lack of significant correlations between mean time on correct problems and the remaining clinical classifications. Therefore complete predictive capability of the multiple choice exam cannot remain an unchallenged conclusion. However, within the computer paradigm itself the manifestation of higher mean times on problems answered incorrectly is further identification of a stable, invariant characteristic of medical inexperience.

Globally however, there was not a prominent disparity between individual performance in the two task environments. Individually, subjects performed similarly in both environments. There was a general overall similarity of behaviour in the two frameworks.

Presumably, efficiency will rise with increasing educational level as previous studies have shown. The descriptive process of clinical reasoning was expanded upon and linked to overt behavior by medical students. It must not be forgotten that the goal among medical teachers and

evaluators is to delineate whether or not the student is clinically competent. The latter must supersede the somewhat abstract inquiries of complex cognitive theories. It should be noted however that the latter theories form the basis for the medical problem solving framework and may in fact aid in identifying superior performance.

The two mechanisms of computer based multiple choice exams and clinical evaluations should not be antithetical to one another. There should not be an adversary relationship. Absolute conclusions vis-a-vis the efficacy of the two approaches should be avoided. Clearly, a computerized multiple choice exam can never function as the sole predictor of clinical competence. However, if similar processes appear to be in evidence in both contexts, then it would appear that both systems have good evaluative properties. Evaluation of medical students should be multidisciplinary in scope, the purpose of which would be for the identification of both strengths and weaknesses.

The discovery of specific processes that pertain to medical problem solving (high backtracking, high mean time spent on problems with lower clinical scores) could in fact alter the approaches employed to teach medicine. Teaching strategies must key in to the mode in which the individual thinks. The medical problem solving paradigm constitutes an attempt to codify these processes. The research project completed has illustrated however that clinical evaluative components must always serve as an adjunct to medical problem solving investigations. Aspects of the two

approaches do covary with one another.

REFERENCES

Barbee, R., Feldman, S., and Chosy, L. The quantitative evaluation of student performance in the medical interview. Journal of Medical Education, 1967, 42, 238-243.

Barrows, H.S., Norman, G.R., Neufeld, V.R., and Feightner, J.W. The Clinical Reasoning (Problem-Solving) Process of the Physician, 1981. Unpublished.

De Groot, A.D. Perception and memory versus thought: Some old ideas and recent findings. In B. Kleinmuntz (Ed.) Problem Solving: Research, Method and Theory. New York: John Wiley and Sons, 1966.

Elstein, A.S., Shulman, L.S., and Sprafka, S.A. Medical Problem Solving: An analysis of clinical reasoning. Cambridge, Mass.: Harvard University Press, 1978.

Ericsson, K.A. Instruction to verbalize as a means to study problem solving processes with the 8-puzzle: A preliminary study. Stockholm, Sweden: University of Stockholm, 1975.

Fleming, P.R., Manderson, W.G., Matthews, M.B., Sanderson, P.H., and Stokes, J.F. Evolution of an examination: MRCP (UK) British Medical Journal, 1974, 2, 99-107.

Greeno, J.G. Natures of problem-solving abilities. In W.K. Estes (Ed.) Handbook of Learning and Cognitive Processes, 1978, 5, 239-269.

Groen, G.J., Dauphinee, W.D., and McQueen, M.M. Time and Strategies on multiple choice questions. Montréal, Québec: McGill University, 1981, Unpublished.

Harden, R.M. How to Assess Clinical Competence - An Overview. Medical Teacher, 1979, 1, 289-296.

Helfer, R.E., and Hess, J.W. An experimental model for making objective measurements of interviewing skills. Journal of Clinical Psychology, 1970, 26, 327-331.

Hinz, C. Direct observation as a means of teaching and evaluating clinical skills. Journal of Medical Education, 1966, 41, 150-161.

Hubbard, J.P., Levit, E.J., Schumacher, C.G., and Schnabel, T.C. An objective evaluation of clinical competence. New England Journal of Medicine, 1965, 272, 1321-1328.

Lindsay, P.H., and Norman, D.A. Human Information Processing: An introduction to psychology. New York, N.Y.: Academic Press, 1972.

Matarazzo, R.G., Phillips, J.S., Wiens, A.N., and Saslow, G. Learning the art of interviewing: A study of what beginning medical students do and their patterns of change. Psychotherapy, 1965, 2, 40-50.

McCaffrey, C. Performance checklists: An effective method of teaching, learning and evaluating. Nurse Educator, 1978, 3, 11-13.

McGaghie, W.C. Medical Problem Solving: A Reanalysis. Journal of Medical Education, 1980, 55, 912-921.

McQueen, M.M. Expert and novice medical problem solving: Medical student and resident performance on multiple choice questions. Montréal, Qué.: McGill University, 1981 Unpublished.

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

Neufeld, V.R., Norman, G.R., Feightner, J.W., and Barrows, H.S. Clinical Problem-Solving by Medical Students: A Cross-Sectional and longitudinal analysis. Medical Education, 1981, 15, 315-322.

Neufeld, V.R., Norman, G.R., Feightner, J.W., and Barrows, H.S. Clinical Methods of medical students: A cross-sectional analysis. In the Proceedings of the annual conference on Research in Medical Education. Washington, D.C.: 1975, 149-153.

Newble, D.I., Elmslie, R.G., and Baxter, A. A problem-based criterion-referenced examination of clinical competence. Journal of Medical Education, 1978, 53, 720-726.

Newell, A. On the analysis of human problem solving protocols. In P.N. Johnson-Laird and P.C. Wason (Eds.), Thinking: Readings in cognitive science. New York, N.Y.: Cambridge University Press, 1977.

Newell, A., and Simon, H.A. Human Problem Solving. Englewood Cliffs, N.J.: Prentice Hall, 1972.

Norman, G.R., and Feightner, J.W. A comparison of behavior on simulated patients and patient management problems. Medical Education, 1981, 15, 26-32.

Norman, G.R., Jacoby, L.L., Feightner, J.W. and Campbell, E.J.M. Clinical Experience and the Structure of Memory. Proceedings ofthe Eighteenth Annual Conference on Reserach in Medical Education, Association of American Medical Colleges (AAMC), Washington, November 6-7, 1979, 21-25.

O'Donohue, W.J. Jr., and Wergin, J.F. Evaluation of Medical Students during a clinical clerkship in internal medicine. Journal of Medical Education, 1978, 53, 55-58.

Simon, H.A. Information-Processing Theory of Human Problem Solving, in W.K. Estes (Ed.), Handbook of Learning and Cognitive Processes, 1978, 5, 271-295.

Simon, H.A., and Gilmartin, K. A simulation of memory for chess positions. Cognitive Psychology, 1973, 5, 29-46.

Stillman, P.L., M.D. Arizona Clinical Interview Medical Rating Scale. Medical Teacher, 1980, 2, 248-251.

Stillman, P., Sabers, D., and Redfield, D. The use of paraprofessionals to teach interviewing skills. Pediatrics, 1976, 57, 769-774.

PHASE II

CLINICAL EVALUATION FORM

DEPARTMENT OF MEDICINE

McGILL UNIVERSITY MEDICAL SCHOOL

STUDENT NAME: _____

HOSPITAL: _____

EVALUATOR: _____

☐ ATTENDING STAFF ☐ HOUSESTAFF

REPORT COVERS: FROM _____ TO _____ 198__

HISTORY

☐ NOT OBSERVED	☐	☐	☐	☐
	INCOMPLETE OR INACCURATE, OMITS ESSENTIAL POINTS; CLUTTERS HISTORY WITH SOME IRRELEVANT NEGATIVES, AND/OR OMITS RELEVANT ONES	COMPLETE AND ACCURATE, BUT OCCASIONALLY MISSING IMPORTANT INFORMATION; SOME IRRELEVANT AND UNNECESSARY NEGATIVES	COMPLETE AND ACCURATE; IMPORTANT INFORMATION IS INCLUDED; FEW MINOR OMISSIONS	COMPREHENSIVE; INFORMATION THOROUGH AND PRECISE INCLUDING ALL RELEVANT NEGATIVES, SYSTEMS REVIEW, PAST HISTORY AND FAMILY HISTORY

COMMENTS _____

PHYSICAL EXAMINATION

☐ NOT OBSERVED	☐	☐	☐	☐
	INCOMPLETE EXAM; FAILS TO FOLLOW UP OBVIOUS LEADS; EMPHASIZES MINOR FINDINGS; MAJOR DEFICIENCIES IN TECHNICAL QUALITY (CLUMSY OR USES INSTRUMENTS INCORRECTLY)	GENERALLY COMPLETE, OCCASIONALLY FAILS TO FOLLOW UP ON OBVIOUS LEAD; NEEDS TO IMPROVE SPEED AND ACCURACY	THOROUGH EXAMINATION; ALL IMPORTANT AREAS FOLLOWED UP; TECHNICALLY SOUND; OBTAINS ACCURATE DATA	THOROUGH EXAMINATION; GATHERS DETAILED INFORMATION NECESSARY TO OBTAIN DIFFERENTIAL DIAGNOSIS; PERFORMS TECHNICALLY SOUND, EFFICIENT EXAMINATION EVEN WITH DIFFICULT PROBLEM; OBTAINS ACCURATE DATA.

COMMENTS _____

DIFFERENTIAL DIAGNOSIS/PROBLEM LIST

☐ NOT OBSERVED	☐	☐	☐	☐
	FREQUENT DIFFICULTY USING DATA TO FORMULATE PROBLEM LIST, SOME SIGNIFICANT PROBLEMS NOT INCLUDED; NON-EXISTENT PROBLEMS MAY BE INCLUDED AND NOT RELATED TO CUES IN DATA BASE; USUALLY FAILS TO PROVIDE AN ASSESSMENT OR EXPLANATION OF THE PROBLEM, OR FREQUENTLY GIVES ERRONEOUS AND UNJUSTIFIED ANALYSIS	OCCASIONAL DIFFICULTY USING AVAILABLE DATA TO FORMULATE PROBLEM LIST, MOST, BUT NOT ALL THE PROBLEMS ARE INCLUDED; SOME FORMULATED PROBLEMS/DIAGNOSES NOT ENTIRELY APPROPRIATE FOR THE CUES IN THE DATA BASE; PROVIDES AN ANALYSIS FOR SOME BUT NOT ALL PROBLEMS; OR GIVES ONLY PARTIALLY JUSTIFIED AND INCOMPLETE ASSESSMENT OF THE PROBLEMS	EVALUATES AVAILABLE DATA TO FORMULATE PROBLEM LIST WHICH IS USUALLY COMPLETE WITH GOOD ASSESSMENT OF PRIORITIES	EFFICIENTLY ANALYZES AVAILABLE DATA; SYNTHESIZES INFORMATION WELL TO ARRIVE AT A CONCISE SUBSTANTIVE DIFFERENTIAL DIAGNOSIS/PROBLEM LIST WITH GOOD REASONABLE ALTERNATIVE EXPLANATIONS FOR MOST PROBLEMS

COMMENTS: _____

INVESTIGATION

☐ NOT OBSERVED	☐	☐	☐	☐
	PLAN FOR DIAGNOSTIC TESTS IS INCOMPLETE OR INEFFICIENT; IMPORTANT TESTS FREQUENTLY OVERLOOKED; DIFFICULTY INTERPRETING TEST RESULTS	PLAN FOR DIAGNOSTIC TESTS SOMEWHAT INCOMPLETE OR INEFFICIENT; OCCASIONALLY IMPORTANT TESTS OVERLOOKED; SOMETIMES HAS DIFFICULTY INTERPRETING TEST RESULTS	PLAN FOR DIAGNOSTIC TESTS COMPLETE AND EFFICIENT; IMPORTANT TESTS ARE INCLUDED; INTERPRETS RESULTS CORRECTLY	PLAN FOR DIAGNOSTIC TESTS COMPLETE AND MAXIMIZES INFORMATION GAIN; EFFICIENTLY PLANS ALTERNATIVE DIAGNOSTIC PROCEDURES, INTERPRETS RESULTS CORRECTLY

COMMENTS: _____

CASE REPORTS

☐ NOT OBSERVED	☐	☐	☐	☐
	WRITE UPS POOR; IMPORTANT INFORMATION MISSING OR IRRELEVANT MATERIAL INCLUDED; DISCUSSIONS OVERLY LONG AND NOT RELEVANT	CASE REPORTS COULD BE IMPROVED; IRRELEVANT MATERIALS SOMETIMES INCLUDED, OR IMPORTANT DATA SOMETIMES MISSING; OCCASIONALLY LATE COMPLETING REPORTS, USUALLY LEGIBLE WITH ATTEMPT AT VISUAL ORGANIZATION; DISCUSSION SOMETIMES TO LENGTHY OR IRRELEVANT	CASE REPORTS GOOD; REPORTS USUALLY COMPLETE AND RELEVANT; ALL IMPORTANT PROBLEMS LISTED; COMPLETED PROMPTLY; LEGIBLE WITH GOOD VISUAL ORGANIZATION; DISCUSSION RELEVANT TO PATIENT'S PROBLEMS	CASE REPORTS OUTSTANDING; NOTES ARE CONCISE, THOROUGH AND RELEVANT, LEGIBLE AND VISUALLY WELL ORGANIZED USING DIAGRAMS WHERE APPROPRIATE; ALL IMPORTANT PROBLEMS INCLUDED AND PRIORITIES ESTABLISHED FOR EACH DISCUSSION IS RELEVANT TO PATIENT'S PROBLEMS WITH APPROPRIATE REFERENCES

COMMENTS _____

KNOWLEDGE

☐ NOT OBSERVED	☐ HAS DIFFICULTY RECALLING BASIC SCIENCE AND CLINICAL INFORMATION AND RELATING IT TO CASES	☐ OCCASIONALLY HAS MINOR DIFFICULTY RELATING BASIC SCIENCE AND CLINICAL INFORMATION TO CASES	☐ ABLE TO RELATE BASIC SCIENCE AND CLINICAL INFORMATION TO CASES	☐ APPLIES BROAD BASE OF PERTINENT BASIC SCIENCE AND CLINICAL INFORMATION TO CASES

COMMENTS _____

ORAL PRESENTATIONS

☐ NOT OBSERVED	☐ CASE PRESENTATIONS DISORGANIZED, POORLY INTEGRATED AND INCOMPLETE	☐ CASE PRESENTATIONS USUALLY ORGANIZED, BUT VERBOSE WITH MINOR OMISSIONS	☐ CASE PRESENTATIONS ORGANIZED AND COMPLETE	☐ CASE PRESENTATIONS COMPLETE, ORDERLY AND POLISHED

COMMENTS _____

COMMUNICATION SKILLS

☐ NOT OBSERVED	☐ GENERALLY LACKS COMMUNICATION SKILLS; DOES NOT EXPLAIN THINGS WELL TO PATIENTS; OFTEN FAILS TO LISTEN; APPEARS ILL AT EASE WITH PATIENTS	☐ TRIES TO COMMUNICATE AND EXPLAIN PROBLEMS, BUT ATTEMPTS ARE OFTEN SUPERFICIAL	☐ COMMUNICATES WELL - AT EASE WITH PATIENTS, LISTENS AND EXPLAINS WELL	☐ SENSITIVE TO PATIENT NEEDS IN ALL SITUATIONS; PUTS THE PATIENT FULLY AT EASE; IS CONCERNED AND FUNCTIONS WELL EVEN WITH DIFFICULT PATIENTS

COMMENTS _____

SELF EDUCATION

☐ NOT OBSERVED	☐ FAILS TO DEMONSTRATE KNOWLEDGE OF REQUIRED READING; SELDOM PARTICIPATES IN SESSIONS	☐ DEMONSTRATES KNOWLEDGE OF REQUIRED READING; PARTICIPATES OCCASIONALLY IN SESSIONS	☐ DEMONSTRATES KNOWLEDGE OF SUPPLEMENTARY, AS WELL AS REQUIRED READING; PARTICIPATES WELL IN GROUP SESSIONS	☐ INTELLECTUALLY AGRESSIVE; ACTIVELY PARTICIPATES IN SESSIONS; DEMONSTRATES KNOWLEDGE OF SIGNIFICANT READING

COMMENTS _____

ATTITUDINAL CHARACTERISTICS

☐ NOT OBSERVED	☐ APPEARS UNINTERESTED; LACKS MOTIVATION; FREQUENTLY LATE OR MISSES SEVERAL SESSIONS	☐ NEEDS PROMPTING TO COMPLETE ASSIGNMENTS; ATTENDS MOST SESSIONS; REASONABLE MOTIVATION	☐ VERY ENTHUSIASTIC; WELL MOTIVATED; ATTENDS ALL SESSIONS; TAKES A LEADERSHIP ROLE IN GROUP SESSIONS; EFFICIENT; ALL ASSIGNMENTS COMPLETED ON TIME	

COMMENTS _____

COMMENTS: