UNIVERSITY OF CALGARY

Scheme Utilization by Medical Students and Experts: from Theoretical to Clinical Advantages

by

Dr. Sylvain Paul Coderre

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF MEDICAL SCIENCE

CALGARY, ALBERTA

JANUARY, 2001

ABSTRACT

Cognitive psychology research supports the notion that experts use mental frameworks or "schemes", both to organize knowledge in memory and for resolving clinical problems. The two main research questions to be addressed are as follows. Do students who utilize "schemes" for learning and solving clinical problems perform better in examinations than students who use other strategies? Do the standard five-option Multiple-choice, Extended-matching and Elimination-type questions elicit a measurable difference in the utilization of "schemes"? Think-aloud protocols were collected to determine the problem solving strategy used by experts and non-experts in answering Gastroenterology questions. The results indicate a clear correlation between problem solving strategy, and specifically "scheme" utilization, and examination success. The three examination formats did not differ in eliciting "scheme" utilization. These results support the strength of teaching "schemes" as an educational strategy to be incorporated in the University of Calgary "Clinical Presentation" curriculum.

ACKNOWLEDGEMENTS

I would like to extend my sincere gratitude to the following individuals, who all played a very important role in the completion of this thesis.

1) Dr Henry Mandin, my thesis supervisor, and Drs Peter Harasym, J.G. DesCôteaux, Lloyd Sutherland, the members of my thesis committee. I thank them for their long hours spent carefully reviewing and providing constructive criticism of the thesis, as well as their constant support and positive reinforcement throughout these three years.
2) Dr. John Baumber, head of the medical education program, for his guidance and advice during my years in the program.
3) Dr. Ron Bridges and my colleagues at the Peter Lougheed Centre, for bringing me to Calgary and supporting my time away from clinical practice to complete this thesis.
4) Dr. Keith McCannell, Department of Gastroenterology, and Dr. Eldon Shaffer, Department of Medicine, for their active involvement in my career development and their support of my leave of absence required to complete this thesis.
5) Drs. Henry Mandin and Allan Jones, for serving as role models for a career in education.
6) The UME office, and specifically Bev Duregon and Shirley Marsh, for their invaluable assistance and patience with me, in putting this thesis together.
7) The Medical Council of Canada, the Calgary Gastroenterology Research and Education Foundation, and the UCMG, for their financial support of my study, my Masters program and leave of absence, respectively.
8) The 44 study participants, students and Gastroenterologists in Calgary, who were integral to all aspects of this thesis.

DEDICATION

I would like to dedicate my work to my wife Heidi. and my two sons Paul and Marcus. for giving me the love. motivation and support so critical to the completion of this lengthy project.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE: INTRODUCTION

The University of Calgary embarked on a process of curricular renewal between 1988 and 1991, at which point the "Clinical Presentation" curriculum was introduced [1]. This pedagogue encourages the use of "schemes" that provide learners with both frameworks upon which knowledge can be built, as well as a starting point for approaching and resolving clinical problems. The use of schemes is consistent with theories of how medical experts store knowledge, i.e., in the form of "semantic networks" [2]. Schemes, as used in the Clinical Presentation curriculum, divide diagnoses into large groups (or "small worlds" [3]), which are further subdivided into progressively smaller categories[4]. Examples of such schemes for the clinical presentation of dysphagia and acute renal failure are presented in Appendix A and B, respectively.

The main research question to be addressed concerns the utilization of schemes by medical students and experts. Do students who utilize schemes for learning and solving clinical problems perform better in examinations than students who use other strategies? Do the standard five-option Multiple-choice, Extended-matching and Elimination-type questions elicit a measurable difference in the utilization of schemes? Before proceeding, definitions and explanations of the relevant terminology used require elaboration. As well, some of the background for the project, which specifically focuses on the area of clinical reasoning in medicine, needs to be amplified. Although schemes are central to the research question, the goal of medical education is learning and the demonstration of behavioral changes by students of having acquired cognitive, psychomotor and professional behavior skills that ultimately are expected of competent physicians. For the

students to reach that goal. they must have the skills defined. experiences provided that promote acquisitions of those skills. and then be assessed by evaluation tools that provide feedback that they have accomplished their objectives. Consistent with this notion that the pillars of education are curriculum. learning experiences and evaluation [5]. this dissertation will detail a curriculum based on clinical presentations. learning experiences founded on organizational schemes. and evaluation tools consistent with both curriculum and learning experiences.

A) Schemes: Definition and Their Integration into the Clinical Presentations Curriculum

1) Definition of a "scheme"

Clinical reasoning in medicine is the central area of focus of this paper. and specifically. the concept of scheme utilization by medical students and experts in clinical reasoning. What is the meaning of "scheme"?

The term "schemata" has been used over many decades in cognitive psychology. dating back to the 1920s and Piaget's landmark writings. Given the longevity of the term. it is not surprising that over time. the usage and definitions of the term have proliferated. Fundamentally. schemata. as viewed by West [6] can be divided into two broad categories:

i) Data *schemata:* used for data storage and retrieval in "bundles" of information.

ii) *Process schemata:* used in active procedures of processing and organizing information.

Data schemata represent the knowledge "of or about" a certain disease or syndrome. while process schemata represent "how" to use data schemata in executing a

clinical task. The concept of information bundling present in the above definition of data schemata is an important one, as it allows for efficient storage of information in memory. Bundling of knowledge into memory can be equated to the technique of information chunking. Chunking is an important process schemata, used increasingly with expertise, which can serve the dual purpose of efficient storage of knowledge into memory in discrete packets, while also neatly separating categories of diagnoses for active problem solving.

The term "scheme" or "schemes" is used in this thesis as the anglicized form of "schema" or "schemata", and is an attempt to put into writing the cognitive processes involved in information bundling (data schemata) and diagnostic chunking (process schemata). The "scheme" is written in essence to recreate, on paper, the major divisions (or chunks) used by clinicians, typically of increasing expertise, to both store their knowledge into memory and solve problems. Therefore, the scheme incorporates some elements of both data and process, and can exert two potential functions, which are first to organize information (data schemata) and second to process information (process schemata). The second function of schemes, used in the clinical setting to solve problems is perhaps the function that is first thought of when discussing the utility of these schemes. When faced with a clinical presentation (defined as the "common and important ways in which a person, group of people, community or population present to a physician") [1] such as chest pain, for example, a medical problem solver may use a scheme or process schemata to arrive at a solution for the problem. In this sense, a scheme is analogous to an "approach" or "algorithm". With expertise, schemes become

very tightly organized and efficiently sorted through by key medical findings. a point that will be revisited later (an example of a scheme for the clinical presentation 'dysphagia' is provided in Appendix A). It is important to note the evidence that shows that for schemes to be used in solving clinical problems. they must be an integral part of the learning process. Unless schemes have been previously utilized as organizing scaffolding for information prior to learning. students resist using them for problem solving[7]. A summary definition of schemes. which incorporates their two potential functions of information organization and problem solving. is found in a paper by Mandin et al [4]: " the term scheme is a mental categorization of knowledge that includes a particular organized way of understanding and responding to a complex situation".

This distinction between medical knowledge organization in memory and practical diagnostic problem solving has also been made by Bordage [8]. and has important implications for education. Schemes are organizational structures to aid representation of knowledge in memory. and are also processing structures to solve problems. These functions are more likely used in conjunction with one another. especially in expert problem solvers (the importance of scheme use in expert problem solving will be re-visited later). If the problem solver is a non-expert. the possibility exists that use of schemes is made in aid of knowledge organization. but not necessarily in practice. to solve clinical problems. The importance of this dichotomy lies in undergraduate education. where the argument can be made that the primary purpose of encouraging the use of a scheme is not to turn these students into expert problem solvers. but rather to help them solidify and organize knowledge in memory.

II) Overview of the Clinical Presentations Curriculum

In the last section, we defined the term scheme, which is a central feature of the main research question in this thesis. However, schemes remain an educational tool, which are utilized for learning the Clinical Presentation curriculum, which we have adopted in Calgary. Therefore, before proceeding any further into exploring the main research question of scheme utilization in experts and non-experts, it is important to discuss the Clinical Presentations curriculum, and the cognitive psychology principles that underlie this curriculum.

The curriculum at the University of Calgary was organized around 125 +/- 5 clinical presentations, a number felt to represent the finite totality of medical disease manifestations [1,4]. Other medical schools have in fact adopted a similar curriculum [9,10]. The strongest arguments for organizing the curriculum in this manner originated from ideas promoted by LaDuca. The core of professionals' practical knowledge constitutes their capacity to identify a job to be done, to know how to go about doing it, and to recognize when it has been done appropriately. LaDuca promoted the assembly of knowledge databases that rely on descriptions of professionals' activity, with emphasis on performance in the context of professional situations [11]. In other words, curricular content should be derived from an analysis of the situations deemed to comprise the profession's domain. Thereafter, the knowledge essential for safe practice needs to be defined by analysis of the professional situations constituting the practice model. In essence, this is how the Clinical Presentation curriculum was constructed.

In the construction of this curriculum, the writing of curricular objectives by faculty experts uncovered the manner in which the knowledge was organized in the memory of these experts, which in turn led to the identification of schemes. A scheme was attached to virtually all clinical presentations. If a scheme did not spontaneously become evident as the organizing scaffolding for the written objectives, faculty were requested to construct a scheme. What was provided served as starting points around which students were to organize their knowledge related to clinical presentations. These schemes were only starting points, since students received strong recommendations to modify or personalize these schemes. These modifications are in keeping with the modes of learning which West[6] calls assimilation or tuning. Students can certainly create their own schemes, however there are some dangers associated with this. Firstly, it is more difficult for a non-expert to distinguish truly key items leading to simplified schemes: this distinction of crucial features comes with expertise. Secondly, if students create faulty schemes, then subsequent modifications of these erroneous schemes will not be as straightforward as assimilation or tuning, but would involve the accommodation mode of learning, a much more significant and challenging rearrangement of knowledge. The natural fit between the Clinical Presentations curriculum and the educational tool of schemes is one reason for its adoption. The strength of schemes as an instructional strategy will be discussed further in the next section (section B). Another important reason for the Clinical Presentation curriculum is that it adheres to other important cognitive psychology principles, which will now be presented.

III) The Cognitive Psychology Underlying the Clinical Presentation Curriculum

The evolution of curricula in recent years, and specifically the advent of Problem-Based Learning (PBL), has occurred in response to a number of developments in the cognitive psychology field. PBL was in fact founded on essentially five sound principles of cognitive psychology[12], which are important in any curriculum and certainly are incorporated into our Clinical Presentation curriculum:

i) Knowledge is structured into semantic networks

ii) Learning is improved when prior knowledge is activated.

iii) Elaboration of study material enhances memory retrieval.

iv) Context specificity enhances memory retrieval.

v) Intrinsic motivation improves achievement.

These principles essentially are aimed at improving the organization, storage and retrieval of information in human memory. This is important given the evolution of cognitive psychology away from seeing expertise as the refinement of a generic problem solving technique (a movement promoted by PBL), towards seeing expertise as highly knowledge dependant, and thus relying on the development of the "expert memory". Thus, becoming an expert requires the acquisition and organization of knowledge in an area, and not acquisition of a general problem solving process. Support for this is provided in part by the concept of case specificity [13] that sees problem solving as highly case-dependent. Success in solving one problem, even in the context of a specific specialty, does not guarantee success in another case, given that successful resolution of a problem is knowledge-dependent. Having said this, no study has been able to show a

positive correlation between knowledge and the ability to solve a particular problem[14,15]. These papers suggest that undefined factors other than quantity of knowledge may be involved in the phenomenon of case specificity, and perhaps this is knowledge organization or mastery. Therefore, the principles of cognitive psychology discussed in the next pages will focus on memory, with semantic networks being the way we organize our memory, activating prior knowledge helping with storage into memory, while elaboration of information and context specificity aid in retrieval of information from memory.

IV) Organization of Memory into Semantic Networks

The first important concept from cognitive psychology is that of the semantic network, defined as "meaningful sets of connections among abstract concepts and/or specific experiences" [2].This network is how we organize our knowledge into long-term memory. This concept is similar to the previous discussion surrounding the definition of schemes, with the addition of more personal experiences to the concept. In defining schemes, the process schemata of chunking was introduced, a technique by which experts organize their memory into information bundles, which is useful for both structuring data as well as problem solving. The chunking of large amounts of information essentially forms the scaffolding of the semantic networks, onto which specific details of information (i.e. data schemata) or experiences can be added. The evolution of these networks from their chaotic nature in novices to their highly structured form in experts relies on the addition of information and experiences with significant meaning to that individual. Therefore, each individual's semantic network around a given topic is unique.

varied in its structure and organization depending on level of expertise, and deeply dependent on a person's experiences, opinions and ideas. Increasing experience fine-tunes these networks, which in turn have been demonstrated to be stronger in the more accurate diagnosticians. [7,16,17]

The key implication of this concept is that for a medical student to start forming these networks, subject matter must be meaningful, and thus relate directly to patients. Subjects such as basic sciences taught in isolation of patient cases, as done in some traditional curricula may not have meaning to a student and therefore have the potential to be unhelpful and forgotten by their inability to enter a semantic network. Therefore, the basic sciences should be incorporated into problem solving of real cases in order for them to be both more effectively remembered, and to enhance the comprehension of clinical medicine. A Clinical Presentations curriculum, by putting the main focus onto real patients and how they actually present to physicians, creates intrinsic meaning into the program and therefore satisfies the principles behind the semantic network concept.

V) Information Storage into Memory: the Activation of Prior Knowledge

The second cognitive psychology concept is that of activation of prior knowledge, which is important in determining storage of information into memory. It is well described in the educational literature that the superior ability of some people to remember texts is likely on the basis of activation of specific pre-existing structures in that individual's memory. For written texts, comprehension is hypothesized to involve interactions between prior knowledge and the information in the text [18,19]. Said otherwise, this concept views learning not as the filling of an "empty cup", but rather that learning

involves active thinking and constant restructuring of information. This restructuring occurs on a foundation of prior knowledge, which in essence predetermines the amount of new information that can be learned. This is a similar concept to West's[6] assimilation and tuning modes of learning, which are the easiest modes of learning in that they involve only minor restructuring of existing information. Activating of prior knowledge has been shown to improve medical student recall of information[20]. A Clinical Presentation curriculum, which structures a mixture of teaching formats, with both lectures covering key concepts and subsequent small groups aimed at discussing these concepts in the context of real cases, fulfills the psychological requirement of activating prior knowledge to improve information storage.

VI) Information Retrieval: Elaboration of Knowledge and Encoding Specificity

The next two concepts, elaboration of knowledge and encoding specificity, both aim to improve retrieval from memory. Adding meaning, as previously discussed, is the other, perhaps most important component of improving retrieval from memory[2]. Elaboration of knowledge is defined [21] as " the process of adding to the information being learned", or "adding related knowledge to the new knowledge". It implies repetition of information, and can be accomplished through writing notes, teaching peers and group discussion. It has the potential to promote redundancy in memory retrieval pathways.

Encoding specificity also promotes retrieval from memory. The theory is that human memory works best when the conditions for retrieval from memory match the conditions for encoding. Encoding specificity has two components. The first is context specificity, which states that retrieval is improved if the subject is taught in the same

context as it will be subsequently used. The second is process specificity, which states that how one stores information will be important for how it will be retrieved. Therefore, with the small group teaching found in a Clinical Presentation curriculum, elaboration of knowledge can take place, with encoding specificity being satisfied by teaching around meaningful real clinical presentations and cases. As stated in Regehr's paper in cognitive psychology, forgetting is more likely a failure of retrieval of information rather than decay in the information trace in memory [2]. If forgetting is due to a failure of retrieval, or an inaccessibility of information, then it stands to reason that measures aimed at the improvement of retrieval will be beneficial to promoting long-term memory, or learning.

VII) The Importance of Creating Interest in Education

The last principle relates to interest or epistemic curiosity (i.e. intrinsic motivation). It is intuitively evident that the more a person is interested in the educational process, the more time will be invested in the process and hence in the end more knowledge should be retained in memory. This is a principle that is pervasive in the adult education literature [22,23]. Adherence to the cognitive psychology principles outlined above should also lead to increased epistemic curiosity [24]. The Clinical Presentation curriculum has adhered to the above principles and hence has been well received by the University of Calgary students. The students have shown a favorable response to schemes, as well as decreased stress levels and improved, balanced workload over the previous curriculum.[25,26]

Therefore, the University of Calgary had adopted the Clinical Presentation curriculum for a number of reasons, including, as previously outlined, the pivotal work

by DaLuca, as well as its adherence to the key cognitive psychology principles on which PBL was founded. The curriculum also fits well with usage of instructional strategies such as teaching schemes for knowledge organization and problem solving. Therein lies a major difference between PBL and the Clinical Presentations curriculum. The PBL founders assumed a generic problem solving process, which is not supported by cognitive psychology research, in that problem solving has been demonstrated to be heavily content-specific[27]. Furthermore, the generic problem solving process fostered by PBL is hypothetico-deductive reasoning [27,28], which the literature does not support as the expert problem solving method. The Clinical Presentation curriculum fosters schemes or forward reasoning strategies, which are the strategies used by experts in problem solving. The literature supporting this latter view will now be presented in the following section.

B) The Strength of Using Schemes in Knowledge Structuring and Problem Solving

The preceding has provided the basic definition of a scheme, as well as presented the rationale for the Clinical Presentations curriculum, and suggested the incorporation of teaching schemes into the curriculum as a potentially powerful instructional method. From this, is it now possible to make a case that schemes are a strong method of problem solving, that experts use schemes, and thus there is ample justification for the presentation of schemes as powerful educational tools? The search for such answers necessitates that the evolution of theories in expert problem solving be traced, starting with the hypothetico-deductive theory (backward reasoning), to forward reasoning, schemes and "small worlds", and finally pattern recognition.

I) Expert Problem Solving: Hypothetico-deductive Reasoning

Initially, problem solving by expert physicians was assumed to proceed in a to-and-fro fashion known as hypothetico-deductive reasoning[29] or "search-and-scan" problem solving (see Appendix B for examples of the three major reasoning strategies). In this method, a hypothesis, defined as a "verbal statement about a situation that may be either true or false", [30] is tested and modified based on outcome of various tests including the history and physical examination. This method was long considered the sole method for solving problems. The initial notion in favor of this approach was the fact that hypothetico-deductive reasoning was the major process used by the scientific community for basic research. From this, the method became known as the "scientific method". From its infancy with the famous mathematician Polya and psychologist Piaget, the theory was applied to medical education through the work, amongst others, of Elstein et al [29]. They demonstrated that physicians, very early in patient encounters, generate hypotheses that lead to the subsequent generation of data. Kassirer also supported this notion of problem solving in an article describing a teaching methodology[31] aimed at practicing the hypotheses-driven method of problem solving. As he describes it, this teaching method is based on the method "that physicians actually employ in practice". In this article, Kassirer describes the hypothetico-deductive method as a concept supported extensively by research in cognitive psychology, essentially citing Newell and Simon's work (1972) done in the context of artificial intelligence. Other work in this area, done by Neufeld et al (1981) [32] describes the diagnostic hypotheses advanced early in the patient encounter as

a "central feature of clinical problem-solving", not only of physicians but also of medical students.

II) The Challenge to the Hypothetico-deductive Model of Reasoning

More recently, cognitive psychologists have questioned the assumption that this "scientific method" of clinical reasoning is the only strategy used by medical experts to solve problems. One of the first and most important studies questioning the assumptions surrounding hypothetico-deductive reasoning was the work of Chase and Simon who repeated and expanded upon previous work by De Groot. They principally examined expert chess players' memory for chess positions, as compared to novices [33].Chase and Simon's work on chess experts led to conclusions principally on knowledge representation by experts. They found that experts used more complex knowledge structure than novices did, with more efficient encoding and "chunking" of information. The evidence for such chunking has been reproduced in the domains of algebra, as well as physics[34]. Contrary to the work of Neufeld, this data suggests that there are fundamental differences between experts and novices. In these studies, the concept of chunking predominates, a concept that is also described by West et al[6] as an important organizational strategy aimed at better management of data. In essence, chunking is a process schemata, analogous to our definition of schemes, suggesting that experts may in fact use these advanced organizational strategies and do so more readily than novices.

Differences between experts and novices with regards to chunking have also been demonstrated in medical studies. In a paper published in 1986, Patel et al [35] demonstrated that with simple recall of propositions (i.e. verbal structures that demonstrate chunking)

experts and novices are very similar. However, when analyzing the number of inferences (i.e. number of transformations of the text, which requires a higher cognitive level) made on typical cases, it appears that the experts made much more inferences of high relevance propositions, and much less inferences on low relevance propositions. Therefore, while the chess studies demonstrated evidence for chunking as a memory organizational method, this study supports the notion that experts use chunking as a problem solving process, by actively searching for key pieces of information, and likewise filtering out irrelevant pieces of information, in order to effectively chunk in or out potential diagnostic groups. As with the chess studies, differences between experts and non-experts in both critical cue recall and diagnostic accuracy become less evident when information is presented in an unstructured manner [19].

Grant and Marsden also found differences in the structure of expert memory. In their paper,[36] the term "forceful feature" is described, and defined as " personally important pieces of information that act as a key to particular memory structures which in turn give rise to the clinical interpretation." These forceful features are therefore clinical data that "forced" the diagnostician towards a diagnosis, and thus are somewhat analogous to the critical cues described in the previous study, which serve a potential chunking role. These authors found no differences in overall numbers of diagnoses or forceful features between non-experts and experts, but did find important differences in the actual nature of these differential diagnoses and forceful features. This also corroborates the notion of differences in expert and non-expert interpretation of clinical data.

Similarly, Joseph and Patel [37] demonstrated expert use of chunking for memory organization and problem solving in a 1990 paper. These authors confirmed that overall, experts and non-experts select the same amount of cues from a text, but that experts focus more on critical information than non-experts do. This study, however, looked at two different aspects of the problem solution. Firstly, they examined the generation of links between cues, and found that experts generated both a greater number of links as well as more specific links to relate cues, including linking of critical and relevant cues. Secondly, the experts started generating their hypotheses earlier and stopped generating hypotheses earlier than the non-experts did. This implies that expert physicians proceed, from a set of initial hypotheses, to systematically narrow uncertainty, while non-expert subjects increased uncertainty as they progressed.

III) The Move to Forward Reasoning and Small Worlds

The evidence thus far presented opens the door to viewing expert problem solving in a different light than the traditional hypothetico-deductive method of reasoning. Hypothetico-deductive reasoning is a strategy for eliminating hypotheses one by one that can be seen as potentially inefficient in clinical medicine. By its to-and-fro nature of constantly returning to the initial problem to test a new hypothesis, hypothetico-deductive reasoning has also been dubbed "backward reasoning". This method is not only an inefficient way of problem solving, it also does not yield a useful way of organizing knowledge into memory. Furthermore, there is evidence suggesting an increased tendency to generate errors when using this method [28]. We have previously seen that in both medicine and other fields, experts do reason differently than non-experts, and rely

on chunking of information for knowledge representation, as well as problem solving, by relying on critical pieces of information to separate chunks of diagnostic possibilities. In addition, the last paper shows us that experts seem to proceed from a set of initial hypotheses, and systematically eliminate these hypotheses by using critical cues and specific links. This is the type of "forward reasoning" that has been attributed to expert reasoning by many authors, having been described in both medical and non-medical literature. Larkin et al[34] refer to this method in a study assessing the solution of problems by expert and novice physicists. They found that while experts used the given quantities of a problem and moved forward to its solution, the novices moved from proposed solutions to the problem back to the given quantities (i.e. hypothetico-deductive reasoning). Joseph and Patel, in the paper presented above, refer to the "two-step framework" of solving clinical problems, which reflects this forward reasoning hypothesis. Greeno and Simon[37] also suggest that experts use the given data to move forward towards a conclusion, an approach requiring small manageable sets of information that can be readily accessed. Similar conclusions have been drawn in medical studies where protocols of cardiologists explaining a case of bacterial endocarditis [38] have yielded evidence to suggest the use of predetermined rules to draw conclusions from the cues in the case. This is especially true for easier cases. For more complicated cases, experts may use more of the biomedical knowledge underlying these rules, with the addition of methods more consistent with backward reasoning [35]. In a review of a number of their own experiments, Patel et al reiterate their conclusions that expertise in medicine

implies preponderance of forward reasoning.[18] with backward reasoning existing as a method used by novices or occasionally experts in complicated cases.

Newell [39] provides further elaboration on this issue by dividing problem solving techniques into two broad categories: weak and strong methods. Weak methods are general reasoning strategies that can be applied in situations where the knowledge base is lacking. These weak methods are used characteristically by novices, and are generally inefficient and frequently misleading strategies. In contrast, strong methods are those based on a sound knowledge base, and thus typically used by experts. These methods are more efficient and precise. As indicated in Groen and Patel's 1985 paper[30], hypothetico-deductive reasoning is a weak method of problem solving, which is used by experts only in ill-structured or difficult problem solving situations (i.e. when they can no longer behave like experts).

The most appropriate interpretation of information from the domain of cognitive psychology is that forward reasoning, including the notion of expert chunking and expert early presentation of hypotheses that are tested with precise, critical cues, is indeed a strong method of problem solving. Another strong method of problem solving is "pattern recognition", which is a means of reaching a diagnosis, typically used by experts, involving more rapid labeling of the condition based upon prior experience [4]. Kushniruk et al [3] has expressed forward reasoning in very elegant terms. They propose the "small worlds" theory of expertise reasoning, which essentially summarizes the information reviewed above. The main points of this theory are:

i) Expert physicians organize their knowledge into subsets of logically related diseases (i.e. "small worlds"). This is analogous to the previously described chunking.

ii) These subsets are distinguished from one another by the presence or absence of limited key medical findings. These critical cues, analogous to Grant's "forceful features",[36] have been shown to be clearly sought by experts in the process of problem solving.

iii) For a given case, the expert physician can efficiently recognize these key medical findings and enter the appropriate "small world".

This notion of "small worlds" exemplifies what we have described as "schemes". For example, for the clinical presentation "dysphagia", some experts will initially organize or "chunk" diagnostic hypotheses into 'oropharyngeal dysphagia' (see Appendix A) or 'esophageal dysphagia'. The appropriate hypothesis set will be entered with one key piece of information, or proposition " do you have trouble swallowing the food or does the food stick after you swallow?" Assuming this is a case of esophageal dysphagia, two more small worlds or diagnostic hypothesis sets are created, i.e. motility or mechanical, which are again separated by one proposition " is the trouble with solids alone or solids/liquids?" A few more questions would then narrow the options until a diagnosis is arrived, as presented in Appendix A. The creation of schemes, such as the one for dysphagia, is based on sound cognitive psychology principles, which are in turn derived from cognitive psychology literature. It accurately reflects the findings described in relation to how experts organize their memory, use chunking, forward reasoning, and

solve problems as described by Kushniruk, by dividing large domains into small worlds consisting of linked diagnostic hypothesis sets. Norman et al.[40] summarize the above discussion very well. They state that: "experts 'cast a broad net' of initial diagnostic hypotheses- which captures all the significant values- and then use the relations among these variables, frequently complex, to arrive at a conclusion. The additional tests, when requested, are highly specific to rule in or out a particular competitor". It is noteworthy that in this paper, a dramatic difference in expert vs. non-expert diagnostic accuracy (91% vs. 25%) was found. This superior diagnostic accuracy of experts, who presumably used strong methods of problem solving to reach these diagnoses, led to our questioning the relationship between diagnostic success and the specific problem solving strategy used by experts and non-experts. This question thus became one of the main research goals of this project.

V) Using Schemes in Medical Education

The studies cited above provide the rationale for the belief that schemes are important and potentially powerful teaching tools. However, although schemes are an attractive and sound way to organize information for students, a number of counter-arguments to their use can be made. For example, in reality there are certain clinical presentations (ex: chronic abdominal pain) for which a condensed and useful scheme is difficult to find. Thus, this clinical presentation, as well as others, remains without ideal organizational and problem solving schemes. Secondly, since learning styles differ, some students will benefit more from this strategy than others will. Moreover, some students resist using schemes as a means of organizing learning, or will resist the suggested

scheme in preference to one created on their own. This latter form of resistance, the creation of personalized schemes, is one to be recommended and encouraged. It should be made clear to the students that expert schemes are only suggested schemes. They need to be encouraged to adapt or completely change these schemes and thus to render them more personal. In fact, remodeling of schemes to a more personal level addresses another concern raised against scheme use: the profound importance that experience plays in the experts' memory and their clinical decision-making. If all experts used a similar scheme to solve a given problem, one would expect considerable overlap in the forceful features and the knowledge they use to solve problems. To the contrary, two studies by Grant and Marsden[16,36] reveal an enormous variability between experts in the knowledge and forceful features used in clinical situations. This suggests that teaching schemes in a rigid fashion does not accurately reflect the actual diversity present amongst experts in real practice. This diversity, essentially created by each individual's unique set of personal clinical experiences, leads to unique knowledge organization, unique schemes, and perhaps overriding of any scheme based on a given experience. As a consequence, the notion of a scheme needs to be tied in with Bordage's previously described concept of semantic networks[8], which are a set of connections between abstract concepts and clinical experiences. Schemes can address the rich network of knowledge held by experts, but do not account for the evolution of individual experiences which clinical practice layers onto the expert knowledge network.

In summary, therefore, the cognitive psychology literature supports the use of schemes as a potentially useful way to provide a novice learner with the background

scaffolding of knowledge used by experts. For schemes to achieve their full potential as a teaching method, they must be presented with enough flexibility to account for individual learning style variability, individual remodeling to a more personalized scheme, as well as eventual incorporation of clinical experiences into this network of knowledge, in keeping with the notion of semantic networks.

C) Evaluation Methods and Their Significance

Having implemented a curriculum based on clinical presentations and with schemes as a learning and problem solving strategy, it becomes essential to devise an evaluation method consistent with the curriculum. More specifically, fully aware of the power of evaluations on student learning [13], the question arose of whether there was any relationship between various types of examination questions and the problem strategy students would utilize. If schemes are powerful instruments for learning and problem solving, is there an evaluation method that tests or perhaps promotes utilization of schemes. This became one of the main research goals of this project.

D) Assessment of Problem Solving Using Pencil-and-paper Evaluation Methods

In a curriculum structured around clinical presentations that are organized according to higher-order or expert methods of memory representation and problem solving, it seems reasonable that an evaluation method be found that assesses scheme utilization in problem solving. In this section, previous research on evaluation of problem solving is presented, with specific focus on evaluating diagnostic skills. As well, a new

format of evaluation questions, termed Elimination-Type (E-type) questions,[41] is described (see Appendix C for example).

I) Use of Pencil-and-paper Methods Over Other Evaluation Formats

Ideally, if the overall aim is to assess problem solving skills, and specifically diagnostic skills, a case can be made for using free-response questions or oral examinations. This argument lies in the assumption that these formats may yield a better global picture of a candidate's clinical reasoning, especially given the oral examination's ability to assess a more "real life" clinical scenario and opportunity for interactive question and answer sessions. There are, however, problems with such evaluations. The first surrounds the problem of case specificity [13], which predicts that success in solving one clinical presentation does not predict success in another. As a consequence, a reliable and valid examination is dependent on a broad sampling of problems. Furthermore, content validity also depends on a representative and adequate sampling of problems[42]. Such extensive sampling is more easily done with pencil-and-paper type of tests, since the other forms of testing are resource-intensive. Furthermore, questions of the free-response types can be fraught with ambiguity, can be unfocused, and have the potential for subjectivity in grading[43]. Similar criticisms were raised for the once popular Patient Management Problems (PMPs), which have now been abandoned by licensing bodies because of lack of reliability [44], as well as a failure of construct validity. [45]

II) MCQ Criticisms: Advent of Key-Feature and Extended-matching Questions

For all of the reasons mentioned above, three pencil-and-paper evaluation formats were selected for this research. The first type of question selected is the classic single

answer, five-option Multiple-choice question (MCQ: see Appendix C for an example). Although MCQs have always been considered an efficient and reliable testing tool, they were not considered ideal for the evaluation of higher-order thinking skills such as problem solving. The conventional wisdom is that MCQs assess lower levels of knowledge such as recall of isolated facts. To the extent that some clinicians question whether MCQs can test actual clinical tasks, these questions tend to have low face validity [46]. Newble best summarizes the prevailing thoughts on MCQs by stating that they measure " a combination of what the student knows, partially knows, can guess, or is cunning enough to surmise from any cues in the questions"[47]. Stated otherwise, this cueing effect can lead to bias, and students with finely honed test-taking skills may be favored[48]. As well, trivialization is felt to occur with the MCQ format[47].

To alleviate the fear that the MCQ format lacks the ability to assess clinical decision-making, examination formats were created in the hope that problem solving could be assessed in a more appropriate fashion. One such format, developed for the Medical Council of Canada Qualifying Examination (MCCQE)[42,49] is called the "key-feature" format. Key-feature questions aim to assess problem solving by focusing on a given problem's most crucial elements for resolution, i.e. the key-features. These problems can focus both on key diagnostic or management aspects of a given case, and constitute a mixture of both write-in problems as well as choosing an answer from a short menu (average 15-20 options). These questions have demonstrated good psychometric properties with regards to reliability, face and construct validity and are thought to assess higher cognitive levels of problem solving.

Extended-matching formats (EMQs: see Appendix C for examples) were introduced in the 1990s in both the NBME and USMCE, amongst others. Case and Swanson[43] have been instrumental in the development of these questions, which are defined as any matching format with more than the five options traditionally used by MCQs. From its conception, the question preparation of the EMQs have been very careful in designing stems that test higher cognitive levels such as problem solving. The first study that examined the psychometric features of Extended-matching[43] showed that Extended-matching items were more difficult, more discriminating, had higher reliability and needed significantly less testing time to achieve reproducible scores than traditional MCQs. Other studies have shown that EMQs, by increasing the number of options used, increased mean item difficulty as well as, perhaps by reducing guessing, provided improvement in item discrimination over the five option MCQ[50]. By increasing item discrimination, EMQs offer comparable levels of reproducibility with 30% less items than the MCQ with five options[50]. Reliability coefficients were also markedly higher with Extended-matching[43]. Fenderson et al[48] administered EMQs, uncued examinations, MCQs and true/false pathology questions to medical students, and found slightly higher total reliability scores (0.90 vs. 0.83) and discrimination value (0.29 vs. 0.25) with EMQs over MCQs, although p values were not stated. In the same paper, Fenderson examined the potential effects on learning of each format, by investigating academic achievement on subsequent comprehensive examination. No difference was found.

The Extended-matching format has also been applied to diagnostic pattern recognition examinations, which are similar to the questions devised by Case and

Swanson, but with shorter stems and restriction of the task to making diagnoses. Gruppen et al[51] confirmed Case and Swanson's previous work in pattern recognition testing by showing that the format had high reliability and validity. The reliability and validity were also found to be quite stable from one medical school to another. Dunn and Woolliscroft also showed favorable psychometric properties with a surgical pattern recognition examination[52].

Two other studies have used EMQs with standardized patient examinations. Blackwell et al[53] compared short-answer questions (SAQs) to EMQs in the second part of paired OSCE stations. They found that both types were similar for eliciting physical findings, diagnoses and treatment plans, as well as overall scores. The distribution of scores did vary, however, in that EMQs yielded a greater percentage of students in the upper quartile, perhaps related to cueing effects. However, the weaker students were distributed equally with both techniques. Overall, EMQs were considered to be psychometrically equivalent, but superior with respect to timesaving for the examiners. The other study by Solomon[54] used EMQs as a tool for assessing students' ability to identify key features when faced with patient problems.

As outlined by Case and Swanson, there are other potential advantages of EMQs [43]. From a test preparation point of view, Case and Swanson feel that these questions flow well with course objectives, are easy to prepare, can be prepared without technical flaws, and are less concerned with examiners "guessing" what distracters would be most appealing to the student. This latter function is one that examiners have been shown to be unable to do consistently. Of particular interest for the protocol of the study about to be

proposed. they suggest that EMQs have a potential advantage over MCQs in testing clinical problem solving.

III) Comparison of MCQs. EMQs and the New Elimination-type Questions

The proposed study for this thesis aims to test this latter statement. No study has formally used think-aloud protocols to assess whether a well-written MCQ differs from EMQs (or any other format. such as "Key-features". for that matter) in challenging examinees to problem solve. There is little doubt that poorly written MCQs can encourage students to learn isolated facts by rote. However. as acknowledged by Case and Swanson. well constructed MCQs could challenge students to problem solve. Maguire et al also recognized that MCQs can yield valid information of clinical reasoning skills. providing that stems and alternatives are well constructed.[55]. Evidence does exist that MCQs have predictive value for more recognized problem solving tasks[56] and can elicit higher order problem solving such as forward reasoning[57]. In fact. all available evaluation methods potentially yield information on clinical reasoning if the content is appropriate. suggesting that content is more important than question type [45].

The present study proposes to compare performance and the problem solving strategies (specifically scheme utilization) by medical students and experts. and evaluate whether MCQs and EMQs have any influence on either. This will be accomplished by directly assessing the examinees' cognition through think aloud protocols. A third question format will be included in the analysis. the Elimination-type (E-type) questions. A preliminary study by Mandin and Harasym[41] yielded results for this new format that were encouraging. This study showed that the E-type format was more reliable and

required fewer items to attain the standard alpha of 0.80. The E-type format was specifically designed in the hope that it would encourage students to use schemes for problem solving. Thus far, it has been shown to provide a diagnostic task that is more difficult and different from the Extended-matching format[41]. The E-type has two parts (see Appendix C), one which is similar in structure to the EMQs, and a second part consisting of an additional question that asks why a certain answer, present in another "branch" of a given scheme, was excluded by the student. This exclusionary question is asked first, before the students are asked the second part of the question, which requires the selection of the most likely diagnosis amongst an extended list of options. The student must explain the exclusion of one possible diagnosis, and it is hoped that it is equivalent to exclusion of that "branch" of a scheme. To do this, they identify which items in the stem that forced them to exclude the diagnosis. The items are the bulleted numbers that follow every piece of information in the stem. The assumption is that the E-type questions, by forcing the student to justify the exclusion of another branch, will add a different and new component to the assessment of problem solving skills by further reducing the possibility of guessing, a concern which remains in the EMQs. Secondly, the E-type questions, by asking examinees why a diagnosis in another arm or branch of a scheme was excluded, will potentially oblige the student to utilize an approach or scheme to a problem, a factor that is believed to be advantageous cognitively.

E) Think-aloud Techniques in Evaluating Cognitive Processes

Researchers in a number of fields have accomplished the difficult task of analyzing cognitive processes by asking subjects to "think-aloud" when solving problems. Think-aloud has been used extensively in cognitive psychology as well as in health care related fields including psychology, physiotherapy, nursing and medicine.[58,59,60] A number of technological devices have been used, including audiotape, videotape and computer simulations. There have been a number of variations in the way this method is accomplished. The main categories of think-aloud are protocol analysis, propositional analysis and methods combining these two broad techniques.

Protocol (i.e. statement) analysis is a method primarily used in the context of concurrent data rather than in studies using retrospective analysis. It is based on certain assumptions, including that verbalized data is a subset of our underlying cognitive processes, and that information currently used by a subject is accessible through verbal data[61]. These methods do not tap into the multitude of possible cognitive processes not expressed verbally. According to Fonteyn, protocol analysis is the division of subjects' verbalizations into segments. The segments are first coded to allow for assertional analysis, which examines the formations of relationships between the verbalized concepts, such as relationships of meaning, significance or cause/effect. Secondly, script analysis is performed, which scrutinizes the reasoning processes used and illustrates the relative importance attributed to certain protocols by the subject. Finally, the actions taken by the subject and the rationale for such actions are analyzed. This method provides

rich and complex data, which is particularly useful for analyzing dynamic or ongoing changes in a subject's information processing and representation of a situation[37].

The second method of cognitive analysis is propositional analysis, a method originated by Kintsch (1974) and Fredericksen (1975)[30]. This strength of propositional analysis, which rectifies one of the weaknesses of protocol analysis, is to capture complex relations between the statements and representations in memory. A proposition is defined as [3]:

> " The smallest unit of meaning that underlies the surface structure of a text".

One of the outcomes of the analysis of these propositions or "units of meaning" is seeking and identifying propositions that provide "chunking" of stimuli. Stated otherwise, the method assumes that "propositions form manageable units of knowledge representation"[18].Once identified, these "chunking" propositions are sought typically in the transcripts of domain experts vs. non-experts. Recall and inferences (i.e. transformations) made on these propositions are then compared between the two groups, as well as whether the propositions identified by the subjects were relevant or irrelevant [30]. When done on concurrent text, this method is termed discourse analysis.

Since the present study primarily addresses global representations of clinical presentations in memory, think-aloud and an adaptation of propositional analysis will be used. For each clinical presentation, certain propositions have been determined to be key "chunking" propositions. For example, with dysphagia, the proposition "dysphagia to solids alone vs. solids/liquids" is a key chunk separating mechanical versus motility

causes of dysphagia. Therefore, a subject who attaches critical importance to this proposition will have been deemed to show evidence of scheme, (i.e. chunking, i.e. "small world") use. The analysis will be done based on recall following the written examination. In the literature, recall protocols have been used mainly in the context of asking subjects to transcribe recalled propositions from cases[29,32]. This has been criticized for assessing a perceived recollection of a case rather than the true cognitive process occurring at the time of the problem solving task. However, for the present study, the method is appropriate since exact numbers of recalls or specific inferences made from recalled texts are not the outcomes of interest. What is of interest is a global description of representations in memory, which makes use of recall appropriate[37]. Fonteyn describes recall as possibly providing a "more complete description about one's reasoning strategies". Furthermore, in order to assist the recall process, subjects are required to briefly describe in writing, after answering each question, how they arrived at an answer for a given problem. Subsequently, this written discourse is used as a guide to the think-aloud discussion.

F) Summary of the Main Research Questions

Therefore, with these five concepts in mind, our study will be carried out with the purpose of answering the following four questions:

Research question #1: Is there a correlation between the reasoning process used in problem solving, specifically scheme utilization, and diagnostic success?

Research question #2: Is there a measurable difference in the scheme utilization elicited by standard five-option Multiple-choice, Extended-matching and Elimination-type questions?

Research question #3: What are the psychometric properties of the three question formats used in the study, both with a standard dichotomous scoring of 0-1, and a partial dichotomous scoring system? (the latter gives a partial mark for a wrong final answer that is, however, in the correct "arm" of a scheme).

Research question #4: Does a method of assessing problem solving exist that is more economical than the standard think-aloud analysis?

CHAPTER TWO: METHODS

A) Pilot study

In early 1999, a pilot study was conducted. The pilot study was primarily aimed at establishing whether a method could be devised to shortcut the cumbersome method of cognitive analysis known as "think-aloud" for the purpose of establishing the strategy used for diagnostic problem solving. In case a shortcut was not possible, a second reason for the pilot study was to determine the extent of inter-observer agreement for the identification of problem solving processes used in the think-aloud method of cognitive assessment.

The methodology for the pilot study was as follows:

1) Examination construction:

An examination was constructed around four clinical problems in Gastroenterology: mechanical dysphagia, motility dysphagia, acute diarrhea, and chronic diarrhea. The examination consisted of twelve questions, with three questions created for each of the four clinical problems.

The three stems created for each of the four clinical problems were randomly assigned to one of three examination formats (see Appendix C for examples of the three examination formats). All of the questions asked for a single best answer, most likely diagnosis, to be chosen from the clinical information in the stem. The three formats were:

i) Standard 5-option Multiple-choice Question (=MCQ)

ii) Extended-matching question (=EMQ), with a list of 10-16 options

iii) Elimination-type question (=E-type), consisting of two parts: one part identical to the EMQ, with the other part asking to identify two or three "features" (numbered pieces of information in question stem) which led to the exclusion of a certain diagnosis, present in an entirely different branch of a scheme.

Thus, of the twelve questions, four were MCQs, four were EMQs, and four were E-types.

2) Subjects:

The examination was administered to four experts in Gastroenterology as well as four non-experts, final year medical students at the University of Calgary. Candidates were considered experts if they were specialists who spent more than 80% of their clinical time in the practice of Gastroenterology.

3) Data Collection:

Once the examinee had completed the twelve questions, a panel of two judges, one specializing in Gastroenterology and one in Nephrology, interviewed the examinee using a think-aloud technique. The examinee, after first revealing his/her final diagnosis, was asked to discuss how he/she arrived at the answer for each question. The examinees had been encouraged, prior to starting the examination, to make notes at the end of each question (prior to proceeding to the next question) on the manner the answer was derived. This recommendation was made in order to help with the think-aloud process. The examinees were then asked to think-aloud, with as little prompting as possible, on the manner each answer was derived. The judges assigned one of four reasoning strategies

for each question: guessing, hypothetico-deductive reasoning, scheme, or pattern recognition. The discussions were audiotaped or videotaped (after informed consent was obtained) for review in case of disagreement between the two judges as to the reasoning process used. Eventually, a consensus on the problem solving strategy was obtained between the two judges.

Subsequently, the reasoning strategy found by think-aloud was compared to:

i) Result from " BBQ method": the examinees were asked, after the think-aloud, to read a scenario of four BBQ repairpersons approaching a defective BBQ in order to repair it (see Appendix D). Each of the repairpersons was representative of one of the four reasoning strategies outlined in the previous paragraph. The idea was to see whether the examinees could identify with one of these repairpersons as to the reasoning strategy they used for each question, and whether this self-analysis correlated with the think-aloud.

ii) Result from a panel of judges examining the written material and determining the reasoning strategy from the examinees' writings. Again, this was to be correlated with the think-aloud, but this method was quickly abandoned since the written materials were not appropriate for cognitive interpretation by judges, as they were frequently quite sketchy, especially in the last half of the examination.

These are the main conclusions of the pilot study:

i) Think-aloud provides very useful and reliable insights into global representations of problems in memory. The judges were generally in agreement after the first hearing of the think-aloud. and could easily arrive at a consensus when necessary. using the audiotape or videotape.

ii) The "BBQ method" was not a reliable shortcut into cognition. with only a 45% correlation with the "gold standard" think-aloud. Problems with the method included examinee confusion over the task asked of them. confusion over the BBQ terminology. and fatigue after a long (over 1 hour) testing session.

iii) Examinees showed several patterns of problem solving process:

a) pattern recognition alone

b) strict adherence to a structured and accurate scheme as a sole strategy

c) use of scheme in combination with another method such as hypothetico-deductive reasoning

d) use of an ill-structured or even faulty scheme with another method such as hypothetico-deductive. guessing or examination savvy

e) absence of any scheme (hypothetico-deductive. exam savvy. guessing)

iv) Because of the small numbers. statistical analysis was not attempted.

Additional observations:

a) All question formats were associated with higher success rates in experts over non-experts (hence construct validity).

b) Experts used schemes 68.9% of questions (24.4% pattern recognition), compared to 33.3% use of schemes in non-experts (47.6% pattern recognition).

c) Elimination-type questions, in experts only, had relatively more scheme use and less pattern recognition than the other question formats did.

B) Main study: Experimental design

1) Examination construction:

An examination was constructed around four clinical presentations in Gastroenterology: dysphagia, chronic diarrhea, nausea and vomiting, and elevated liver enzymes. The examination consisted of twelve questions, with three questions created for each of the four clinical presentations.

The three stems created for each of the four clinical presentations were randomly assigned to one of three examination formats (see Appendix C for examples of the three examination formats). All of the questions asked for a single best answer, most likely diagnosis, to be chosen from the clinical information in the stem. Two diagnoses were accepted for question #11 after it became clear during the study that both were equally acceptable answers. The three formats were identical to those in the pilot study.

Thus, of the twelve questions, four were MCQs, four were EMQs, and four were E-types. The twelve questions used are presented in Appendix C.

2) Subjects:

The examination was administered, from February to May 2000, to twenty experts in Gastroenterology in two centers, Calgary (15) and Ottawa (5), as well as

twenty non-experts, final year medical students at the University of Calgary. Candidates were considered experts if they were specialists who spent more than 80% of their clinical time in the practice of Gastroenterology.

3) Data Collection:

There were two levels of examination scoring:

a) Assessment of cognitive process:

This assessment was conducted in a fashion identical to that described above for pilot study. As with the pilot study, a consensus between the two observers could easily be obtained after review of the tapes. The determination of the problem solving strategy used was different from the pilot study in the following way. The verbal discourse was analyzed using a modified propositional analysis. This analysis consisted of searching the examinees' discourse for key predetermined propositions that linked small worlds and thus provided evidence for chunking (i.e. scheme use). These propositions were as follows:

Table 1: Propositions Demonstrating Evidence of Chunking, for Each Clinical Presentation.

Clinical Presentation	Key Chunking Propositions
Dysphagia	Oropharyngeal vs. esophageal Mechanical vs. motility
Elevated Liver Enzymes	Hepatocellular vs. cholestatic Intrahepatic vs. extrahepatic cholestasis
Nausea and Vomiting	GI vs. non-GI causes GI vs. metabolic vs. cns vs. drugs
Diarrhea	Small bowel vs. large bowel Steatorrhea (malabsorption) vs. non-steatorrhea Osmotic vs. secretory vs. inflammatory vs. motility

From the propositional analysis, the judges assigned a grade to the cognitive process on a scale from 0 to 4, reflecting the degree to which a "strong" or "expert"

problem solving process was used. The observations made during the previously described pilot study made possible the development of this scale. This cognitive process scale will hereafter be referred to as the "Process scale". Scores from the "Process scale" were subsequently divided by 4 in order to match the 0-1 range of the dichotomous scores described in the next paragraph.

Table 2: Correlation of Process Score with Problem Solving Strategy

Process Score	Problem Solving Strategy
0	Examinee relies on one of the following three "weak" methods: -hypothetico-deductive reasoning -guessing -examination savvy
1	Scheme utilized was either unstructured or inaccurate. -significant reliance on one of the three "weak" methods mentioned above
2	Scheme utilized generally structured and accurate. -some use of one of the three "weak" methods

3	Scheme utilized was very structured and accurate. -no use of any of the "weak" methods
4	Pattern recognition: no other method used

b) Grading of answers

Two scores were assigned for each question:

i) a dichotomous score (hereby referred to as "Score1"): mark of 0 or 1 for the correct diagnosis on the MCQS or EMQs. For the E-type. if the diagnosis was incorrect. a score of 0.5 was assigned if the examinee had identified the key exclusionary feature(s) for each question.

ii) a partial dichotomous score (hereby referred to as "Score 2"): for each question with an incorrect diagnosis. a mark of 0.5 was given if the examinee's answer was at least considered in the correct "branch" of a scheme (ex: 'cholestatic' branch for liver enzymes. 'mechanical' branch for dysphagia. 'small bowel' branch for diarrhea. 'metabolic' branch for nausea and vomiting).

4) Data analysis:

Table 3: Four Main Research Questions and Corresponding Statistical Analysis

Research Question	Statistical Analysis
#1: Is there a correlation between the reasoning process used in problem solving, specifically scheme utilization, and diagnostic success?	2 ('score1', 'score2') X 3 (categories of 'process scores') X 2 (expert, non-expert) MANOVA analysis
#2: Is there a measurable difference in the scheme utilization elicited by standard five-option Multiple-choice, Extended-matching and Elimination-type questions?	Comparison of global 'process score' means for each of the three examination formats using one-way ANOVA analysis
#3: What are the psychometric properties of the three question formats used in the study, both with a standard dichotomous scoring of 0-1, and a partial dichotomous scoring	Reliability of each format will be calculated using a Cronbach's alpha. Discrimination of each format will be calculated using a standard index.
#4: Does a method of assessing problem solving exist that is more economical than the standard think-aloud analysis?	See following paragraph

Others:

i) Comparison of global process score means for each clinical presentation will be accomplished using one-way ANOVA analysis.

ii) Comparison of process scores means for each format within the clinical presentation (i.e. process score means for each question) will be accomplished by repeated measures analysis.

4) Pilot testing of "crying baby" method of assessing cognitive processes

As was previously discussed in the results of the pilot study, the "BBQ method" seemed to be somewhat confusing to the examinees, because of a lack of clarity in the instructions as well as unfamiliar terminology with regards to the workings of a BBQ.

Therefore, we modified the BBQ method to an easier scenario involving crying babies that we hoped would be more identifiable to the participants and thus correlate higher with the "gold standard" think-aloud method. Once the examinees had completed the examination, they were presented with the scenario and instructions as found in Appendix D. However, the analysis was stopped after 5 candidates for a number of reasons, mainly the inability to complete all the parts of the examination in an acceptable time frame for the candidates. Also, the method had low face validity, as well as a low correlation (55%) with the gold standard think-aloud in those five candidates.

CHAPTER THREE: RESULTS

The results of the study will be presented in conjunction with their related research question. Since the psychometric data is crucial to the interpretation of the remaining results, these (and thus research question #3) will be presented first. (NOTE: details of raw scores, all the ANOVA calculations and other tables not deemed critical results are all found in Appendix E)

A) Results for Research question #3: What are the psychometric properties of the three question formats used in the study, both with a standard dichotomous scoring of 0-1, and a partial dichotomous scoring system? (the latter gives a partial mark for a wrong final answer that is, however, in the correct "arm" of a scheme).

The reliability coefficients and discrimination indices of the examination and the three formats, as well as the scores, are presented in Tables 4a-c.

Table 4a: Overall Cronbach alpha reliabilities for expert subjects. non-expert subjects. and all subjects combined.

Subjects	# of Items Used	Zero Variance	Cronbach's Alpha coefficient
Overall	36	-	.8711
Expert	24	12	.5703
Non-Expert	36	-	.7356

Table 4b: Cronbach alpha reliabilities based on format scores over all subjects.

Scores	Number of Items		Alpha
	Used	Zero Variance	
Dichotomous Score1	12	-	.8074
Dichotomous Score2	12	-	.7429
Process Score Score 3	12	-	.8346

Table 4c: Cronbach alpha reliabilities and Discrimination Indices based on question format over all subjects.

Question format	Number Of Items Used	Alpha	Disc. Index: Dichotomous Score1	Disc. Index: Process Score
Multiple-choice	12	.7582	.625	.6375
Extended-matching	12	.6586	.575	.5313
Elimination-type	12	.6525	.4625	.5625

Construct validity of the formats is seen in Figures 1a-i which demonstrates an overall superiority of the experts over the non-experts across most formats, except when dealing with the nausea and vomiting clinical presentation.

47.

Figure 1a: Average Multiple-choice Dichotomous Score1 across Clinical Presentations

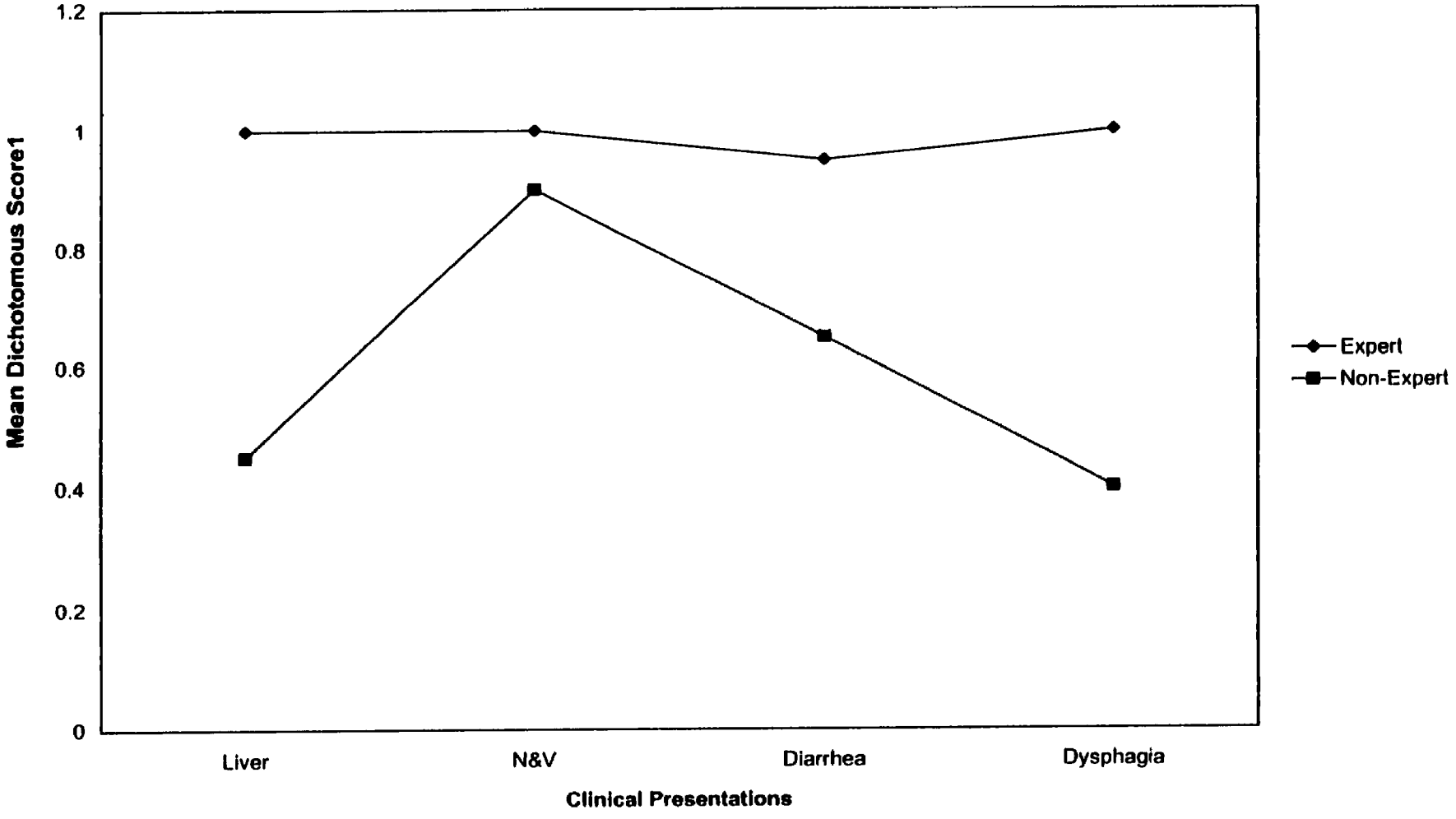


Figure 1b: Average Multiple-choice Dichotomous Score2 across Clinical Presentations

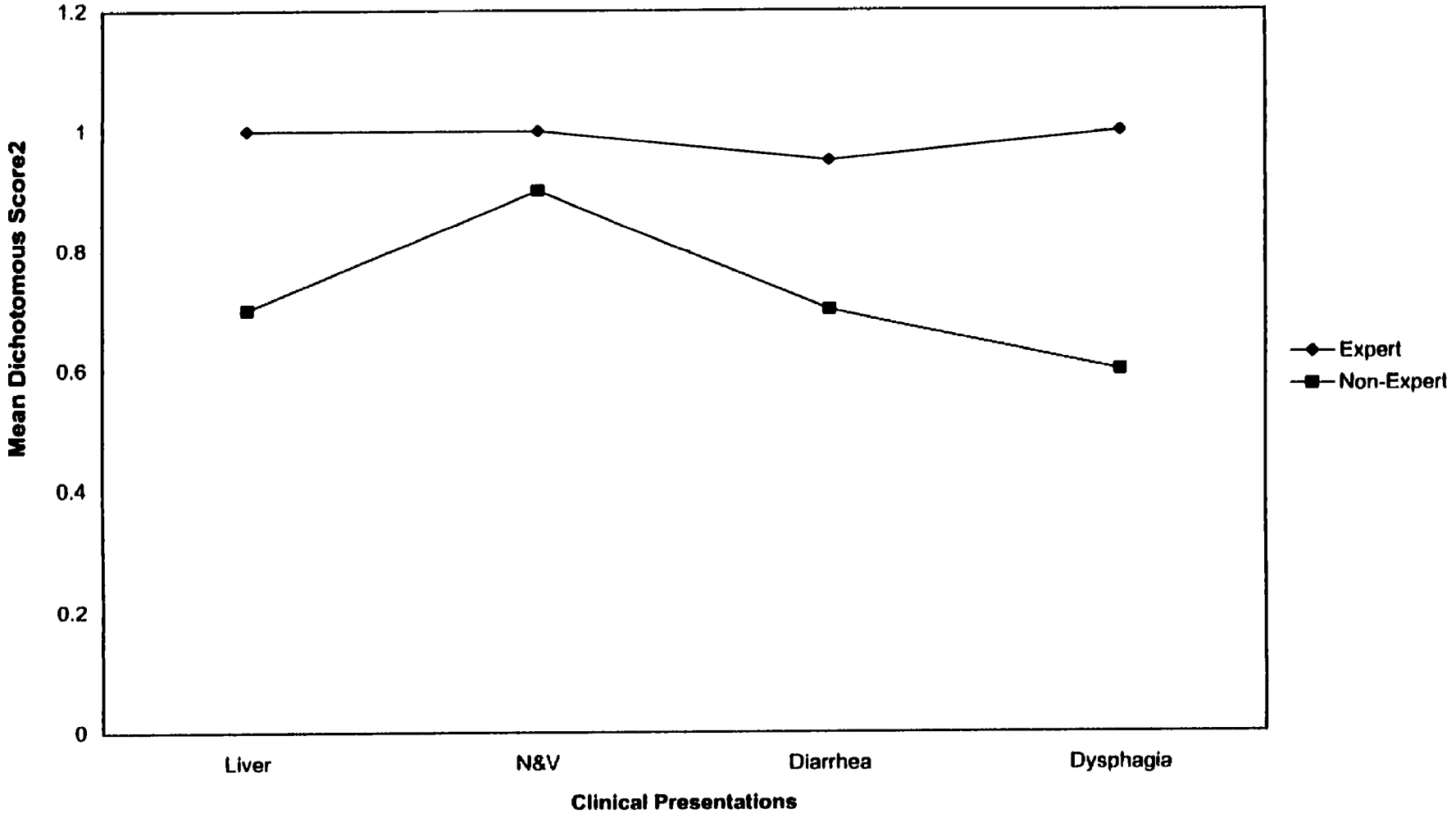


Figure 1c: Average Multiple-choice Process Score across Clinical Presentations

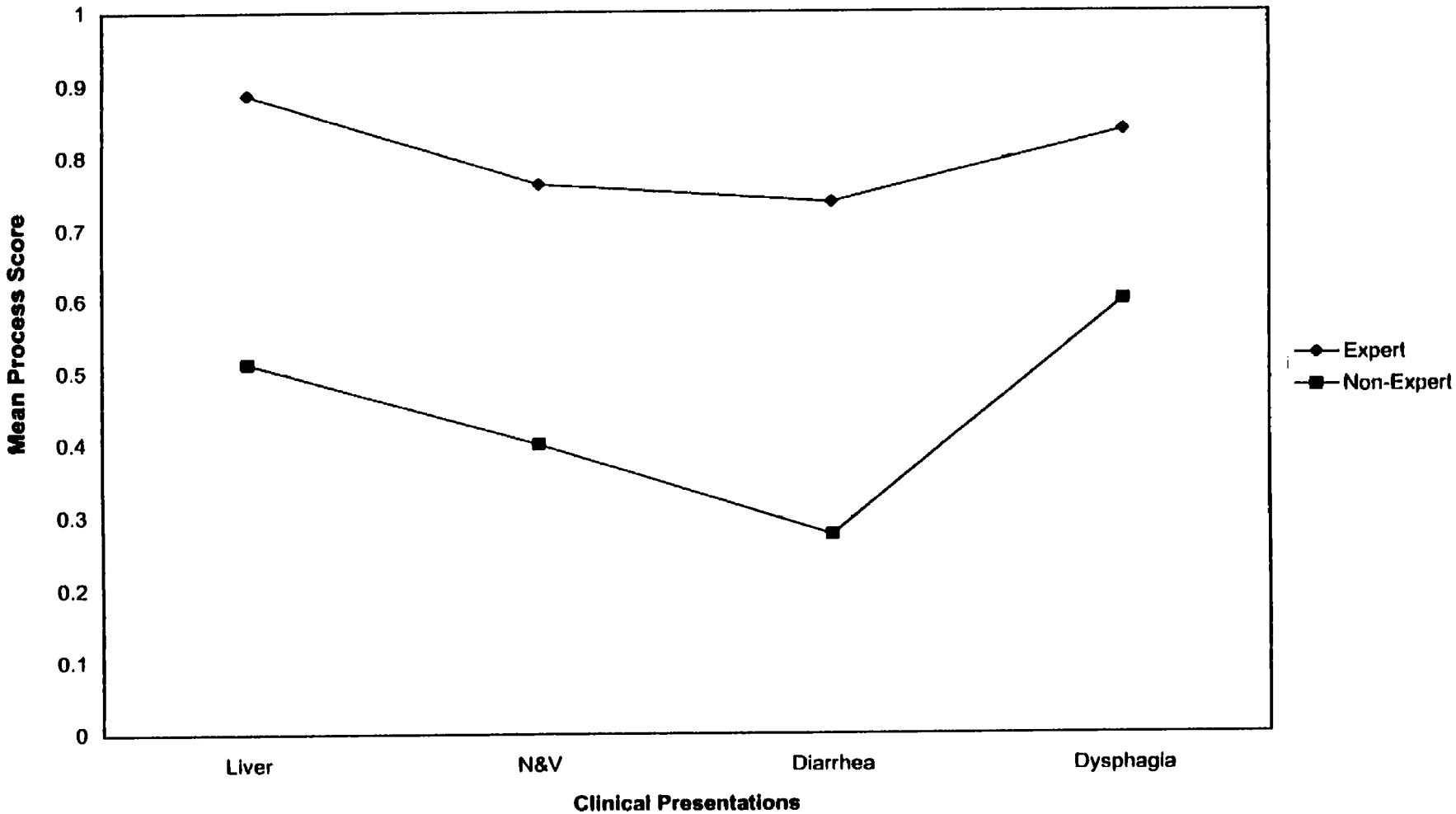


Figure 1d: Average Extended-matching Dichotomous Score1 across Clinical Presentations

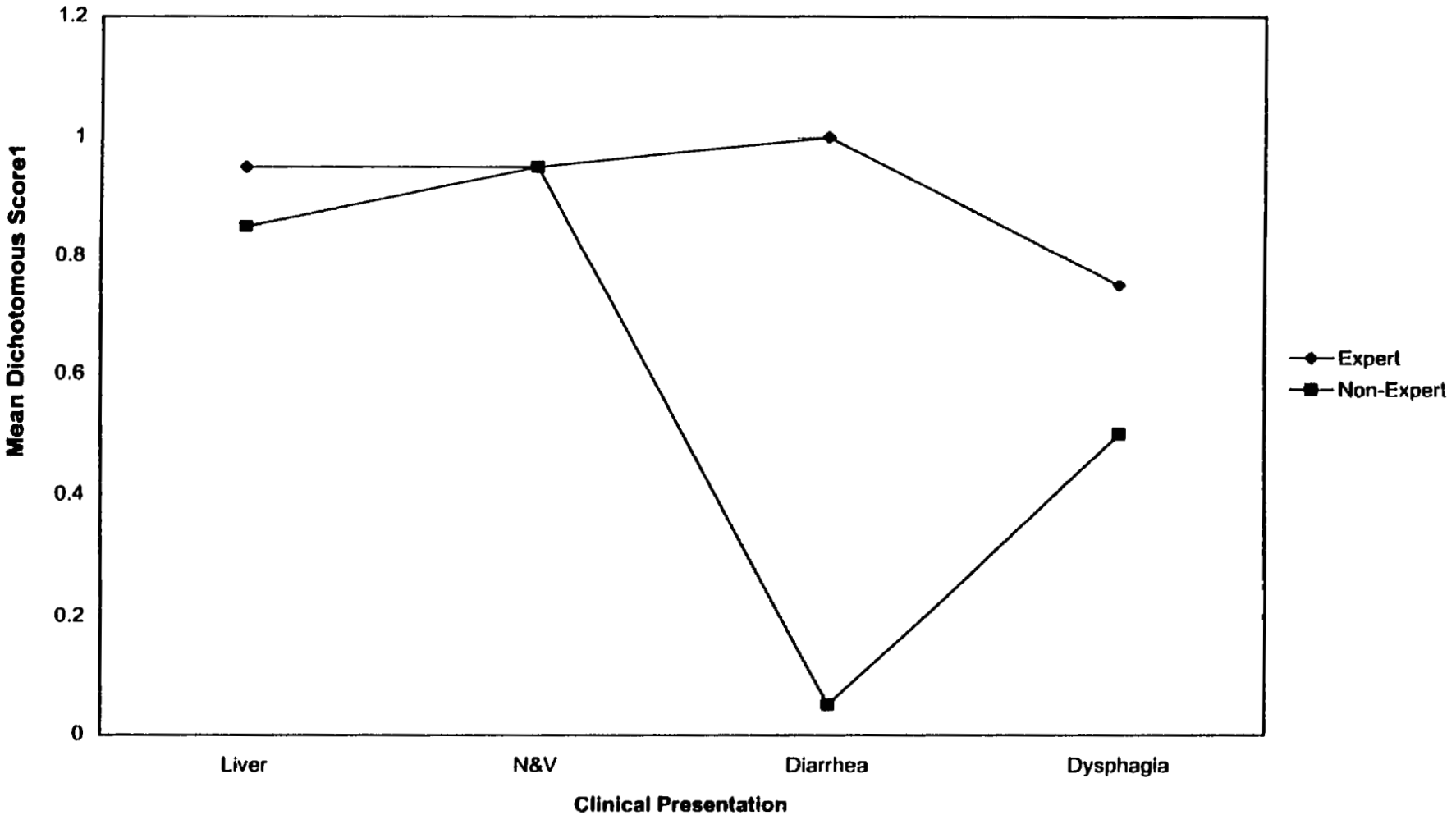


Figure 1e: Average Extended-matching Dichotomous Score2 across Clinical Presentations

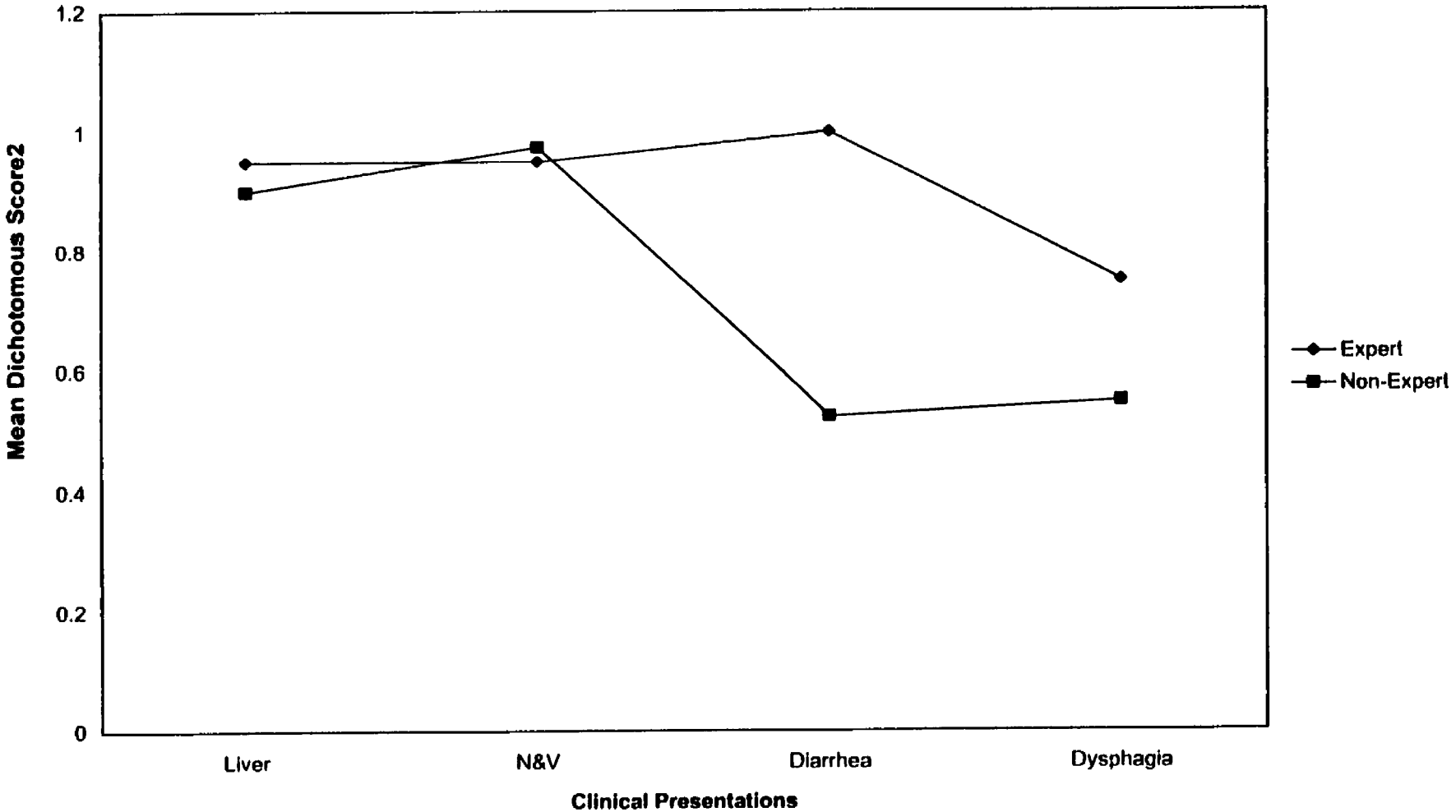


Figure 1f: Average Extended-matching Process Score across Clinical Presentations

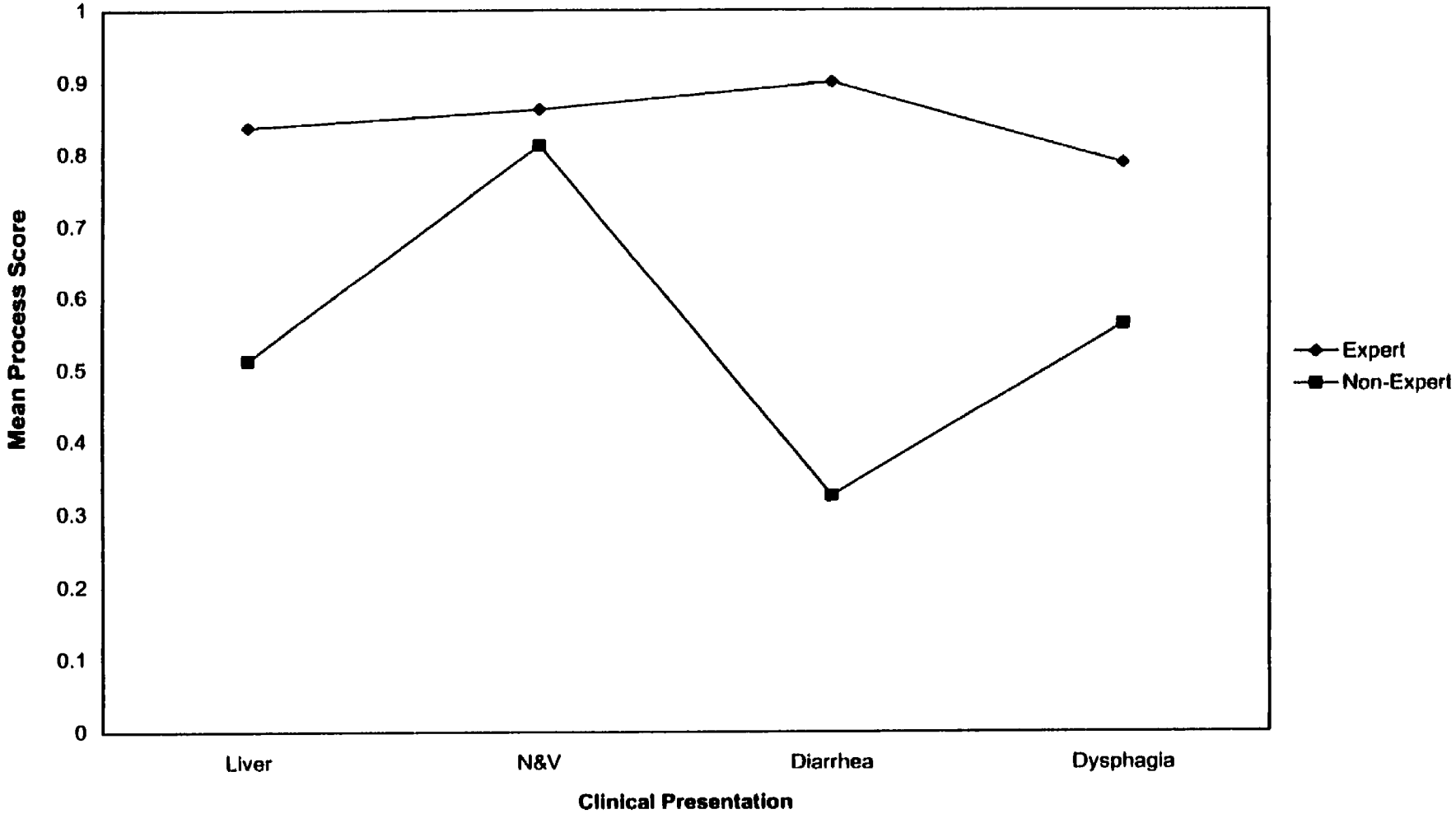
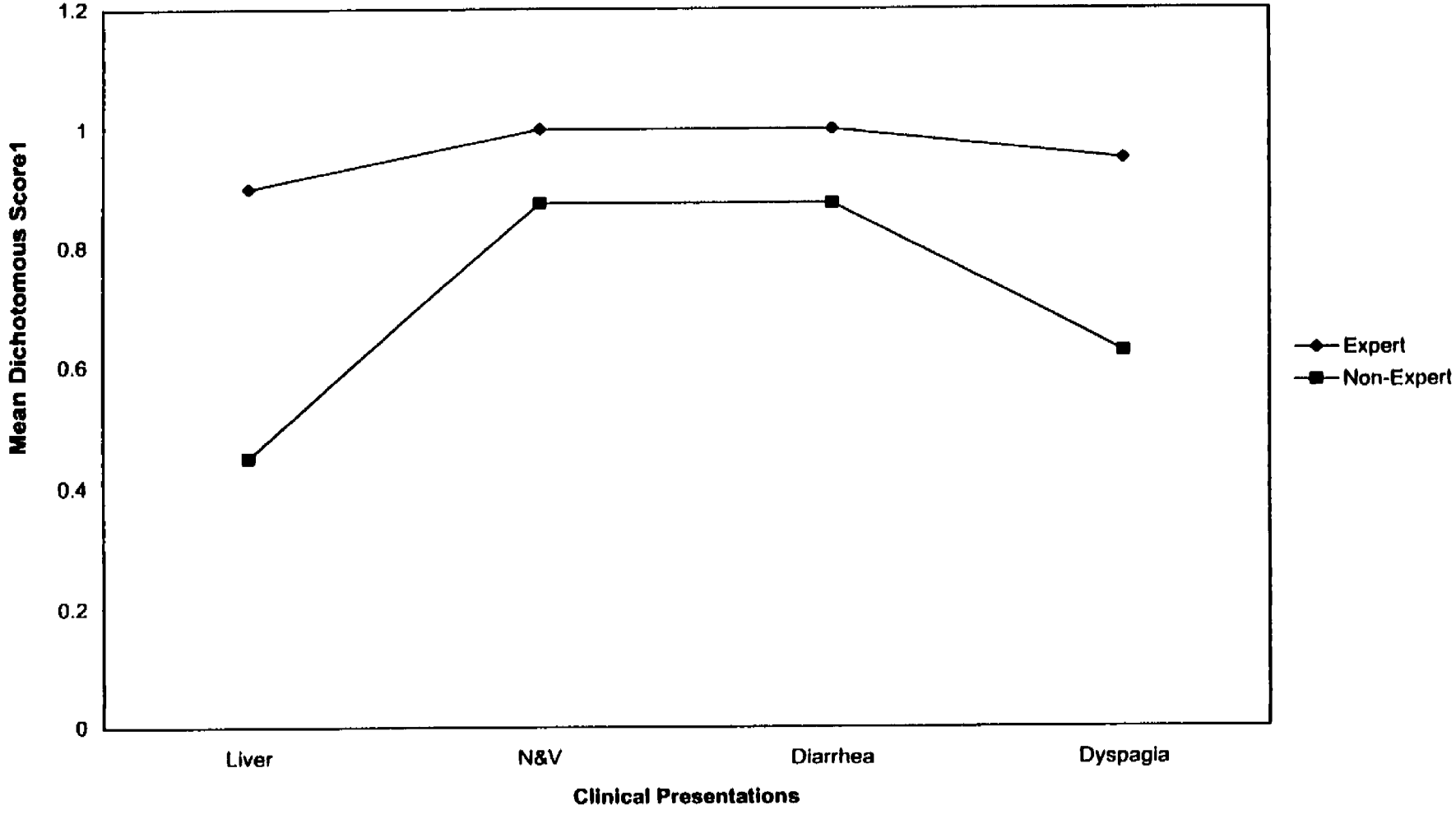


Figure 1g: Average Elimination Dichotomous Score1 across Clinical Presentations



54.

Figure 1h: Average Elimination Dichotomous Score2 Across Clinical Presentations

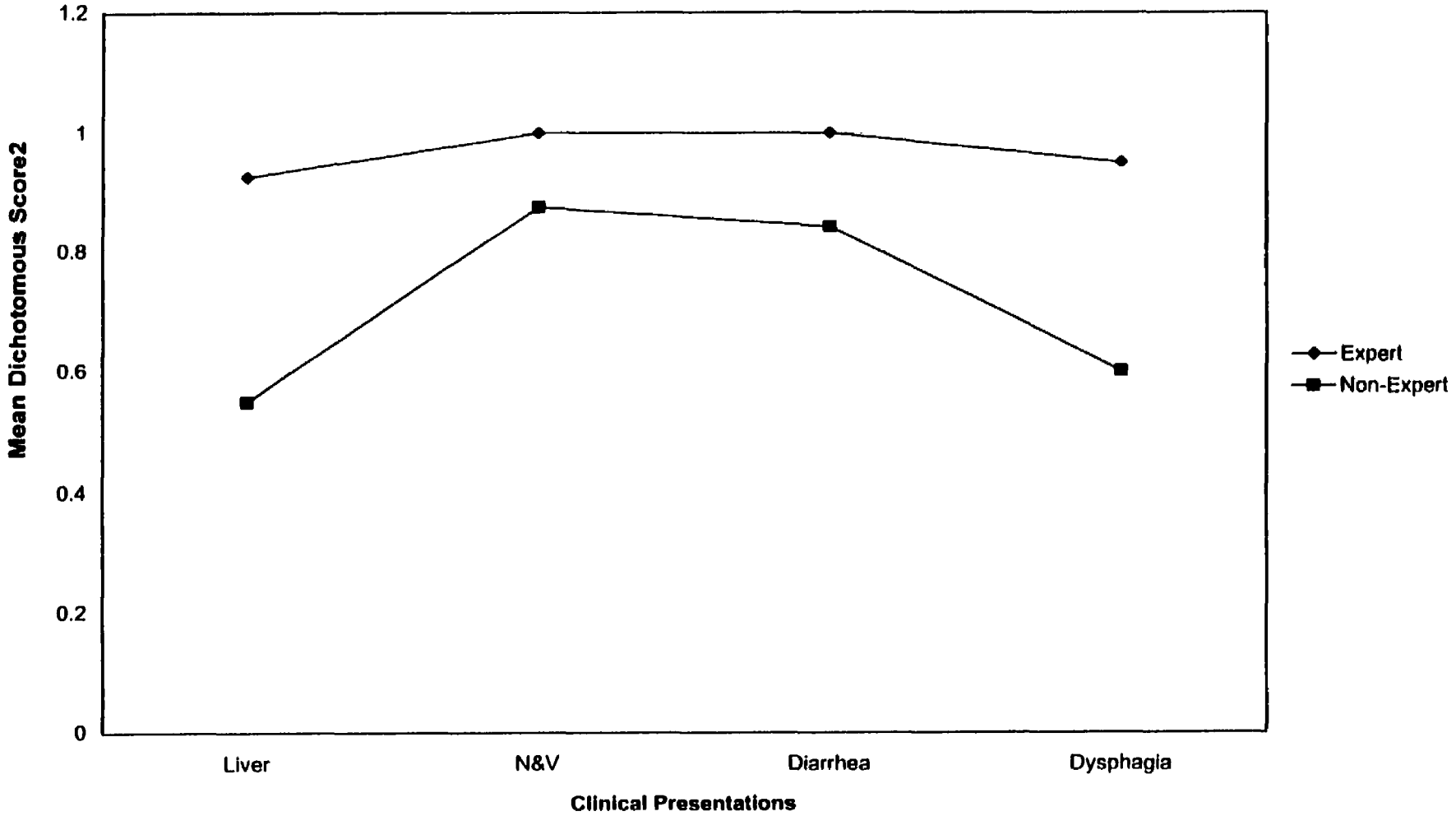
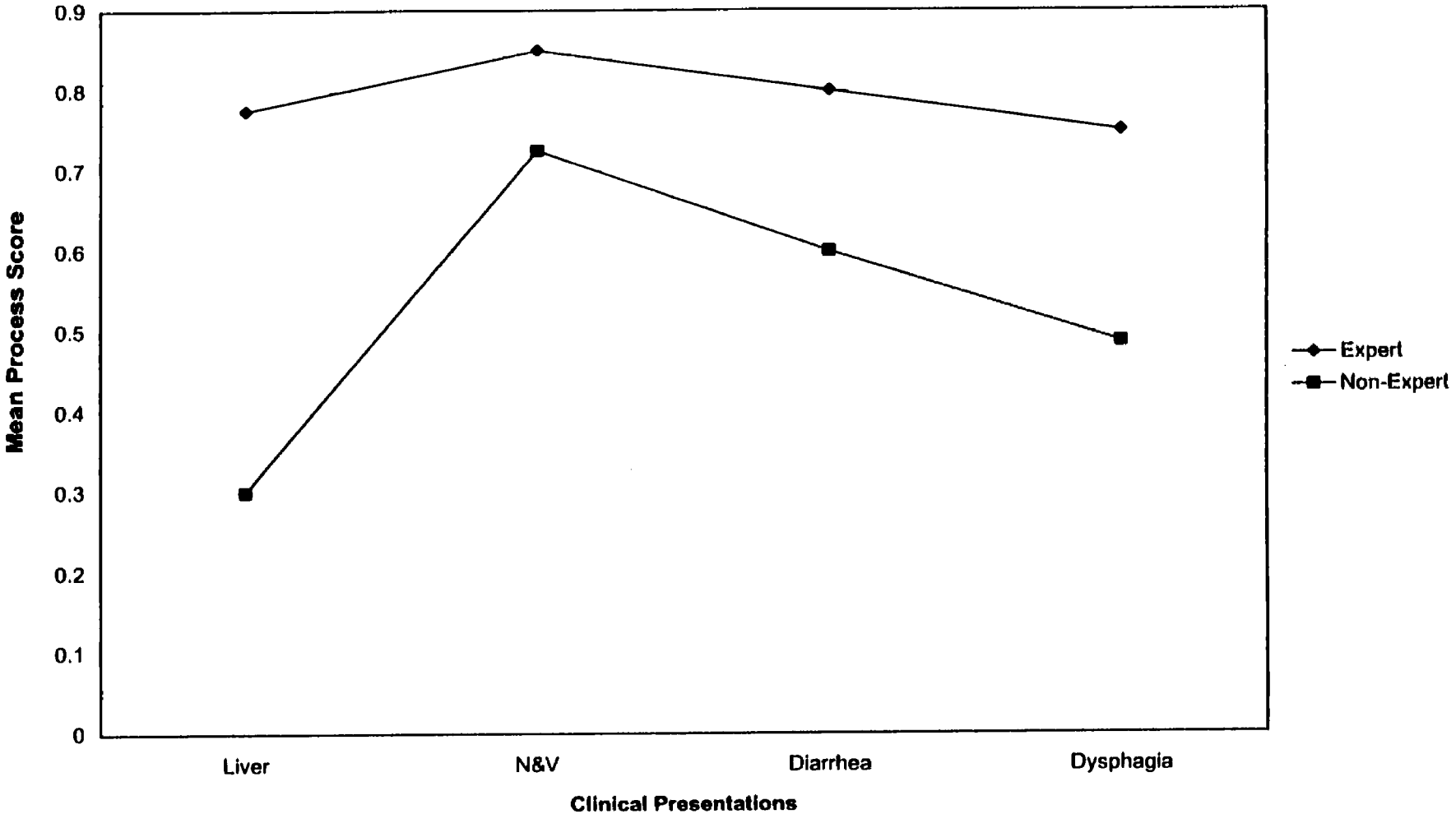


Figure 1i: Average Elimination Process Score across Clinical Presentations



B) Results for Research Question #1: Is there a correlation between the reasoning process used in problem solving, specifically scheme utilization, and diagnostic success?

The next portion of the study was aimed primarily at determining whether the cognitive process used to solve a problem had any impact on examination success. Figures 2a-l demonstrate a trend for a positive correlation between problem solving strategy and examination success, especially in the non-expert group. For these figures, the process scores for each examination question were grouped into three categories: 'process 1', which groups the process scores 0 and 1, and is analogous to hypothetico-deductive reasoning. The 'process 2' category groups the process scores 2 and 3 and is analogous to scheme utilization, and 'process 3' which groups the process score 4 and is analogous to pattern recognition. These groups were justifiable from a cognitive strategy point of view as well as from the experience of the process score judges, who found that the most troublesome distinctions were between the 0 and 1 process scores and the 2 and 3 process scores. A mean dichotomous score was found for each of these three 'process groups', for each question, and plotted into Figures 2a-l. With the exception of the dysphagia Extended-matching question (where one non-expert was in category 3 and answered incorrectly) and diarrhea Extended-matching question (where all but one non-expert answered incorrectly), a linear relationship generally holds true between process and dichotomous scores. (NOTE: dichotomous score 2 figures resembled dichotomous score 1 figures very closely and therefore were not included.)

Figure 2a: Liver Multiple-choice - Dichotomous Score1 by Process Category

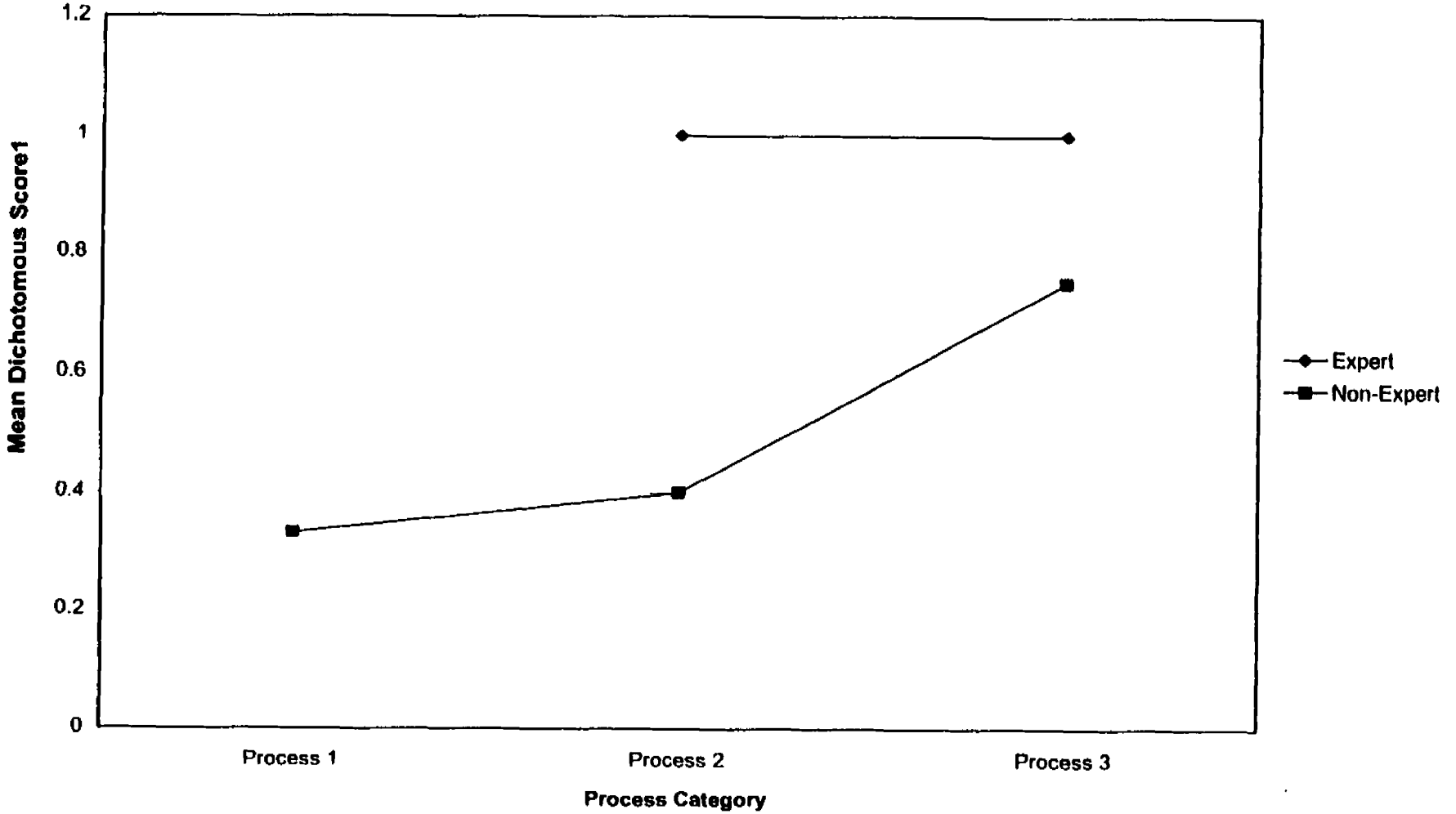


Figure 2b: Liver Extended-matching - Dichotomous Score1 by Process Category

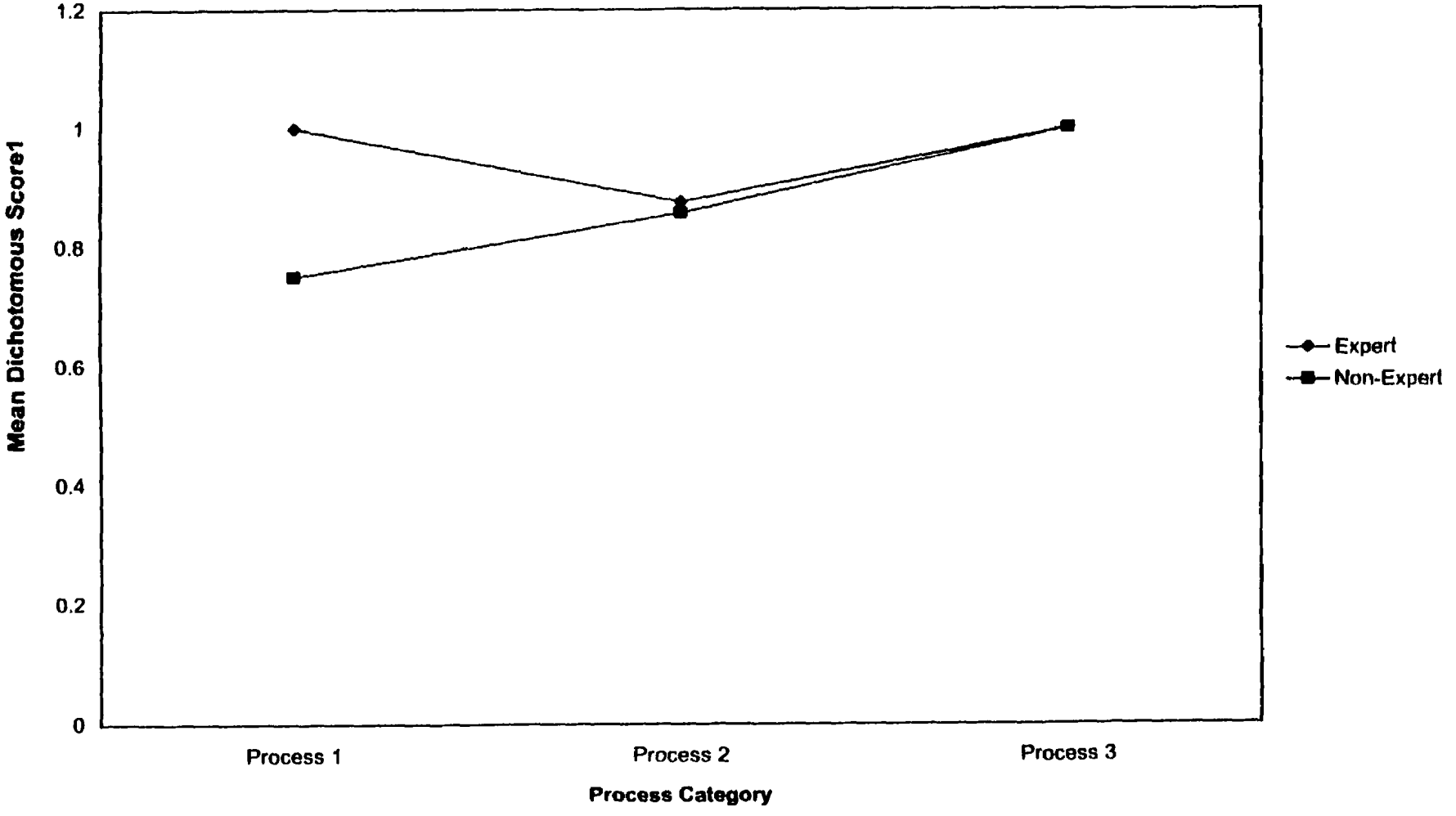


Figure 2c: Liver Elimination - Dichotomous Score1 by Process Category

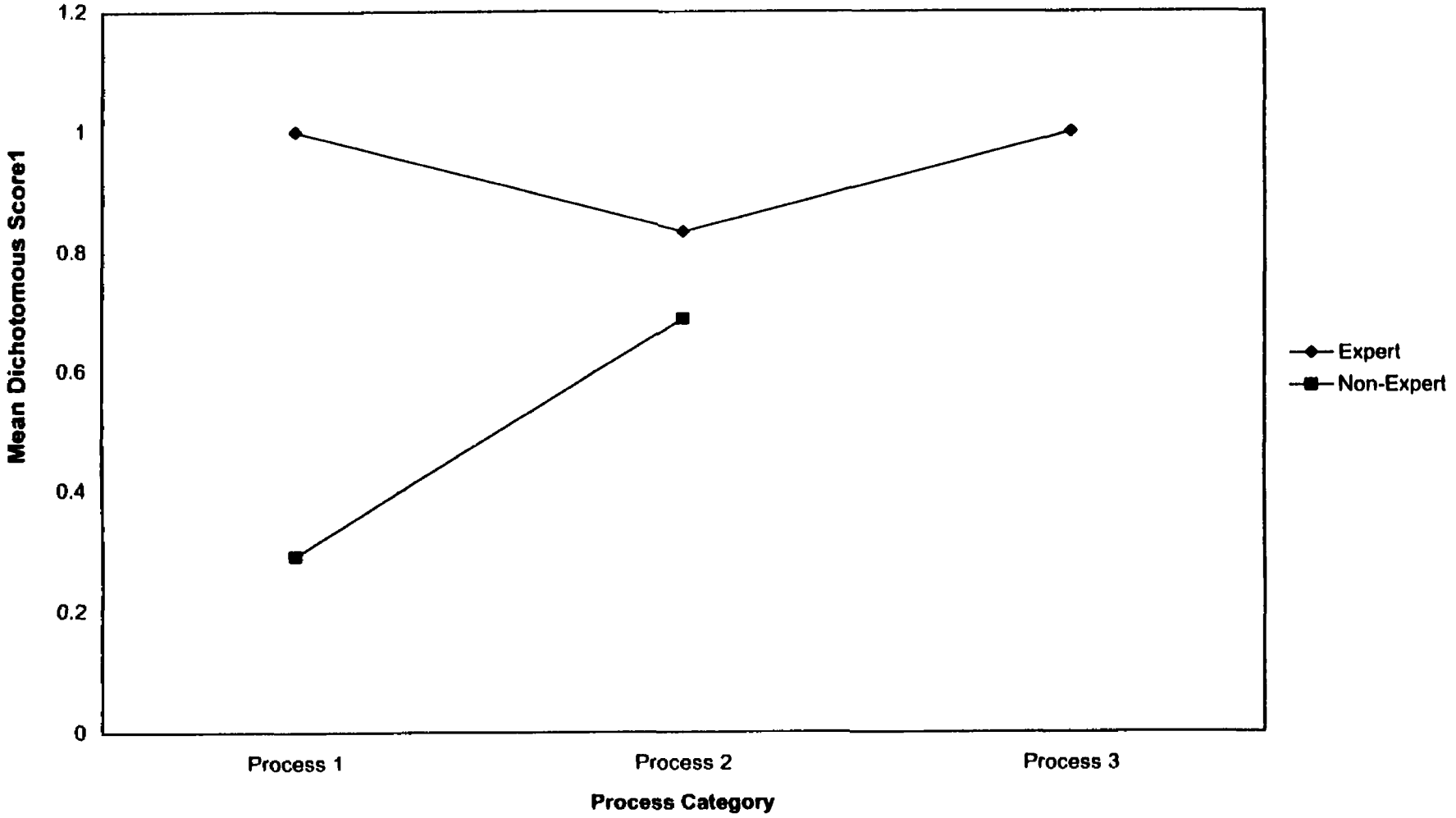


Figure 2d: Nausea & Vomiting Multiple-choice - Dichotomous Score1 by Process Category

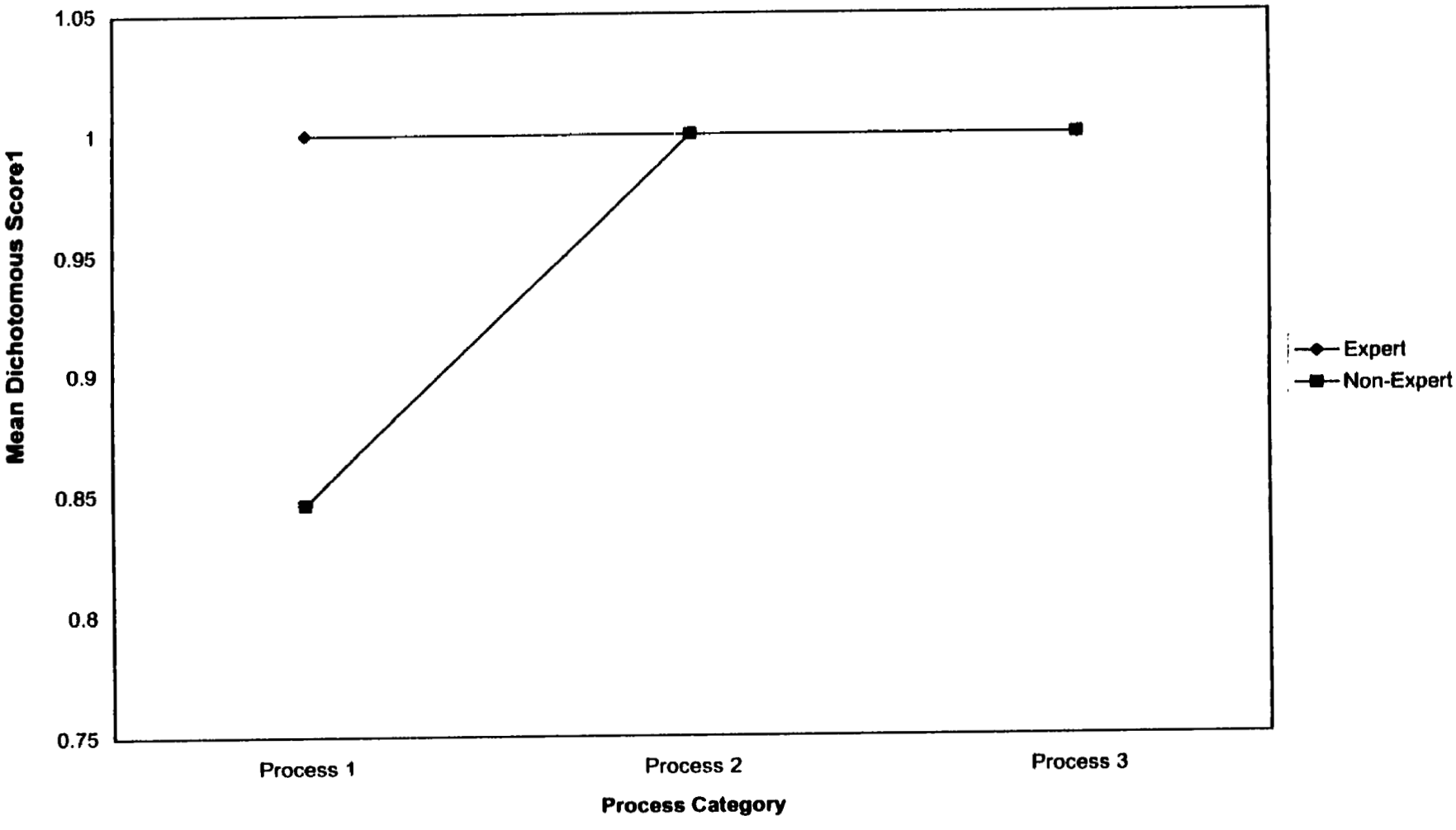


Figure 2e: Nausea & Vomiting Extended-matching - Dichotomous Score1 by Process Category

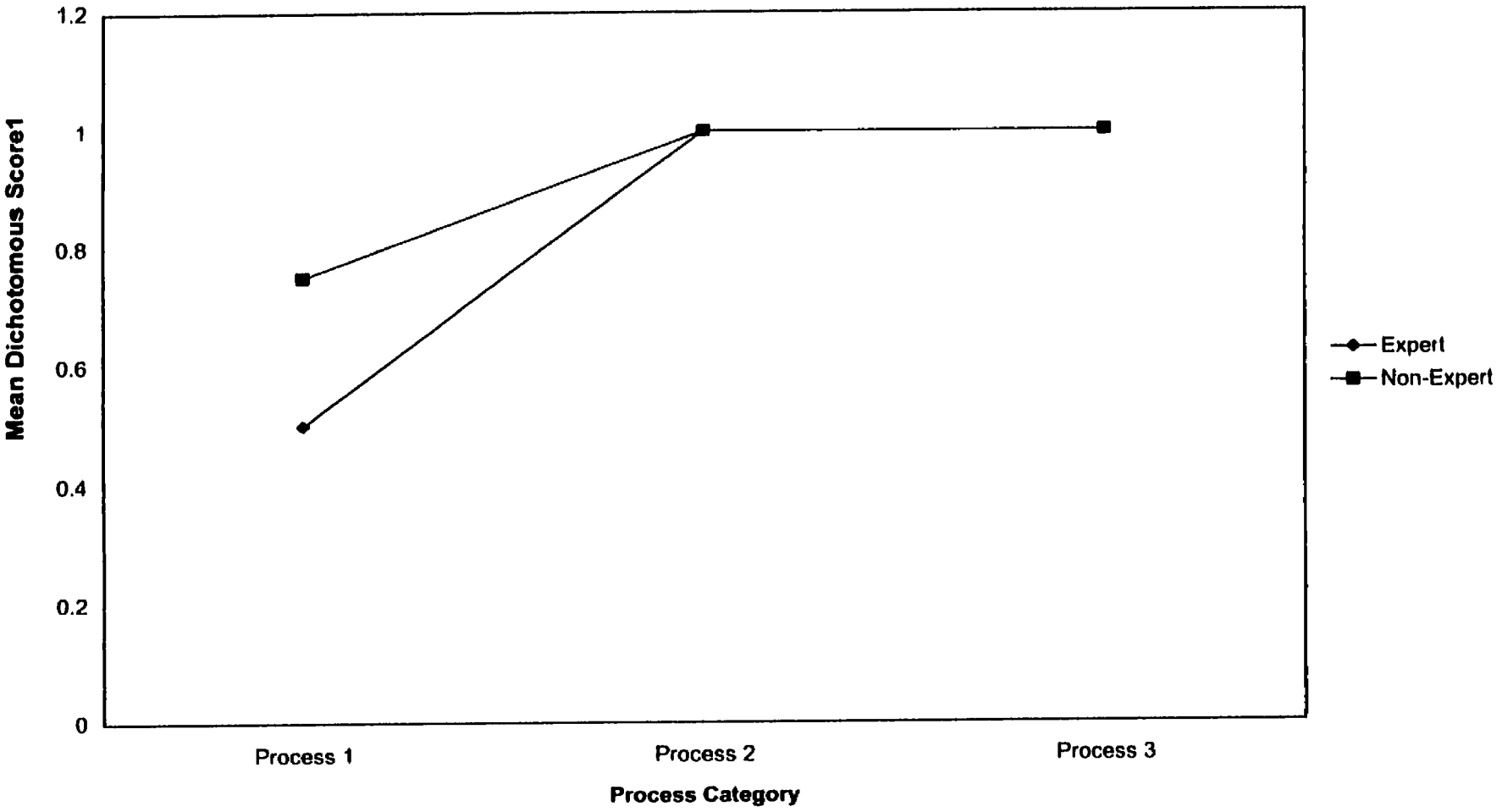


Figure 2f: Nausea & Vomiting Elimination - Dichotomous Score1 by Process Category

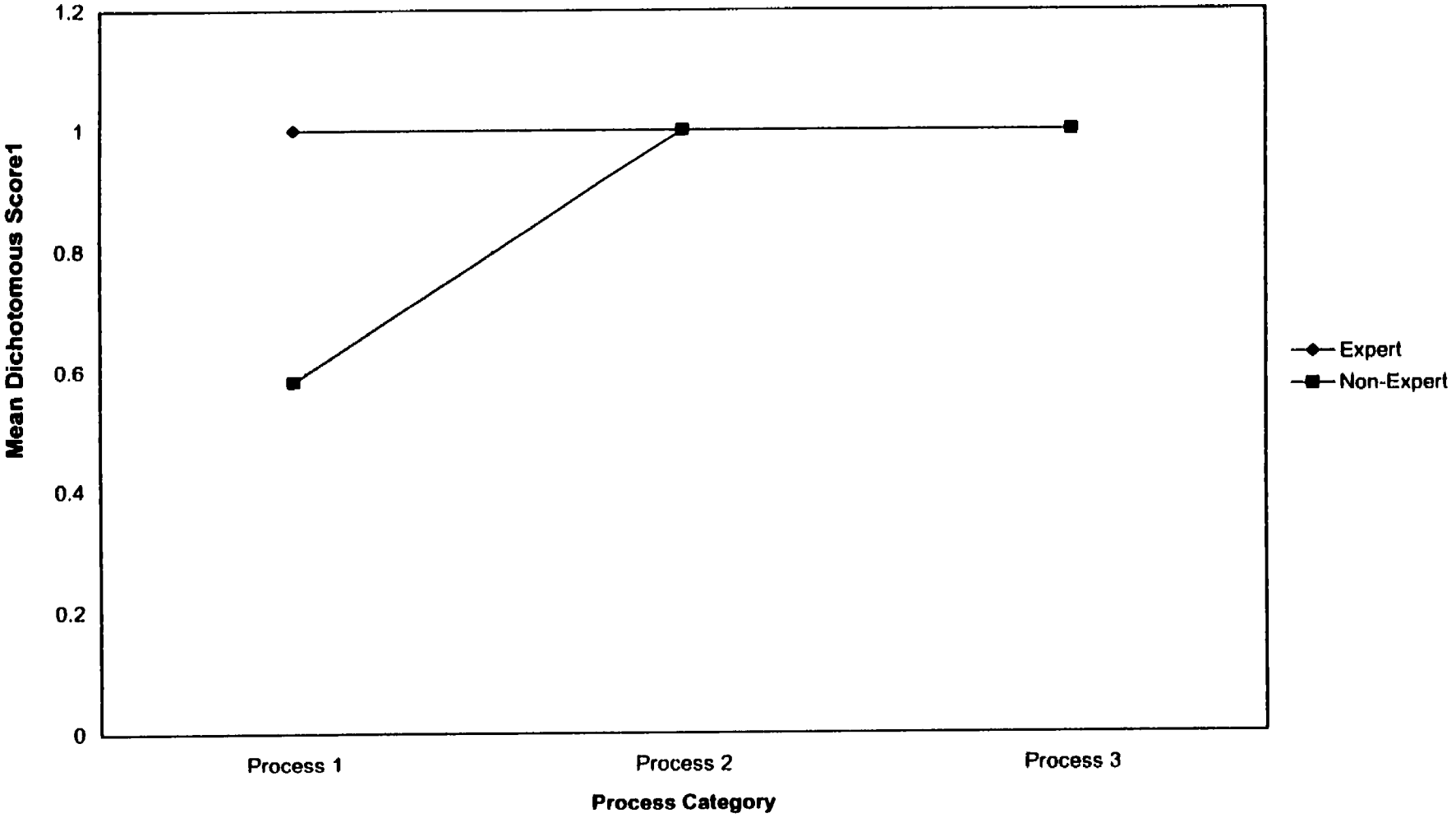


Figure 2g: Diarrhea Multiple-choice - Dichotomous Score1 by Process Category

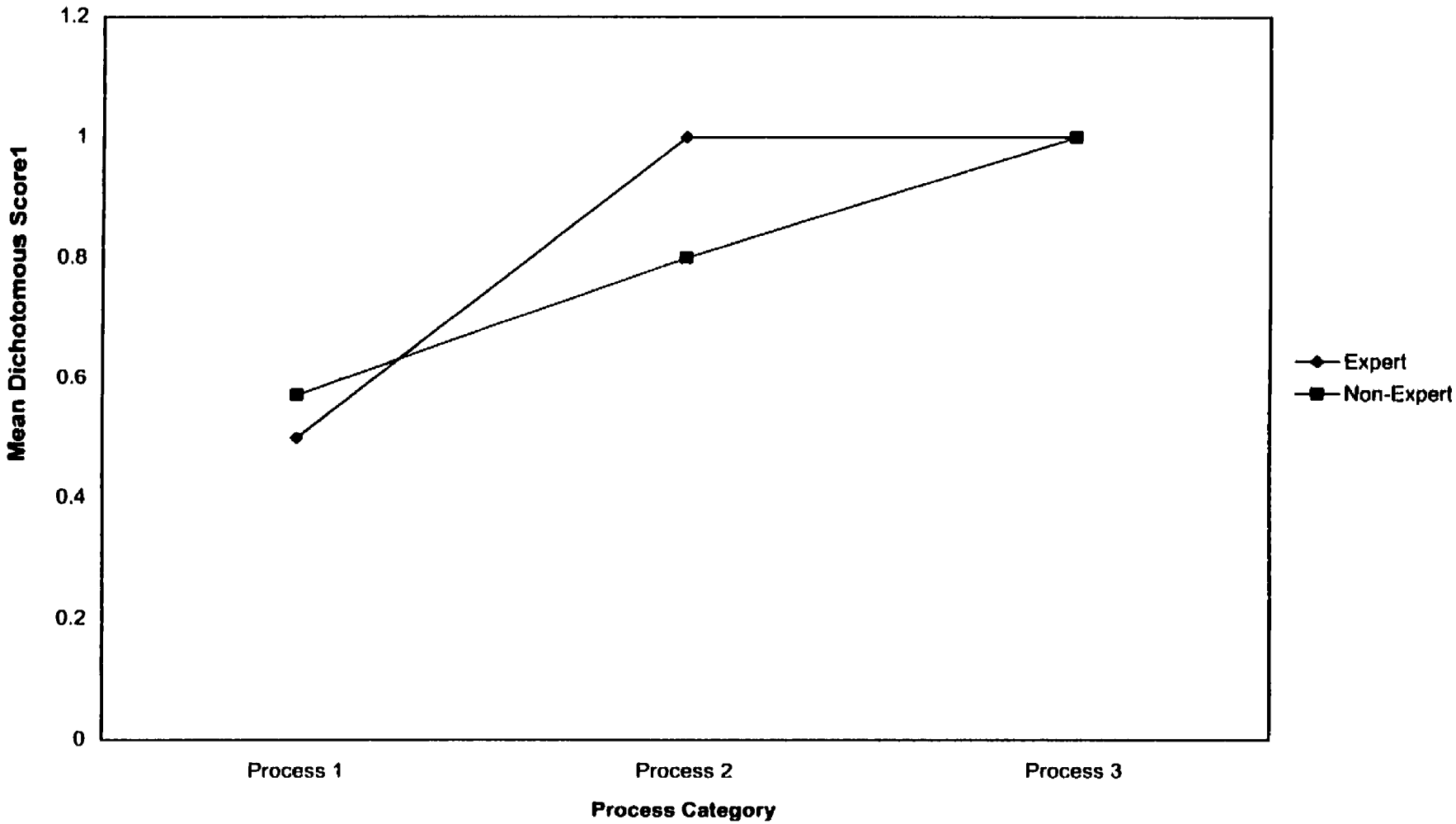


Figure 2h: Diarrhea Extended-matching - Dichotomous Score1 by Process Category

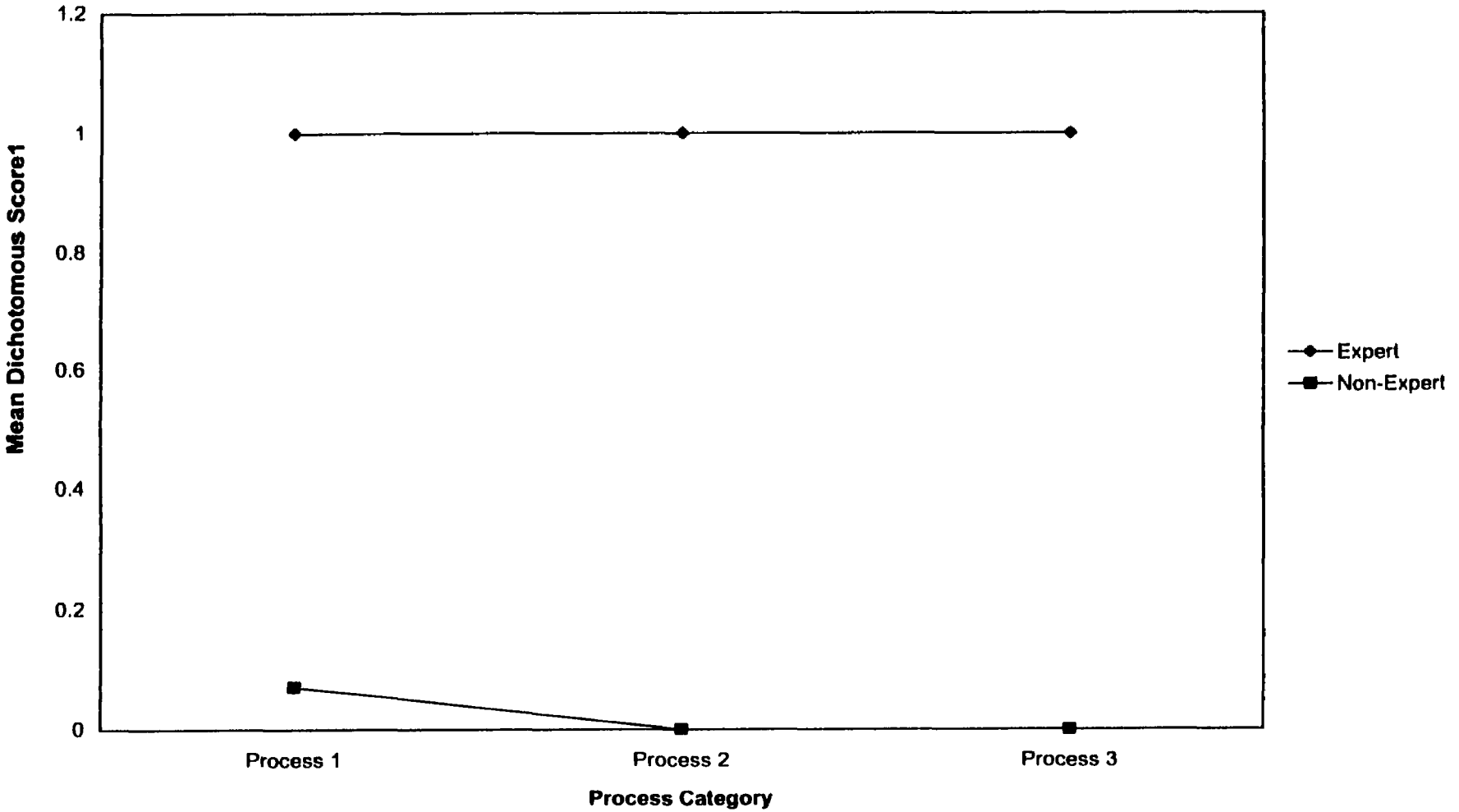


Figure 2i: Diarrhea Elimination - Dichotomous Score1 by Process Category

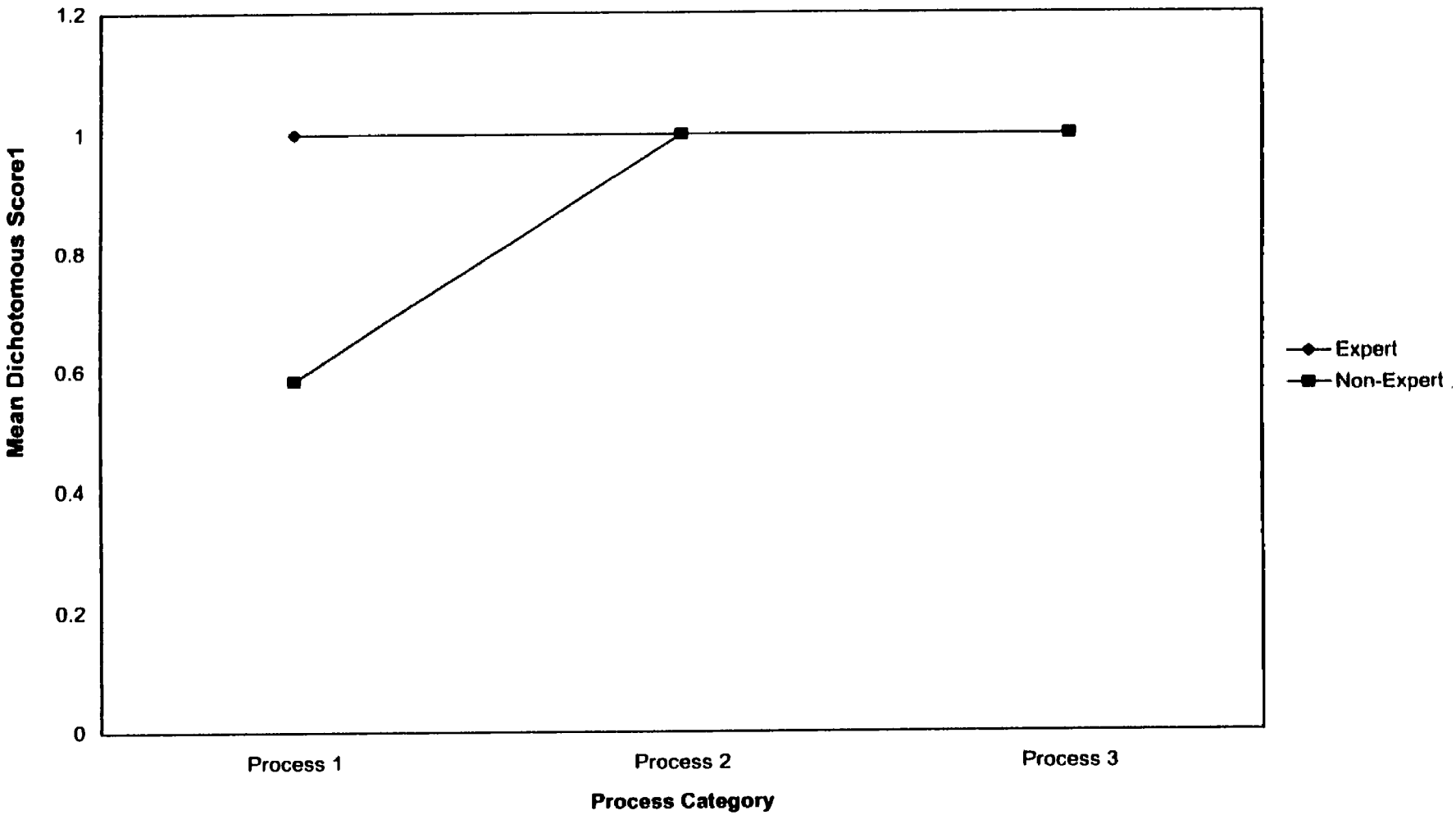


Figure 2j: Dysphagia Multiple-choice - Dichotomous Score1 by Process Category

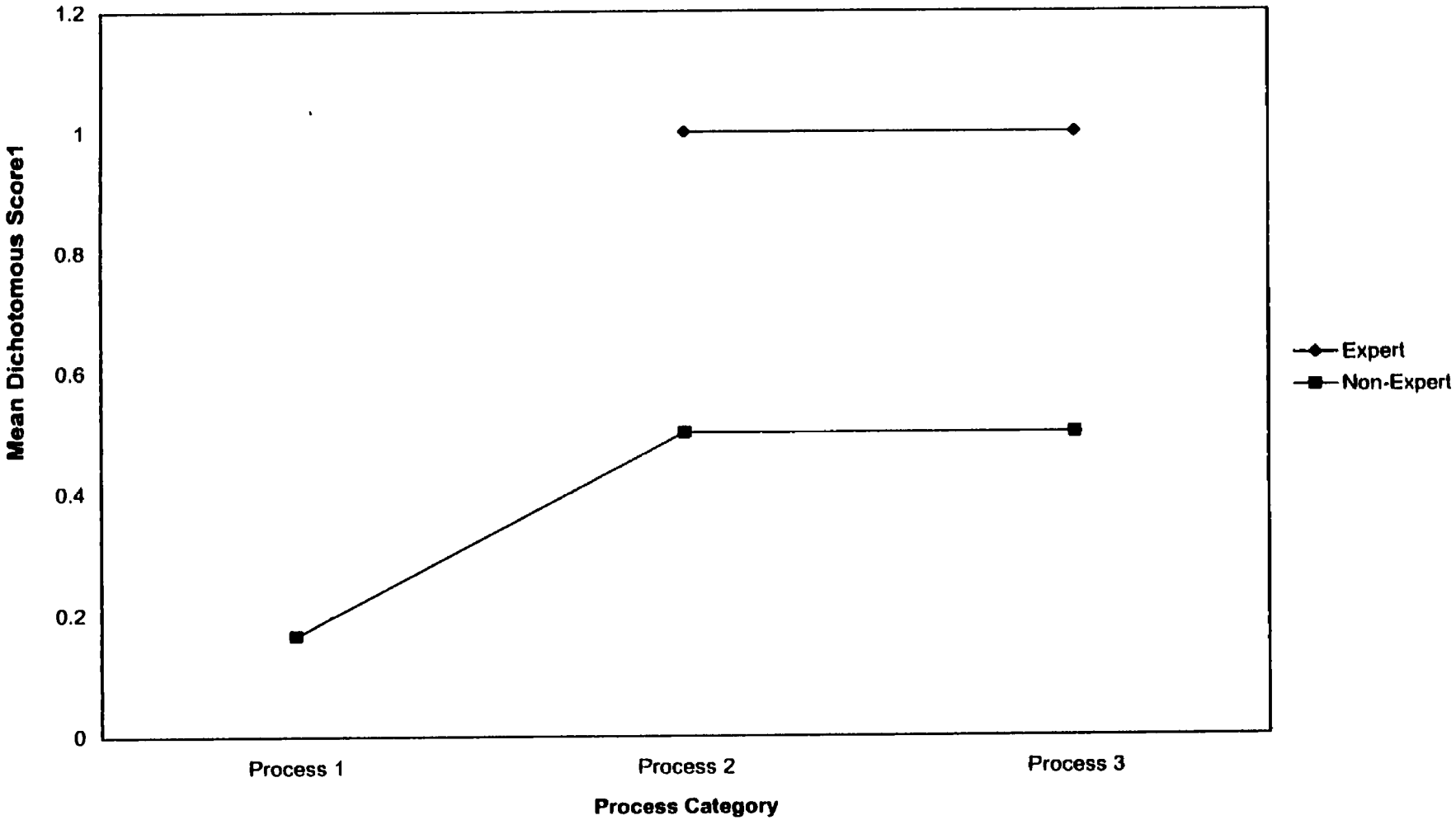


Figure 2k: Dysphagia Extended-matching - Dichotomous Score1 by Process Category

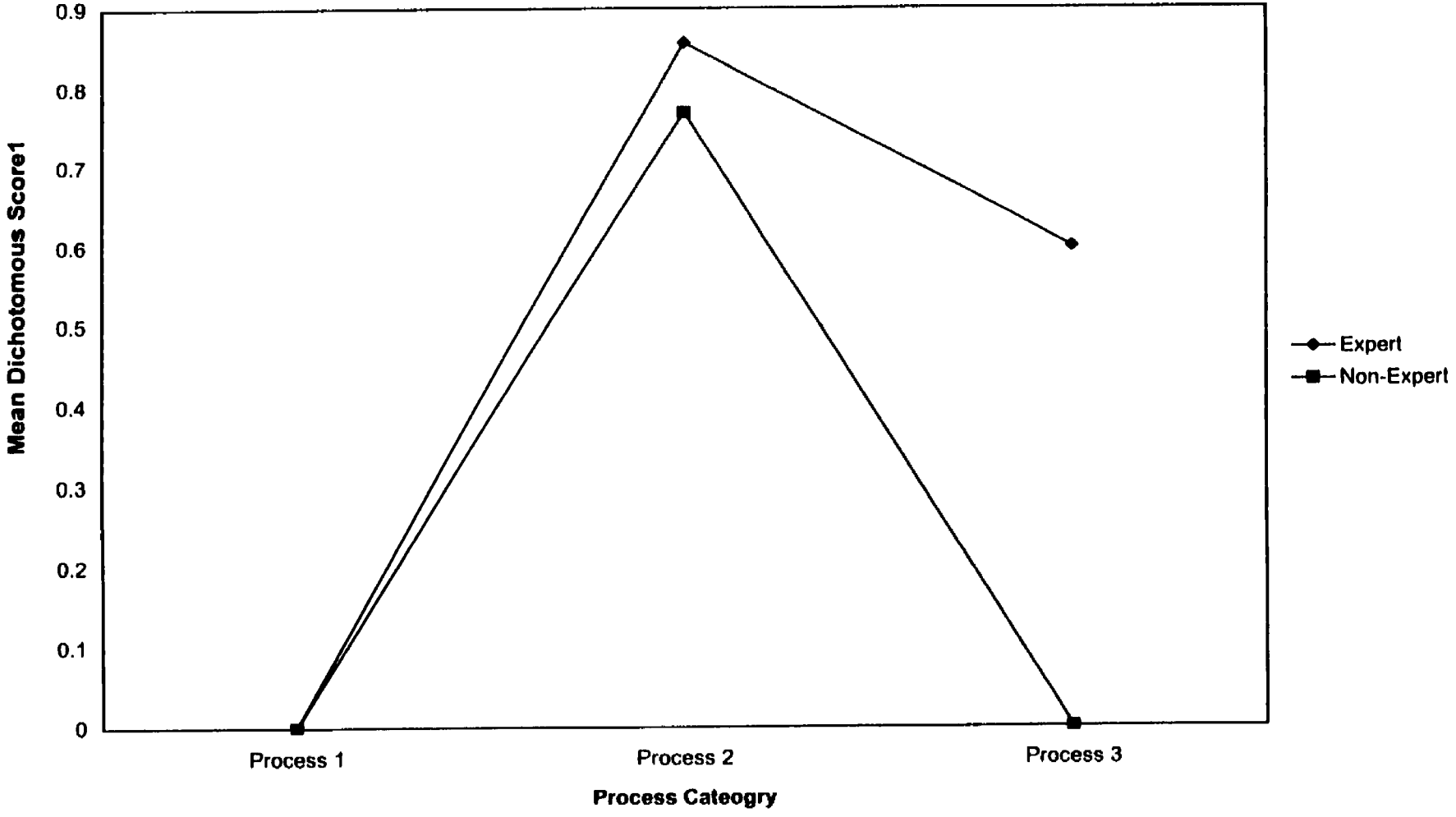
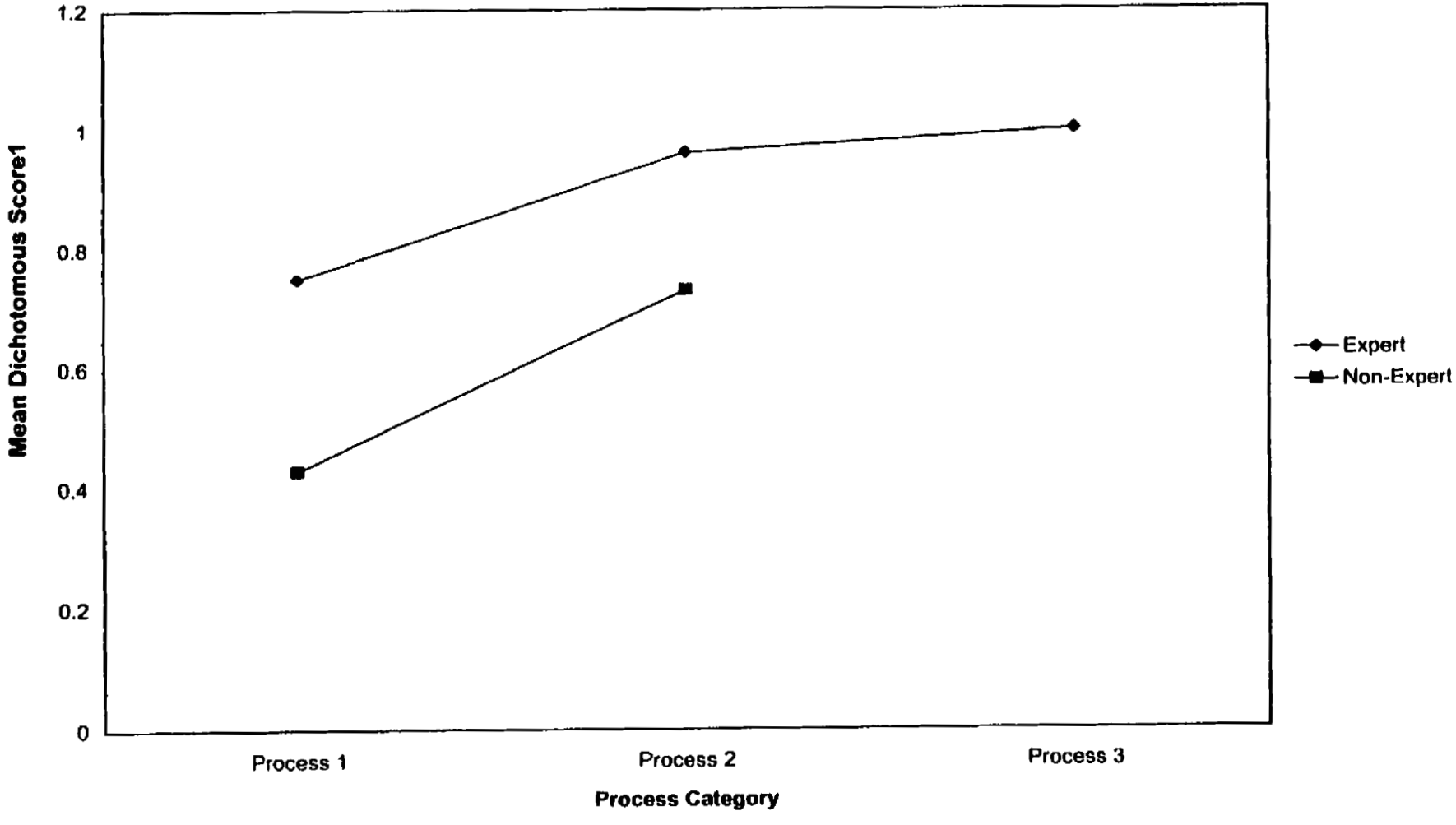


Figure 2II: Dysphagia Elimination - Dichotomous Score1 by Process Category



The trend for cognitive process correlating with examination success was confirmed in Table 5. where a two (experts/non-experts) by three (process score categories described above) by two (dichotomous scores 1 and 2) MANOVA was utilized to determine the presence or absence of effect. A "scheme" effect on examination scores was seen in five of the twelve cells.

Table 5: 2x3x2 MANOVA Comparing Expert and Non-expert Subjects and Scheme Use on Dichotomous Scores for All Formats and Clinical Presentations

	Elevated Liver Enzymes	Nausea & Vomiting	Diarrhea	Dysphagia
Multiple-choice	✔ Score Effect (.000) Expert Effect (.004) Score x Expert Interaction (.003)	✔ No Sig Effects	✔ No Sig Effects	✔ Score Effect (.001) Expert Effect (.002) Process Effect (.026) Score x Expert Interaction (.005)
Extended-matching	✔ No Sig Effects	✔ Process Effect (.004)	✔ Expert Effect (.000) Score Effect (.000) Expert x Score Interaction (.000)	✔ Process Effect (p=.001)
Elimination	✔ No Sig Effects	✔ Expert Effect (.034) Process Effect (.005) Expert x Process Interaction (.005)	✔ Expert Effect (.021) Process Effect (.034) Expert x Process Interaction (.034)	✔ Expert Effect (.043)

This therefore demonstrates that the process used, especially for non-experts, has an impact on examination success. Other variables were also demonstrated to impact examination success, including expertise and the specific question posed, the latter analogous to case specificity. The impact of expertise on examination scores, an intuitive notion, is also demonstrated in Table 5, with the expert effect seen in six of the twelve cells of the MANOVA, but also in Figures 1a-i, which demonstrate a relatively consistent superiority of experts across the questions in Dichotomous Scores 1 and 2. These same figures demonstrate the importance of the specific question in occasionally causing a very high non-expert success (ex: Fig 1d, N/V) or very low non-expert success (ex: Fig. 1d.diarrhea).

Results for Research question #2: Is there a measurable difference in the scheme utilization elicited by standard five-option Multiple-choice, Extended-matching and Elimination-type questions?

This question was answered using the process scores of each examinee, as agreed upon by the two judges after the think-aloud analysis, and averaging these scores for each format. Therefore, an average process score for each format (thus collapsed across the clinical presentations) was found for each examinee, and then a global mean was calculated for each of the formats for the twenty experts, then twenty non-experts. These global means for each format are found in Tables 6a and 6b. One-way ANOVA was performed for both the expert and non-expert groups, and no statistically significant

difference was found, suggesting that one format did not encourage scheme utilization more so than another format.

Table 6a: Process Score Means of Multiple Choice, Extended Matching, and Elimination-type Questions Collapsed over Clinical Presentation, Non-Expert Subjects (n=20); p value by one-way ANOVA.

Multiple-choice	Extended-match	Elimination-type	P value
.447	.469	.528	.196

Table 6b: Process Score Means of Multiple Choice, Extended Matching, and Elimination-type Questions Collapsed over Clinical Presentation, Expert Subjects (n=20); p value by one-way ANOVA.

Multiple-choice	Extended-match	Elimination-type	P value
.806	.823	.794	0.692

Therefore, the process score did not appear to be influenced by the examination format for both the experts and non-experts.

Two other potential influences on the process scores were then analyzed: the clinical presentation and specific question, or case. Tables 7a,b and 8a,b reveal that non-experts' process scores do vary significantly with the clinical presentation and specific case involved, while experts' scores are stable across these two factors.

Tables 7a and 7b present the findings for the influence of the clinical presentation on the process scores. For this analysis, a mean process score was initially found for each examinee in each of the four clinical presentations (thus collapsed over question format), from which a global mean for each of the four clinical presentations was calculated, for both experts and non-experts. A one-way ANOVA was performed to test for statistically significant differences. In the case of the non-experts, a statistically significant difference was found between the process score means across clinical presentations, with a post hoc Scheffe test demonstrating that the difference lies between the nausea/vomiting and diarrhea clinical presentations, and the nausea/vomiting and liver enzyme clinical presentations. A significant difference was not found in the expert group with regards to the clinical presentation impact on process scores.

Table 7a: Process Score Means of Clinical Presentation Collapsed over Question Format. Non-Expert Subjects (n=20): p value found by one-way ANOVA.

Liver Enzymes	Nausea and Vomiting	Diarrhea	Dysphagia	P value
.442	.646	.400	.550	0.001

Table 7b: Process Score Means of Clinical Presentation Collapsed over Question Format. Expert Subjects (n=20): p value found by one-way ANOVA.

Liver Enzymes	Nausea and Vomiting	Diarrhea	Dysphagia	P value
.833	.825	.813	.792	0.780

Tables 8a and 8b present the findings for the influence of the question on the process scores. in other words. the mean process scores of the three examination formats within each of the four clinical presentations. For this analysis. a global mean process score was found for all experts and all non-experts for each of the twelve questions. A 3x1 repeated measures was performed to test for statistically significant differences. In the case of the non-experts. a statistically significant difference was found between the process score means of the three formats within the clinical presentations of nausea and vomiting. elevated liver enzymes and diarrhea. A post hoc Scheffe test demonstrated that the differences lied within the Extended-matching and the Multiple-choice questions in the nausea/vomiting clinical presentation. and the Elimination-type question with the other two formats in the diarrhea clinical presentation. A significant difference was not found in the expert group with regards to the question or case impact on process scores.

Table 8a: Process Score Means of Multiple Choice. Extended Matching. and Elimination Questions within each Clinical Presentation. Non-Expert Subjects (n=20); p value by 3X1 repeated measures.

	Multiple-choice	Extended-match	Elimination-type	P value
Liver Enzymes	0.513	0.513	0.300	0.029
NauseaVomiting	0.400	0.813	0.725	0.011
Diarrhea	0.275	0.325	0.600	0.002
Dysphagia	0.613	0.563	0.488	0.187

Table 8b: Process Score Means of Multiple Choice, Extended Matching, and Elimination Questions within each Clinical Presentation. Expert Subjects (n=20): p value by 3X1 repeated measures.

	Multiple-choice	Extended-match	Elimination-type	P value
Liver Enzymes	0.888	0.838	0.775	0.302
NauseaVomiting	0.788	0.863	0.850	0.713
Diarrhea	0.738	0.900	0.800	0.151
Dysphagia	0.838	0.788	0.750	0.218

D) Results for Research question #4: Does a method of assessing problem solving exist that is more economical than the standard think-aloud analysis?

The "crying baby" method was abandoned early in the study, for a number of reasons outlined at the end of the "methods" section. The conclusion therefore at this point in time is that cognitive processes are much too complex to be analyzed by a self-assessment tool. Think-aloud remains the most reliable way of tapping into complex cognitive structures.

E) Other Results

Table 9a: Frequency Table for Process Score, Expert Subjects (n=20)

Question Format	Process Score	Liver	N&V	Diarrhea	Dysphagia
Multiple-Choice	0	0	3	2	0
	1	0	0	0	0
	2	0	1	3	0
	3	9	3	7	13
	4	11	13	8	7

Extended-Matching	0	1	2	0	0
	1	0	0	1	1
	2	1	0	0	0
	3	7	3	5	14
	4	11	15	14	5
Elimination-type	0	1	2	1	1
	1	0	0	0	1
	2	2	0	1	0
	3	10	4	10	13
	4	7	14	8	5

Table 9b: Frequency Table for Process Score. Non-Expert Subjects (n=20)

Question Format	Scheme	Liver	N&V	Diarrhea	Dysphagia
Multiple-choice	0	5	9	8	2
	1	1	3	5	4
	2	6	1	5	1
	3	4	1	1	9
	4	4	6	1	4
Extended-matching	0	4	2	5	2
	1	4	2	9	4
	2	3	0	3	2
	3	5	1	1	11
	4	4	15	2	1
Elimination-type	0	9	3	3	3
	1	3	3	3	4
	2	3	0	4	4
	3	5	1	3	9
	4	0	13	7	0

CHAPTER FIVE: DISCUSSION

This study by its analysis of forty think-aloud protocols, has provided intriguing quantitative and qualitative information. To summarize the quantitative findings, in relation to each research question, this study found:

i) Research question #3: What are the psychometric properties of the three question formats used in the study, both with a standard dichotomous scoring of 0-1, and a partial dichotomous scoring system? (the latter gives a partial mark for a wrong final answer that is, however, in the correct "arm" of a scheme).

Psychometric properties (reliability, discrimination, construct validity) of the three question formats were shown to be similar and quite acceptable (Figures 1a-i, Tables 4a-c).

ii) Research question #1: Is there a correlation between the reasoning process used in problem solving, specifically scheme utilization, and diagnostic success?

Several factors influence the examination (dichotomous) scores:

i) the cognitive process (i.e. problem solving strategy) used (Figures 2a-l, Table5)

ii) level of expertise (Figure 1a-i, Table 5)

iii) examination question or case (Figure 1a-i) (the effect of the clinical presentation as well as examination question format).

iii) Research question #2: Is there a measurable difference in the scheme utilization elicited by standard five-option Multiple-choice, Extended-matching and Elimination-type questions?

There is no significant difference in the process scores (i.e. problem solving strategy) in the three types of question tested. This is true for both the experts and non-experts (Tables 6a, 6b). In summary, there exists no overall impact of question format on scheme utilization.

Two other factors did however appear to impact scheme utilization in non-experts. Significant differences were found in the process scores (i.e. problem solving strategy) of non-experts across the four clinical presentations (Tables 7a, 7b). Significant differences were also found in the process scores of non-experts, across the three formats tested, when analyzed within each clinical presentation (Tables 8a, 8b). These results suggest that for non-experts, the clinical presentation and the specific question (or case) posed have an impact on whether a scheme or some other problem solving strategy is utilized. Experts' selection of problem solving strategy is similar across all clinical presentations and questions or case. These differences may be explained by differences in the teaching of the schemes to the students for each clinical presentation, with some teachers emphasizing schemes more than others, and thus students learning some schemes and not others. Experts, on the other hand, have organized their knowledge into the scheme framework across all clinical presentations and thus remained uninfluenced in their scheme use by the specific clinical presentation or case.

A) Discussion Surrounding Research Question #3

The first point of discussion relates to the important psychometric properties of the three examination formats. All three formats showed good reliability, with a slight advantage to the Multiple-choice format. The partial scoring method, attributing a partial score for incorrect answers that were in the correct branch of a scheme, did not improve reliability. Experts achieved generally higher scores than non-experts across all formats, thus indicative of construct validity of all three formats (Figures 1a-i). It is interesting to note that the scheme use score achieved a reliability of .83. This is indicative of some degree of consistency in the process, such as schemes, which the examinees use to answer the twelve questions. This also demonstrates that such a scale can potentially distinguish between the top and bottom students, presumably by distinguishing between students who use strong vs. weak methods of problem solving.

The discrimination indices in Table 4c were also quite high for all three formats, for the dichotomous score, with a slight advantage once again to the Multiple-choice format. It is interesting to note that the discrimination index for the process score (i.e. clinical reasoning strategy) was in general higher than the index for the dichotomous score, demonstrating the capability of this score to distinguish the top from the bottom students, by analyzing and scoring their clinical reasoning.

B) Discussion surrounding Research Question #1

Another aim of the study was to examine whether there is any correlation between examination success and the use of schemes. Intuitive evidence for a positive correlation

is evident in Figures 2a-l. which certainly demonstrate a correlation between examination success and process scores. especially in the case of non-experts. There is an increase in dichotomous scores as the process scores increase. The lines for the experts are generally straighter. reflective of the general success rate for the experts in the examination. The Extended-matching. diarrhea case is an exception only because among the non-experts. all but one answered incorrectly. The Extended-matching dysphagia case also differs for the non-experts. but only because the single student who used pattern recognition for that case answered incorrectly.

When further analyzed. using the MANOVA in Table 5. a scheme or process score effect becomes manifest in five of the twelve questions. This implies the clear presence of cognitive process as an influence on the variance of the dichotomous score means. and thus corroborating the trend seen in Figure 2a-l. There is a positive correlation between scheme utilization and examination success.

Other factors were also demonstrated to influence examination dichotomous scores. including expertise and the specific question. or case (=case specificity). Table 5 demonstrates that as would be expected. expertise seems to play an important role in the variance of the dichotomous examination success. Figures 1a-i demonstrate a trend for superior results for experts across all the questions. with a narrow difference in the nausea and vomiting presentation which is explained in part by the relative ease of these questions for the non-experts. in that they were general questions rather than sub-specialized to Gastroenterology. Case specificity is shown in Figure 1a-i which show distinctive success in the non-expert group with certain cases. such as the nausea and

vomiting case of diabetic ketoacidosis (Figure 1d), and the distinctive failure of the non-expert group in certain cases, such as the diarrhea bacterial overgrowth case (Figure 1d). The concept of case specificity is complex, and within it differences may be attributed to the clinical presentation of the case, perhaps the format presented, as well as the specifics of the stem construction. This stem construction includes carefully placed and reasonable distracters that, interestingly in our study, never swayed the experts but did sway some non-experts into wrong diagnoses.

C) Discussion surrounding Research Question #2

One of the principal goals of this research protocol was to investigate whether the three examination formats used in this study, the standard five-option Multiple-choice, Extended-matching and Elimination-type questions, elicited different cognitive processes (i.e. problem solving strategy), specifically in regards to scheme utilization. The observation from the data is that all three formats, when constructed with problem solving tasks in mind, can in fact evoke higher levels of cognitive processes such as chunking or scheme utilization. Evidence for this lies in the process scores for the three formats, which overall averaged to 3.27 for the experts (equivalent to the use of well-structured schemes according to the process score scale in Table 2) and 2.04 for the non-experts (equivalent to using a structured scheme in combination with another method). These scores provide evidence that pencil-and-paper tests can be used to test problem solving tasks. Furthermore, these three formats are capable of evoking strong methods of

problem solving, such as pattern recognition and scheme utilization, as demonstrated by the generally high process scores.

These process scores (reflecting problem solving strategy), combined with the underlying detailed analysis of think-aloud protocols, provide both further qualitative and quantitative evidence that differences exist between experts and non-experts in the processes of problem solving utilized. The data is contrary to statements recently made in the literature[15] but consistent with the view that experts differ from non-experts not only in the quality of their knowledge bases in memory, but also in the processes of problem solving utilized. Experts use strong methods of problem solving such as scheme utilization and pattern recognition. This difference in problem solving process holds true across three examination formats and four clinical presentations (Figure 1a-i). Tables 9a and 9b also demonstrate that the experts had a strong predilection for the scheme and pattern recognition problem solving methods, and only infrequently used weaker methods such as hypothetico-deductive reasoning. Non-experts on the other hand revealed a more even spread across the problem solving methods. It is interesting to note that the experts who did use weak methods did so in the nausea and vomiting case. This case, with a final diagnosis of a metabolic condition, was not directly related to the domain of Gastroenterology, the domain where the gastroenterologists are truly experts.

In the same clinical presentation, nausea and vomiting, novices used a disproportionate amount of pattern recognition. This may be at least in part attributed to the fact that the questions were relatively easier in that clinical presentation. The more interesting possibility is that the two judges of the think-aloud process may have

misinterpreted the process used. Retrospectively, instead of labeling the process as "pattern recognition" (with a process score of '4'), perhaps a more appropriate label might have been "novice pattern recognition" (with a process score of '1'). This mislabeling may have altered some of the study's results. These labels will be further discussed below.

When the three question formats were directly compared with respect to whether schemes were being promoted, as reflected by the mean process score (i.e. problem solving strategy) collapsed over all four clinical presentations, no differences were found for both the expert and non-expert group (Tables 6a and 6b). Explanation of this result may be found in the view raised by several authors [45,55] that it is not the examination format itself that dictates the cognitive level of the testing, but rather the specific construction of the question stems. We have demonstrated that a well-constructed Multiple-choice question, designed specifically to target problem solving, can achieve this purpose as well as any other format currently available. Critics of the Multiple-choice format, who believe that it only tests recall of isolated facts, need to consider altering the construction of the stems rather than the format. On the other hand, several expert and non-expert examinees did comment that Extended-matching questions made it more difficult to go through the list of options prior to answering the question. One of the goals of a pencil-and-paper test is to challenge the examinee to answer the question by solving the problem described in the stem, prior to considering the options available. Based upon the subjective comments of the examinees, the Extended-matching and Elimination-type

formats, because of the inherent difficulty of reading through an extended option list, appeared to provide a better challenge than the Multiple-choice format.

While the examination format did not appear to influence the examinees' process scores and thus their problem solving strategy, two other factors did appear to exert an influence on the process scores of the non-expert group. The first factor that appears to exert an influence on the process scores of the non-experts is the clinical presentation, as found in Table 7a. The mean process scores in the nausea and vomiting clinical presentation were significantly higher than the diarrhea and elevated liver enzymes clinical presentations. This appears to indicate an influence of the clinical presentation on non-expert problem solving strategy, but not in the expert group.

The second factor influencing non-expert process scores was the specific question or case. In Table 8a, process scores were compared for the three formats within each of the clinical presentations, and statistically significant differences were found between the three formats for three of the clinical presentations. In the elevated liver enzyme presentation, the Elimination-type format had a lower process score than the other two formats. Scheffe post hoc testing (appendix E) demonstrated that the Extended-matching had significantly higher process scores than the Multiple-choice format in the nausea and vomiting presentation, while Elimination-type had significantly higher scores than the other two formats in the diarrhea presentation. These differences are likely explained by the well-described phenomena of case specificity. However, these differences of the examination formats within each clinical presentation raises the possibility of differences

existing between the three examination formats, despite the results presented in Tables 6a and 6b.

D) Other Points of Discussion

Some observations from this study can be made with regards to the linking of the stage theory of knowledge structuring, as proposed by Schmidt and Bordage, with problem solving strategy. Work by Bordage and Schmidt has categorized the evolution of knowledge organization and diagnostic reasoning according to the progressive increase in training from novice to expert.[62,63] Bordage's first category was called *reduced knowledge*, meaning that very little knowledge is available and thus no real problem solving exists. If some knowledge exists but is poorly organized, it is termed *dispersed knowledge*. The next category is called *elaborated knowledge*, and is found in the work of both Bordage and Schmidt. This knowledge network is a rich and elaborate causal network, with considerable reliance on pathophysiological explanations. The *compiled stage*, also present in both papers, is a more advanced stage, achieved through extensive and repeated application of acquired knowledge. This stage contains the elaborate knowledge base, which becomes compiled into simplified causal models that contain only higher-level concepts. Schmidt describes two additional stages, *illness and instance scripts*. These two levels are richly embedded with clinical experience and specific cases in memory, similar in concept to Bordage's "prototypes"[64]. These prototypes are an expert way of categorizing knowledge in memory around key cases or clear examples. It is understood that all of these knowledge structures can exist at the same time in the memory of a single individual.

Observations from the study's qualitative data lend themselves to our hypothesis that the evolution of knowledge-structure from novices to experts is associated with a corresponding evolution of problem solving strategies. Figure 1a-l demonstrate that in general, experts' process scores are generally quite high and higher than those of the non-experts. The non-experts in the study were at the stage of final year medical students who had actually completed their clerkship but were in the process of studying for their licensing examination. This is an interesting group to study given that they do have some clinical experience but not enough to be classified into Bordage's "compiled" category, which aims to a higher level of clinical experience. Therefore, the best classification for this group is the "elaborated" stage, which may be expected to demonstrate a somewhat intermediate stage of clinical reasoning, with use of a mixture of hypothetico-deductive and scheme utilization. Subjectively this group in fact did demonstrate such a mixture, which is reflected in their frequency distribution in Table 9b and overall mean scheme score of 2.04, a value indicative of some scheme utilization in conjunction with hypothetico-deductive reasoning. This intermediate stage best fits the label of using "primitive schemes". The experts would be considered at the compiled/script levels, and in fact demonstrated both scheme use and pattern recognition as major problem solving strategies in Table 9a, as reflected by an overall mean scheme score of 3.24. These findings raise the possibility of a defined relationship between knowledge-structure and problem solving strategy. It must be emphasized these are only observations leading to theory at this point, as the study was not directly designed to clearly establish the knowledge structure with problem solving strategy relationship.

Table 10: Relationship of Knowledge Structure with Problem Solving Strategy

Knowledge structure	Problem solving strategy
Reduced	Little or no problem solving
Dispersed	Hypothetico-deductive
Elaborated	Primitive Scheme
Compiled	Scheme
Illness scripts and instance scripts	Pattern recognition

The left-hand portion of Table 10 is a compiled version of work by Bordage and Schmidt [62,63]. The right-hand portion of Table 10 is a proposed model based in part on the findings of this paper and previous work by Mandin[65].

The next point to be extracted from this study is also a qualitative one, and a feature of problem solving which was observed in our 40 think-aloud protocols, which was not apparent to us at the beginning of the study. From listening to our subjects' verbalizations of their problem solving process, it appears that pattern recognition occurs in two very distinct forms, which we will label "expert pattern recognition" (EPR) and "novice pattern recognition" (NPR). EPR, when examined deeper by our think-aloud, appeared distinctly as a very rapid utilization of a scheme converging to a very certain diagnosis. On the contrary, NPR was clearly not supported by a scheme structure, but was merely a sketchy and uncertain attempt at piecing together a few pieces of

information. The best example for this was frequently provided by Question #5, a case of a patient with apparent ulcerative colitis (UC), who presents with a cholestatic liver enzyme profile. Universally, experts jumped at the diagnosis of primary sclerosing cholangitis (PSC), a disease with cholestatic liver enzyme elevation that is an extraintestinal manifestation of inflammatory bowel disease. Even though the experts jumped at this diagnosis, when asked to explain their answer further they then invariably took us through the steps of the classic scheme for elevated liver enzymes, that is cholestatic vs. hepatocellular, extrahepatic vs. intrahepatic cholestasis based on ultrasound, etc. They had essentially, without realizing it, gone through this algorithm with the additional piece of information of the ulcerative colitis providing a more rapid route to the end diagnosis. A large part of this form of pattern recognition lies in the clinical experiences of the experts in this area, which allows for additional rapid association of the written case with real life cases, a process described by Schmidt as "instance scripts". EPR therefore is a rapid progression through a scheme, aided by these "instance scripts", a concept analogous to the semantic network which bridges experience with facts in a very rich manner in experts.

NPR, on the other hand is not supported by the underlying scheme or clinical experience when examined at a deeper level. It is a very superficial linking of the known association between UC and PSC, and is frequently not reasoned any deeper than this or even a very certain diagnosis, as opposed to the experts who are invariably absolutely certain of their diagnosis. This argument is analogous to the discussions occurring in the literature between Regehr and Patel [66,67]. Regehr's commentary related to his experience

of viewing a champion tic-tac-toe player who was in fact a chicken, and raised potential analogies between medical expertise and the poultry tic-tac-toe expert. Patel argued that tic-tac-toe problem solving, in contrary to the more complex medical problem solving, is a well-structured task with self-evident options. Both sides of this debate provide some potential insight into our proposed separation of expert from novice pattern recognition. The chicken, although a tic-tac-toe grandmaster, can be argued to be recognizing patterns and reacting in a rapid, unreflecting manner, and not by carefully exploring possibilities at a deeper level. The NPR is in our experience, analogous to the way in which this chicken superficially recognizes patterns without a deeper algorithm or approach, in contrary to what medical experts do when they pattern recognize cases. Another way of looking at it is that NPR consists of, as Patel describes, "low-road transfers" which are essentially automated behaviors, while EPR consists of "high-road transfers", which occur by intentional mindful abstraction. This being said, it must be noted that on occasion, so-called novices can still demonstrate expert pattern recognition, as clearly exemplified by one examinee who, in Question #2, pattern recognized multiple myeloma in a confident, expert way, based largely on a very similar clinical scenario she had experienced. Likewise some experts demonstrated novice pattern recognition, especially when our Gastroenterologists dealt with the nausea and vomiting scenarios, answers to which were essentially metabolic causes and outside of their immediate domain of expertise. Perhaps an earlier recognition of this phenomena may have altered our data, by actually assigning a process score of '1' rather than a '4' to the "apparent" pattern recognition displayed by some novices and on occasion experts.

This area of pattern recognition types opens the door for consideration of future studies. Firstly, a qualitative project assessing this distinction between NPR and EPR would provide useful knowledge with regards to whether this is a real phenomenon, and whether EPR truly does reflect a fast progression through an underlying expert scheme. A similar think-aloud protocol could be used with pencil-and-paper testing, with specific construction of questions aimed at a description of prototypical cases that both novices and experts could pattern recognize relatively easily. A detailed think-aloud of underlying deeper structures or schemes would have to be performed to get at whether this distinction is in fact a real phenomenon. If the distinction between EPR and NPR does exist, then our study could be repeated with this distinction in mind, in which case as per our process scale EPR would be a '4', while NPR would be a '1'. Such a clearer and improved scaling system may allow a clearer correlation between process and examination scores, and could in fact reveal some differences between various examination formats. Along those lines, we have come to the realization that the Elimination-type format, by asking to eliminate a specific diagnosis, may in fact encourage hypothetico-deductive or NPR type of reasoning. These questions, in further study, should be modified by asking the examinee, for example, "Why did you eliminate the motility causes of dysphagia". This, and perhaps the creation of newer formats, may succeed to a greater extent at matching examination questions to our curricular structure.

E) Study Limitations and Final Summary

There are a number of limitations of this study that need to be pointed out. Firstly, as already indicated, the observers of the think-aloud may have mistakenly labeled novice

pattern recognition as a '4' in the process scale, which may have modified the results. Secondly, the think-aloud judges arrived at their process score by mutual agreement, and thus the initial individual scores of the judges were not used or analyzed. These judges were not blinded to the examinees' final diagnosis, which does have the potential to alter their view of the cognitive process in think-aloud. Also, though the stems designed for the three formats in each clinical presentation were almost identical, there were subtle but potentially important differences in the key pieces of information in these questions, which could alter the results, especially the comparisons between formats. Finally, the positive results seen with scheme utilization and increased diagnostic success need to be clearly interpreted as a correlation and not causal effect at this point. There may be other hidden factors that could very well explain this positive effect.

Therefore, in summary, the goal of this paper was to explore the field of clinical reasoning in medicine, with an initial presentation of the literature behind the University of Calgary's adoption of a clinical presentations curriculum and their associated schemes. For this curriculum to succeed, a method of evaluation that specifically tests for the utilization of schemes needs to be found. Although our three formats did not differ significantly, it is encouraging to note that all three formats did in fact achieve a measure of higher cognition and specifically scheme utilization testing. The formats also achieved acceptable psychometric properties. Also encouraging is that the traditional, efficient and reliable five-option Multiple-choice format, when carefully constructed, can achieve the desired goal of testing for scheme utilization and therefore will remain the method of

choice in our curriculum. with perhaps more experimenting with the other two formats or even development of newer strategies.

Scheme utilization for knowledge representation and problem solving remains well founded in cognitive psychology. supported in the literature. and should be continued as a potentially very powerful means of education for both undergraduate and postgraduate training. This study provides direct evidence of a correlation between examination success and use of schemes in problem solving. This study also provides added evidence that experts do in fact use schemes and strong pattern recognition instead of hypothetico-deductive reasoning. and that non-experts also use schemes but not as strongly as the experts. Non-experts. as opposed to experts. are influenced by the clinical presentation and specific case in their scheme utilization. For experts. scheme utilization is almost universal. but we cannot ignore that experience is a strong component of expert problem solving. and in a few examples our experts chose diagnoses that were contrary to their actual scheme. because of recent clinical experiences. Scheme utilization is not necessarily always a clear-cut phenomenon given the complexities and variability of human character and personal experiences. Therefore. in presenting schemes in a curriculum. it is important to allow flexibility in the schemes presented and encourage personal modifications of these schemes. given that medical students can benefit from this cognitive elaboration of their own schemes. as well as could potentially resent being "told what to do". This was exemplified best to us by one of our student examinees. who at the very beginning of the think-aloud said rather defiantly " I don't use schemes!" However. as we proceeded with the think-aloud. it became apparent to us that she did in

fact use schemes. When this was pointed out to her, she said "well I don't use **your** schemes", which was in fact true. Therefore, although the approach of teaching schemes is sound and advantageous in our opinion, it needs to be accomplished with careful consideration of properly evaluating for scheme use, as well as taking into account the need for personal decision-making by the students in their utilization.

REFERENCES

1. Mandin. H.. Harasym. P.. Eagle. C. Watanabe. M. Developing a "Clinical Presentation" Curriculum at the University of Calgary. Academic Medicine. **70** (1995): 186-193.

2. Regehr. G. and Norman. G. Issues in Cognitive Psychology: Implications for Professional Education. Academic Medicine. **71** (1996): 988-1001.

3. Kushniruk. A.. Patel. V.. Marley. A. Small Worlds and Medical Expertise: Implications for Medical cognition and Knowledge Engineering. International Journal of Medical Informatics. **49** (1998): 255-271.

4. Mandin. H.. Jones. A.. Woloschuk. W.. Harasym. P. Helping Students Learn to Think Like Experts When Solving Clinical Problems. Academic Medicine. **72** (1997): 173-193.

5. Suter. E.. Mandin. H.. Small. P. The Sciences in the Education of Physicians. Basic Science Educator. **8** (1998): 7-9.

6. West. C.. Farmer. J.. Wolff. P. *Instructional Design: Implications for Cognitive Science*. Prentice Hall. Englewood Cliffs. New Jersey. 1991. P1-35.

7. Mandin. H.. Woloschuk. W.. Tarrant. M.. Harasym. P. Problem Solving During a Family Medicine Clerkship. Advances in Medical Education. (1997): 622-624.

8. Bordage. G. and Lemieux. M. Semantic Structures and Diagnostic Thinking of Experts and Novices. Academic Medicine. **66** (1991 supp): S70-S72.

9. O'Neill. P.. Metcalfe. D.. David. T. The Core Content of the Undergraduate Curriculum in Manchester. Medical Education. **33** (1999): 121-129.

10. Harden. R.. Davis. M.. Crosby. J. The New Dundee Medical Curriculum: a Whole That is Greater Than the Sum of the Parts. Medical Education. **31** (1997): 264-271.

11. LaDuca. A. Validation of Professional Licensure Examinations. Evaluation and the Health Professions. **17** (1994): 178-197.

12. Schmidt. H. Foundations of Problem-Based Learning: Some Explanatory Notes. Medical Education. **27** (1993): 422-432.

13. Newble. D.. Dawson. B.. Page. G.. Dauphinee. D.. Macdonald. M.. Mulholland. H.. Swanson. D.. Thomson. A.. van der Vleuten. C. Guidelines for Assessing Clinical Competence. Teaching and Learning in Medicine. **6** (1994): 213-220.

14. Norman. G.. Tugwell. P.. Feightner. W.. Muzzin. L.. Jacoby. L. Knowledge and Clinical Problem-solving. Medical Education. **19** (1985): 344-356.

15. Eva. K.. Neville. A.. Norman. G. Exploring the Etiology of Content Specificity: Factors Influencing Analogic Transfer and Problem Solving. Academic Medicine. **73** (1998 supp.): S1-S5.

16. Grant. J. and Marsden. P. Primary Knowledge, Medical Education and Consultant Expertise. Medical Education. **22** (1988): 173-179.

17. Bordage. G.. Grant. J.. Marsden. P. Quantitative Assessment of Diagnostic Ability. Medical Education. **24** (1990): 413-425.

18. Patel. V.. Evans. D.. Groen. G. Biomedical Knowledge and Clinical Reasoning. *Cognitive Science in Medicine.* A Bradford Book. 1989. P53-111.

19. Coughlin. L. and Patel. V. Processing of Critical Information by Physicians and Medical Students. Journal of Medical Education. **62** (1987): 818-828.

20. Schmidt. H. Problem-based Learning: Rationale and Description. Medical Education. **17** (1983): 11-16.

21. Gagne. E. *The Cognitive Psychology of School Learning.* Little Brown and Company. Boston. 1995. P 83.

22. Zemke. R. and Zemke. S. Adult Learning: What Do We Know For Sure? Training. June 1995: 31-40.

23. Griffin. V. Holistic Learning/Teaching in Adult Education: Would You Play a One-string Guitar? *The Craft of Teaching Adults.* Culture Concepts Inc. Toronto. 1988. P105-129.

24. Berkson. L. Problem-Based Learning: Have the Expectations Been Met? Academic Medicine. **68** (1993 supp): S79-S88.

25. Woloschuk. W.. Harasym. P.. Mandin. H.. Jones. A. Use of Scheme-based Problem Solving: an Evaluation of the Implementation and Utilization of Schemes in a Clinical Presentation Curriculum. Medical Education. **34** (2000): 437-442.

26. Woloschuk. W.. Harasym. P.. Mandin. H. Implementing a Clinical Presentation Curriculum: Impact on Student Stress and Workload. Teaching and Learning in Medicine. **10** (1998): 44-50.

27. Papa. F. and Harasym. P. Medical Curriculum Reform in North America. 1765 to the Present: a Cognitive Science Perspective. Academic Medicine. **74** (1999): 154-164.

28. Patel. V.. Groen. G.. Norman. G. Effects of Conventional and Problem-based Medical Curricula on Problem solving. Academic Medicine. **66** (1991): 380-389.

29. Elstein. A.. Shulman. L.. Sprafka. S. *Medical Problem-Solving: An Analysis of Clinical Reasoning.* Harvard Univ. Press. 1978. P.250.

30. Groen. G. and Patel. V. Medical Problem-solving: Some Questionable Assumptions. Medical Education. **19** (1985): 95-100.

31. Kassirer. J. Teaching Clinical Medicine by Iterative Hypothesis Testing: Let's Preach What We Practice. New England Journal of Medicine. **309** (1983): 921-923.

32. Neufeld. V.. Norman. G.. Feightner. J.. Barrows. H. Clinical Problem-solving by Medical Students: a Cross-sectional and Longitudinal Analysis. Medical Education. **15** (1981): 315-321.

33. Chase. W. and Simon. H. Perception in Chess. Cognitive Psychology. **1** (1973): 55-81.

34. Larkin. J.. McDermott. J.. Simon. D.. Simon. H. Expert and Novice Performance in Solving Physics Problems. Science. **208** (1980): 1335-1342.

35. Patel. V.. Groen. G.. Frederiksen. C. Differences Between Medical Students and Doctors in Memory for Clinical Cases. Medical Education. **20** (1986): 3-9.

36. Grant. J. and Marsden. P. The Structure of Memorized Knowledge in Students and Clinicians: an Explanation for Diagnostic Expertise. Medical Education. **21** (1987): 92-97.

37. Joseph. G. and Patel. V. Domain Knowledge and Hypothesis Generation in Diagnostic Reasoning. Medical Decision-Making. **10** (1990): 31-46.

38. Patel. V. and Groen. G. Knowledge Based Solution Strategies in Medical Diagnosis. Cognitive Science. **10** (1986): 91-116.

39. Newell. A. Artificial Intelligence and the concept of mind. *Computer Models of Language and Thought.* Freeman. Oxford. 1973. P 52.

40. Norman. G.. Trott. A.. Brooks. L.. Kinsey. E.. Smith. M. Cognitive Differences in Clinical Reasoning Related to Postgraduate Training. Teaching and Learning in Medicine. **6** (1994): 114-120.

41. Mandin. H. and Harasym. P. Psychometric Properties of "Exclusion" Questions for Assessment of Clinical Problem Solving. *Annual Meeting of the Association for Medical Education in Europe.* (1998).

42. Page. G. and Bordage. G. The Medical Council of Canada's Key Features Project: A More Valid Written Examination of Clinical Decision-making Skills. Academic Medicine. **70** (1995): 104-110.

43. Case. S. and Swanson. D. Extended-Matching Items: A Practical Alternative to Free-response Questions. Teaching and Learning in Medicine. **5** (1993): 107-115.

44. Norman. G. Striking the Balance. Academic Medicine. **69** (1994): 209-210.

45. Van der Vleuten. C. and Newble. D. How Can We Test Clinical Reasoning? The Lancet. **345** (1995): 1032-1034.

46. Elstein. A. Beyond Multiple-choice Questions and Essays: The Need for a New Way to Assess Clinical Competence. Academic Medicine. **68** (1993): 244-249.

47. Newble. D. A comparison of Multiple-choice and Free-response Tests in Examination of Clinical Competence. Medical Education. **13** (1979): 263-268.

48. Fenderson. B.. Damjanov. I.. Robeson. M.. Veloski. J.. Rubin. E. et al. The Virtues of Extended Matching and Uncued Tests as Alternatives to Multiple Choice Questions. Human Pathology. **28** (1997): 526-532.

49. Page. G.. Bordage. G.. Allen. T. Developing Key-feature Problems and Examinations to Assess Clinical Decision-Making Skills. Academic Medicine. **70** (1995): 194-201.

50. Case. S. and Swanson. D. Comparison of Items in Five-option and Extended-matching Questions in Assessing Diagnostic Skills. Academic Medicine. **69** (1994 supp): S1-S4.

51. Gruppen. L.. Grum. C.. Fincher. R.. Parenti. C.. Cleary. L.. Swaney. J.. Case. S.. Swanson. D.. Woolliscroft. J. Multi-site Reliability of a Diagnostic Pattern-recognition Knowledge-assessment Instrument. Academic Medicine. **71** (1994 supp): S65-S67.

52. Dunn. M. and Woolliscroft. J. Assessment of a Pattern-recognition Examination in a Clinical Clerkship. Academic Medicine. **69** (1994): 683-684.

53. Blackwell. T.. Ainsworth. A.. Dorsey. N.. Callaway. M.. Rogers. L.. Collins. K. A comparison of Short-answer and Extended-matching Question Scores in an OSCE. Academic Medicine. **69** (1991 supp): S40-42.

54. Solomon. D.. Speer. A.. Perkowski. L.. DiPette. D. Evaluating Problem Solving Based on the Use of History Findings in a Standardized-patient Examination. Academic Medicine. **69** (1994): 754-757.

55. Maguire. T.. Shakun. E.. Harley. C. Setting Standards for Multiple-choice Items in Clinical Reasoning. Evaluation and the Health Professions. **15** (1992): 434-452.

56. Boshuizen. H.. van der Vleuten. C.. Schmidt. H.. Machiels-Bongaerts. M. et al. Measuring Knowledge and Clinical Reasoning Skills in a Problem-based Curriculum. Medical Education. **31** (1997): 115-121.

57. Shakun. E.. Maguire. T.. Cook. D. Strategy Choices in Multiple-choice Items. Academic Medicine. **69** (1994 supp): S7-S9.

58. Fisher. A. and Fonteyn. M. An Exploration of an Innovative Methodological Approach for Examining Nurses' Heuristic Use in Clinical Practice. Scholarly Inquiry for Nursing Practice. **9** (1995): 263-276.

59. Ryan-Wenger. N. and Lee. J. The Clinical Reasoning Case Study: A Powerful Teaching Tool. The Nurse Practitioner. **22** (1997): 66-70.

60. Davison. G.. Vogel. R.. Coffman. S. Think-aloud Approaches to Cognitive Assessment and the Articulated Thoughts in Simulated Situations Paradigm. Journal of Consulting and Clinical Psychology. **65** (1997): 950-958.

61. Fonteyn. M.. Kuipers. B.. Goebe. S. A Description of Think Aloud Method and Protocol Analysis. Qualitative Health Research. **3** (1993): 430-441.

62. Bordage. G. Elaborated Knowledge: A Key to Successful Diagnostic Thinking. Academic Medicine. **69** (1994): 883-885.

63. Schmidt. H.. Norman. G.. Boshuizen. H. A Cognitive Perspective on Medical Expertise: Theory and Implications. Academic Medicine. **65** (1990): 611-621.

64. Bordage. G. and Zacks. R. The Structure of Medical Knowledge in the Memories of Medical Students and General Practitioners: Categories and Prototypes. Medical Education. **18** (1984): 406-416.

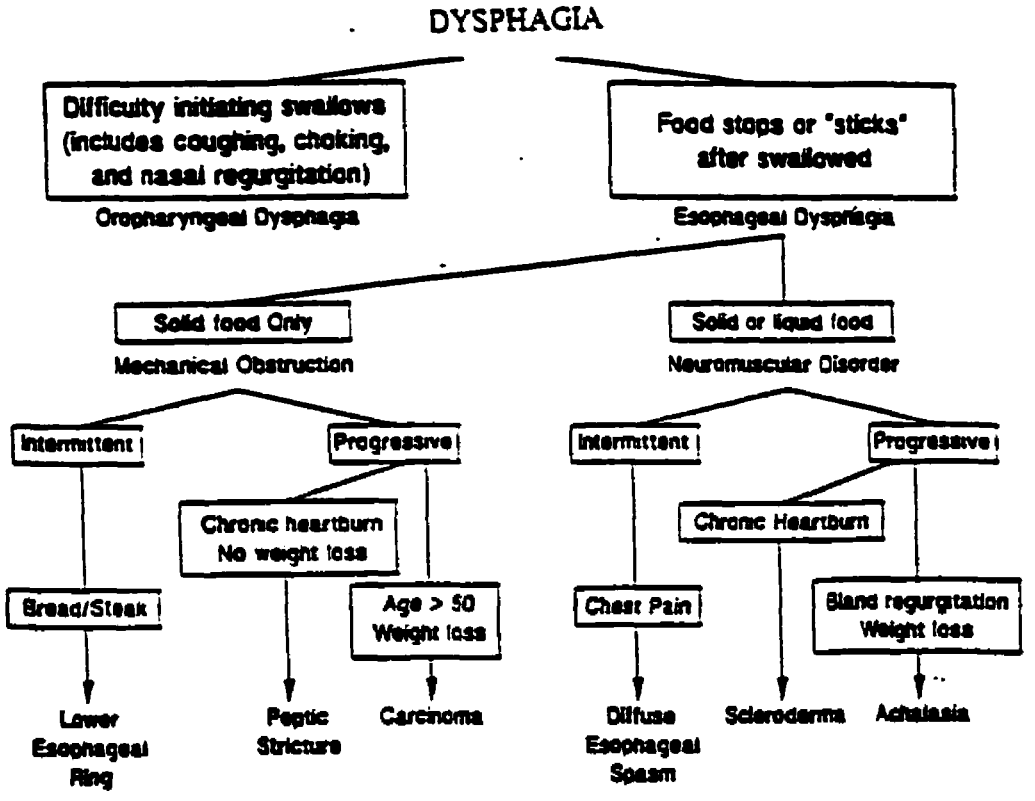
65. Mandin. H. Evaluation: the Engine that Drives us Forward or Back. Journal of Clinical and Investigative Medicine. **23** (2000): 70-76.

66. Regehr, G. Chickens and Children Do Not an Expert Make. Academic Medicine. **69** (1994): 970-971.

67. Patel, V. and Kaufman, D. On Poultry Expertise, Precocious Kids, and Diagnostic Reasoning. Academic Medicine. **69** (1994): 971-972.

68. Sleisenger, M. Fordtran, J. *Gastrointestinal Disease: Pathophysiology, Diagnosis, Management.* W.B. Saunders and Co. Philadelphia. P 334.

APPENDIX A: Example of a Scheme for the Clinical Presentation "Dysphagia" [68]



Essentially, five questions will lead to one of the six major diagnoses:

1) Difficulty initiating the swallow or "sticking" once it has passed the throat?

2) Dysphagia to solids alone or solids and liquids?

3) Intermittent dysphagia or progressive?

4) Is there associated weight loss?

5) Is there associated heartburn?

APPENDIX B: Comparison of Problem Solving Strategies [4]

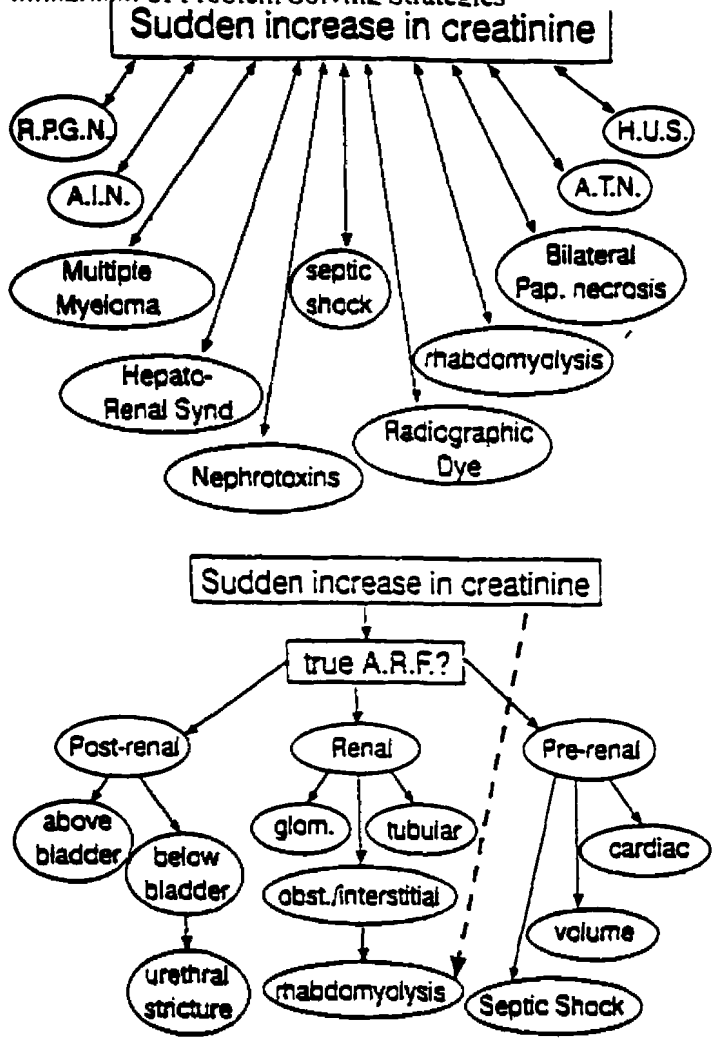


Figure 2. Comparison of the search-and-scan (top panel) and scheme-driven (bottom panel) forms of inquiry applied to a patient with an acute rise in serum creatinine. In the top panel, various diagnostic possibilities are the result of hypothesis generation; the two-headed arrows indicate that each hypothesis is tested in turn. The bottom panel shows an example of the scheme appropriate to the problem being solved. Such schemes are useful for students' information storage and learning but are also used to solve diagnostic problems, since for each arrow an inquiry is possible that will direct subsequent inquiries in the appropriate direction. The dashed arrow represents "pattern recognition," a means of reaching a diagnosis that does not involve the cognitive process of problem solving shown in either panel but instead represents experts' (common) or students' (rare) labeling of the condition based upon prior experience.

(with permission from the authors)

APPENDIX C: Twelve Questions Used in the Study, Grouped Per Clinical Presentation

Question #1: Elevated Liver Enzymes FORMAT: Elimination-type question

A 28 year-old woman[1] is found, on a routine physical examination for insurance purposes, to have elevated liver enzymes. The laboratory work and ultrasound report is sent with her and is found below. She has not had her enzymes checked before as far as she knows. She complains of intense itchiness throughout her body[2], but is otherwise asymptomatic[3]. Specifically, she has no complaints of jaundice[4], abdominal pain[5], weight loss[6], diarrhea or blood in her stools[7], nor does she have a history of fever/chills or night sweats[8]. She has never had hepatitis[9]. She drinks one or two drinks of alcohol per week[10] but denies high-risk sexual behaviour or IV drug use[11]. She has not had a tattoo or a blood transfusion[12]. She has no history of skin changes[13], nor any neurological[14], cardiovascular[15] or respiratory diseases[16].
Her past medical history is significant for hypothyroidism, occasional migraines and cholecystectomy[17]. She is on Advil PRN for headaches, Synthroid, but no other medications including OTCs and herbals[18]. She smokes ½ ppd[19]. She is married with two children and works as a lawyer. She has recently purchased a cat for her daughter[20]. There is no family history of liver disease[21]

Physical examination is unremarkable[22]. She looks well, with no skin changes/jaundice. There are no stigmata of chronic liver diseases[23]. Head and neck, chest, cardiovascular and abdominal exam are completely normal.

LABS: 1) CBC, electrolytes, creatinine, albumin, INR, bilirubin normal[24]
2) ALK PHOS, 555 (30-130), GGT 390 (11-63), ALT 74 (1-60), AST 64 (1-55)[25]
3) ULTRASOUND: Liver, bile ducts are normal. Other structures are normal. Gallbladder has been removed [26]

1) What two features might lead you to specifically eliminate option J) (from the list below) as a possible diagnosis: _____

2) Select the most likely diagnosis from the list below: _____
A) Alcoholic liver disease
B) Alpha-1 antitrypsin deficiency
C) Autoimmune hepatitis
D) Cholangiocarcinoma
E) Choledocholithiasis
F) Chronic Hepatitis B
G) Chronic Hepatitis C
H) Congenital biliary atresia

Please describe how you arrived at these answers

I) Drug-induced cholestasis
J) Genetic hemochromatosis
K) Gilbert's syndrome
L) Lymphoma of liver
M) Pancreatic cancer
N) Primary biliary cirrosis
O) Primary sclerosing cholangitis
P) Wilson's disease

Question #5 Elevated Liver Enzymes **FORMAT: Multiple-choice question**

A 38 year-old man is found, on a routine physical examination for insurance purposes, to have elevated liver enzymes. The laboratory work and ultrasound report is sent with him and is found below. He has not had his enzymes checked before as far as he knows. He is completely asymptomatic currently, other than his usual four or five loose stools per day, which are sometimes bloody. Specifically, he has no complaints of itchiness, jaundice, abdominal pain or weight loss, nor does he have a history of fever/chills or night sweats. He has never had hepatitis. He drinks three to four drinks of alcohol per week, but denies high-risk sexual behaviour or IV drug use. He has not had a tattoo or a blood transfusion. He has no history of skin changes, nor any neurological, cardiovascular or respiratory diseases.

His past medical history is significant for cholecystectomy, as well as a diagnosis of "colitis" 10 years ago that was treated with two months of a "cortisone-type" drug. He takes Tylenol PRN for fibromyalgia, but no other medication including OTCs and herbals. He smokes 1 ppd. He is a single pipe fitter, who has been exposed to asbestos in the past. There is no family history of liver disease.

Physical examination is unremarkable. He looks well, with no skin changes/jaundice. There are no stigmata of chronic liver diseases. Head and neck, chest, cardiovascular and abdominal exam are completely normal.

LABS: 1) CBC, electrolytes, creatinine, albumin, INR, bilirubin normal
2) ALK PHOS, 678 (30-130), GGT 393 (11-63), ALT 88 (1-60), AST 95 (1-55)
3) ULTRASOUND: Liver, bile ducts are normal. Other structures are normal. Gallbladder has been removed.

1) What is the most likely diagnosis for this patient?

A) Primary sclerosing cholangitis
B) Choledocholithiasis
C) Chronic Hepatitis C
D) Primary biliary cirrhosis
E) Drug-induced cholestasis ANS:_____

Please describe how you arrived at this answer

Question #9: Elevated Liver Enzymes FORMAT: Extended-matching question

A 55 year-old man is found. on a routine physical examination to have elevated liver enzymes. The laboratory work and ultrasound report is sent with him and is found below. He has not had his enzymes checked before as far as he knows. He is completely asymptomatic. with no complaints of itchiness. jaundice. abdominal pain or weight loss. He has no diarrhea. blood in his stools. fever/chills or night sweats. He has never had hepatitis. He drinks no alcohol whatsoever. and denies high-risk sexual behavior or IV drug use. He has not had a tattoo or a blood transfusion. He has no history of skin changes. nor any neurological. cardiovascular or respiratory diseases.
His past medical history is significant for a cholecystectomy. He has chronic schizophrenia treated with chlorpromazine. and takes Aspirin occasionally for leg cramps. He had a pneumonia four weeks ago. treated with 14 days of Amoxicillin He is on no other medications. including OTCs and herbals. He smokes 2 ppd. He lives in a group home. is unemployed. and has been exposed to other residents with a flu recently. There is no family history of liver disease.

Physical examination is unremarkable. He looks well. with no skin changes/jaundice. There are no stigmata of chronic liver diseases. Head and neck. chest. cardiovascular and abdominal exam are completely normal.

LABS: 1) CBC. electrolytes. creatinine. albumin. INR. bilirubin normal
2) ALK PHOS. 432 (30-130). GGT 690 (11-63). ALT 98 (1-60). AST 92 (1-55)
3) ULTRASOUND: Liver. bile ducts are normal. Other structures are normal. Gallbladder has been removed.

1) Select the most likely diagnosis from the list below: _____

A) Alcoholic liver disease
B) Alpha-1 antitrypsin deficiency
C) Autoimmune hepatitis
D) Cholangiocarcinoma
E) Choledocholithiasis
F) Chronic Hepatitis B
G) Chronic Hepatitis C
H) Congenital biliary atresia
I) Drug-induced cholestasis
J) Genetic hemochromatosis
K) Gilbert's syndrome
L) Lymphoma of liver
M) Pancreatic cancer
N) Primary biliary cirrosis
O) Primary sclerosing cholangitis
P) Wilson's disease

Please describe how you arrived at this answer

Question #2: Nausea and Vomiting **FORMAT: Multiple-choice Question:**

A 62 year-old male presents to the Emergency Department with a two-week history of nausea and vomiting. Prior to this, he had never had any such symptoms. He denies any history of heartburn, dysphagia or abdominal pain. He feels hungry at times, and has taken in extra nutritional supplements to maintain his weight. He denies any diarrhea, blood in his stools, and has no night sweats, fevers, headaches or diplopia.
His family history is unremarkable. His past history is significant for back pain, diagnosed as osteoporosis-related vertebral fractures, hypothyroidism (on Synthroid), and depression (on Prozac). He has no other medical/surgical history, and takes no medications other than Advil as needed for back pains. He is a non-smoker, non-drinker and is a retired carpenter. His physical examination reveals some pallor, but is otherwise completely normal.

LABS: 1) HB: 88 (137-180), normocytic; WBC, platelets normal
2) Electrolytes, bicarbonate normal, glucose normal; creatinine 185 (was 168 two months ago); albumin 20 (35-40); calcium 2.60 (2.0-2.50); lipase, magnesium are normal.

1) What is the most likely diagnosis for this patient?

A) Peptic ulcer disease
B) Hypercalcemia from multiple myeloma
C) Acute pancreatitis
D) Diabetic ketoacidosis
E) Adrenal insufficiency ANS:_____

Please describe how you arrived at this answer

Question #6: Nausea and Vomiting **FORMAT: Elimination-type Question:**

A 58 year-old male[1] presents to the Emergency Department with a two-week history of nausea and vomiting[2]. Prior to this, he had never had any such symptoms.[3] He denies any history of heartburn[4], dysphagia[5] or abdominal pain[6]. He feels hungry at times, and has taken in extra nutritional supplements to maintain his weight[7]. He denies any diarrhea[8] blood in his stools[9], and has no night sweats, fevers[10], headaches or diplopia[11] His family history is unremarkable[12].

His past history is significant for emphysema.[13] He was discharged from hospital seven weeks ago on Ventolin/Becloforte puffers and 8 tablets of prednisone[14]. His breathing improved three weeks ago and he stopped all of his medications[15]. He is a chronic schizophrenic (on Haloperidol) but has no other health problems[16]. He smokes 2 ppd for 40 years, but does not drink alcohol[17] He is a retired welder. His physical examination[18]is consistent with emphysema (barrel chest, purse-lipped breathing, decreased lung sounds throughout) but is otherwise completely normal.

LABS: 1) HB, WBC, platelets normal[19]
2) Sodium 120 (135-145), potassium 5.8 (3.5-5.0) bicarbonate 20 (25-30); chloride 90; glucose normal, creatinine normal[20]
3) Albumin, calcium, lipase, magnesium are all normal[21]

1) What two features might lead you to specifically eliminate option B) (from the list below) as a possible diagnosis: _____

2) Select the most likely diagnosis from the list below: _____

A) Acute renal failure
B) Acute pancreatitis
C) Adrenal insufficiency
D) Biliary colic
E) Colonic carcinoma
F) Crohn's disease
G) Diabetic ketoacidosis
H) Drug-induced vomiting
I) Hypercalcemia from multiple myeloma
J) Peptic ulcer disease
K) Pituitary apoplexy
L) Pontine glioma
M) Psychogenic vomiting
N) Reflux esophagitis
O) Small bowel obstruction

Please describe how you arrived at these answers

Question #10: Nausea and Vomiting **FORMAT: Extended-matching Question:**

A 54 year-old male presents to the Emergency Department with a two-week history of nausea and vomiting. Prior to this, he had never had any such symptoms. He has vague upper abdominal pain, but denies any history of heartburn or dysphagia. He feels hungry at times, and has taken in extra nutritional supplements to maintain his weight. He denies any diarrhea, blood in his stools, and has no night sweats, fevers, headaches or diplopia. His family history is unremarkable.

His past history is significant for a "slight stroke" three years ago (no residual deficits), at which time he was found to be a diabetic. He has changed his diet for his diabetes but is on no medications other then Valium as needed for anxiety attacks. He has no other health problems. He is a non-smoker and does not drink alcohol. He is a retired lawyer. His physical examination is normal except for dehydration.

LABS: 1) HB, WBC, platelets normal
2) Sodium 132 (135-145), potassium 3.4 (3.5-5.0) bicarbonate 8 (25-30); chloride 90; glucose 19.5, creatinine 105; albumin, calcium, magnesium, lipase are all normal.

Select the most likely diagnosis from the list below: _____

A) Acute renal failure
B) Acute pancreatitis
C) Adrenal insufficiency
D) Biliary colic
E) Colonic carcinoma
F) Crohn's disease
G) Diabetic ketoacidosis
H) Drug-induced vomiting
I) Hypercalcemia from multiple myeloma
J) Peptic ulcer disease
K) Pituitary apoplexy
L) Pontine glioma
M) Psychogenic vomiting
N) Reflux esophagitis
O) Small bowel obstruction

Please describe how you arrived at this answer

Question #3: Chronic Diarrhea FORMAT: Multiple-choice Question:

A 35 year old woman presents with a one year history of diarrhea. She describes her stools are 10 - 12 profuse, watery bowel movements per day, with no blood in her stools. She is eating well but has lost 15 lbs. over the last year. She has no abdominal pain. She is unsure if her stools are oily, but they are difficult to flush. She is otherwise perfectly well, with no previous surgeries. She smokes ½ pack a day but does not drink alcohol. She has never traveled, camped or drank well water. Her family history reveals an aunt with ulcerative colitis. Examination is unremarkable except for pallor. Stool, C & S, O & P and C.diff. are all negative. Laboratory work shows a microcytic anemia (Hb 95, mcv 63), with low ferritin, but normal B12 and folate levels.

1) What is the most likely diagnosis for this patient?

A) Celiac disease
B) Crohn's colitis
C) Villous adenoma of rectum
D) Pancreatic insufficiency
E) Bacterial overgrowth ANS:_____

Please describe how you arrived at this answer

Question #7: Chronic Diarrhea FORMAT: Extended-matching Question:

A 33 year old woman presents with a one year history of diarrhea. She describes her stools as 10 - 12 profuse, water bowel movements per day with no blood in her stools. She is eating well, but has lost 20 lbs. over the last year. She has no abdominal pain. Her sometimes sees oil droplets in her stool, and they are very difficult to flush. She had a surgery for stomach ulcers at age 20, and had a repeat surgery five years later for "bile gastritis". She is otherwise healthy. She smokes ½ pack per day but does not drink alcohol. She has not drank well water, and has not traveled or gone camping recently. Her family history is significant for two cousins with Crohn's disease. Examination is unremarkable. Stool, C & S, O & P and C.diff. are all negative. Her CBC shows a macrocytic anemia (hb 108, mcv 110) with a normal ferritin, but low B12 and elevated folate levels.

Select the most likely diagnosis from the list below: _____

A) Bacterial overgrowth
B) Celiac disease
C) Collagenous colitis
D) Crohn's colitis
E) Crohn's ileitis
F) Colonic carcinoma
G) Factitious diarrhea
H) Giardiasis
I) Ischemic colitis
J) Irritable bowel syndrome
K) Lactose intolerance
L) Pancreatic insufficiency
M) Shigella dysentery
N) Villous adenoma of rectum
O) Viral gastroenteritis

Please describe how you arrived at this answer

Question #11: Chronic Diarrhea FORMAT: Elimination-type Question:

A 39 year old woman[1] presents with a one year history of diarrhea.[2] She describes her stools as 10 - 12 profuse, watery bowel movements per day.[3] with no blood in her stools.[4] She is not eating well because of abdominal pain after eating[5] and has lost 25 lbs. over the last year.[6] She says that she sees oil droplets in her stool, and they are difficult to flush.[7] She is otherwise well,[8] with no previous surgeries.[9] She smokes ½ pack per day[10] and admits to drinking 10oz of rye per day since her teenage years.[11] She has not traveled, drank well water or gone camping.[12] Her family history reveals a brother with Crohn's disease[13]. Examination reveals some mild periumbilical tenderness[14]. Stool, C & S, O &P and C.diff are all negative[15]
LAB: CBC shows a macrocytic anemia (hb 110, mcv 108), with a normal ferritin and folate but low B12 levels.[16]

1) What three features might lead you to specifically eliminate option D) (from the list below) as a possible diagnosis: _____

2) Select the most likely diagnosis from the list below: _____

A) Bacterial overgrowth
B) Celiac disease
C) Collagenous colitis
D) Crohn's colitis
E) Crohn's ileitis
F) Colonic carcinoma
G) Factitious diarrhea
H) Giardiasis
I) Ischemic colitis
J) Irritable bowel syndrome
K) Lactose intolerance
L) Pancreatic insufficiency
M) Shigella dysentery
N) Villous adenoma of rectum
O) Viral gastroenteritis

Please describe how you arrived at these answers

Question #4: Dysphagia **FORMAT: Elimination-type Question:**

A 58 year old male① presents with a one year history of food "sticking" retrosternally after he swallows.② This occurs only with solid foods.③ This does not occur with every meal, and in fact he describes 5 - 6 episodes over the last year.④ He has had no heartburn.⑤ or weight loss. ⑥ He has not had any chest pains.⑦ His past history reveals Hodgkin's disease treated when he was 22 years old and apparently cured. He has no other health problems, is on no medications, and is a lifetime non-smoker.⑧ Physical examination is unremarkable.⑨

1) What two features might lead you to specifically eliminate option A) (from the list below) as a possible diagnosis: _____

2) Select the most likely diagnosis from the list below: _____

A) Achalasia
B) Amyotrophic lateral sclerosis
C) Diffuse esophageal spasm
D) Esophageal cancer
E) Lower esophageal ring
F) Nutcracker esophagus
G) Peptic stricture
H) Psychogenic dysphagia
I) Scleroderma
J) Zenker's diverticulum
K) Extrinsic esophageal compression

Please describe how you arrived at these answers

Question #8: Dysphagia FORMAT: Multiple-choice Question:

A 62 year old male presents with a one year history of food "sticking" retrosternally after he swallows. This occurs only with solid foods. He feels that it is getting increasingly more frequent and noticeable. He says that he has had bad heartburn for the last 3 years, partially helped with ranitidine. His weight has been steady, and he has had no chest pain. His past history reveals a partial thyroidectomy 20 years ago for a benign adenoma. He has had no other health problems, is on no medications and is a lifetime non-smoker. Physical examination is unremarkable.

What is the most likely diagnosis for this patient?

A) Achalasia
B) Scleroderma
C) Esophageal cancer
D) Lower esophageal ring
E) Peptic stricture

Please describe how you arrived at this answer

Question #12: Dysphagia FORMAT: Extended-matching Question:

A 38 year old man presents with a one year history of food "sticking" retrosternally after he swallows. From the onset, this has occurred with both solids and liquids. Initially, this was occurring once a week, but now it is occurring with every meal. He has lost 10 lbs. since this started. He experiences heartburn only rarely, and gets infrequent vague, mild chest pains. His past history is negative except for a "spot" on his lungs on chest x-ray, investigated and felt to benign. He is on no medications, and smokes 2 ppd for 20 years. Physical exam is unremarkable.

Select the most likely diagnosis from the list below: _____

A) Achalasia
B) Amyotrophic lateral sclerosis
C) Diffuse esophageal spasm
D) Esophageal cancer
E) Lower esophageal ring
F) Nutcracker esophagus
G) Peptic stricture
H) Psychogenic dysphagia
I) Scleroderma
J) Zenker's diverticulum
K) Extrinsic esophageal compression

Please describe how you arrived at this answer

APPENDIX D: Methods Attempted to Replace the Think-aloud

For both these methods, a scenario was presented to the examinees, from which they were asked: "which person (Gary, Paula, Henry, Sally) most resembles how you arrived at your diagnosis for each particular question". For this, Gary represents the problem solving strategy of 'guessing', while Henry represents 'hypothetico-deductive reasoning', Sally 'scheme use' and Paula 'pattern recognition'.

The two scenarios that were tried are as follows.

1) The "BBQ" Method (used in the pilot study)

On a hot summer long weekend, four barbecue (BBQ) repairpersons from the same repair company were asked to assess four different dysfunctional BBQs, all of which had the same problem: the burner will not light. After assessing the situation, all four find (and fix) the exact same problem on all four BBQs: spiders blocked one of the Venturi tubes. These Venturis are the L-shaped tubes connecting the BBQ's burner to the valve outlet (i.e. control knobs): the valve outlet itself receives the gas from the gas cylinder. A diagram is attached to explain the BBQ parts mentioned.

i) Gary: "I really was not entirely certain what the problem was, but it seemed to me that the Venturis would be the most likely culprit".

ii) Paula: " I have seen spiders clog up these tubes so many times at this time of year, that I knew it had to be the Venturis".

iii) Henry:" I first looked for a problem with the gas cylinder, then went to the gas hose. When I didn't find the problem there, I moved through to the valve outlet, then the Venturis, where I found the problem".

iv) Sally: "I let the BBQ's hissing sound guide me: if it is there, then the problem is at or after the valve: if it is absent, the problem is before the valve, either the hose or cylinder. This BBQ was hissing, so I looked at areas after the valve, and found the clogged Venturi".

2) The "Crying Baby" Method

A group of four parents, each with his/her baby, were meeting to discuss issues relating to their neighborhood. Suddenly, all four babies began to cry. Quickly, each baby's parent assessed the crying, and eventually all concluded that their baby was crying because of fatigue. The babies were put to bed, and the crying stopped. When asked how they arrived at the conclusion that their baby was crying due to fatigue, the four parents explained:

i) Gary: " I really had no idea why my boy was crying, but took my best guess that he was tired"

ii) Paula: " The pitch and intensity of my girl's crying allowed me to directly recognize that this could be due to nothing else but fatigue"

iii) Henry: " I proceeded by a process of elimination:

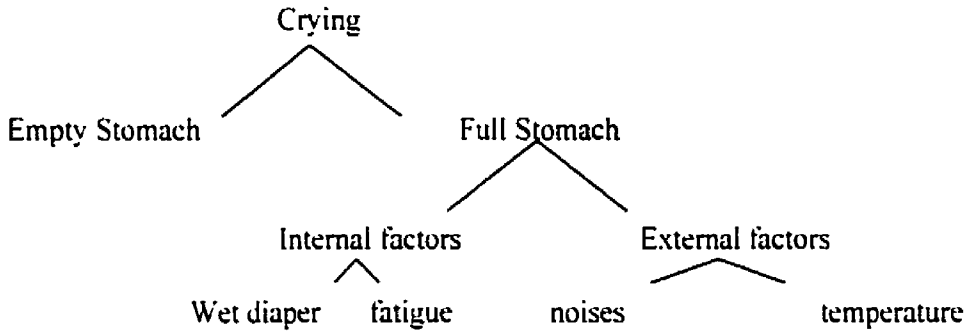
-Is she hungry? When feeding her didn't work, I moved on to option #2

-Is her diaper wet? Changing her didn't work, so I moved on to option #3

-Is she warm/cold? She didn't feel warm or cold, so I moved to option #4

-Is she tired: by process of elimination, I put her to bed and the crying stopped."

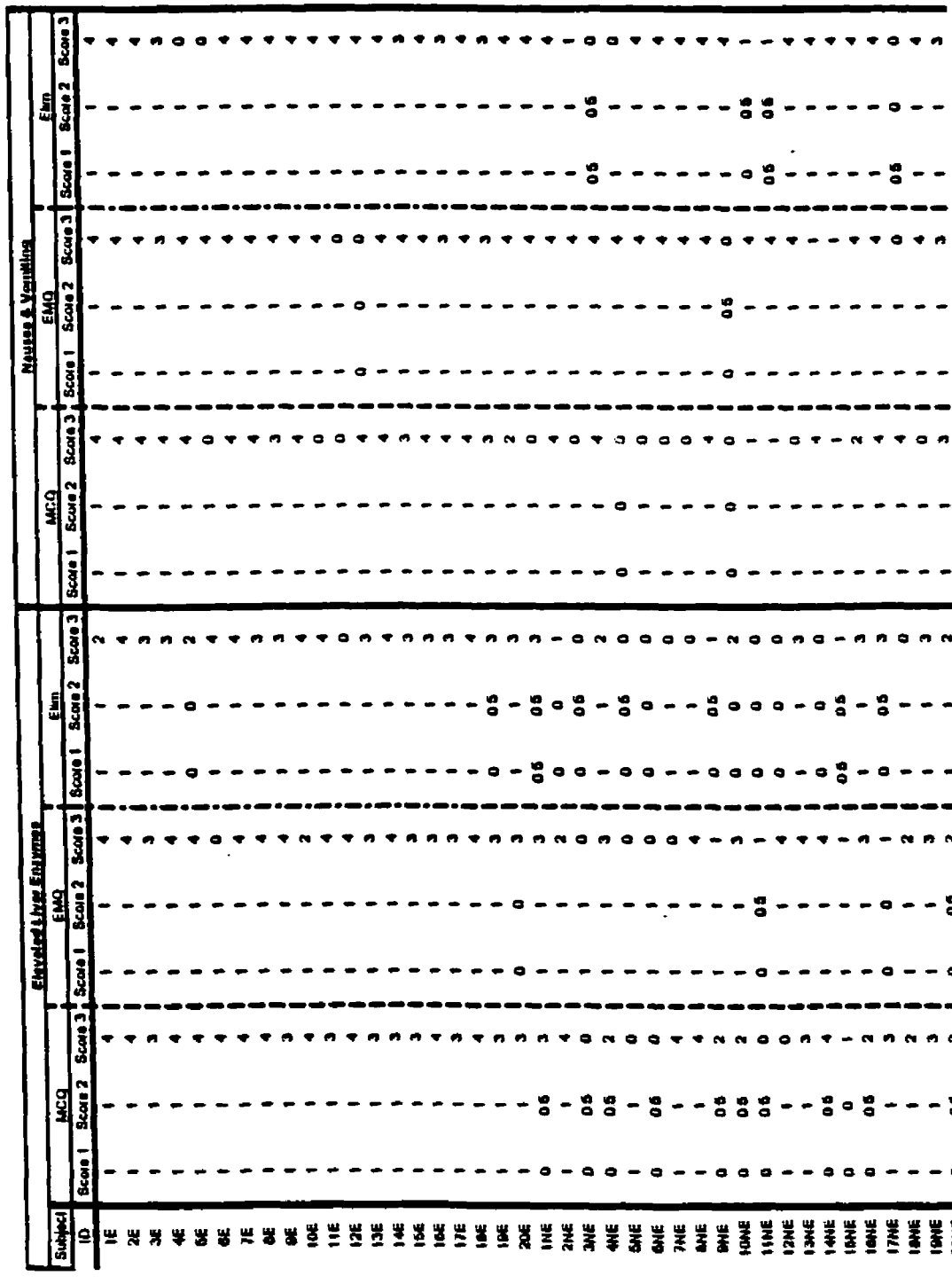
iv) Sally: " When my baby cries. I use the following algorithm:



Knowing I had just fed my baby. I proceeded to the "full stomach" arm. There were clearly no external factors. so the answer had to lie in the "internal factors". The diaper was dry. so fatigue had to be the answer"

APPENDIX E: Details of Data and Statistical Analyses

1) Initial Raw Data Set



	Diarrhea									Dysphagia								
Subject	MCQ			EMQ			Ekm			MCQ			EMQ			Ekm		
ID	Score 1	Score 2	Score 3	Score 1	Score 2	Score 3	Score 1	Score 2	Score 3	Score 1	Score 2	Score 3	Score 1	Score 2	Score 3	Score 1	Score 2	Score 3
1E	1	1	3	1	1	3	1	1	3	1	1	3	0	0	1	0.5	0.5	1
2E	1	1	3	1	1	3	1	1	3	1	1	3	1	1	3	1	1	3
3E	1	1	4	1	1	3	1	1	3	1	1	3	0	0	3	1	1	3
4E	1	1	3	1	1	4	1	1	4	1	1	3	1	1	3	1	1	3
5E	1	1	2	1	1	4	1	1	4	1	1	4	0	0	4	1	1	0
6E	1	1	4	1	1	1	1	1	4	1	1	3	1	1	3	1	1	4
7E	1	1	4	1	1	4	1	1	4	1	1	3	1	1	3	1	1	3
8E	1	1	4	1	1	4	1	1	4	1	1	3	1	1	3	1	1	3
9E	1	1	3	1	1	3	1	1	4	1	1	3	1	1	3	1	1	3
10E	1	1	4	1	1	4	1	1	0	1	1	4	0	0	4	1	1	4
11E	1	1	3	1	1	4	1	1	3	1	1	3	1	1	3	0.5	0.5	3
12E	1	1	0	1	1	4	1	1	3	1	1	4	1	1	4	1	1	3
13E	0	0	0	1	1	4	1	1	3	1	1	3	1	1	3	1	1	3
14E	1	1	2	1	1	4	1	1	2	1	1	4	1	1	4	1	1	4
15E	1	1	2	1	1	3	1	1	3	1	1	3	1	1	3	1	1	3
16E	1	1	4	1	1	4	1	1	4	1	1	4	1	1	3	1	1	4
17E	1	1	4	1	1	4	1	1	3	1	1	3	1	1	3	1	1	3
18E	1	1	4	1	1	4	1	1	3	1	1	4	1	1	4	1	1	4
19E	1	1	3	1	1	4	1	1	3	1	1	3	0	0	3	1	1	3
20E	1	1	3	1	1	4	1	1	4	1	1	4	1	1	3	1	1	3
1NE	1	1	0	0	0.5	2	1	1	2	0	0.5	3	1	1	3	1	1	3
2NE	1	1	2	0	0.5	1	1	1	2	0	0.5	3	1	1	3	1	1	3
3NE	0	0.5	2	0	0.5	1	1	1	3	0	0	0	0	0	1	0	0.5	0
4NE	0	0	1	0	0.5	1	1	1	3	1	1	3	0	0	3	0.5	0.5	3
5NE	1	1	0	0	0.5	0	1	1	2	0	0.5	4	0	0	1	0.5	0.5	0
6NE	0	0	1	0	0.5	1	1	1	4	0	0	3	0	0	3	0	0	3
7NE	1	1	0	1	1	0	1	1	2	1	1	3	1	1	3	1	1	1
8NE	1	1	0	0	0.5	0	1	1	1	0	0.5	4	0	0	0	1	1	1
9NE	1	1	2	0	0.5	4	0	0	0	0	0.5	3	1	1	3	0.5	0.5	3
10NE	0	0	0	0	0.5	4	1	1	4	0	0.5	1	0	0	1	0.5	0.5	1
11NE	1	1	1	0	0.5	0	0.5	0.5	0	0	0.5	0	0	0	0	1	1	0
12NE	1	1	4	0	0.5	1	1	1	4	1	1	1	1	1	3	0.5	0	2
13NE	1	1	2	0	0.5	2	1	1	4	1	1	4	0	0	4	0	0	1
14NE	1	1	0	0	0.5	1	1	1	4	1	1	4	1	1	2	0.5	0.5	3
15NE	0	0	1	0	0.5	1	0.5	0	1	0	0	2	1	1	3	1	1	2
16NE	1	1	0	0	0.5	2	1	1	1	1	1	3	1	1	3	1	1	3
17NE	0	0.5	1	0	0.5	1	1	1	4	0	0	1	0	0.5	1	0	0	1
18NE	1	1	0	0	0.5	0	0.5	0	0	0	0.5	1	0	0.5	2	1	1	2
19NE	0	0	2	0	0.5	1	1	1	4	1	1	3	1	1	3	0.5	0	2
20NE	1	1	3	0	0.5	3	1	1	3	1	1	3	1	1	3	1	1	3

2a) Raw Data Means and Standard Deviations by Score (20 expert subjects)

Clinical Presentation	Question Format	Score	Mean Score	Standard Deviation
Elevated Liver Enzymes	Multiple-choice	Score 1	1.000	0
		Score 2	1.000	0
		Score 3	.8875	.1276
	Extended-matching	Score 1	.9500	.2236
		Score 2	.9500	.2236
		Score 3	.8375	.2470
	Elimination –type	Score 1	.9000	.3078
		Score 2	.9250	.2447
		Score 3	.7750	.2420
Nausea & Vomiting	Multiple-choice	Score 1	1.000	0.000
		Score 2	1.000	0.000
		Score 3	.7875	.3652
	Extended-matching	Score 1	.9500	.2236
		Score 2	.9500	.2236
		Score 3	.8625	.3086
	Elimination-type	Score 1	1.000	0.000
		Score 2	1.000	0.000
		Score 3	.8500	.3078
Diarrhea	Multiple-choice	Score 1	.9500	.2236
		Score 2	.9500	.2236
		Score 3	.7375	.3086
	Extended-matching	Score 1	1.000	0.000
		Score 2	1.000	0.000
		Score 3	.9000	.1885
	Elimination-type	Score 1	1.000	0.000
		Score 2	1.000	0.000
		Score 3	.8000	.2379
Dysphagia	Multiple-choice	Score 1	1.000	0.000
		Score 2	1.000	0.000
		Score 3	.8375	.1223
	Extended-matching	Score 1	.7500	.4443
		Score 2	.7500	.4443
		Score 3	.7875	.1678
	Elimination-type	Score 1	.9500	.1539
		Score 2	.9500	.1539
		Score 3	.7500	.2433

2b) Raw Data Means and Standard Deviations by Score (20 non-expert subjects)

Clinical Presentation	Question Format	Score	Mean Score	Standard Deviation
Elevated Liver Enzymes	Multiple-choice	Score 1	.4500	.5104
		Score 2	.7000	.2991
		Score 3	.5125	.3670
	Extended-matching	Score 1	.8500	.3663
		Score 2	.9000	.2616
		Score 3	.5125	.3670
	Elimination-type	Score 1	.4500	.4840
		Score 2	.5500	.4261
		Score 3	.3000	.3204
Nausea & Vomiting	Multiple-choice	Score 1	.9000	.3078
		Score 2	.9000	.3078
		Score 3	.4000	.4472
	Extended-matching	Score 1	.9500	.2236
		Score 2	.9750	.1118
		Score 3	.8125	.3616
	Elimination-type	Score 1	.8750	.2751
		Score 2	.8750	.2751
		Score 3	.7250	.4128
Diarrhea	Multiple-choice	Score 1	.6500	.4894
		Score 2	.7000	.4413
		Score 3	.2750	.2913
	Extended-matching	Score 1	.0500	.2236
		Score 2	.5250	.1118
		Score 3	.3250	.3046
	Elimination-type	Score 1	.8750	.2751
		Score 2	.8250	.3726
		Score 3	.6000	.3752
Dysphagia	Multiple-choice	Score 1	.4000	.5026
		Score 2	.6000	.3839
		Score 3	.6125	.3292
	Extended-matching	Score 1	.5000	.5130
		Score 2	.5500	.4840
		Score 3	.5625	.2910
	Elimination-type	Score 1	.6250	.3932
		Score 2	.6000	.4168
		Score 3	.4875	.2865

3)Statistical Details Related to Research Question #1

3a) Groupings of the Process Scores: Frequency Tables

i) Experts (n=20)

Question Format	Scheme	Liver	N&V	Diarrhea	Dysphagia
Multiple-choice	0-1 (1) Guess/Deductive	0	3	2	0
	2-3 (2) Scheme	9	4	10	13
	4 (3) Pattern Recognition	11	13	8	7
Extended-matching	0-1 (1) Guess/Deductive	1	2	1	1
	2-3 (2) Scheme	8	3	5	14
	4 (3) Pattern Recognition	11	15	14	5
Elimination-type	0-1 (1) Guess/Deductive	1	2	1	2
	2-3 (2) Scheme	12	4	11	13
	4 (3) Pattern Recognition	7	14	8	5

ii) Non-Experts (n=20)

Question Format	Scheme	Liver	N&V	Diarrhea	Dysphagia
Multiple-choice	0-1 (1) Guess/Deductive	6	12	13	6
	2-3 (2) Scheme	10	2	6	10
	4 (3) Pattern Recognition	4	6	1	4
Extended-matching	0-1 (1) Guess/Deductive	8	4	14	6
	2-3 (2) Scheme	8	1	4	13
	4 (3) Pattern Recognition	4	15	2	1
Elimination-type	0-1 (1) Guess/Deductive	12	6	6	7
	2-3 (2) Scheme	8	1	7	13
	4 (3) Pattern Recognition	0	13	7	0

3b) Details of MANOVA Calculations Related to Table 5

i) 2x3x2 Multiple Analysis of Variance Comparing Expert and Non-Expert Subjects and Process Use on Dichotomous and Partial Dichotomous Multiple Choice Scores on a Nausea & Vomiting Clinical Presentation

	Df	MS	F	P
Expert/Non-Expert	(1, 34)	.0394	.4018	NS
Process	(2,34)	.0463	.4718	NS
Score	(1,34)	-.0000	-	-
Expert x Process	(2,34)	.0463	.4718	NS
Expert x Score	(1,34)	.0000	-	NS
Process x Score	(2,34)	-.0000	-	-
Expert x Process x Score	(2,34)	-.0000	-	-

ii) 2x3x2 Multiple Analysis of Variance Comparing Expert and Non-Expert Subjects and Process Use on Dichotomous and Partial Dichotomous Multiple Choice Scores on a Diarrhea Clinical Presentation

	Df	MS	F	p
Expert/Non-Expert	(1, 34)	.0251	.0973	NS
Process	(2,34)	.5613	2.179	NS
Score	(1,34)	.0038	.5835	NS
Expert x Process	(2,34)	.2342	.9093	NS
Expert x Score	(1,34)	.0038	.5835	NS
Process x Score	(2,34)	.0015	.2340	NS
Expert x Process x Score	(2,34)	.0015	.2340	NS

iii) 2x3x2 Multiple Analysis of Variance Comparing Expert and Non-Expert Subjects and Process Use on Dichotomous and Partial Dichotomous Multiple Choice Scores on a Liver Enzyme Clinical Presentation

	Df	MS	F	P
Expert/Non-Expert	(1, 34)	.1.472	9.547	.004*
Process	(2,34)	.269	1.741	NS
Score	(1,34)	.272	16.381	.000*
Expert x Process	(1,34)	.250	1.619	NS
Expert x Score	(1,34)	.164	9.851	.003*
Process x Score	(2,34)	-.0000	.847	NS
Expert x Process x Score	(2,34)	-.0000	1.670	NS

iv) 2x3x2 Multiple Analysis of Variance Comparing Expert and Non-Expert Subjects and Process Use on Dichotomous and Partial Dichotomous Multiple Choice Scores on a Dysphagia Clinical Presentation

	Df	MS	F	p
Expert/Non-Expert	(1, 34)	1.866	11.551	.002*
Process	(2,34)	.655	4.054	.026*
Score	(1,34)	.232	13.789	.001*
Expert x Process	(2,34)	-.0000	.004	NS
Expert x Score	(1,34)	.149	8.876	.005*
Process x Score	(2,34)	-.0000	.194	NS
Expert x Process x Score	(2,34)	-.0000	.355	NS

v) 2x3x2 Multiple Analysis of Variance Comparing Expert and Non-Expert Subjects and Process Use on Dichotomous and Partial Dichotomous Extended Matching Scores on a Elevated Liver Enzymes Clinical Presentation

	Df	MS	F	P
Expert/Non-Expert	(1, 34)	.0410	.2735	NS
Process	(2,34)	.0877	.5856	NS
Score	(1,34)	.0046	.7077	NS
Expert x Process	(2,34)	.0458	.3059	NS
Expert x Score	(1,34)	.0046	.7077	NS
Process x Score	(2,34)	.0018	.2783	NS
Expert x Process x Score	(2,34)	.0018	.2783	NS

vi) 2x3x2 Multiple Analysis of Variance Comparing Expert and Non-Expert Subjects and Process Use on Dichotomous and Partial Dichotomous Extended Matching Scores on a Nausea & Vomiting Clinical Presentation

	Df	MS	F	P
Expert/Non-Expert	(1, 34)	.0881	1.625	NS
Process	(2,34)	.5425	10.005*	.0004*
Score	(1,34)	.0035	1.278	NS
Expert x Process	(2,34)	.1121	2.067	NS
Expert x Score	(1,34)	.0035	1.278	NS
Process x Score	(2,34)	.0045	1.626	NS
Expert x Process x Score	(2,34)	.0045	1.626	NS

vii) 2x3x2 Multiple Analysis of Variance Comparing Expert and Non-Expert Subjects and Process Use on Dichotomous and Partial Dichotomous Extended Matching Scores on a Diarrhea Clinical Presentation

	Df	MS	F	P
Expert/Non-Expert	(1, 34)	4.610	150.0*	.0000*
Process	(2,34)	.0022	.0706	NS
Score	(1,34)	.5123	150.0*	.0000*
Expert x Process	(2,34)	.0022	.0706	NS
Expert x Score	(1,34)	.5123	150.0*	.0000*
Process x Score	(2,34)	.0002	.0706	NS
Expert x Process x Score	(2,34)	.0002	.0706	NS

viii) 2x3x2 Multiple Analysis of Variance Comparing Expert and Non-Expert Subjects and Process Use on Dichotomous and Partial Dichotomous Extended Matching Scores on a Dysphagia Clinical Presentation

	Df	MS	F	P
Expert/Non-Expert	(1, 34)	.3126	1.074	NS
Process	(2,34)	2.518	8.652*	.0009*
Score	(1,34)	.0029	.4567	NS
Expert x Process	(2,34)	.2313	.7949	NS
Expert x Score	(1,34)	.0029	.4567	NS
Process x Score	(2,34)	.0007	.1144	NS
Expert x Process x Score	(2,34)	.0007	.1144	NS

ix) 2x3x2 Multiple Analysis of Variance Comparing Expert and Non-Expert Subjects and Process Use on Dichotomous and Partial Dichotomous Elimination Scores on a Nausea & Vomiting Clinical Presentation

	Df	MS	F	P
Expert/Non-Expert	(1, 34)	.1681	4.900*	.0337*
Process	(2,34)	.2172	6.330*	.0046*
Score	(1,34)	-.0000	.0.000	1.000
Expert x Process	(2,34)	.2182	6.330*	.0046*
Expert x Score	(1,34)	.0000	0.000	1.000
Process x Score	(2,34)	-.0000	0.000	1.000
Expert x Process x Score	(2,34)	-.0000	0.000	1.000

x) 2x3x2 Multiple Analysis of Variance Comparing Expert and Non-Expert Subjects and Process Use on Dichotomous and Partial Dichotomous Elimination Scores on a Diarrhea Clinical Presentation

	Df	MS	F	P
Expert/Non-Expert	(1, 34)	.2997	5.823*	.0214*
Process	(2,34)	.1936	3.761*	.0335*
Score	(1,34)	.0083	1.698	NS
Expert x Process	(2,34)	.1936	3.761*	.0335*
Expert x Score	(1,34)	.0083	1.698	NS
Process x Score	(2,34)	.0049	1.097	NS
Expert x Process x Score	(2,34)	.0049	1.097	NS

xi) 2x3x2 Multiple Analysis of Variance Comparing Expert and Non-Expert Subjects and Process Use on Dichotomous and Partial Dichotomous Elimination Scores on a Liver Enzyme Clinical Presentation

	Df	MS	F	P
Expert/Non-Expert	(1, 34)	.908	3.511	NS
Process	(2,34)	-.0000	.084	NS
Score	(1,34)	-.0000	1.090	NS
Expert x Process	(2,34)	.422	1.632	NS
Expert x Score	(1,34)	-.0000	.524	NS
Process x Score	(2,34)	-.0000	.016	NS
Expert x Process x Score	(2,34)	-.0000	.302	NS

xii) 2x3x2 Multiple Analysis of Variance Comparing Expert and Non-Expert Subjects and Process Use on Dichotomous and Partial Dichotomous Elimination Scores on a Dysphagia Clinical Presentation

	Df	MS	F	P
Expert/Non-Expert	(1, 34)	.733	4.431	.043*
Process	(2,34)	.245	1.405	NS
Score	(1,34)	-.0000	.002	NS
Expert x Process	(2,34)	-.0000	.004	NS
Expert x Score	(1,34)	-.0000	.002	NS
Process x Score	(2,34)	-.0000	.906	NS
Expert x Process x Score	(2,34)	-.0000	1.517	NS

4) Statistical Details Related to Research Question #2

4a) Details of ANOVA calculation, for Tables 6a and 6b

i) Table 6a: MAIN EFFECT: QUESTION FORMAT (score 3 - non-expert) averaged over clinical presentation

	Sum of Squares	df	Mean Square	F	p-level
Effect	.070703	2	.035352	1.703318	.195686
Error	.788672	38	.020755		

ii) Table 6b: MAIN EFFECT: QUESTION FORMAT (score 3 – expert) averaged over clinical presentation

	Sum of Squares	df	Mean Square	F	p-level
Effect	.008565	2	.004282	.371614	.692099
Error	.437905	38	.011524		

4b) Details of ANOVA calculation for Tables 7a and 7b

Table 2a: MAIN EFFECT: CASE (score 3 - non-expert) averaged over question format

	Sum of Squares	df	Mean Square	F	p-level
Effect	.736372	3	.245457	6.558042	.000692
Error	2.133420	57	.037428		

Scheffe test: variable Var.1 (score 3 - non-expert)

Probabilities for Post Hoc Tests
MAIN EFFECT: CASE

		{1} .4416667	{2} .6458333	{3} .4000000	{4} .5500000
1	{1}		.016481*	.926297	.379624
2	{2}	.016481*		.002480*	.489320
3	{3}	.926297	.002480*		.123648
4	{4}	.379624	.489320	.123648	

1= liver enzymes. 2= nausea and vomiting. 3= diarrhea. 4= dysphagia

Table 7b: MAIN EFFECT: CASE (score 3 – expert) averaged over question format

	Sum of Squares	df	Mean Square	F	p-level
Effect	.019705	3	.006568	.362041	.780633
Error	1.034115	57	.018142		

4c) Details of 3x1 repeated measures – question format analysis for each clinical presentation for Tables 8a (non-expert) and 8b (expert)

i) Table 8a

1. Elevated Liver Enzyme

MAIN EFFECT: FORMAT (score 3 - non-expert)

	Sum of Squares	df	Mean Square	F	p-level
Effect	.602083	2	.301042	3.891566	.029006
Error	2.939583	38	077357		

Scheffe test; variable Var.1 (score 3 - non-expert)
Probabilities for Post Hoc Tests
MAIN EFFECT: FORMAT

		{1} .5125000	{2} .5125000	{3} .3000000
1	{1}		1.000000	.066198
2	{2}	1.000000		.066198
3	{3}	.066198	.066198	

1=Multiple Choice
2=Extended Matching
3=Elimination

2. Nausea & Vomiting

MAIN EFFECT: FORMAT (score 3 - non-expert)

	Sum of Squares	df	Mean Square	F	p-level
Effect	1.889583	2	.944792	5.139576	.010579

Error	6.985417	38	.183827

Scheffe test: variable Var.1 (score 3 - non-expert)
Probabilities for Post Hoc Tests
MAIN EFFECT: FORMAT

		{1}	{2}	{3}
		.4000000	.8125000	.7250000
1	{1}		.015891*	.068878
2	{2}	.015891*		.812927
3	{3}	.068878	.812927	

1=Multiple Choice
2=Extended Matching
3=Elimination

3. Diarrhea

MAIN EFFECT: FORMAT (score 3 - non-expert)

	Sum of Squares	df	Mean Square	F	p-level
Effect	1.225000	2	.612500	7.487936	.001813
Error	3.108333	38	.081798		

Scheffe test: variable Var.1 (score 3 - non-expert)
Probabilities for Post Hoc Tests
MAIN EFFECT: FORMAT

		{1}	{2}	{3}
		.2750000	.3250000	.6000000
1	{1}		.858813	.003856*
2	{2}	.858813		.015961*
3	{3}	.003856*	.015961*	

1=Multiple Choice
2=Extended Matching
3=Elimination

4. Dysphagia

MAIN EFFECT: FORMAT (score 3 - non-expert)

	Sum of Squares	df	Mean Square	F	p-level
Effect	.158333	2	.079167	1.752427	.187072
Error	1.716667	38	.045175		

ii) Table 8b:

1. Elevated Liver Enzyme

MAIN EFFECT: FORMAT (score 3 - expert)

	Sum of Squares	df	Mean Square	F	p-level
Effect	.127083	2	.063542	1.234292	.302445
Error	1.956250	38	.051480		

2. Nausea & Vomiting

MAIN EFFECT: FORMAT (score 3 - expert)

	Sum of Squares	df	Mean Square	F	p-level
Effect	.064583	2	.032292	.340659	.713451
Error	3.602083	38	.094792		

3. Diarrhea

MAIN EFFECT: FORMAT (score 3 - expert)

	Sum of Squares	df	Mean Square	F	p-level
Effect	.268750	2	.134375	1.991064	.150544
Error	2.564583	38	.067489		

4. Dysphagia

MAIN EFFECT: FORMAT (score 3 - expert)

	Sum of Squares	df	Mean Square	F	p-level
Effect	.077083	2	.038542	1.586908	.217815
Error	.922917	38	.024287		