EXPERTISE IN NURSES' CLINICAL JUDGMENTS:
THE ROLE OF COGNITIVE VARIABLES AND EXPERIENCE

by

LYNDA A. CHRISTIE

BSN, The University of British Columbia, 1967
MA (Education), Simon Fraser University, 1986

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
in
THE FACULTY OF GRADUATE STUDIES
DEPARTMENT OF EDUCATIONAL PSYCHOLOGY
AND SPECIAL EDUCATION

We accept this dissertation as conforming to
the required standard

UNIVERSITY OF BRITISH COLUMBIA
August 1996
© Lynda A. Christie, 1996
In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Educational Psychology and Special Education

The University of British Columbia
Vancouver, Canada

Date August 29, 1996
ABSTRACT

Many researchers have failed to find a relationship between experience and judgment accuracy. In this study the purpose was to understand the relationship between experience and expertise in clinical judgment. Common sense suggests that experienced subjects make better quality judgments, compared to novices. Clinical judgments, however, are ill-structured and characterized by uncertainty; they take place in a dynamic context, with delayed or nonexistent feedback and are difficult to learn.

Cognitive operations that translate "cues" (such as risk factors, signs, and symptoms) into judgments are not fully understood. Cognitive constructs (conceptual structure, sensitivity to patterns in data, and judgment process) and individual differences in age, education, and experience were explored to identify their relationship to judgment expertise. Indicators of judgment quality were: accuracy, consistency, latency, confidence, calibration, and knowledge accessibility.

In phase 1 of this study, cues were identified that best predicted healing time for 258 surgical patients with abdominal incisions. In Phase 2, the subjects were 36 nurses with a range of experience caring for surgical patients. Generating both quantitative and qualitative data, subjects made judgments about incisional healing on the basis of information from actual patients. Multidimensional scaling was used to reveal conceptual structure, and lens modeling was applied to assess sensitivity to broad patterns. An information board task with think-aloud protocols demonstrated judgment process. The selection of tasks was based on their analysis- or intuition-inducing features, using K. R. Hammond's (1990) cognitive continuum theory.

Experience accounted for a only a small proportion of variance in performance, whereas confidence in judgment was more strongly related to experience. Taken together, these findings replicated previous research. Protocol data showed that
metacognition, knowledge accessibility, and reflectivity increased with experience. Conceptual structure predicted judgment accuracy under intuitive conditions. Support was found for Dreyfus and Dreyfus' (1986) hypothesized transition in cognition, from deliberate processing of discrete cues, to intuitive processing of patterns of cues encoded in memories for specific cases.

This study has theoretical significance by adding to knowledge about clinical judgment, and by increasing understanding of cognitive changes associated with expertise. This study has practical significance in providing direction for the development of teaching methods aimed to increase learning from experience in probabilistic contexts.
# TABLE OF CONTENTS

Abstract ........................................ ii  
Table of Contents ................................ iv  
List of Tables .................................... ix  
List of Figures .................................... xii  
Acknowledgements ................................ xiii  
Dedication ....................................... xvi  

CHAPTER 1 INTRODUCTION TO THE STUDY ................. 1  
Background ...................................... 2  
The Context ..................................... 3  
The Problem Relating Experience and Judgment Accuracy .......... 4  
Explanations for the Lack of Relationship ..................... 5  

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conceptual Structure</td>
<td>6</td>
</tr>
<tr>
<td>Sensitivity to Patterns in Data</td>
<td>7</td>
</tr>
<tr>
<td>Judgment Process</td>
<td>10</td>
</tr>
<tr>
<td>Assessment of Experience</td>
<td>11</td>
</tr>
</tbody>
</table>

Summary of the Problem ................................ 12  
Purposes of the Study .................................. 13  
Significance of the Study ................................ 13  
Research Questions .................................... 14
CHAPTER 2 REVIEW OF THE LITERATURE .............................................. 16

Section A: Expertise In Clinical Judgment ............................................. 16

Characteristics of Expertise ............................................................. 16
Dreyfus' Model of Expertise ............................................................. 21
Characteristics of Clinical Judgment .................................................. 25
Lens Model for Clinical Judgment ....................................................... 29

Section B: Summary of Relevant Research .......................................... 32

Judgment Accuracy as an Indicator of Expertise ................................... 32
Consistency of Clinical Judgment ....................................................... 35
Additional Indicators of Expertise ...................................................... 36

Section C: Findings from the Literature .............................................. 42

Conceptual Structure ....................................................................... 42
Sensitivity to Patterns in Data ............................................................ 44
Judgment Process .............................................................................. 51
Individual Differences in Education and Age ...................................... 61
Individual Differences in Experience ................................................ 62

Section D: Additional Factors Influencing Judgment Results .............. 65

Heuristics and Cognitive Biases .......................................................... 65
Experience and the Use of Heuristics .................................................. 70
Research Assumptions and Design Factors ......................................... 71

Section E: Summary ........................................................................... 78

CHAPTER 3 RESEARCH QUESTIONS, METHODS, AND
ANTICIPATED PATTERNS OF RELATIONS ........................................... 80

Section A: Research Questions, Design, and Procedures .................... 80

Overall Research Question .................................................................. 80
Specific Research Questions ............................................................... 80
Design and Procedures ....................................................................... 81
Sample Size ....................................................................................... 83
<table>
<thead>
<tr>
<th>Section and Subsection</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Judgment Process</td>
<td>158</td>
</tr>
<tr>
<td>Age, Education, and Experience</td>
<td>170</td>
</tr>
<tr>
<td>Section D: Results Related to Multivariate Analyses</td>
<td>174</td>
</tr>
<tr>
<td>Results Related to Research Conditions</td>
<td>174</td>
</tr>
<tr>
<td>Results Related to Methods for Confidence Assessment</td>
<td>178</td>
</tr>
<tr>
<td>Section E: Summary</td>
<td>178</td>
</tr>
<tr>
<td>Prediction of Judgment Expertise</td>
<td>178</td>
</tr>
<tr>
<td>Summary of Patterns of Relations</td>
<td>181</td>
</tr>
<tr>
<td><strong>CHAPTER 5 DISCUSSION</strong></td>
<td>183</td>
</tr>
<tr>
<td>Section A: Discussion of Major Findings</td>
<td>183</td>
</tr>
<tr>
<td>Research Question 1</td>
<td>183</td>
</tr>
<tr>
<td>Research Question 2</td>
<td>189</td>
</tr>
<tr>
<td>Research Question 3</td>
<td>199</td>
</tr>
<tr>
<td>Research Questions 4 and 5</td>
<td>200</td>
</tr>
<tr>
<td>Research Question 6</td>
<td>203</td>
</tr>
<tr>
<td>Section B: Discussion Related to Methods</td>
<td>204</td>
</tr>
<tr>
<td>Findings Related to Methods</td>
<td>204</td>
</tr>
<tr>
<td>Findings Related to Research Traditions</td>
<td>209</td>
</tr>
<tr>
<td>Limitations of the Study</td>
<td>213</td>
</tr>
<tr>
<td>Section C: Contribution to Theory Development</td>
<td>215</td>
</tr>
<tr>
<td>Section D: Implications for Education and Practice</td>
<td>223</td>
</tr>
<tr>
<td>Section E: Summary</td>
<td>224</td>
</tr>
<tr>
<td>Summary Related to Research Questions</td>
<td>224</td>
</tr>
<tr>
<td>Recommendations for Further Research</td>
<td>227</td>
</tr>
<tr>
<td>Conclusion</td>
<td>232</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>234</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table 3-01: Means and Standard Deviations for the Dependent Variable, Healing Time, and Regression Model R-Squared for Normal Order, Reverse Order, and Cross-Validation Cases .................. 101

Table 3-02: Outline of Tasks, Measures, and Associated Constructs ............... 112

Table 4-01: Intercorrelations Between Significant Predictors and Log Transformation of Healing Time ........................................ 123

Table 4-02: Indices of Variable Importance for Significant Predictors of Healing Time ................................................................. 124

Table 4-03: Classifications of Subjects According to Experience Level, Age Category, and Education .................................................... 127

Table 4-04: Means and Standard Deviations for Accuracy Assessed by Error Variance and Percent Correct for Normal Order and Reverse Order Cases, by Experience Level ........................................ 130

Table 4-05: Means and Standard Deviations for Average Confidence (Concurrent Method) and Overall Confidence (Retrospective Method) for Normal Order and Reverse Order Cases, by Experience Level .................................................. 131

Table 4-06: Nonparametric Summaries of Consistency for Repeated Judgments of Healing Time and Repeated Judgments of Confidence (Concurrent) for Normal Order and Reverse Order Cases, by Experience Level ........................................ 132

Table 4-07: Means and Standard Deviations for Two Methods of Assessing Judgment Calibration for Normal and Reverse Order Cases, by Experience Level .................................................. 136

Table 4-08: Means and Standard Deviations for Two Methods of Assessing Judgment Quality for Normal Order and Reverse Order Cases, by Experience Level .................................................. 138

Table 4-09: Means and Standard Deviations for Confidence Averaged Over Correct and Incorrect Judgments and for Confidence-Accuracy Correlations for Normal Order and Reverse Order Cases, by Experience Level .................................................. 139
Table 4-10: Means and Standard Deviations for Latency (in Seconds) Averaged Over Cases for Normal Order and Reverse Order Cases, by Case Sequence and by Experience Level .......... 141
Table 4-11: Means and Standard Deviations for Knowledge Accessibility Scale, by Experience Level ................................. 142
Table 4-12a: Intercorrelation Matrix for Experience and Indicators of Expertise Based on Performance on Normal Order Cases .......... 143
Table 4-12b: Intercorrelation Matrix for Experience and Indicators of Expertise Based on Performance on Reverse Order Cases .......... 144
Table 4-13: Means and Standard Deviations for Conceptual Coherence Scale from Protocol Data, by Experience Level ................................. 146
Table 4-14: Means and Standard Deviations for Average Distance Between Concept Pairs in Selected Clusters, by Experience Level ............................................................ 150
Table 4-15 Nonparametric Measures of Performance on the Lens Model Task for Normal Order and Reverse Order Cases, by Experience Level ............................................................ 153
Table 4-16: Means and Standard Deviations for Measures Derived from Lens Model Analyses for Normal Order and Reverse Order Cases, by Experience Level ............................................................ 154
Table 4-17: Means and Standard Deviations for Accuracy of Cross-Validation of Judgment Policies (Based on Error Variance Measures and Percent Correct Measures), by Experience Level ............................................................ 157
Table 4-18: Means and Standard Deviations for Data Search Measures on the Information Board Task and Number of Subjects with Correct Judgments, by Experience Level ............................................................ 160
Table 4-19: Means and Standard Deviations for Judgment Process Scale, by Experience Level ............................................................ 161
Table 4-20: Means and Standard Deviations for Measures of Calibration of Healing Time and Omega-Squared from the Orthogonal Judgment Task, by Experience Level ............................................................ 166
Table 4-21: Means and Standard Deviations for Measures of Configural Cue-Use, by Experience Level ............................................................ 167
Table 4-22a: Intercorrelations Among Measures of Configural Cue-Use, Accuracy Measures, and Experience, for Normal Order Cases ................................................................. 168

Table 4-22b: Intercorrelations Among Measures of Configural Cue-Use, Accuracy Measures, and Experience, for Reverse Order Cases ................................................................. 169

Table 4-23a: Intercorrelations Between Experience Measures, Age, Education, and Judgment Accuracy for Normal Order Cases ................................................................. 171

Table 4-23b: Intercorrelations Between Experience Measures, Age, Education, and Judgment Accuracy for Reverse Order Cases ................................................................. 172

Table 4-24: Correlation Coefficients Between Cognitive Measures and Measures of Judgment Accuracy for Normal Order and Reverse Order Cases ................................................................. 173

Table 4-25: Multivariate Analysis of Variance for the Effects of Slide Condition and Paragraph Order (Normal Order and Reverse Order) on Judgment Accuracy (Error Variance Measure) ................................................................. 175

Table 4-26: Multivariate Analysis of Variance for the Effects of Slide Condition and Paragraph Order (Normal Order and Reverse Order) on Judgment Accuracy (Percent Correct Measure) ................................................................. 176

Table 4-27: Means and Standard Deviations for Judgment Accuracy (Percent Correct for all cases -- Normal Order and Reverse Order), by Slide Condition and Experience Level ................................................................. 177

Table 4-28a: Intercorrelations for Predictor Variables for Accuracy for Normal Order Cases (Percent Correct Measure), and Experience ................................................................. 179

Table 4-28b: Intercorrelations for Predictor Variables for Accuracy for Reverse Order Cases (Error Variance Measure), and Experience ................................................................. 180

Table 4-28c: Intercorrelations for Predictor Variables for Accuracy for Reverse Order Cases (Percent Correct Measure), and Experience ................................................................. 181
LIST OF FIGURES

Figure 1:
Brunswik's Lens Model, showing conceptual relations .................. 30

Figure 2:
Brunswik's Lens Model, showing mathematical relations .................. 88
ACKNOWLEDGEMENTS

I wish to acknowledge and thank many individuals and groups for the help I was given with this research. First, I want to thank my Committee members, Dr. Patricia Arlin, Dr. Nand Kishor, and Dr. Janet Jamieson, for assisting me to bring this study to completion. They helped me develop and clarify my ideas, and express them more concisely. They also provided useful feedback on an earlier draft of this dissertation. In particular, Dr. Arlin's influence as a researcher, as a mentor, and as a friend was very much appreciated.

I also want to thank Dr. Arleigh Reichl for his assistance with applying the calibration equations, and for helping me run the various SPSS computer programs.

In the summer of 1993, I was fortunate to receive an Educational Leave. I thank the administration at Langara for that privilege. In addition, I thank the Langara Research Committee for the grant I received.

I thank the staff at the Registered Nurses' Association of BC for providing assistance in advertising for experienced nurses as subjects, and for giving me a research bursary.

I want to thank the surgical patients and the nurses who volunteered for this study. Without their participation, the research would not have been possible. I acknowledge the assistance given by Dr. Chris Bradley, Director of Research and Evaluation, at Vancouver Hospital, and Mariette Carmen, Enterostomal Therapist, at St. Pauls' Hospital. I also thank Dr. Diane Cooper, clinical nurse scholar from the University of Texas, for her input regarding the wound healing concepts.

To all my colleagues from the Langara Nursing program, I owe a very heartfelt "thanks" for their encouragement and support. In the early years, Karen helped
by sharing clinical groups, so I could attend classes during the day, Anna was always asking how I was coming along, and Alice listened to my ideas for a proposal. Jane and Gail were in the clinical area providing encouragement as I began the long process of finding 281 surgical patients who would be willing to volunteer for my study. More recently, Ken and Wanda tested out my tasks, Glenda and Maureen helped me to better organize myself, and Jack reminded me to "focus". Lately, I appreciate the fact that many colleagues stopped asking how I was coming along. Audrey, Pam, Mary, and Barb wished me well as I prepared for my oral defense.

I need to say a special thanks to Roberta for the many ways she has facilitated this research. During the term, we did some joint planning and team teaching -- Roberta helped me keep on track, as I combined teaching and research. In preparation for attending AERA in New York, Roberta re-coded several of the verbal protocols so I could calculate Cohen's Kappa, and we worked together constructing the poster. She was a resource for spelling, grammar, and APA style (including the removal of anthropomorphisms); she taught me to use WordPerfect, patiently listened to my many problems, and also shared in the exciting times. I am hoping my motivation for dissertation research might be contagious, and I can do the same for her one day.

Many people provided technical support. Ted Morris from Ubiquitous Software developed the computer programs that were used to present two of the tasks to subjects, and Sue transcribed tapes of nurses' dialogue. Tom did great work with the lens model diagram, and Joe constructed the chart showing the sequencing of tasks. Blair transformed my numbers into neat, readable tables, and Atsuko from the Langara library obtained papers by FAX from Ohio and Alberta in
order for me to meet my deadlines. Finally, my son, Tim, caught errors as he proofread.

I want to express my sincere appreciation to Dr. Robert M. Hamm, Director of the Program in Clinical Decision Making at the University of Oklahoma College of Medicine. Dr. Hamm was the external reviewer for this dissertation. I thank him for reviewing the entire document so thoroughly, and making many excellent recommendations. I also value his suggestions regarding future publications.

Last, but not least, I must thank my family for their patience with my total absorption in this dissertation over the years. I thank John for letting me print out many drafts at his store, and for not noticing how late it was or how much paper I used. I realize that while I have been involved in this research, I have not had time to be a traditional "Mom"; I hope my children -- Ruth, Dan, Tim, and Jeannene -- can understand how much it has meant to me to be able to pursue this goal and to finally reach it.

I thank everyone who helped me make the completion of this research a reality.
I wish to dedicate this dissertation to the memory of my father, Eric Schwab, who passed away in August, 1988, at the age of 74. He encouraged me from an early age to love learning, to seek academic challenges, and to complete what I started. He was the best father anyone could have had.

I also dedicate this dissertation to my mother, Joyce Schwab, who has been a constant source of support and inspiration during all the time I have spent on this research. She has listened patiently to my difficulties, and encouraged me to move ahead a little each week. She also shared in my happiness as each chapter was completed. I know I was always in her prayers, and for that I am thankful.
I. INTRODUCTION TO THE STUDY

Making judgments is a fundamental human activity in which cognitive operations are used to integrate separate elements of information (cues) into a unitary response. This response is a person's attempt to weigh the importance of information and combine relevant cues in an optimal manner. Examples of judgments include predicting future events, estimating the value of a product or idea, and evaluating the relative merits of various courses of action in a particular situation. Stating the aggregate meaning of a number of equivocal informational cues such as test results, history, risk factors, and signs and symptoms constitutes acts of judgment common in educational and medical contexts.

People sometimes reserve the term 'judgment' for those assessment situations where processing of information can be consciously recalled. Such a restriction is limiting. A person in a familiar context may synthesize cues into judgments without necessarily being able to articulate how such a task is achieved. In fact, studies have revealed that the mind uses nonconsciously acquired (or tacit) knowledge to make judgments (Kihlstrom, 1987; Lewicki, Czyzewska, & Hoffman 1987; Mitchell & Beach, 1990).

Over the past 25 years, considerable judgment research has been carried out (Hogarth, 1980; Slovic, Fischhoff, & Lichtenstein, 1977; Slovic & Lichtenstein, 1971; Tversky & Kahneman, 1974). The literature includes many examples where judgments have been modeled (or policies were captured"). Examples of such studies include the prediction of children's reading achievement (Cooksey & Freebody, 1985, 1987), the selection of students for graduate school (Dawes & Corrigan, 1974), the identification of students' preferences for jobs (Einhorn, 1971),

1 Terms peculiar to the literature on judgment are included in the glossary in Appendix A.
and physicians' judgments of severity of congestive heart failure (LaDuca, Engel, & Chovan, 1988).

**Background**

In the health-care field, professionals must process vast quantities of data when making judgments. In this dynamic setting, both normal variations and pathological manifestations are monitored. Making judgments can be extremely challenging. In the past, difficulty with clinical judgment has been attributed to the lack of information. Currently, however, problems seem more often related to an abundance of data, together with the well-known limitations imposed by information-processing constraints. Furthermore, difficulty exists because clinical judgment tasks are ill-structured, and characterized by uncertainty: most data have a probabilistic relationship to the judgment (Christensen & Elstein, 1991; Eddy, 1988).

Not all aspects of clinical judgment are fully understood. In particular, how a clinician weighs and combines various cues has been a source of unresolved inquiry. One reason for the difficulty in studying clinical judgments is that actions taken based on judgment may alter subsequent events. Also, it is difficult to simulate judgment tasks in the laboratory; yet, very often, conducting a judgment study in the clinical setting poses ethical and practical problems.

The study of clinical judgment has a long history as an important area of research (Bieri et al., 1975; Einhorn, 1974; Meehl, 1954; Oskamp, 1962; Sarbin, Taft, & Bailey, 1960). In addition, recent studies aimed to capture physicians' judgment policies are available (Deber, 1986; Poses & Anthony, 1991; Speroff, Connors, & Dawson, 1989; Wigton, 1988).
The Context

Clinical judgment takes place in many settings with a variety of clinicians. Nurses were the subjects in this study, and their judgments about wound healing in surgical patients were investigated. As health care professionals, nurses make clinical judgments on a regular basis. Many studies of nurses' judgments are found in the literature (Benner, 1984; Benner, Tanner, & Chesla, 1992; Corcoran, 1986b; Gordon, 1980; Gordon, Murphy, Candee, & Hiltunen, 1994; K. R. Hammond, 1964, 1966a; K. R. Hammond, Kelly, Schneider, & Vancini, 1966; Itano, 1989; Kelly, 1964).

The context of this study was nurses' judgments related to wound healing. Cooper (1990a, 1990b), a clinical nurse scholar with expertise in this field, viewed assessment of wound healing as fundamental to caring for surgical patients. Cooper (1990b) argued that "only with a clear grasp of the intricacies of the healing trajectory are clinicians able to provide optimal, goal-directed wound care" (p. 168). Judgments of wound healing are based on physiological theory that reveal connections between the cues and the time that incisions take to heal (Cruse & Foord, 1973, 1980; Irvin, 1981; P. L. Jones & Millman, 1990; Viljanto, 1991). This connection with theory is hypothesized to provide a basis for causal schemas that link particular data to healing time. In addition, cues that are diagnostic of problems associated with healing should, over time, become perceived as a pattern.

Considerable research in cognitive psychology has been focused on understanding performance in problem solving, judgment, and decision making (Groen & Patel, 1991; Kahneman, 1991; Pitz & Sachs, 1984; Shulman & Elstein, 1975). Research on judgment quality within a clinical judgment context has potential to reveal cognitive changes that contribute to the understanding of expertise in general (Elstein & Bordage, 1988; Norman, Patel, & Schmitt, 1990). Making
judgments can be considered as a type of unstructured problem solving where no algorithm exists to calculate a correct response. Selected studies related to problem solving, as well as research directly related to judgment performance, therefore, have been utilized. Although nursing judgments are substantively different from medical judgments, they are similar in process; pertinent literature on judgment from both nursing and medical perspectives has been reviewed.

The Problem Relating Experience and Judgment Accuracy

In this study, the problem being investigated is that of understanding the relationship between experience and expertise in clinical judgment performance. Common sense suggests that experienced nurses will make better quality judgments, compared to novices. Nurses interact with a variety of patients where exposure to patterns of clinical data occurs repeatedly. It seems logical to expect that, with experience, nurses would learn to weigh and integrate cues in a manner that reflected the clinical ecology.

Many studies on clinical judgment, however, fail to demonstrate a positive relationship between experience and judgment accuracy. For example, Goldberg (1968), in his thorough review, concluded that "the amount of professional training and experience does not relate to judgmental accuracy" (p. 484). More recently, Garb (1989, 1992) provided evidence that supported Goldberg's claim. In addition, Camerer and Johnson (1991) claimed "experts, paradoxically, are no better at making predictions than novices" (p. 210).

Literature in clinical medicine revealed the same findings; for example, Dawson, Connors, Hsiue, and Shaw (1987) found no relationship between experience and judgment accuracy of patients' hemodynamic status. Studies
consistently fail to reveal what is generally assumed to be true: that clinicians increase their clinical judgment accuracy with experience.

**Explanations for the Lack of Relationship**

Brehmer (1980b), Dawes (1989), and Einhorn (1980) examined the difficulties involved in learning from experience in probabilistic environments. These authors explained the absence of a relationship between experience and judgment accuracy shown in previous studies: when outcomes are dependent on the judgments made, clinicians are prevented from obtaining accurate feedback needed to learn optimal decision rules. In addition, Einhorn and Hogarth (1978) proposed that prolonged clinical experience can induce high *confidence* in the accuracy of judgments, exceeding that warranted by judgment validity; such confidence functions to maintain an *illusion of validity*.

These explanations, however, may not fully account for previous results. The possibility exists that previous findings represented type II errors (a positive relationship could exist, but researchers have failed to reveal the true state). Schmidt (1992, 1996) explained that researchers in psychology and other social sciences have traditionally relied on a variety of statistical tests of significance in interpreting the meaning of data. The prevailing decision rule has been as follows: if the statistic (for example, *t* or *F*) is *significant*, there is an effect or a relation. If the statistic is *not significant*, then the conclusion is made that there is no effect or relation. When traditional procedures for interpretation are used, the focus is on controlling type I error. Cohen (1990, 1994) and Schmidt (1992, 1996) claimed that beta levels (type II errors) are frequently in the 50% to 80% range.

Asher (1993) made the point that meaningfulness must be assessed on the basis of theory. Serlin (1987) stated that "it is only on the basis of theory that one
can decide . . . whether the experimental results can be generalized" (p. 365). In
the present study, the theoretical conceptions included the ideas that the relationship
between experience and judgment expertise revealed through laboratory research is
influenced by the conditions under which judgment performance is assessed and
how experience is defined. In addition, cognitive factors may interact with indicators
of expertise and with the conditions. The particular cognitive factors examined are
conceptual structure, sensitivity to patterns in data, and judgment process. It is
possible that small, but nevertheless meaningful relations (that is, relations that are
meaningful within the context of a particular theoretical perspective), are present that
are not revealed through traditional means.

**Conceptual Structure**

Conceptual structure refers to one's mental organization of domain-specific
conds, including the inter-relationships among concepts. This structure
significantly influences the interpretation a person makes in response to
combinations of clinical cues. Several researchers have identified particular
changes in conceptual structure with expertise (Bordage & Lemieux, 1991; Homa,
Rhoads, & Chambliss, 1979; Medin, 1989).

It is not the configurations of cues, in themselves, that function as causal
stimulus for judgment performance, but rather, the meaning of cue configurations
constructed within the conceptual structure. It seems reasonable, therefore, to
propose that conceptual structure could be an important variable to consider when
studying judgment performance.

One rationale for previous judgment studies failing to demonstrate a
relationship between clinical judgment performance and experience is that in these
studies the fact that certain changes in clinicians' conceptual structure may be
essential for expert judgment to be demonstrated may not have been considered. Changes in conceptual structure do not occur automatically with experience; in previous studies, subjects with expert-like conceptual structures may not have been adequately represented.

**Sensitivity to Patterns in Data**

Reber (1989) claimed that "when a stimulus environment is structured, people learn to exploit that structure" (p. 221). Experts demonstrate a high degree of sensitivity to patterns in the data. One rationale for previous investigators failing to find a relationship between experience and judgment performance is that typical patterns in clinical data may not have been included in experimental contexts; these omissions would tend to reduce experience-related differences obtained from performance in laboratory tasks.

Birnbaum (1982) pointed out the importance of the structure of the laboratory task. If the structure designed for research is different from that of the clinical ecology, intuitive (tacit) knowledge acquired from experience may not be useful as a guide to judgment performance. Clinicians who make judgments in such a setting would not benefit from using the typically effective cognitive strategies and heuristics that they normally used (Tversky & Kahneman, 1974).

Two categories of cues examined in this study were referred to as context cues and individuating cues. Context cues included global information about patients (such as age, gender, diagnosis, and surgery) that facilitated the construction of an appropriate context. Individuating cues included specific patient data (such as surgery time, complications, and blood loss). In making judgments, clinicians frequently use the context cues as an anchor to make a rough estimate,
while the other cues are used for making precise adjustments (Tversky & Kahneman, 1974).

In this study, the order of cue-presentation was manipulated. K. R. Hammond (1990) claimed that task factors influence the type of cognition that subjects employ: Hammond considered tasks that involved rapid judgments as intuition-inducing, whereas he regarded tasks that required slow deliberation as analysis-inducing. If expertise in judgment were contingent upon clinicians' sensitivity and responsiveness to patterns of cues, then altering the sequence of cue presentation was theorized to have a selective influence on performance accuracy for experienced subjects. Performance for these subjects was expected to decrease when cases were presented in an atypical order as this manipulation was anticipated to interfere with intuitive processing.

Structural differences between the cues represented in research tasks and those naturally available in life are often unavoidable. In clinical settings, information is not formatted in terms of discrete cues, but is an integral aspect of phenomena and events in a continuous environment. Friedman, Howell, and Jensen (1985) stated:

Policy-capturing studies generally use numerical values as cues -- in essence, providing the judge with data that have already been processed to a degree. In their natural habitat, of course, decision problems are not conveniently decomposed into these elements (p. 666).

Because life does not come "pre-packaged" so precisely (Kolers, 1977, 1979a), the presentation of labelled cues for research has potential to alter task structure. Experienced clinicians may be prevented from exhibiting the degree of excellence of which they are capable in making judgments from the "raw" sensory input. Tasks structured for research purposes often contain little ambiguity; reality is
translated into symbolic description. With such reduction, the difficulty for novices decreases, and an artificial ceiling for experienced clinicians is created. Any relationship between experience and expertise becomes attenuated. Recently, Lamond, Crow, Chase, Doggen, and Swinkies (1996) raised this concern about the presentation format of information with respect to research in nursing.

To address these problems, the researcher may consider employing alternative design features such as conducting studies in naturalistic settings with authentic patients; such a solution, however, poses numerous ethical and practical difficulties. A possible approach to minimize the effects of the difference in cue-structure between low-fidelity tasks and actual judgments that occur in the clinical setting, may be to prime subjects' memories for previous cases. Showing slides of critical stimuli to subjects as part of the research method may be sufficient to act as a "prime" for subjects who have had relevant experience. If experts implicitly use their rich store of memory-based cases to facilitate judgment performance, this deliberate triggering of mental context may be a way to tap into implicit, or tacit, knowledge that guides expert judgment. In consumer research, Bettman and Sujan (1987) and Herr (1989) found differential priming effects on judgment for high- and low-knowledge subjects.

In this study, case-based memories are theorized to be part of an internal context, or schema, that influences performance (Groen & Patel, 1986; Patel, Groen, & Frederiksen, 1986). If presenting slides is successful in instantiating this mental context, judgment performance in the laboratory may be selectively enhanced. Subjects who have implicit knowledge of cue-criterion relations that reflect the clinical ecology, should have greater facilitation of judgment performance, compared to other subjects.
Judgment Process

A third explanation for the lack of a relationship between experience and judgment performance in previous studies is that traditional methods to assess performance under research conditions may not have been designed to capture the judgment process used by expert clinicians in the clinical setting. During the time when clinicians learn to make particular judgments, information processing is theorized to be initially an application of proposition-like rules (declarative knowledge, or "knowing that"). Later the rules become transformed into a dynamic network of relations (procedural knowledge, or "knowing how"). Unless research tasks used to reveal expertise have sufficient flexibility to allow for such processing differences, expertise-experience relationships as measured in the laboratory may not be a valid indicator of actual relationships.

It is not clear from the judgment literature how patterns of information are processed, and how such processing varies with experience. The topic of pattern recognition has been of interest to cognitive psychologists for many years (Campbell, 1966; Neisser, 1976; Reed, 1972). When clinicians work in an area for a period of time, they acquire an implicit understanding of the patterns of cues in the ecology which is reflected in their judgment process. This knowledge may be used procedurally as a basis for processing the data from each case, not as separate pieces, but interactively (or configurally). Perceptual recognition is enhanced by the meaningfulness of cue patterns (Reicher, 1969).

Experienced judges often claim that they process cues in terms of their patterns (Brehmer, 1969; Brooks, Norman, & Allen, 1991; Goldberg, 1969; Hoffman, 1968). Meehl (1950), over 45 years ago, stated:
One of the most important words in the vocabulary of clinicians is the word 'pattern'. We speak of ourselves as thinking in terms of totalities, organizations, configurations; and we look upon case material in this patterned, non-atomistic way (p. 165).

In the literature, there are major differences in interpretation of research on the role of configural processing in judgment performance. The authors of many studies using regression methods demonstrated that patterned processing of cues accounted for only a small proportion of variance in judgment performance (K. R. Hammond, Hursch, & Todd, 1964; K. R. Hammond & Summers, 1965). Other authors, such as N. H. Anderson (1969, 1972) and N. H. Anderson and Shanteau (1977) claimed that configurality in judgment is common, but cannot be measured by regression methods.

To the extent that experts process cues configurally as a means of attaining accuracy, researchers who use approaches which are insensitive to configural processing may fail to find difference with experience. Methods of assessing configurality in addition to regression have been explored.

**Assessment of Experience**

A fourth possibility for failure to find a relationship between experience and judgment performance is that experience may not have been assessed optimally. In this study, experience was measured in three ways. First, the number of years since nursing graduation was obtained, providing a time-based measure with good variability. The second index of experience was the number of months caring for surgical patients; this method had a possible advantage over the first in that the relevance of experience to the judgment was taken into account. The third approach was to obtain an estimate of the number of times the subject viewed an incision during his or her career.
Age and education were important variables to include in order to make correct interpretations of the relations between experience and judgment performance. The experienced nurse is typically older, but may, or may not, have had more education compared to the inexperienced nurse. Any difference in expertise associated with experience, therefore, may be more reflective of the effects of age and/or education than length of work experience. By obtaining data on both age and education, and using three indices of experience, a more complete understanding of the relationships may be attained.

Summary of the Problem

The problem that has been addressed in this research is summarized as follows: several researchers have searched for, but have failed to find, a relationship between experience and clinical judgment performance. On one hand, the conclusion that no relationship exists may be valid, with explanations provided by Brehmer (1980b), Dawes (1989), Einhorn (1980), and Einhorn and Hogarth (1978) for why accuracy in judgment performance does not increase with experience. On the other hand, these findings may represent only partial truth; a positive relationship may in fact exist, but for many reasons, it is difficult to demonstrate in the laboratory context. A deeper understanding of the patterns of relationships among cognitive factors, individual differences, and clinical judgment performance is necessary to reveal the nature of the relationship between judgment expertise and experience.

It may be that it is not experience assessed in years or months that is directly related to judgment performance, but rather the cognitive impact that such experience has had. Such impact varies with differences in cognition as demonstrated through conceptual structure, sensitivity to patterns in data, and judgment process. In this study these explanations were explored.
Purposes of the Study

The three purposes of this study were:

1) to demonstrate differences in accuracy of performance in a clinical judgment task and identify patterns of relations between indicators of expertise from the literature (including judgment accuracy) and experience;

2) to assess whether cognitive constructs (conceptual structure, sensitivity to patterns in data, and judgment process), and individual differences in age, education, and experience were predictors of clinical judgment performance;

3) to assess whether conditions (order of cue-presentation condition and memory-priming condition) revealed patterns in clinical judgment performance.

Significance of the Study

The study has both theoretical and practical importance. First, it has potential to add to knowledge about how people search for, weight, and integrate cues to make clinical judgments in a probabilistic context. Because there is considerable evidence that such judgments are difficult to make, and may not be made very accurately, this is an important aspect of judgment research to pursue. The second aspect of theoretical significance is that this study has potential to contribute to knowledge about changes in cognition with expertise. An attempt to identify possible reasons why many clinical judgment studies have failed to find a relationship between judgment accuracy and experience ought to illuminate some of the means by which judgment expertise occurs.

The major practical significance of this study is that better understanding of the reasons for previous findings will provide some direction for educators in the
health care field. These educators are responsible for developing curricula and teaching clinical judgment to student health care professionals. The knowledge gained about the influence of cognitive factors may be useful to assist educators in encouraging ways to increase students' learning from experience. In addition, the findings may constitute evidence that for some judgment contexts, experience does not automatically bring about increased judgment performance. This study may provide an impetus for the development of methods to enhance appropriate learning from experience.

Another more specific aspect of this study that has potential for practical importance is that related to judgment confidence. From the literature, many experienced practitioners have high confidence in the accuracy of their judgments that may be unwarranted. As Schwartz and Griffin (1986) pointed out, clinical judgments are risky, with costs associated with errors of both types (misses and false alarms); this research into the influence of cognitive variables and experience is a preliminary step towards identifying ways to improve judgment calibration.

In summary, a judgment study is important in educational psychology. Lewis and Anderson (1985) claimed: "As educators we would like to make the acquisition of expertise more efficient; as psychologists, our interests lie in defining the processes which underlie complex skill acquisition" (p. 26-27). Because making judgments is an attempt to go "beyond the data" in a way that leads to accurate outcomes, research in transition in judgment may reveal human reasoning at its best.

**Overall Research Question**

What are the patterns of relationships among measures of selected cognitive constructs (conceptual structure, sensitivity to patterns in data, and
judgment process), individual difference variables (age, education, and experience), task conditions, and performance in a clinical judgment task?

**Specific Research Questions**

In a domain-specific, probabilistic clinical judgment task, where outcomes are only moderately predictable:

1. What are the patterns of relationships among various measures of judgment performance (indicators of expertise) and experience for a group of subjects in a clinical judgment task?

2. What are the patterns of relationships among measures of conceptual structure, sensitivity to patterns in data, judgment process, and performance in a clinical judgment task?

3. What are the patterns of relationships among individual differences in age, education, and experience and performance in a clinical judgment task?

4. To what extent does cue-presentation condition (context cues followed by individuating cues, or the reverse), reveal patterns of relationships in performance in a clinical judgment task?

5. To what extent does memory-priming condition (exposure to relevant domain-specific visual stimuli, versus no exposure), reveal patterns of relationships in performance in a clinical judgment task?

6. Of all the measures included in the study, which measures are most predictive of clinical judgment performance?
II. REVIEW OF THE LITERATURE

Chapter 2 is divided into five sections. Section A is an introduction to the major concepts of expertise and clinical judgment from a cognitive perspective. In section B, findings from judgment studies related to various indicators of judgment quality, including accuracy and consistency are reported. In section C, findings from the literature associated with both expertise and clinical judgment are described. Additional factors (such as cognitive biases) that influence judgment performance are examined in section D. The last section is a summary of concepts from the literature that have provided a theoretical base for the present study.

Section A: Expertise in Clinical Judgment

Expertise has become an intriguing subject for investigation across many disciplines. Studies have been conducted that contrast novice and expert performance in many fields. Examples include: physics (Chi, Feltovich, & Glaser, 1981; Larkin, McDermott, Simon, & Simon, 1980); nursing (Benner, 1984; Tanner, 1984); mathematics (Schoenfeld & Herrmann, 1982); teaching (Arlin, 1993; Clarridge & Berliner, 1991); and accounting (Choo, 1989; Ashton, 1991).

The discussion in this section includes characteristics related to expertise, a description of the Dreyfus model, characteristics of judgment performance, and the use of the lens model for clinical judgment.

Characteristics of Expertise

Several authors have described characteristics of expertise (Benner, 1984; Chan, 1982; Glaser, 1986, 1989; Hampton, 1994; Kennedy, 1987; Rabinowitz & Glaser, 1985; Shanteau, 1987, 1988). The following is an overview and synthesis of many accounts from the literature.
**Expertise requires extensive domain-specific knowledge.** Glaser (1986), Posner (1988) and others have identified that knowledge is critical to expertise; experts must have extensive, up-to-date content knowledge in their field. Biomedical knowledge is fundamental to medical reasoning for medical students and physicians (Patel, Evans, & Groen, 1989; Patel & Groen, 1986). Expertise varies dramatically as a function of the elaborateness and richness of the connections between units of knowledge. An integrated network of domain-specific information is required.

Benner and Tanner (1987) described what constitutes domain-specific knowledge in nursing:

The language of disease is a language of pathology and tests to rule out or confirm hypotheses. The language of illness is a human language with emotions and lived experience. . . .

Nurses deal with both illness and disease, but they do not limit their knowledge of patients to disease facts or physiological states. They have learned, over time, to find a "grasp" or "understanding" of patients' illnesses . . . Personal histories and the contexts of the illness are as necessary to them as the traditional signs and symptoms of the disease (p. 25).

Nurses need a level of biomedical knowledge as well as knowledge of human science to be able provide holistic care to patients in a safe and caring manner.

Clinical knowledge has a high degree of domain-specificity (Elstein, Shulman, & Sprafka, 1978; Palchik, Wolf, Cassidy, Ike, & Davis, 1990). Often knowledge is associated with memories of patients who may have been encountered years, or even decades, earlier (Allen, Norman, & Brooks, 1992).

**Expertise requires competent use of knowledge.** Simple possession of knowledge is insufficient for expertise: there is, in addition, skilled use of knowledge to address practice problems. Patel and Groen (1991) suggested that knowledge progression occurred in three-stages: the first involved the development of adequate
representations of knowledge, the second stage involved learning to distinguish between relevant and irrelevant data, and the third stage required using relevant representations in an efficient manner.

Competent behavior requires that relevant information is readily available when needed. Schön (1983) described this knowledge as tacit, or implicit, in patterns of action. It is a specialized form of procedural knowledge, or "knowing how", as distinguished from merely "knowing that" (Ryle, 1949). This "knowing-in-action" is apparent in applied professions such as teaching and nursing. Jenks (1993) provided an example of this "knowing-in-action" as she described the ways nurses use personal knowledge of patients in their care.

**Expertise involves well-developed metacognition.** Metacognition can be defined as "thinking about thinking, or knowledge related to self-appraisal and self-regulation of one's thinking and actions" (Paris & Ayres, 1994, p. 167). In studies of expertise, researchers have reported higher levels of metacognition in experts compared to novices (VanLehn, 1989). Chi, Glaser, and Rees (1982), for example, found that experts were more accurate than novices at rating the difficulty of physics problems. More recently, Eteläpelto (1993) claimed that experts were superior to novices in their metacognitive knowledge, task-specific awareness, and cognitive monitoring. One form of metacognition that is particularly important in relation to expert performance is conditional knowledge. Such knowledge is an aspect of metacognition which is informative about the appropriate conditions and situations where knowledge and skills are best applied; conditional knowledge helps people to know when, where, and why to apply their strategies (Paris & Ayres, 1994). Paris and Byrnes (1989) stated that conditional knowledge may be fundamental for the spontaneous transfer of appropriate strategies to new situations, a skill that is critical to expert performance.
**Expertise involves skilled attention.** Considerable research has been carried out with respect to attention (Nosofsky, 1987a, 1987b). In a study of expert medical practitioners and novices, Norman, Brooks, and Allen (1989) found differences in attention between the two groups, but only under implicit memory conditions; under explicit conditions, no differences were revealed. Barroso (1985) investigated how attentional strategies of medical students were modified by experience; he reported that experts learned to discriminate relevant and irrelevant data, and focused their attention accordingly. Krogstad, Ettenson, and Shanteau (1984) and Ettenson, Shanteau, and Krogstad (1987) found that two groups of expert auditors focused attention on one cue, whereas the novices used several different cues, with no group consensus. Such behavior was viewed as *skilled omission* rather than a cognitive limitation.

Clarridge and Berliner (1991), in a study of novice and expert teachers, reported differences between the teacher groups in ability to pay attention and notice unacceptable classroom behaviors. Experts were able to focus attention on aspects that non-experts either overlooked or were unable to see. When already extracted information is presented to novices (as is often the case in judgment studies), these subjects are often capable of making judgments that are nearly as good as those of experts.

Benner and Wrubel (1982) described the ways in which nurses' attention and perceptual abilities become honed; expert nurses learn salient qualitative distinctions, and achieve a perceptual grasp of whole situations, which often include what action to take. "Expert nurse clinicians who have spent many hours observing such subtle patient changes as those related to progress in labor, septic shock, or
wound healing, become discriminating judges of these states" (Benner & Wrubel, p. 12).

**Expertise involves sensitivity to meaningful patterns.** Research carried out by Chase and Simon (1973) revealed that expert chess players could be distinguished from those who were less skilled by the experts' ability to correctly reproduce meaningful patterns of chess pieces. Memory performance for these subjects for randomly placed chess pieces was only average. This finding led to the notion of "chunking", which helped account for performance that would exceed non-experts' working memory capacity. Shanteau (1987, 1988) argued that experts are able to see patterns that novices cannot, and make use of these patterns in judgment.

Bennner and Tanner (1987) described pattern recognition from a nursing context. These authors claimed that skilled pattern recognition is intuitive knowledge based on background understanding and memories of skilled clinical observations of past whole situations. For clinicians with expertise, pattern recognition is useful in understanding and possibly resolving unstructured problem situations.

**Expertise often involves fast performance.** Speed associated with expertise can be accounted for, in part, by the compilation of separate components of a skill into larger units as automatization occurs. This process has been described by Shiffrin and Schneider (1977). Such automatized processing is fast, effortless, and not available to awareness, and thus, experts are often inarticulate in describing judgment strategies. Berliner (1986) attributed difficulty with articulation to automatization. Not all aspects of judgment performance, however, become faster with expertise: experts take longer to acquire a qualitative understanding of a
novel problem (Chi et al., 1981), but the time from problem representation to solution is usually rapid.

**Dreyfus Model of Expertise**

Holyoak (1991) described the progression of research with respect to expertise. First generation research began with the early insights of Newell and Simon (1972) as one of the earliest teams to investigate differences in cognition between novices and experts. These researchers assumed that humans operated as limited information processing systems, and argued that problem solving took place by search in a "problem space". Newell and Simon contended that heuristic methods, such as means-end analysis, could be applied to a broad range of domains. The expert was viewed as someone particularly skilled at general solution methods. These researchers focused on well-defined problems such as criptarithmetic and the Tower of Hanoi problem.

Later it became evident that detailed, domain-specific knowledge was essential to expertise. General problem solving methods were recognized as more characteristic of *novice* rather than *expert* performance. The second generation of theories emphasized the power of highly automatized knowledge, which involved compiling separate components of a skill into larger units. Chase and Simon's (1973) hypothesis that expertise involved the development of large integrated "chunks" of related and meaningful knowledge became understood as an instantiation of compilation. The expert was someone who had practiced a task and was able to solve problems and make good judgments efficiently.

It is possible that a third generation of expertise theories has begun (Holyoak, 1991; Patel & Groen, 1991; Schneider, 1987). Second generation theories focused on routine expertise where experts were outstanding at speed and accuracy. In
contrast, performing well in unpredictable situations required what Hatano and Inagaki (1986) referred to as adaptive expertise. They realized there was more to expertise than simply a high skill level. An adaptive expert is one who handles novel situations well, has insight and imagination, and can reason from basic principles.

Kennedy (1987) proposed that expertise goes beyond technical skill: she summarized the additional characteristics of expertise as the ability to apply general principles, to know when to change the rules, and to engage in interactive (dialectical) relationships between analysis and action. This type of expert is sensitive to the meanings of patterns of data and can see fundamental problems and also envision new possibilities, where others do not. Kennedy's account of expertise is similar to Arlin's (1990) description of wisdom in terms of problem finding. Arlin proposed that "having a sense of taste for problems that are of fundamental importance" (p. 235) may be an aspect of problem finding that is associated with wisdom.

Links between people's thinking and the quality of their problem-solving and judgment ability have also been identified in earlier literature (Arlin, 1986). Using a developmental perspective, Arlin examined formal and postformal thinking and related performance of young adults in ill-defined (or unstructured) problem situations. Arlin characterized adult thinking as possibly representing both contractions of logical systems and expansions of those same systems. The use of faulty logic and simplified representations of reality (both of which compromise judgment and decision making) constitute examples of the former. The ability to think in relativistic terms in a manner that is receptive to new information is an example of the latter. Relativistic thinking may be particularly helpful to the development of expertise in judgment, because good judgments are based on the appropriate integration of probabilistic cues.
Theories with flexibility to account for both routine and adaptive expertise are needed, including descriptions of the thinking and performance at various levels, and the ways in which the development of expertise takes place.

The model of expertise developed by Dreyfus and Dreyfus (1986) was selected for this study. This model is based on the progression of expertise in terms of cognitive changes. Bechtel (1988) and Holyoak (1991) have elaborated on these changes associated with expertise, and have claimed that with expertise comes a qualitative change in cognition from the deliberate focus on declarative knowledge governed by analytic rules, to the more automatic processing of domain-specific patterns of data. Researchers in nursing (Benner, 1984) and in teaching (Berliner, 1988; Clarridge & Berliner, 1991) have used this model to investigate expertise in their respective domains. Empirical support for the model has been found: experience appears necessary but not sufficient for expertise; the changes in cognition are critical. A summary of the model follows, based on Dreyfus and Dreyfus (1986), Benner (1984), and Benner, Tanner, and Chesla (1992).

**Novice level.** The novice employs precise rules which apply to objectively specifiable circumstances that are recognizable without experience. Because the rules are context-free (independent of other aspects of the situation), behavior is relatively inflexible. The novice lacks a coherent sense of the overall task, and often evaluates performance by how well the learned rules were followed.

**Advanced beginner level.** The advanced beginner starts to recognize the role of context and modifies the rules in some situations. These exceptions are specified in terms of previously encountered situations. The advanced beginner melds book knowledge with on-the-job experience, building up case or episodic knowledge in association with semantic knowledge. Thus, although this individual's
performance becomes more flexible than that of the novice, it remains slow, uncoordinated, and laborious.

**Competent level.** The competent performer of a skill is distinguished by the development of goals that facilitate the coordination of rules and known facts. Rules are applied not simply because they are the rules, but because they are perceived to be helpful in reaching goals. The person actively directs efforts rather than responds passively to events. An individual at this mid-range level of expertise displays rational planning, conscious assessment of elements that are salient with respect to the plan, and analytical, rule-guided choice of action.

**Proficient level.** The proficient performer moves beyond the deliberate mode of reasoning and begins to rely on recalling previous events similar to the current one. This recognition is based not on specific features, but on holistic similarity. Once the recollection occurs, the proficient performer may proceed analytically as would a competent performer. Where the proficient performer exceeds the one who is competent is in the ability to bring relevant past situations to bear, and use these in establishing goals and applying rules.

**Expert level.** The expert no longer exhibits the deliberativeness of competent performance; the whole process of responding becomes smooth and fluid. The expert sees the situation and sees what to do, and often cannot articulate reasons behind the judgment. The expert responds intuitively. Having enough experience with a variety of situations, the mind of the expert decomposes the class of situations into subclasses that share the same goal and decision or tactic. Thus, a situation, when seen as similar to members of this class, is not only thereby understood, but simultaneously the associated decision or tactic presents itself.
Dreyfus and Dreyfus (1986) argued for a qualitative difference between the cognitive activity of individuals who have attained competent performance, compared to that of people who have achieved either of the two highest levels. Using nursing subjects, Benner (1984) found a transformation in thinking and performance was required for nurses to progress beyond the competent level. The use of holistic recognition of similarity is a crucial element of this change. Experts have the ability to recognize situations as similar to those encountered previously and to categorize them appropriately (Dreyfus & Dreyfus, 1986).

Although there is considerable support in nursing for Benner's (1984) research on the progression of novice to expert based on Dreyfus and Dreyfus (1986), there is also some criticism. English (1993) contested that the definitions Benner used for intuition and expertise were inadequate. He also disputed the idea that expert nurses use intuition in making clinical judgments. Using a cognitive psychology perspective, English provided alternative interpretations of Benner's examples of intuition demonstrated by expert nurses.

**Characteristics of Clinical Judgment**

The literature related to reasoning processes reveals considerable overlap in the use of terms such as problem solving, judgment, and decision making (Elstein, 1992). Within a medical context, Deber and Baumann (1992) used the term *problem solving* to refer to the search for the best solution to a problem, and *decision making* to refer to a situation in which a choice must be made from several alternatives. In practice, problem solving is often used to determine what has led to a current situation that is in some respects not ideal, and decision making is used to identify what is likely the best approach to take, given the situation.
Siegel-Jacobs and Yates (1996) defined judgment as "an opinion about the status of some event in the real world" (p. 4). These authors considered judgments to be more fundamental than decisions. Judgment usually refer to constructing a response where none are ready-made, whereas decision making implies selecting alternatives from a set. Abernathy and Hamm (1995) defined judgment as "the mental ability to combine simultaneously information from multiple sources, to go beyond well-defined, sharply bounded categories by interpolating and extrapolating" (p. 217-218).

Problem solving, judgment, and decision making all can take place at varying levels. For example, with one global problem, the goal may be to make an overall decision about what action to take based on judgments of the relative values of outcomes, given varying circumstances; there may, however, be many micro-decisions to execute about which data to collect, and micro-judgments about the meaning of the data. There even may be micro-problems to solve in the course of obtaining the data.

Clinical judgments can be characterized as follows:

**Clinical judgments are based on cues.** Hogarth (1980) stated that predictive judgments are based on cues; a prediction is the extrapolation of a relationship between a set of cues and a target event. Examples of target events include making judgments about pain management for palliative care patients (Corcoran, 1986b), diagnosing pulmonary embolism from clinical data (Wigton, Hoellerich, & Patil, 1986), making judgments about patients who have suspected bacteremia (Poses & Anthony, 1991), and predicting outcomes in coronary disease (K. L. Lee et al., 1986).
Elstein (1976) defined *clinical* as "any of the artful, informal, qualitative, or not explicitly quantitative strategies generally employed by the clinician for [judgment] tasks" (p. 696). Clinical judgment involves detecting subtle patterns and weighing conflicting evidence. Making judgments can be viewed as responding optimally to a particular set of domain-specific cues where there is no algorithm to derive an absolutely correct answer. The cues contain only some of the information about the clinical state being judged; frequently, there are both redundant cues and missing cues.

**Clinical judgments are probabilistic.** Medical judgments are characterized as "risky" and "probabilistic" (N. H. Anderson & Shanteau, 1970; Schwartz & Griffin, 1986). Garb (1989) claimed that in order for clinicians to learn from their experience to make good judgments, they needed to think in probabilistic terms. Compared to learning deterministic rules, such thinking is very difficult. Christensen-Szalanski and Bushyhead (1981) reported an example of probabilistic information processing where physicians showed sensitivity to the predictive value of the symptoms of pneumonia, but overestimated its probability. Elstein (1976) claimed "the relationship between the disease states and symptoms . . . is one of probability rather than logical necessity" (p. 697).

Many of the clinical aspects of patient care in nursing involve processing probabilistic information (Crow & Spicer, 1995). Tanner (1984) described the environment for nursing judgment as one where cues were fallible, low-validity information was redundant, situations were dynamic, and new data were constantly arriving.

**Clinical judgments contain uncertainty.** Peterson and Pitz (1988) distinguished between confidence and uncertainty. These authors defined
confidence as a person's subjective probability that his or her judgments were accurate. In contrast, they defined uncertainty as the person's beliefs about the variability of possible outcomes. If beliefs about a set of these outcomes can be conceptualized as forming a subjective probability distribution, then uncertainty corresponds to the variance of that distribution. Uncertainty is a function of the range of different outcomes that appear plausible to a judge. The greater this range, the more uncertainty is perceived in a task. Experienced clinicians are often more aware of the uncertainty which is actually present, whereas novices tend to perceive much less uncertainty in the same situation.

An example of modeling medical judgments under conditions of uncertainty was presented by Boreham (1989); this author illustrated physicians' judgments about appropriate drug dosages of phenytoin for a variety of patients with epilepsy. Uncertainty was demonstrated by the uniqueness of each patient, and the unpredictability of responses to medication.

**Clinical judgment depends upon skilled use of knowledge.** In order to make good clinical judgments, it is not enough to have textbook knowledge. The meaning of the cues as a set, within the particular context, must be understood at a deep level. Much of the domain-specific knowledge is linked to previous cases where some particular aspect was seen as salient. Skill in clinical judgment, however, goes beyond the mere possession of specific knowledge. Clinicians must have a sense of what the appropriate patterns are, and yet be receptive to perceive what is actually happening as events unfold; they must simultaneously ignore what does not matter, and tune in to what is critical as they construct particular meanings that represent a synthesis of the relevant data.
Clinical judgments constitute both art and science. Elstein (1976) regarded expert clinical judgment as more of an art than a science. Brehmer (1976) considered clinical judgments to be inductive inferences where the state of a criterion variable is inferred on the basis of a set of information cues such as signs and symptoms. Expert judgments, however, are not always logical, rule-governed inferences of this type. Such a definition excludes implicit judgments that demonstrate procedural knowledge where the rules cannot be articulated. Dreyfus and Dreyfus (1986) stated: "The expert is simply not following any rules! He [or she] is . . . recognizing thousands of special cases" (p. 108). Brehmer's definition may apply more to judgments made by novice clinicians than by those made by experts. Dreyfus and Dreyfus claimed that expertise may consist not so much in being able to carry out complex inferences better than novices do, but in making the most plausible responses.

Lens Model for Clinical Judgment

The lens model was developed by Brunswik (1955, 1956) and described by K. R. Hammond (1955, 1966b, 1972, 1990), Hammond, Stewart, Brehmer, and Steinmann (1975) and Hammond and other colleagues. Brunswik called his model the "lens model" because its pictorial representation resembled the focusing process of a lens. Brunswik believed that a person interacted and made judgments in an external environment (an "ecology") in which the actual state of affairs (or true meanings) are hidden by the presence of numerous ambiguous (proximal) cues that have a probabilistic relationship with the actual (distal) state. Thus, in the judgment situation, a person must aggregate, integrate, or re-combine into a single response multiple fallible cues from the environment, including those which are separated in time. Hammond (1990) suggested that this conception of the judgment process gave rise to the use of a lens as an analogy because of its condensing features.
See Figure 1, which characterizes the conceptual aspects of the model for a single person. The convergent properties of the lens are illustrated by the lines in the figure.

Figure 1. Brunswik's lens model, showing conceptual relations.

K. R. Hammond (1966b) stated that the lens model was the fundamental theoretical basis for Brunswik's view of psychology. In the present study, the lens model was employed as one way to conceptualize the components of the judgment task: the cues, the criterion to be judged, and the judgment of the criterion. The basic idea is that the clinician weights cue dimensions and then integrates them into a unitary judgment. The mathematics related to this model are described in chapter 3. The conceptual aspects are described here to illustrate the relationships between the cues and the criterion state being judged.
One assumption of the lens model is that judgment performance is a function of the characteristics of both the person as well as that of the task system, or ecology. Each judge comes to the task with unique experience and background, and literally sees the cues in a different way (L. K. Hammond, 1970). Expectations, memories, and educational experiences influence the perception of cues, their weighting, and their integration. The notion of the "lens" can be considered in a metaphoric way, in the sense that each person engaging in an interaction in an ecology perceives the cues through his or her own perspective, or "lens". Examining this source of difference is essential in understanding how individual differences in cognition influence judgment performance.

The circles in the center of Figure 1 represent the informational cues upon which the judgments are based. The criterion state is the actual clinical state (in this study, healing time), represented by the circle on the left. The lines from the cues to the criterion state vary in thickness, indicating that some cues have stronger relations to the criterion than others; that is, they have more predictive potential. A dotted line reflects no cue-criterion relationship. Some, but not all, information about the judged state is contained in the probabilistic cues. Thus, the lack of perfect predictability of a criterion is a reflection of the uncertainty that is inherent in the task.

The judgments (predictions) made about the criterion state are represented by the circle on the right side of the diagram. Lines with varying thickness that extend from the cues to the judgment represent a clinician's weighting of the cues. The weighting of relevant cues according to their informativeness in the actual ecology constitutes evidence for sensitivity to overall clinical patterns in the ecology.

Applying the lens model is one way to examine subjects' sensitivity to these broad, clinical patterns. This model provided several measures of clinical judgment
performance, including overall scores that described how closely subjects' judgments were related to the criterion state (lens model achievement), a set of cue-weights for each judge, and an index that assessed the overall degree to which the subjects' weighting of cues matched their importance in the ecology. Hammond et al. (1964) found that this latter index had "promise as an important component of clinical inference" (p. 446).

Section B: Summary of Relevant Research

Research question 1 is:

What are the patterns of relationships among various measures of judgment performance (indicators of expertise) and experience for a group of subjects in a clinical judgment task?

Research has been selected for review that is relevant to judgment accuracy and consistency (and other indicators of expertise) with the goal of clarifying the relationship between experience and expertise.

Judgment Accuracy as an Indicator of Expertise

Thompson, Ryan, and Kitzman (1990), writing from a nursing perspective, maintained that the basis upon which expertise is defined varies considerably among studies. Some researchers select clinicians who have been working in the field for many years as "experts". The problem with this approach is that, as Faust et al. (1988), Garb (1989), and Ridderikhoff (1991) argued, longevity in a domain does not automatically qualify one as an expert. Shanteau and Stewart (1992) and Shanteau (1992) acknowledged that the definition of expert is an obvious precondition to any analysis of expertise. Shanteau recommended that people in a domain be allowed to define the experts. In his research, Shanteau defined experts
as "those who have been recognized within their profession as having the necessary skills and abilities to perform at the highest level" (Shanteau, p. 255). In some studies, expertise is defined by credentials (P. E. Johnson, Hassebrock, Duran, & Moller, 1982; Schvaneveldt et al., 1985). This method is effective in fields such as chess; in other fields, however, Clarridge and Berliner (1991) argued that social criteria often lack validity for research purposes.

Implications of the definition of expert are illustrated in a study conducted by Patel and Groen (1986). These researchers investigated the diagnostic reasoning processes of 7 cardiologists with a simulated case of acute bacterial endocarditis. Four of these experts made a correct diagnosis on this task, and 3 did not. All of these cardiologists, however, were considered to be using expert reasoning. Moskowits, Kuipers, and Kassirer (1988) analyzed the decision-making processes of three expert physicians (pulmonary specialists), and described the problem-solving strategies and knowledge representations that these subjects employed to make decisions. The ways in which these subjects dealt with risks and difficult tradeoffs were reported. Large individual differences were noted. The authors pointed out that the cognitive procedures employed likely contributed to errors in decision making under uncertainty. Both of these studies are excellent examples of research in cognitive processing. The authors, however, have predetermined that individuals who are deemed "experts" on the basis of credentials will (by definition) employ expert reasoning in a given judgment task. This practice may not be as useful for the purpose of identifying process-outcome links as defining expertise on the basis of performance criteria.

A further example of the importance of the definition of expertise is found in a study of novices and expert mathematics teachers. Berliner et al. (1988) developed an innovative predictive task, using multiple choice mathematics questions
(representative of those used in the American National Assessment Tests) as stimuli. While thinking aloud, subjects predicted the percentage of students who correctly responded to such questions. This task revealed interesting processing differences between teacher groups, but no differences in accuracy. Thus, it is not safe to assume that experienced professionals are more skilled on a particular task, compared to less experienced professionals.

In the present study, the criterion for expertise in clinical judgment for a particular task was stipulated to be judgment accuracy for that task. Ericsson and Smith (1991) explained the expertise approach:

[This] approach is an attempt to describe the critical performance under standard conditions, to analyze it, and to identify the components of the performance that make it superior. . . . Whereas other approaches can use social indicators as criterion variables, . . . [this] approach requires the design of . . . tasks wherein the superior performance can be demonstrated (p. 8).

In order to make comparisons about judgment performance, it is important to examine judgments made by subjects who have a similar professional role. M. U. Smith (1990) pointed out that in some research, professors are identified as the experts, and students are the novices. The professors, however, may not necessarily demonstrate the same cognitive processing as expert practitioners in the field, and thus comparisons may be limited. For example, in investigating a range of expertise levels, Murphy and Wright (1984) studied patient descriptions from subjects with different background preparation (practicing psychologists, counsellors for children, and novice undergraduates). Comparison of these descriptions would have limited usefulness in illuminating the development of cognition along an expertise continuum.
Similarly, because clinical judgment demands certain domain-specific knowledge, unless all subjects in a study have at least a base-line knowledge level, any performance differences identified likely should be attributed to these knowledge differences. In one study conducted by Dawson, Zeitz, and Wright (1989), the novices were undergraduates selected on the basis of their lack of knowledge of children, and the experts were experienced supervisory staff at a treatment center for disturbed children. The experts had advanced degrees in psychology, special education, or social work. Although differences in judgment performance were found, they most likely resulted from the large differences in knowledge.

**Consistency of Clinical Judgment**

Consistency is a necessary (but insufficient) criterion for judgment accuracy. Goldberg and Wertz (1966) identified various types of consistency (or reliability) as important in clinical judgment, including consensus over judges (with the same data), and convergence over data sources (with the same judge). A review of the early literature on judgment reliability has revealed that both of these types of consistency are low (Hunt, Jones, & Hunt, 1957; Watts, 1980).

Koran (1975) presented evidence of a lack of reliability of judgment in clinical medicine and surgery. For example, 3 surgeons assessed the clinical progress of patients being evaluated for surgery. Over several days, each physician made a total of 72 examinations on the same patients. Assessments for dehydration and abdominal rigidity could not be reliably judged; only 51% of the time was agreement reached on the judgment of whether patients were improving or getting worse.

Koran (1975) also documented that agreement between clinical and radiological judgments of liver enlargement was no greater than chance. Eddy (1988) reported that in a study of 13 pathologists making judgments on more than
1,000 biopsy specimens, inter-judge agreement was 51%; (intra-judge agreement, however, was found to be moderately good, at 68%). In a study of nursing judgment, K. R. Hammond et al. (1966a) found that the nurses studied did not employ cues in a consistent fashion. In other investigations where judgment performance has been assessed (Brenner & Howard, 1976; Hoffman, Slovic, & Rorer, 1968), the same findings emerge.

Not all inter-judge variation, of course, can be assumed to be unreliability. There are times when actual differences in the phenomena being judged can account for what may initially appear to be a lack of consistency. For example, Tape, Heckerling, Ornato, & Wigton (1991) reported that differences in physicians' judgments about the prevalence of pneumonia reflected regional variation in incidence.

Some judgment studies show adequate reliability. Einhorn (1974) investigated the judgments of 3 experienced pathologists for 193 biopsy slides and found inter-judge reliability of .69. Einhorn attributed special importance to the fact that these slides were authentic; interactions and context effects could play whatever role they normally did. Another investigation demonstrating consistency is a policy-capturing study in which 67 nurses made performance evaluations (Zedeck & Kafry, 1977). Ettenson et al. (1987) reported good reliability with a group of auditors, with mean intra-judge consistency of .83 for experts and .66 for novices; inter-judge reliability was .74 and .41 for experts and novices, respectively.

**Additional Indicators of Expertise**

Additional indicators of expertise used in this study were confidence, appropriateness of confidence, judgment latency, and knowledge accessibility.
Conceptual aspects of these indicators are presented in this section; descriptions of their methods of measurement are presented in chapter 3.

**Confidence.** Confidence referred to a stated belief about the probability that one's judgment was correct. In the present study, confidence was seen as metacognitive knowledge; it was viewed as a belief about the degree to which one's judgments were accurate. Confidence in clinical judgment has been of interest for many years (O'Connor, 1989; Oskamp, 1962; Ryback, 1967). Paese and Sniezek (1991) examined factors that influenced confidence in judgment and found that increased practice and effort, and an abundance of relevant data tended to increase confidence, without necessarily bringing about an increase in competence.

**Appropriateness of confidence.** Oskamp (1962) used the term *appropriateness of confidence*, which he recommended as a major criterion for expertise in clinical judgment. He proposed that expert clinicians should be capable of distinguishing when they were apt to be right from when they were apt to be wrong. A measure of *confidence in relation to accuracy*, therefore, was suggested as a way to supplement the use of accuracy alone. Garb (1986) and Oskamp reported appropriateness of confidence was significantly related to professional experience. In contrast, Grebstein (1963) and Wedding (1983) found no relationship between subjects' accuracy of neuropsychological judgments and confidence.

One way to validate degrees of confidence is to examine the calibration of a set of confidence statements. Overconfidence occurs when, over a series of judgments, stated probabilities exceeded the actual proportion of correct judgments (Paese & Sniezek, 1991). A robust finding from the judgment literature is that of overconfidence (Fischhoff & MacGregor, 1982; Fischhoff, Slovic, & Lichtenstein, 1977; Lichtenstein & Fischhoff, 1977). Oskamp (1965) found increasing confidence
as the amount of information increased, without increases in judgment accuracy. Einhorn and Hogarth (1978) reported that even though experienced clinicians demonstrated a substantial lack of clinical ability, they had great confidence in their fallible judgment. These authors stated: "Neither the extent of professional training nor the amount of information available to clinicians necessarily increases predictive accuracy" (Einhorn & Hogarth, p. 395).

Not many studies focusing on confidence and accuracy of judgment have been conducted in nursing. An early study was one by K. R. Hammond et al. (1966a) in which confidence ratings of nurses were reported as moderate, with little variation over 100 cases. More recently, Baumann, Deber, and Thompson (1991) found overconfidence in a group of critical care nurses.

Although there is a preponderance of evidence for the finding of overconfidence in the literature, Vreugdenhil and Koele (1988) found underconfidence in subjects when predicting future events. An interesting study on judgment confidence is one conducted by Gigerenzer, Hoffrage, and Kleinbolting (1991). These researchers investigated confidence from a Brunswikian perspective; they contended that, contrary to arguments otherwise, people are good judges of the accuracy of their knowledge, provided that knowledge was representatively sampled from a specified reference class.

Gigerenzer et al. (1991) made the assumption that if subjects cannot make a judgment based on information known directly, they constructed what these authors referred to as a "probabilistic mental model" (PMM). "A PMM connects the specific structure of the task with a probability structure of the corresponding natural environment (stored in long-term memory)" (p. 507). This PMM is the functional equivalent of Newell and Simon's (1972) "problem space." Gigerenzer and
colleagues predicted confidence based on this theoretical perspective, and obtained data to support their position.

Similarly, Justlin (1994) proposed that a researcher's choice of stimuli or questions may contribute to the extent of bias observed in some studies. He found that nonrepresentative items may lead subjects to overconfidence, and argued for taking an ecological approach to item selection when conducting research in confidence. Erev, Wallsten, and Budescu (1994) also presented evidence that analyses contributed to results in making conclusions about over- and under-confidence. These authors stated:

In the revision-of-opinion literature, subjective probability (SP) judgments have been analyzed as a function of objective probability (OP) and generally have found to be conservative, that is, to represent under-confidence.

In the calibration literature, analyses of OP (operationalized as relative frequency correct) as a function of SP have led to the opposite conclusion, that judgment is generally overconfident. ... Both results can be obtained from the same set of data, depending on the method of analysis (Erev et al., p. 519).

The literature on judgment confidence reveals that more research is needed to understand how accuracy and confidence relate in particular contexts, using various methods.

Judgment latency. Experts typically are quick to perform judgment tasks in their domain. With reference to psychomotor tasks, Rabbitt and Banerji (1989) reported that very long practice improved decision speed. The literature is not as clear with respect to conceptual tasks. Berliner (1986), Chi et al. (1981), and Glaser (1989) found that experts were slower in constructing mental representations, compared to novices. Experts were very fast, however, with familiar tasks which could be performed intuitively. Dreyfus and Dreyfus (1986) reported that experts can
play chess at the rate of 5 to 10 seconds a move if they relied on intuition and
employed little analysis and comparison of alternatives.

Barrows, Norman, Neufeld, and Feightner (1982) reported that experienced
physicians initiated hypotheses very early in a clinical encounter, and the earliness
of hypothesis generation was associated with correct diagnosis. The total time spent
on the task was not related to diagnostic accuracy. Similarly, Muzzin et al. (1982)
found that expert physicians generated diagnostic hypotheses almost immediately.

With nurses as subjects, Westfall, Tanner, Putzier, and Padrick (1986) found
that clinical inferences were activated quickly; the timing of these judgments,
however, was not related to accuracy. Hogarth (1975b) found the relationship
between decision time and perceived task complexity to be concave; with both
complex and simple tasks, typical decision time was relatively short, whereas with
tasks of intermediate complexity, judgment time was longer. Hogarth attributed this
finding to information processing limitations.

The literature gives good support for including judgment latency as an
indicator of expertise to explore in relation to accuracy. M. U. Smith (1992),
however, warned that it is not safe to assume that the time devoted to a task is a
direct measure of processing time. A source of difficulty in interpreting judgment
latency in research where reading case material is required is that the time taken for
reading will unavoidably be included in the measure. Although it might be possible
to remove subjects' differences in reading rate, such adjustment would not be wise.
As Patel and Frederiksen (1984) and Coughlin and Patel (1987) explained with
respect to medical students and physicians, when subjects read clinical cases, two
aspects of comprehension were found to be important: a "bottom-up" (text-based)
processing of the cues from the case, and a "top-down" (experienced-based)
instantiation of relevant prior knowledge. To make judgments that have potential to reveal differences associated with experience, both of these processes must be considered together. Clinicians construct a representation, while reading, in an interactive manner. Any attempt to remove differences in reading rate could also remove variation essential in achieving differential comprehension and mental representation of cases.

**Accessibility of domain-specific knowledge.** Prawat (1989) defined knowledge access as "the ability to draw on or utilize one's intellectual resources in situations where these may be relevant" (p. 1). Experts have highly accessible domain-specific knowledge. For familiar tasks, experts do not need to employ deliberate attentional resources to retrieve needed information. For example, in describing access, Dreyfus and Dreyfus (1986) stated that "not only situational understandings spring to mind, but also associated appropriate actions" (p. 324). Barrows and Bennett (1972) noted in their study of neurologists, hypotheses seemed to "pop" into the clinician's head, suggesting strong links in memory between salient cues and associations triggered by these cues. Tanner (1984) also found that expert nurses described the immediate access of relevant data as knowledge "popping" into mind.

Kihlstrom (1987), Lewicki et al. (1987); Lewicki, Hill, and Bizot (1988), and Lewicki, Hill, and Czyzewska (1992), have studied information processing with implicit knowledge; these researchers claimed that it is unnecessary to limit the concept of access in relation to one's knowledge to conscious access. Currently, controversy exists in the literature about the nature and significance of the role of unconscious perception and cognition (Kihlstrom, Barnhart, & Tataryn, 1992; Merkile, 1992). Brown (1982) pointed out that the ability to use knowledge flexibly required multiple access.
Access to knowledge and organization of conceptual structure may be related: because of the many links and interconnections, access to appropriate knowledge is enhanced through hierarchical structuring. Brown (1982) believed that the development of reflective access, which involved being able to use knowledge, and to contemplate knowledge as an object of thought, was critically important for expertise.

Section C: Findings From the Literature

In this section, literature related to the second and third research questions are discussed.

Research question 2 is:

What are the patterns of relationships among measures of conceptual structure, sensitivity to patterns in data, judgment process, and performance in a clinical judgment task?

The cognitive constructs are discussed in sequence.

Conceptual Structure

Shavelson (1974) defined structure as "an assemblage of identifiable elements and the relationships between those elements" (p. 231). He defined cognitive structure as a hypothetical construct related to the organization of concepts in memory. The terms cognitive structure and conceptual structure are used interchangeably in this study.

Several researchers have investigated the conceptual structure of individuals (L. K. Hammond, 1970; Larkin et al., 1980; Mitchell & Chi, 1984; Shavelson &
Stanton, 1975; M. U. Smith, 1992). Research has revealed that conceptual organization changes with expertise (Benner, 1984; Boshuizen & Claessen, 1984). From a linear array with little coherence or interconnections, knowledge becomes greatly elaborated and successively transformed into integrated hierarchical units based on fundamental categories. Evidence suggests that these categories are not classical in nature, with sets of necessary and defining features, but rather are prototypical (Bordage & Zacks, 1984; Crow & Spicer, 1995; Grant & Marsden, 1987, 1988; Patel & Frederiksen, 1984). This transformation in knowledge organization has been critical to the evolution of expertise.

Small knowledge units are "chunked" into larger components, thereby allowing a person to perceive patterns in the data as a particular event or phenomenon. These patterns of data are known as schemata, the building blocks of cognition (Rumelhart, 1980). Glaser (1986) defined a schema as "a modifiable information structure that represents generic structures of concepts. . . . Schemata represent knowledge that we experience" (p. 921). Schemata are like prototypes of frequently experienced situations in memory that individuals use to integrate instances of related knowledge. Like internal theories or models, schemata enable individuals to impute meaning and make predictions (Kahneman & Tversky, 1973; Tversky & Kahneman, 1980). One major difference between novices and experts is that the experts have well-developed schemata for domain-specific situations.

As expertise is gained, people modify their knowledge base, which facilitates more advanced thinking. An increase in the hierarchical organization of knowledge makes many implications of the data available as logical inferences. Solving problems and making judgments become a matter of categorizing the specific situations according to basic type (Chi et al., 1981). Using category knowledge as well as the values of specific variables, the expert clinician can generate highly
useful inferences. In contrast, novices know separate facts, but often needed inferences are not generated.

Results of investigations of expertise in judgment performance may be obscured if subjects have large differences in knowledge. Chi, Hutchinson, and Robin (1989) explored knowledge organization and the effect of structure on the use of the information. These researchers selected a task for which novice and expert subjects knew the same number of attributes. These researchers were able to discern difference in knowledge organization that they otherwise would not have done had they not ensured that both groups had a certain knowledge base. The aim of the present study is not to show that knowledge is essential to expert judgment. That point is already known. What is of interest is to investigate differences in judgment performance between novices and experienced subjects, when the novices have basic knowledge.

**Sensitivity to Patterns in Data**

Kolers (1970, 1979b) described three stages of learning to read, which may be analogous to novice to expert changes in perceiving information in a domain. Initially, the beginning reader perceives each separate letter (like each specific cue the novice sees), and then integrates letters to form words. These words are often read mechanistically, without meaning. Much later the skilled reader "reads meaning directly from the words within a language; he [or she] does not read the words themselves" (Kolers, 1970, p. 116). Berliner (1986) also used the reading analogy to describe the expert teacher: "We regard the reading of a classroom, like the reading of a chessboard, [italics added], to be in part a pattern recognition phenomenon based on hundreds and thousands of hours of experience" (p. 11). Possibly clinical experts in some real sense read the clinical situation: they no longer process each
cue separately but are sensitively perceptive to the meanings constructed on the basis of the set of related cues as a whole.

Kolers and Smythe (1984) and Kolers and Brison (1984) argued that perceiving the meaning of cues is not a "given", but is an achievement or a construction. The meaning perceived depended on one's concepts and skill in perceiving as to how the stimulus flux of continuous life events is segmented into units. People do not encode equally all features of objects or events (or their pictures or symbolic description), but encode in relation to what they have learned to be useful. One's skill in perceiving meaning in sets of cues has a large influence on one's representational ability.

Chi et al. (1982) and Chi, Glaser, and Farr (1988) attributed expertise in problem solving, in part, to changes in representational ability. To the extent that judgment performance is similar to problem solving, literature in this area has relevance for the present study. Chi and colleagues claimed that problem representation depended not only on task characteristics, but also upon observers’ conceptual categories brought to the task. These researchers described this interaction as an outcome of both the initial categorization process (arising from "bottom-up" analysis of cues), and generic ("top-down") category knowledge.

Bordage and Lemieux (1991) demonstrated with medical judgments that "the successful diagnosticians . . . are those who use the most diversified sets of abstract relationships and, therefore, who have broader or deeper representations of the problem" (p. 71). Similarly, Lesgold (1984) found that with radiologists, existing schemata formed rules by which new data were interpreted; these schemata functioned as triggers when conditions were satisfied: "The meaning of any given
film feature is determined, in part, by surrounding context" (p. 53). Trigger mechanisms become refined with practice.

Several authors (Evans, 1989; Evans & Gadd, 1989; N. F. Jones, 1957; Lesgold, 1989; Margolis, 1987; Ofir & Lynch, 1984; Rock, Bransford, Maistro, & Morey, 1987) have argued for the importance of context in judgment. Margolis claimed that in making judgments, people who are in a familiar ecology rely on a rich variety of patterned cues. Margolis suggested that people learn to perceive individual aspects of a situation only in the context of a whole, which is often implicit or imputed rather than overt. Many relations of stereotypical patterns are constructed over time (Stelmachers & McHugh, 1964). Initially, however, when novices make what are ostensibly the same judgments, no mental patterns are yet available; these judges must rely on deliberately retrieving the meaning of each cue and estimating its influence without benefit of such context. Novice judgments reflect declarative knowledge, where each piece of information is assessed separately and the judgment consists of mentally combining positive, negative, and neutral cues. In contrast, expert judgments are highly contextual; that is, the cues are interpreted within context, and the meaning constructed on the basis of particular cue-patterns changes with an alteration in context. The expert is sensitive to what is truly relevant in the context at a certain point in time.

Elstein, Shulman, and Sprafka (1990), in their ten-year retrospective review of medical problem solving, argued for a view of situated, context-dependent cognition that is consistent with their earlier findings on case-specificity. Most researchers agree that in studying problem solving and clinical judgment, context must be considered.
Hobus, Boshuizen, and Schmidt (1991) examined the influence of context variables such as age, gender, and previous diseases, on judgment performance. They reported that "the development of expertise seems associated with the development of illness scripts . . . resulting in better diagnostic accuracy" (p. 3).

In this study, the primary judgment task involved reading brief case information about surgical patients and predicting healing time. This dependent variable was selected, in part, on methodological grounds described in chapter 3. The decision to use healing time as the dependent variable also has conceptual rationale. Nurses working in surgical settings encounter hundreds of patients each year. People who are admitted to these units vary in terms of age, general health, diagnoses, life style factors, and in a host of other ways. Over time, there is implicit learning of what patterns of cues tend to be associated with particular outcomes, such as rapid healing, wound infection, or wound separation.

Following the example of Cleeremans and McClelland (1991), a patient's admission to hospital, surgical procedure, recovery period, discharge from hospital, and home recovery can be viewed as an "event sequence". These researchers found that with experience, people learned typical event sequences to which they were frequently exposed. Cleeremans and McClelland's subjects could complete partial patterns; they were, however, not necessarily able to verbalize their knowledge. Such "event sequences" can be compared to "illness scripts" that Hobus et al. (1991) described. In the present study, nurses' encounters with surgical patients were theorized to be important in the development of implicit knowledge of "healing scripts". When presented with case data, nurses with more experience were expected to have developed schematic knowledge needed for relating various cues and completing patterns (making judgments of healing time).
Experts solve problems and make judgments based on the detection of similarity between the given problem or case and relevant situations from experience (Patel et al., 1989). As expertise increases, some studies show that this assessment is related to perceiving similarity more in conceptual relations than in perceptual features. For example, Chi et al. (1982) reported that whereas an expert physicist may identify a problem as one based on "Newton's second law," a novice may label the same problem in literal terms as a "pulley problem."

Because conceptual relations depend on subjects' perceptions of the stimuli, it is not possible for an experimenter to dictate their existence in a research context. In the present study, these relations were determined by the way subjects experienced the stimuli, as Whittlesea (1987) and Whittlesea and Brooks (1988) described.

Chase and Simon's (1973) finding of perceptual chunking of meaningful chess patterns illustrated how sensitivity to patterns varies with expertise. The chess expert and physics expert have the ability to abstract relevant tacit knowledge elicited by external cues. The chess master's expertise is derived, in part, from the ability to impose a cognitive structure on the pattern of chess pieces. Similarly, the expert physicist can "see" the deep structure that underlies the terms in a physics problem. Chi et al. (1981) stated that "even though the same set of key words may be deemed important by subjects of both skill groups, the actual cues used by the experts are not the words themselves, but what they signify" [italics added] (p. 149). It is not the words, but the meaning that functions as cues. This same idea was expressed by L. K. Hammond (1970) in relation to clinical judgment. Novices have a limited ability to see underlying meanings and relations, compared to more experienced clinicians.
Research on concepts has shown that an individual's ability to categorize depends on knowledge which may be encoded in patterns rather than in sets of propositions. Rosch (1975), for example, proposed an approach to semantic categories that treats categories like patterns. Barsalou (1985), as well, argued that familiar concepts are not propositions stored in memory, but are dynamic constructions of continuous, interrelated knowledge, tailored to current needs and goals.

Judgment performance is influenced by the way in which mental representations evolve with expertise. Novices examine each case and make judgments about them as separate entities. Experts, in contrast, no longer think atomistically; they think in terms of a relevant category of similar cases. They perceive the present case within the context of representative cases, based on perceived typicality (Medin & Schafer, 1978). Knowledge of complex interaction is distributed across many special cases, and thus this knowledge is easier to learn (Brooks, 1987). Differentiation is taking place on a wider scale. L. B. Smith (1989) found that similarities were at first perceived globally, and then became more refined with experience. As each situation is compared to previously-experienced situations in memory, individuals construct (or re-construct) meaningfully differentiated subgroups (Hayes-Roth & Hayes-Roth, 1977; Rosch & Mervis, 1975; Rosch, Simpson, & Miller, 1976).

From studies reporting novice-expert differences, it is suggested that representations of problem situations evolve with expertise (B. White & Frederiksen, 1986). Novices construct mental representations with little elaboration; the information related to each cue is initially identified only in binary terms, as a positive or negative impact related to the judgment. Later, mental representations develop so that information for each cue is represented as sets of two-dimensional continua;
thus, the degree to which information is positive or negative can be represented. Finally, mental representations of experts are evolved further, containing information in an integrated, multidimensional form, referred to as mental models. These models include not only knowledge related to the variables of the particular case, but also "distributional" knowledge: that is, knowledge of typicality compared to cases of a similar category and knowledge of correlated features (Mellers, Richards, & Birnbaum, 1992).

If representations that experts construct are distributional in form, it is of interest to determine how expectations derived from past experience might influence judgment performance. In previous nursing research, Tanner (1984) found that expert nurses were sensitive to patterns of cues that did not match expectations. It may be that this "mismatch" signals some anomalous input and guides further reasoning for expert clinicians.

In a series of experiments, Flannagan, Fried, and Holyoak (1986) examined how categories were learned from observation of exemplars. These authors assessed the possible role of prior expectations on the way these exemplars were mentally distributed. They used the category density model to account for their findings. In this model, learning was treated as the acquisition of knowledge about the distribution of category exemplars over a feature space. A central assumption of this model was that the learner used presented instances as a sample to induce a density function over the feature space for the population of potential category members. Category exemplars are represented as configurations of values corresponding to points in a multidimensional feature space. A further assumption was that novel instances would be classified on the basis of distributional knowledge. The probability of classifying an instance into a particular category was viewed as proportional to the relative likelihood that the item was generated by that
category's distribution, compared to that of the alternative categories. Flannagan and colleagues claimed that whenever a stimulus was observed, the frequencies of its attributes were incremented appropriately, building up a mental "record" that, over time, faithfully reflected the distributions of attributes in the observed stimuli. These researchers found that people who had learned a distribution through experience could classify new items more accurately when these new items matched distributional expectations than when they did not.

The research of Flannagan et al. (1986) is relevant for the present study for several reasons. First, if experienced clinicians have cases encoded as multidimensional distributions, this would provide a large advantage when assessing new cases, provided the cases were selected representatively. Not only higher-order category information, but typicality and associated characteristics would become available. Second, this account of categorization, in which the role of experience-related changes in the weights of features (cues) is emphasized, is consistent with an explanation of expertise based on the Dreyfus model. In addition, this approach would help to account for experts' use of the representativeness heuristic, including its functional characteristics in natural continuous environments, and its dysfunction in the laboratory setting where discrete, nonrepresentative judgments are often studied.

**Judgment Process**

Einhorn (1974) stated that "the combining of information lies at the core of expertise" (p. 570). The impact of information processing on judgment performance has been recognized as an important area of study for many years; yet, it is difficult to specify any one processing strategy that is uniformly "best", because what is optimal depends on: (1) task characteristics, (the complexity, the presentation of cue
information, and the nature of the cue-criterion relationships); and (2) characteristics of the judge (ability or skill, familiarity with the task, use of weighting and integrating strategies, and information-processing style).

**Task characteristics.** Several researchers have examined task difficulty as a factor in judgment performance (Abdolmohammadi & Wright, 1987; Corcoran, 1986b). The complexity of a task is viewed as a product of an interaction between task characteristics and subject factors which cannot be completely separated. Such task characteristics include the number of cues, where a greater number tends to increase complexity (Payne, 1976). Task structure is also important in determining complexity; Patel, Groen, and Arocha (1990) attributed difficulty in a medical reasoning task to structural factors. Most tasks become easier with expertise; Adelson (1984), however, found that the difficulty of a task can increase with expertise. The degree to which the structure of simulated tasks used in the laboratory for research matched the structure of actual tasks in the ecology may have influenced the validity of the results of information processing studies.

Hamm (1988a), K. R. Hammond (1987), and Hammond, Hamm, Grassia, and Pearson (1987) have categorized tasks in terms of where they are best located on a task continuum; one end of this continuum are intuition-inducing tasks, and the other end are analysis-inducing, with tasks requiring both forms of cognition in between. Hammond and colleagues argued that judgment performance is enhanced when there is a match between the preferred means of cognition of the clinician and the analysis- or intuition-inducing properties of the task. Structural features of the task significantly influence the type of processing used. If, as Dreyfus and Dreyfus (1986) claimed, cognition changes from analysis to a dialectical relationship of both analysis and intuition, then it may be illuminating if subjects at various skill levels performed tasks that induced intuition as well as tasks that induced analysis.
The nature of the structural relationship between cues and the criterion state in predictive tasks can influence information processing (Connolly, 1977). Some cues have a causal structure, whereas other cues have a reflective structure, in which the cues are indicative of the criterion. One structural feature that has information-processing consequences is dimensionality (Peterson & Scott, 1983). Garner (1978, 1983) and Medin and Schwanenflugel (1981) investigated the influence of integral and separable task dimensions; these researchers argued that a subject may be able to re-define integral stimulus dimensions into a new dimension, particularly when dimensions are correlated. Tasks which encourage the processing of information as holistic or gestalt stimuli have potential to increase the efficiency of the processing and reduce the perceived complexity. This type of transition seems relevant to the change in the perception of tasks that occurs with expertise. Garner (1970) pointed out, however, that not all tasks are equivalent in this regard: only some stimuli form good patterns.

It is not clear from the literature how the presentation of cues (verbal or phenomenal) influences information processing. The use of verbal cues, as opposed to abstract cues that provided no context, have been shown to facilitate accuracy in prediction tasks (Koele, 1980; Miller, 1971; Muchinsky & Dudycha, 1975; Sniezek, 1986); these researchers, however, did not compare verbal cues with phenomenal cues, nor did they investigate how subjects' familiarity with each cue type could influence information processing. Carlström (1989) used both verbal and pictorial cue-presentation formats in a study of army aviators. Results revealed that two cues were used in the verbal format, and only one in the pictorial presentation.

Phelps and Shanteau (1978) also examined cue format. These researchers studied experts who made judgments about the quality of livestock. One format was
verbal (cue-dimensions of verbally-specified attributes relevant to judging livestock); the other format was phenomenal (the livestock displayed as photographs). The results showed that 9 to 11 cue dimensions, compared to three dimensions, were processed in each condition, respectively. No conclusions about the format of cue expression can be reached, however, because the verbal-phenomenal condition was confounded with orthogonal-correlated cue format. The verbal condition displayed cues that were artificially made to be orthogonal. The photographs displayed information that was naturally correlated; only three distinct dimensions were available. In addition, subjects made ratings for eight photographs on two occasions. Multiple regression was used with the 16 ratings of quality as the dependent variable, and 11 predictor variables. Eight photographs may be an insufficient number of stimuli upon which to make a conclusion about the number of dimensions used.

The reason that format of cue presentation is important is that in clinical settings, the cues are patterns of phenomena and events familiar to experienced clinicians. During a clinical encounter, clinicians care for patients who display these cues (such as pain), not as printed words, but as dynamic patterns of particular phenomena in context. Yet, judgment studies are usually conducted with descriptions of cues presented in print. Experienced clinicians who perform in the laboratory may be disadvantaged; novices (who are likely more familiar with verbal cues), may perform better in the laboratory than they would have in the clinical setting.

The nature of cue-criterion relationships has a potential to influence information processing in judgment (Brehmer, 1969; Brehmer & Slovic, 1980). For example, when high validity cues are not available, or when the predictive validity of the cues is low, tasks are seen as having considerable complexity (Brehmer, 1980a).
Two further task features that influence processing are the nature of the relationships between cues and the criterion, and the mathematical function by which cues are combined. When cue-values vary directly or inversely with criterion values, the relationship between a cue and the criterion is described as linear. When cue-values and criterion values are systematically related in a manner that reveals a curved line when the data are plotted, the relationship is described as nonlinear.

Cue integration functions are frequently classified as additive (net effect of the summation of positive, negative and neutral cues, each assessed separately) and nonadditive (all other forms of cue integration, including multiplicative). An essential feature of nonadditive cue integration function is that to some extent, cues are processed configurally. K. R. Hammond et al. (1975) defined configurality as a cue-integration method in which cues were combined in a manner that the use of one cue depended on the value of other cues.

Mellers (1980) pointed out the independence of cue-criterion relations and cue integration function. Although in the literature, the terms linear and additive are often seen together, additive functions can be comprised of cues which are related to the criterion either in a linear or a nonlinear manner. The same is true of nonadditive functions.

Hoffman (1968) reported that a large number of empirical studies have demonstrated that performance frequently can be well-described by linear cue-utilization, and additive integration for both experienced and naive subjects. Phenomena which the researcher has deliberately constructed to relate in a nonlinear way can nevertheless often be re-framed and well-described by a linear
model (Dawes, 1979). Hoffman pointed out that modeling the linearity in a task system can overwhelm any nonlinearity.

The type of cue-criterion relations in a task influences how subjects process information. If expertise is dependent on clinicians' ability to better detect nonlinear relations compared to novices, it is important for judgment tasks that are designed to reveal expertise to contain cues that relate to the criterion in nonlinear ways. Such a precaution, of course, would not preclude other descriptions from being used, but nonlinear modeling would at least be a possibility.

There are many previous studies where researchers have tried to capture the cognitive processing of clinicians with configural tasks (Ogilvie & Schmitt, 1979; Meehl, 1950, 1954; Wiggins & Hoffman, 1968a, 1968b); with most of these attempts to demonstrate configurality, the results have been small.

**Characteristics of the judge.** Characteristics of the judge influence the manner in which information is processed. Having particular knowledge or skill may be critical to processing information in a task. The importance of perceptual skill to many tasks has been recognized in nursing (Benner & Wrubel, 1982) and in teaching (Carter, Cushing, Sabers, Stein, & Berliner, 1988; Sabers, Cushing, & Berliner, 1991). Gilovich (1981) and Tyszka (1986) have found that subject familiarity increased performance on judgment tasks.

Mellers (1980), in her studies on learning in probabilistic tasks, found that subjects' ability to predict the criterion was task dependent; tasks with linear cue-criterion relations and additive cue integration functions were learned more quickly than those tasks characterized by nonlinearity and/or nonadditivity.
There is evidence for differences between experts and novices in their weighting and integration of cues. For example, Wallsten and Budescue (1981), reported that in a task where Minnesota Multiphasic Personality Inventories [MMPI profiles] were judged, experienced subjects used information configurally; in contrast, novices used an additive strategy.

The ability to successfully process cues with nonlinear relations may increase with experience (N. H. Anderson, 1972). E. J. Johnson (1988) found experts used nonlinear cues to increase accuracy of financial judgment. Experts refined their judgments depending on the combined influence of all the relevant cues as a set; the use of this process led to variable weighting of cues. Novices may have known the direction of influence of each cue separately, but they had major difficulty making optimal adjustment for interactive cues or for nonlinear relations. They tended to weight each cue the same, regardless of the values of other cues.

As the amount of information increases, judges tend to use simplifying strategies for integration of information. Einhorn (1970, 1971) reported that the use of interactions and complex configurations are widespread in the judgment process; he pointed out that integration functions that are extremely difficult to model, may be easier to use from a cognitive perspective, compared to less mathematically complex functions. Payne (1976), using a process tracing methodology, found that high task complexity led to differential use of strategies for judgment. Subjects tended to switch from an initial compensatory strategy to one that is noncompensatory, as information load increased.

**Information processing style.** Consideration of information processing style may be important in understanding the progression of expertise. Two particular styles discussed are intuitive processing and analytical processing.
Many researchers investigating expertise identify the importance of intuition in judgment (Abernathy & Hamm, 1995; Benner & Tanner, 1987; Elstein, 1988; Hamm, 1988a, 1988b; Mitchell & Beach, 1990). Glaser (1986) argued that experts develop the ability to perceive meaningful patterns, which are seen in the course of everyday activities. Pattern recognition occurs so rapidly that the phenomenological experience takes on the character of intuitions. Elstein suggested that understanding of intuitive reasoning processes is an important objective: "the most valuable output of both AI [Artificial Intelligence] and behavioral decision research may be to give us insight into how we intuitively deal with complex choices and tradeoffs" (Elstein, 1988, p. 155).

Brooks (1978, 1984) and Jacoby and Brooks (1984) used the term nonanalytic, rather than intuitive processing. These authors claimed nonanalytic processing is based on immediate assessment of holistic similarity of an incoming stimulus to a previously encountered situation retrieved from memory. These authors, as well as Dreyfus and Dreyfus (1986), regarded the transformation from the processing of elements to the processing of the whole situation as critical to the development of expertise. Dreyfus and Dreyfus stated:

[The new level of performance [expert level] coincides with a shift from the logical processing of atomic facts to the recognition without recourse to isolable elements, of the similarity between a current situation and a stored image-like representation of a previous situation it resembles (p. 66).

Similarly, Schmidt, Norman, and Boshuizen (1990) described expert reasoning in medicine as pattern recognition.

[Pattern recognition is] based on the similarity between the presenting situation and some previous patient available from memory. . . . The final stage of expertise is nonanalytic (p. 617-618).
Abernathy and Hamm (1995) acknowledged that expert intuition exceeded any scheme used to describe it; it could not be completely captured by rules. Abernathy and Hamm demonstrated that the experts who they studied responded to situations for which they did not have a script, and seemed to elude the bounds of any analytic framework.

Reber and Lewis (1977) and Reber (1989) related intuition to implicit learning. These authors claimed that this learning represents the epistemic core of intuition; it is the process by which tacit knowledge is acquired. Characteristics of such learning include three critical features: (1) It fosters the construction of tacit knowledge that is abstract and representative of the structure of the environment in which the subjects have been immersed; (2) implicit learning is acquired without conscious attempts to learn; and, (3) this knowledge can be used implicitly to solve problems and make accurate judgments about novel stimulus circumstances. Implicit learning can be distinguished from explicit learning, which consists of the acquisition of declarative knowledge and deliberately-learned rules for application. In contrast, when people solve problems in their practice settings over long time periods, they develop implicit knowledge that allows them to behave in ways that reflect their knowledge of environmental structures and patterns.

Intuition is the end product of an implicit learning experience. Introspectively, it seems compelling and obvious. Yet, Reber (1989) claimed that, from empirical and theoretical perspectives, intuition has not been well understood. Implicit learning provides individuals with a strong sense of what is the appropriate or inappropriate response to make, but people are largely unaware of the reasons for their mental state. The knowledge is "tacit", as Schön (1983) described.
Intuition was defined by Benner and Tanner (1987) as "understanding without a rationale" (p. 22). These authors acknowledged that intuition has seldom been granted legitimacy as a sound approach to clinical judgment; yet, they believed that it distinguished expert judgment from that of beginners.

Intuitive processing, however, is not always accurate (Aspinall, 1976; Bordage, 1984; Kassirer & Kopelman, 1989; LaFortune, 1988; Moskowitz, et al. 1988). Borak and Veilleux (1982) cited many examples where health care professionals (including 85 physicians, 43 of which were statistically sophisticated, and 43 clinical nurses) were assessed for accuracy of their intuitive reasoning. Subjects with statistical knowledge performed best, but all subjects made many errors of intuitive logic. These results constitute evidence of the difficulty of clinical reasoning, and the danger of over-using intuition in contexts where analysis is more likely accurate.

Einhorn and Hogarth (1978) explained that in order to learn to make good intuitive judgments, it is necessary to consider judgments, actions, and outcome feedback together. Yet, these authors claimed that people who make judgments intuitively frequently lack awareness of environmental effects. Wason (1960) pointed out that "in real life there is no authority to pronounce judgment on inferences: the inferences can only [sic] be checked against the evidence" (p. 139). In addition, Wason claimed that people have difficulty in making use of disconfirming information: they are often unwilling to attempt to falsify hypotheses, and thus test those intuitive ideas which carry the feeling of certitude.

In contrast to intuitive processing, analytic processing is the deliberate processing of separate features (such as signs and symptoms) and weighting and combining them to make judgments about specific cases. Explicit learning fosters
analytic processing of information. Declarative knowledge resulting from explicit learning episodes is drawn upon, often in a slow and systematic manner.

Using decision analysis, Corcoran (1986a) presented a good example of an analytic approach to making clinical judgments in a nursing context; this author stated that analysis can be an effective guide "in complex, troublesome situations where there are mutually exclusive courses of action" (p. 149). Benner and Tanner (1987) referred to the type of analytic information processing that proficient and expert nurses used as "deliberative rationality".

The literature related to the third research question is reviewed next. Research question 3 is:

What are the patterns of relationships among individual differences in age, education, and experience and performance in a clinical judgment task?

**Individual Differences in Education and Age**

The influences of individual differences in education and age on judgment performance are difficult to separate. Tanner (1987) summarized clinical judgment studies in nursing from 1966 to 1987. Tanner reported that performance was positively related to the academic degree held; for subjects with more than 6 years experience, however, age and performance were negatively related. Aspinall (1976) found a decline in judgment performance for nurses with more than 10 years experience. In a review of studies of physician performance, Lockyer (1992) reported that older age of physician (50 to 55 years) was associated with lower competence, compared to younger physicians; the specific measures of competence were not described. Similarly, Salem-Schatz, Avorn, and Soumerai (1990) found that older physicians had comparatively low knowledge scores, yet high confidence.
In a study comparing senior nursing students, Brooks and Shepherd (1990) found almost identical mean decision-making scores for students from three types of programs, which differed greatly in length and philosophy. Sanford, Genrich, and Nowotny (1992) also reported similar findings.

**Individual Differences in Experience**

The relationship between experience and judgment accuracy is a major question in this study. Many researchers regard the attribution of expertise to experience as undisputable; however, it is not always clear what it is about experience that makes a difference, or how experience influences cognition and performance. Berliner (1988) stated: "In any domain of expertise, one must learn through experience. . . . Experience seems to change people so that they literally "see" differently" (p. 49).

The research in this area reveals mixed findings: Goldstein, Deysach, and Kleinknecht (1973) found that clinicians, including those who were experienced, performed poorly on judgments of cerebral impairment. Kundel, Nodine, and Carmody (1978), in investigating subjects' judgments of the presence and location of lung nodules, reported no difference in accuracy of performance for subjects with a wide range of experience. Using a policy capturing approach, Borko and Cadwell (1982) demonstrated much individual difference in teachers' decision strategies; no consistent policy could be identified in the group studied. Subjects were 41 elementary school teachers with a wide range of experience. Lesgold (1984) reported that experience beyond what is considered basic for radiologists did not correlate with expertise.

Faust et al. (1988) studied a nationally representative sample of 155 clinical neuropsychologists. Subject made appraisals related to neuropsychological
disorders such as Alzheimer's disease, head injury, or epilepsy. Faust et al. found that (except for a tendency among experienced practitioners to over-diagnose abnormality) "no systematic relations were obtained between training, experience, and accuracy across a series of neuropsychologic judgments" (p. 145-146).

In research in nursing, del Bueno (1990) and J. E. White, Nativio, Kobert, and Engberg (1992) found very little differences in the accuracy of clinical judgment associated with experience. Tanner, Padrick, Westfall, and Putzier (1987) reported that there were basically no differences among subjects in the accuracy of diagnostic hypotheses using simulated patients presented in videotaped vignettes. The three groups of subjects studied (beginning students, senior students, and experienced nurses with baccalaureate degrees) had large differences in both knowledge and experience. K. R. Hammond et al. (1966) studied 6 nurses with a range of experience and found large individual differences in inference patterns that seemed unrelated to experience; the nurses did not discriminate among the cues on the basis of their usefulness in judgment.

Additional studies, for example, those conducted by Dana, Cocking, and Dana (1970) and Elstein et al. (1993) revealed no relation between experience and judgment accuracy.

On the positive side, Papa, Shores, and Meyer (1990) studied 173 subjects with varied clinical experience in medicine and found that the number of months of clinical experience was significantly related to diagnostic accuracy in simulations of patients with chest pain. The subjects for this study had relatively little experience; the results may indicate that experience makes more contribution to judgment performance early in the learning process.
In making diagnostic judgments about dermatologic conditions displayed as slides, Norman, Rosenthal, Brooks, Allen, and Muzzin (1989) demonstrated increased accuracy with experience. Two studies where physicians made accurate clinical judgments about respiratory conditions with an outpatient population were reported (Christensen-Szalanski, Diehr, Bushyhead, & Wood, 1982). Nissila (1992), in a simulation study of 5 experienced orthopedic nurses and 5 inexperienced nurses, found that 2 of the former group reached correct decisions in all three cases; none of the inexperienced nurses were correct on all cases.

One of the difficulties in determining the effects of experience is that experience is defined differently in different studies. For example, del Bueno (1990) defined experienced nurses as those having worked at least three months in a particular clinical area. At the other extreme, Phelps and Shanteau (1978) considered experienced subjects to be individuals with 21 to 25 years of experience in judging livestock. Thus, a "novice" judge in one study may be an "experienced" judge in another study.

A few studies revealed a relation between clinical judgment accuracy and experience, but such a finding is not dependable. It seems that this area of investigation has demonstrated very little development of research knowledge. Schmidt (1996) attributed the slow growth of cumulative knowledge in psychology to the over-reliance on statistical testing. In many studies the magnitude of relations may be small, and thus there is a high probability that traditional methods of data analysis will not reveal them. The consequence is that there is an accumulation of studies in which researchers have concluded (possibly erroneously) that there is no relationship. Schmidt (1996) argued that "traditional methods based on significance testing make it impossible to reach correct conclusions about the meaning of these studies" (p. 118).
It is difficult to know how to synthesize the results of these studies which have been reviewed. Unless synthesis is achieved, however, discovery of the underlying regularities cannot take place (Schmidt, 1996). Such regularities are the foundation for scientific progress. Schmidt criticized what he referred to as the "voting method" in which conclusions were determined by the results of the majority of the studies. Schafer (1993) also criticized this "box score approach" and recommended a qualitative synthesis be carried out.

Section D: Additional Factors Influencing Judgment Results

Research questions 4 and 5 are:

To what extent does cue-presentation condition (context cues followed by individuating cues, or the reverse), reveal patterns of relationships in performance in a clinical judgment task?

To what extent does memory-priming condition (exposure to relevant domain-specific visual stimuli, versus no exposure), reveal patterns of relationships in performance in a clinical judgment task?

In addition to the cognitive variables and individual differences already discussed, the use of cognitive biases and heuristics may influence judgment performance. Assumptions related to research design and methods may have an impact on conclusions reached. These topics are discussed next.

Heuristics and Cognitive Biases

Clinicians make both random and systematic error when predicting a criterion in a judgment context. As has already been documented, reliability figures constitute evidence of inconsistency in judgment. Because cues are related to the criterion
probabilistically rather than deterministically, task uncertainty makes random error
inevitable.

The human mind is subject to cognitive constraints such as limited attentional
resources and finite working memory capacity (Hogarth, 1975a, 1980). Simplifying
cognitive strategies, called heuristics, are used to reduce information-processing
load. Examples of heuristics include availability and anchoring and adjustment
(Kahneman & Tversky, 1982). To the extent that these strategies lead to systematic
error in judgment, they are considered as biases. Thus, a bias in this judgment
context refers to systematic errors that stem from individuals' perceptual processes,
and/or information-processing strategies (Tversky & Kahneman, 1974).

Individuals may use these heuristics in a way that does not lead to error in
specified conditions under which heuristics have potential to be valid. In particular,
he has argued that several biases identified in discrete incidents result from
heuristics that are functional in the natural continuous environment: "Judgment and
choice depend crucially upon the context in which they occur and the cognitive
representation of that context" (Hogarth, p. 213).

The heuristics discussed include: availability and representativeness
heuristics, anchoring and adjustment, failure to consider regression to the mean,
failure to perceive true correlations, and the perception of illusory correlations.

**Availability heuristic.** In a problem solving or judgment context, the ease
with which relevant instances come to mind is known as availability (Tversky and
Kahneman, 1973). When people use availability as the basis for judgments of the
probability of events or the frequency of particular classifications of entities or states
of affairs, they are applying the availability heuristic. In general, availability is
correlated with ecological frequency, but because it is also affected by additional factors, the use of this heuristic in judgment may lead to bias.

One factor with particular influence is the differential salience of available cases (Nisbett & Ross, 1980). Often the extreme cases, (for this study, cases with delayed wound healing), because of their salience, exert greater influence on judgment than their actual frequency warrants. In the present study, experienced clinicians are expected to be more likely to use the availability heuristic. Whether judgment performance in the laboratory becomes biased or is facilitated will depend on the extent to which the presented cases reflect the type of experience the subjects have had.

**Representativeness.** Representativeness is a heuristic which people often employ as they make judgments about the likelihood of uncertain events. Using this heuristic, the subjective probability of an event is determined by the degree to which the event is similar in characteristics to a relevant population (Kahneman & Tversky, 1972). People consistently judge a more representative event as the more likely event.

A problem with this reasoning in terms of judgment accuracy is that sample size, or base rate, is often ignored. Dawes (1986) explained that when the heuristic of representativeness is operating, a schema is accessed automatically, but there is generally no intuition of the schema's prevalence tied to the schema itself.

**Anchoring with adjustment.** This commonly used judgment strategy occurs when a person makes an estimate of a criterion state based on some initial value, and then subsequently adjusts the estimate as additional information is processed. Hogarth (1980) identified one potential risk of this technique as arising from the way
the original anchor is generated. People have been known to be influenced by anchors generated by completely random processes (Tversky & Kahneman, 1974).

Shapiro (1977) proposed that subjects categorized cases according to similar cases from memory. In this study, if subjects categorize in this way, their estimates of average healing times for category members may act as an anchor; whether such an anchor is helpful or not, would depend on the match between clinicians' experience and the cases presented.

A potential source of error common with this heuristic is conservatism, or failure to make sufficient adjustment for the particular case being considered (Kahneman & Tversky, 1972). Considerable adjustment would be required for cases at the periphery of the distributions. Experts may be those subjects who select good anchors and make appropriate adjustments.

**Regression to the mean.** In making clinical predictions, failure to consider regression to the mean may lead to overly extreme, or "nonregressive", predictions (Kahneman & Tversky, 1973). What people sometimes fail to consider is that measures of cues include error. When cues have extreme values, the likelihood is that some cues, if measured without error, are in actual fact, more moderate.

Another reason for predictions that appear nonregressive is that action in the clinical setting is taken based on cue values associated with extreme predicted outcomes; such intervention may make the outcome more moderate than it would have been without the action. The presence of such cues leads to aggressive treatment. Thus, if the treatment is effective, paradoxically, extreme cues may lead to more moderate clinical outcomes than what is suggested, given the extreme predictors. Cases with cue values that are not as extreme do not attract the same attention; hence, they may not receive extra treatment. Einhorn and Hogarth (1978)
stated that because action taken on the basis of particular cue-values alters the outcome, clinicians have difficulty in learning accurate cue-criterion relations in probabilistic environments.

**Perception of actual and illusory correlation.** Chapman and Chapman (1967) defined illusory correlation as a systematic error in perceiving co-occurrences, where, in fact, no actual relationship existed. Such error, as well as the failure to perceive true correlations, may lead to inaccuracies in clinical judgments. With experience, people can become sensitive to correlations among features (Medin, Altom, Edelson, & Freko 1982). Lewicki (1986) demonstrated that the mind can implicitly learn to detect covariation in data. Distinguishing between data-based and theory-based covariation estimates, Jennings, Amabile, and Ross (1982) found the former estimates to be extremely conservative. Only when objective correlations approached .85 did subjects consistently rate relationships as strongly positive. In contrast, theory-based estimates led to reports of correlations consistent with theory that were *not present* in the data; in other words, these were illusory correlations.

Smedslund (1963) reported a study of clinical inference conducted with a group of nursing students; a series of cases were presented to subjects in which the presence and absence of a symptom were associated equally often with a diagnosis; that is, the correlation between the symptom and the disease was zero. The subjects concluded that the correlation was positive and could cite many examples that "confirmed" their hypothesis, a finding which also revealed confirmation bias.

There is research evidence which reveals that accurate covariation detection is difficult (Jennings et al., 1982). With experience, however, people learn to see recurring patterns. For example, Medin et al. (1982) reported that subjects were able to use correlations in symptoms as a cue for certain artificial diagnoses.
Experienced clinicians perceive *patterns* of signs, symptoms, and associated actions which can function as global cues.

**Experience and the Use of Heuristics**

The literature is not clear with respect to the link between experience and the extent of bias in judgment: Shanteau (1978) argued that "bias increases as the *unfamiliarity* of the stimulus increases" (p. 581). Heller, Saltzstein, and Caspe (1992) maintained that bias in medical judgment arising from heuristics *increases* with experience. O'Neil (1994) found, in a sample of 214 community health nurses, that the nurses who were more experienced demonstrated more bias in diagnostic reasoning, compared to less experienced nurses. Christensen and Elstein (1991) claimed that "both experts [who were more experienced] and novices are believed to be equally susceptible to biases" (p. 25).

Richards and Wierzbicki (1990) found anchoring effects varied inversely with confidence, but these researchers' subjects were undergraduate students, and, therefore, the results may not be applicable to clinicians. Christensen, Heckerling, MacKesy, Bernstein, and Elstein (1991) examined medical students, residents, and experienced physicians for the presence of a framing bias. The results are interesting in that whereas students and physicians *did not* show bias, the residents [intermediate level group] demonstrated framing bias on 5 of 11 cases. The authors proposed that the medical students may have had too little knowledge to make the frame relevant, whereas the experienced physicians have highly stable knowledge, not easily altered by framing manipulations.

It seems reasonable to consider that some experience is *necessary* for a clinician to experience certain biases, such as anchoring and adjustment, and
representativeness, because such bias requires a level of schematic knowledge; this knowledge takes at least some experience to develop.

**Research Assumptions and Design Factors**

The factors being addressed here were those which influenced the study design and task selection: contextual factors, implications of representativeness, assumptions about cognition, and assumptions arising from research traditions.

**Contextual factors.** Context has a powerful influence on any research results. Birnbaum (1982) identifies two types of context: the context provided in the laboratory, including the type of environment and tasks, and the directions, and the internal context the subject brings to the laboratory in the form of knowledge, expectations, and memories. Subjects' judgments depend on both types of contexts. This discussion will focus on the internal context.

Birnbaum (1982) argued that when a stimulus is presented in a research task, each subject's mental context from experience is differentially brought to the task. Birnbaum recommended that these effects be considered, not as biases to be eliminated, but as integral aspects to be studied. He used the term "systextual design" to refer to research in which context is systematically manipulated.

In this study, priming was used to influence each subject's internal context in an effort to induce greater availability of prior cases; in addition, cues were presented in two sequences in an attempt to induce either analysis or intuition. Manipulating cue-presentation sequence was anticipated to influence anchoring, similar to research conducted by Friedlander & Stockman (1983).

**Implication of representativeness.** According to K. R. Hammond (1990), a particular set of stimuli or cues that carry a simple meaning to novice subjects may,
in addition, communicate to experienced subjects what are referred to as "secondary" cues. Bouwman (1982) provided a good example of two accountants' protocols that revealed expertise-related differences in response to such cues: based on expectations derived from knowledge of representativeness, the expert used a particular contradiction to guide further exploration; in contrast, the novice, having little experience from which to make expectations, reported the contradiction, but apparently never recognized its informativeness. The mismatch functioned as a secondary cue in the former instance, but did not in the latter.

Knowledge associated with representativeness of the cue configuration can often function as secondary cues. In real-life situations, to those who can detect these secondary cues, the information gained can be useful in judgment. Whether or not secondary cues facilitate performance in a research context, however, depends on the particular design employed.

If representative design is used, such cues may be helpful because the mental context of experienced subjects and the configuration of stimuli used in the study have a high degree of correspondence. In contrast, designs in which cues are orthogonalized (or are otherwise nonrepresentative) could lead to reduced judgment accuracy because the particular composition of presented cases would not match expectations.

The essence of the principle of representative design is that the judge is best studied using cues and cue-weightings that occur in the natural ecology (Brunswik, 1955). Rock et al. (1987), as well, argued for taking an ecological approach to the study of clinical judgment. Experienced subjects learn to consider a range of particular phenomena as normal or typical; phenomena outside this range are viewed as atypical. Natural environments are comprised of ongoing, inter-related
events, rather than a series of static, separate states; Hogarth (1981) argued that "representative design is even more important in assessing human capabilities in continuous as opposed to discrete situations" (p. 212).

Kahneman (1992) described the concept of a cognitive norm as the internal variability of an attribute within a category. For experienced subjects, the relative activation of different values of an attribute is governed more by the rules of memory than by the rules of logic or statistics. This activation forms the basis for various heuristics that are employed to ease the burden of information processing. These heuristics, however, will be effective only in an ecology that is similar to that experienced by the clinician.

Some researchers have maintained that the reason for representative design is to generalize to the clinical ecology. Mook (1983) pointed out that many experimental studies are criticized for lack of external validity, and yet achieving external validity is not always the research goal. With the present study, the goal was to understand clinical judgment at a deep level, and to fairly compare subjects at different experience levels in terms of performance on a task in which various clinical cues are weighed and integrated. To the extent that experts use the representativeness heuristic as a means of achieving their expertise, representative design becomes a necessary element.

One aspect of representativeness in the research setting is the presence of ambiguity corresponding to that which exists in the natural clinical setting. In attempting to understand how human beings accomplish judgment tasks, Brunswik (1956) and K. R. Hammond (1955, 1990) recommended that ambiguity be present in the conditions under which judgment is studied. Tasks are commonly referred to as "ambiguous" if they are not able to be clearly perceived or readily interpreted. In this
study, ambiguity is viewed as an interaction between a subject and a particular case, not as an objectively-defined attribute of a stimulus alone; thus, the perception of ambiguity is contingent upon each subject's unique knowledge and experience. Which stimulus is seen as ambiguous very much depends on who is looking.

Einhorn and Hogarth (1985) distinguished between first- and second-order ambiguity. The former is present where a judge is aware of the uncertainty in the judgment context, and the probabilistic nature of the relationship between the cues and the criterion. The latter is present when the judge is uncertain about the extent and nature of these relations. Experienced clinicians tend towards first-order ambiguity, whereas novices likely experience second-order.

This distinction may be important in determining optimal judgment strategies: second-order ambiguity may induce novices to weight cues equally and use an additive rule. In contrast, first-order ambiguity may tend to elicit configural strategies, exceeding the limits of information processing. This more complex strategy is not always associated with greater accuracy (Camerer & Johnson, 1991). Because outcome feedback is often limited in clinical situations, and confidence tends to increase with experience, clinicians may not realize when their strategies are ineffective.

Assumptions about knowledge. The present study is an investigation of the extent to which cognitive constructs and experience account for variation in performance in a clinical judgment task. The assumptions made about knowledge, therefore, are critical in interpreting the results obtained and the conclusions made. Research cannot reveal what restrictive assumptions have precluded from the outset.
Bechtel and Abrahamsen (1991) examined two assumptions about knowledge that are relevant for this study: one is that knowledge is comprised of symbols under the constraint of rules and is expressed in sentences or propositions that have a truth value. These symbols are assumed to be enduring entities which are stored in, and retrieved from, memory.

The second assumption about knowledge is that it may be expressed in non-propositional formats; that is, in a nonverbal, or non-symbolic form. Researchers with this theoretical perspective who model cognition using this assumption referred to such models variously as "PDP" [parallel distributed processing], connectionist, or neural network models. Recently, there has been much interest in this area of cognitive science (Bereiter, 1991; Caspar, Rothenfluh, & Segal, 1992; Feldman & Ballard, 1982; McClelland & Rumelhart, 1986; Quinlan, 1991). The proponents of these PDP models argued that the mind functions as a network of elementary nodes connected to each other so that active units excite or inhibit each other in a dynamic system. Networks reflect patterns by encoding regularities in weighted connections that are modified by experience. Within a connectionist framework, pattern recognition plays a fundamental role at all levels of cognitive processing. Learning consists of altering the weights of connections between units so that small adjustments in the way in which inputs are processed on subsequent occasions are made. No stored symbols or rules are required.

Both symbolic and connectionist systems are computational systems: in the former, computation involves the transformation of symbols according to rules, whereas in the latter, computation is implemented by units that excite and inhibit one another.
The basic aspects of these views are not altogether new: in fact, Neisser (1963) distinguished between sequential and multiple mental processes; the former involved a step-by-step process useful when there is little uncertainty, and the latter involved perceiving input "as a whole", important in recognizing ambiguous cues and patterns. Similarly, Luria (1966) described the integrative nature of mental functioning in terms of successive (serial) and simultaneous (parallel) processing. Recently, Sloman (1996) supported these earlier views: he stated "The mind has dual aspects, one of which conforms to the associationist view and one of which conforms to the analytic, sequential view" (p. 3).

Researchers are exploring the issue of integrating symbolic and PDP approaches (J. A. Anderson, 1990; Bechtel, 1988; Estes, 1988; Wolters & Phal, 1990). Lesgold (1989) argued that both approaches were useful in medicine:

[S]ymbolic models, as currently developed and tested, are well suited to capturing physicians' accounts of the reasons for their diagnoses, while connectionist models are perhaps better suited to mimicking the behavior of physicians (Lesgold, p. 395).

A comprehensive study of clinical judgment performance demands both of these assumptions. When novices make judgments, cognition can be properly described as symbol manipulation; idea-units and rules for action are apprehended, encoded, and retrieved in verbal codes. Limiting expert judgment to symbol manipulation, however, would not reveal cue-judgment connections that are encoded, not in verbal form, but in domain-specific patterns of information. These assumptions, fundamental to the Dreyfus model, are critical for describing progression in judgment performance from novice to expert.

The assumption related to research on judgment expertise that is often made is that analysis is the predominant type of cognitive processing employed. Brooks et
al. (1991) pointed out that an assumption of many studies in medical education is that skill consists of learning about separate features which are weighed and combined when clinicians make judgments about specific cases. This analytic assumption underlies regression models where researchers attempt to capture expertise with changes in the weights applied to a set of cues; it is assumed that these cues have a linear relationship with the criterion and are combined in an additive manner.

Abernathy and Hamm (1995) have discussed the use of a combination of analysis and intuition with respect to physicians. In addition, there is recognition in the literature in nursing education that nurses require both analysis and intuition (Miller & Rew, 1989; Radwin, 1995). Considerable evidence suggests that the analytic approach and the intuitive approach are used interactively. For example, Dreyfus and Dreyfus (1986) described the way in which analysis and intuition are combined as people progress towards expertise:

[Analysis and intuition work together in the human mind. Although intuition is the final fruit of skill acquisition, analytic thinking is necessary for beginners learning a new skill. It is also useful at the highest levels of expertise where it can sharpen and clarify intuitive insights (p. xiv).

In early medical studies, models employed did not capture the interdependency in problem features and data items, which in all medical fields is critical. An example is the independent cues model referred to by Medin and Smith (1984). The causal relationships among items were not represented, nor was the dynamic nature of illness as a process that was acquired and changed over time. Yet, as Murphy and Medin (1985) argued, it is precisely this type of theoretical knowledge that promotes conceptual coherence. Expert clinical judgment involves being able to perceive theoretically-derived similarity among cases. Perception of such similarity gives direction to judgment that has potential to be much superior to judgments
based on unconnected cues and superficial features. To reveal expertise, the researcher must select tasks that provide opportunity for subjects to demonstrate both analysis and intuition, and as well as the processing of similarity relations.

**Section E: Summary**

Research question 6 is:

Of all the measures included in the study, which measures are most predictive of clinical judgment performance?

Convincing evidence exists that conceptual structure changes with expertise; as knowledge organization becomes more highly elaborated, it is restructured into a hierarchical form, allowing rapid access to (or construction of) patterns of relations and inferences.

As expertise increases, sensitivity to domain-specific patterns becomes greater. With familiar tasks, experts' "chunking" of separate (but related) pieces of data allows cues to be available together in working memory which may encourage pattern recognition by facilitating the comparison of relations among task elements. Mental representations evolve with expertise from a static unidimensional internal "problem space" to a dynamic multidimensional mental model containing distributional information; this transformation is accompanied by corresponding increases in functional ability.

Judgment process also changes with expertise, from a focus on specific cues and linear relations to an emphasis on meaningful patterns of cues. Experts' increased familiarity with the ecology and with typical cases promotes greater automatization of routine aspects of a task, allowing attention to be focused on the incongruent or novel aspects. Because of information processing constraints, and
the need to process a great deal of data, clinicians frequently employ a number of heuristics when making judgments. Whether the use of these mental strategies leads to cognitive bias, or leads to good judgments, depends on the context. In the laboratory setting, there are many examples where heuristics (which are functional in the natural ecology) become dysfunctional.

Researchers investigating judgment performance in the laboratory have often failed to find differences in judgment accuracy with clinicians of varying experience levels. This study is an examination of possible reasons; it may be that the laboratory context is not conducive to revealing expertise, or that the artificial tasks often used do not elicit the same thinking and responses that authentic judgments situations would. Although experts may have greater explicit knowledge compared to novices, it may be in their considerable tacit, or implicit knowledge shown in action, where performance differences will prove to be more detectable.

Revealing patterns of relationships among the indicators of expertise, the cognitive constructs, the individual differences, and the research conditions are matters that cannot be addressed on the basis of the existing literature. In order to identify what measures (if any) are predictive of judgment accuracy, empirical study is required. In the next chapter, the methods used to address these research questions are described.
III. RESEARCH QUESTIONS, METHODS, AND ANTICIPATED PATTERNS OF RELATIONS

This chapter is organized into four sections. The research questions and a description of the design and procedures are in Section A. An explanation of the lens model from a methodological perspective, followed by some critical discussion of linear modeling is in Section B. An outline of the methods employed (research tasks and conditions, and measures of constructs) are included in Section C. A list of anticipated patterns of relations are available in Section D.

Section A: Research Questions, Design, and Procedures

The following overall research question and specific research questions have provided direction to the study.

Overall Research Question:

What are the patterns of relationships among measures of selected cognitive constructs (conceptual structure, sensitivity to patterns in data, and judgment process), individual difference variables (age, education, and experience), task conditions, and performance in a clinical judgment task?

Specific Research Questions:

In a domain-specific, probabilistic clinical judgment task, where outcomes are only moderately predictable:

1. What are the patterns of relationships among various measures of judgment performance (indicators of expertise) and experience for a group of subjects in a clinical judgment task?
2. What are the patterns of relationships among measures of conceptual structure, sensitivity to patterns in data, judgment process, and performance in a clinical judgment task?

3. What are the patterns of relationships among individual differences in age, education, and experience and performance in a clinical judgment task?

4. To what extent does cue-presentation condition (context cues followed by individuating cues, or the reverse), reveal patterns of relationships in performance in a clinical judgment task?

5. To what extent does memory-priming condition (exposure to relevant domain-specific visual stimuli, versus no exposure), reveal patterns of relationships in performance in a clinical judgment task?

6. Of all the measures included in the study, which measures are most predictive of clinical judgment performance?

**Design and Procedures**

In phase 1 of the study, multiple regression analysis was used to derive an equation that characterized the relationships between the cues and healing time in the clinical ecology. The equation was based on clinical data obtained from patients and their medical charts, and from direct reports of healing. The data included: *personal history cues* (age, gender, and occupation); *medical cues* (weight, height, diagnosis, and medical conditions); *surgery cues* (type of surgery, length of surgery, and complications), and *incision cues* (length, approximation, and dressing type). The dependent variable was healing time, assessed in days. The equation derived
from multiple regression was comprised of the set of cues that best predicted healing time in this patient population.

In phase 2 of the study, the subjects were nurses who completed a number of tasks that required judgments about incisional healing in surgical patients. The primary task was the one in which judgments about healing time were made, and policies were captured for each subject. These policies constituted a mathematical description of the relationships between the cues and the judgments. Equations for each subject describing his or her judgment policy were computed and compared to the equations derived from the ecology. This judgment task was also used to elicit judgment confidence and to measure judgment latency.

For the healing time judgment task, two sets of parallel cases were developed; data from representative patients from phase 1 were used, and the sets of cases were constructed to be as equivalent as possible. One set was developed with the paragraph containing the global cues first, followed by the paragraph containing the specific cues. This order was referred to as normal order. The other set of cases was structured with the specific cues first, followed by the global cues. This order was referred to as reverse order.

Nurses who had a range of experience caring for abdominal surgical patients predicted healing times for both sets of cases; sequencing of the administration of the two sets of cases (normal order and reverse order) was counterbalanced. Subjects were informed that the case data used in the tasks were obtained from actual patients. For each case, immediately after making a judgment of healing time, subjects rated their confidence that each judgment of healing time was within the allowable range of the actual healing time. Subjects were individually-tested by the researcher. Directions were given prior to each task, and rest breaks were
scheduled to reduce fatigue. Testing time was about 3 hours. Subjects were paid $50.00 each for their participation.

**Sample Size**

Based on previous research, the assumption was made that in the population a small effect size was most likely. An effect size in this context refers to the magnitude of the relationship being investigated. Cohen (1988) and Olejnik (1984) recommended that the power (the probability of rejecting a false null hypothesis) in a research study should be .70 to .85. The number of randomly sampled subjects needed to attain sufficient power would range from 80 to several hundred, depending on the design, the alpha level selected, and other factors (Cohen, 1988; Olejnik, 1984; Trattner & O'Leary, 1980). There were inadequate resources to conduct a study of such magnitude. In addition, it was not possible to sample subjects randomly. Shaver (1993) argued that "a test of statistical significance used without randomization, . . . does not yield valid information about the probability of a result under the null hypothesis" (p. 299).

The goal of the study was to assess for meaningful patterns between various measures of expertise in judgment performance and subjects' experience. The measures of performance were obtained under two conditions, using tasks which were theorized to induce analysis or intuition. There were theoretically-derived rationales for predicting particular patterns. Thirty-six subjects who varied with respect to their experience in caring for surgical patients volunteered for the study. It was recognized that this size of a sample was inadequate to achieve the level of power recommended. Other studies where researchers have investigated novice-expert differences, however, have had small sample sizes: Hammond et al. (1964) analyzed data for three groups (5 subjects were naive, 5 were semi-sophisticated,
and 5 were sophisticated), and Coughlin and Patel (1987) investigated information processing for 16 medical students and 16 family medicine practitioners. Sample size is small also when data are analyzed by means of protocol analysis; for example, Moskowits et al. (1988) carried out an in-depth analysis on the protocols for 3 subjects.

In the present study, the goal was to assess for patterns that may be meaningful from a clinical and educational perspective. Meaningfulness is not determined by statistical significance (Cohen, 1994; Schafer, 1993; Shaver, 1993). Schmidt (1992, 1996) and Thompson (1987) argued that statistical significance alone does not permit evaluation of the importance of a finding. Schmidt (1992, 1996) stated that even though a study has inadequate power, it may have potential to contain valuable information when combined with similar studies in meta-analysis. This researcher suggested that each small-scale study be considered as a data point to contribute to a later meta-analysis. A sample size of 36 subjects, therefore, was considered to be small, but adequate for the specified purpose.

Section B: The Lens Model

Description of the Lens Model

The lens model is based on research carried out by Brunswik (1955, 1956) and extended by K. R. Hammond (1955), Hammond et al. (1964), Hammond and Summers (1965, 1972), Deane, Hammond, and Summers (1972), and Hursch, Hammond, and Hursch (1964). Brunswik (1955) argued that in order to obtain an adequate analysis of judgment performance, a description of the environment as well as that of the individual subject's response was required. The lens model provided such quantitative description.
The lens model has had much discussion in the literature (Castellan, 1972, 1973; Groner, 1972; J. C. Lee & Tucker, 1962; Petrinovich, 1979; Stewart, 1976; Tucker, 1964). This description represents a synthesis from these readings.

The lens model is both a conceptual model as well as a method to illustrate the judge's weighting and integration of cues that have varying degrees of relevance for a particular judgment. An advantage of using this method is that subjects at various experience levels can make judgments for the same cases, and comparisons in judgment performance can be made. Another advantage is that the cue-weights are derived in an objective way (using a least squares criterion) that can be compared to the judge's subjective weighting of particular cues.

In any task in which a criterion \( Y_e \) (referred to as the criterion state, or distal criterion) must be judged, the subject's response is based on an analysis of informational cues, \( X_i \) (referred to as proximal stimuli). In many judgment tasks, the cues are probabilistically related to the criterion, and the judgment made is probabilistically related to the cues. The basic premise of the model is that the environment contains uncertainty, and that judgments are made about a distal criterion on the basis of proximal cues that lack perfect validity.

When applied in a clinical context, the cues \( (X_i) \) are the pieces of information about patients, such as age, diagnosis, state of health, and surgery, available for consideration.

The lens model can be described in terms of the value of the multiple correlation coefficient, \( R_e \), for the environment, and \( R_s \), for the subject. \( R_e \) describes the relation between the cues and \( Y_e \) (the criterion state, or the variable being judged); \( R_s \) describes the relation between the cues and \( Y_s \) (each subject's response). Conceptually, \( R_e \) reflects the extent to which the criterion values can be
predicted by the cues, or *task predictability*, whereas, $R_s$ refers to the subject's *cognitive control*, or consistency.

Cues vary in ecological validity (correlation with the clinical state), and in utilization validity (correlation with the subject's *judgment* of the criterion state). One assumption of this model is that the criterion state and the judgments can be expressed as a linear function of the cues:

\begin{align*}
Y_e &= b_e' X, \text{ and} \\
Y_s &= b_s' X,
\end{align*}

where $b_e'$ and $b_s'$ are vectors of cue weights (regression coefficients) and $X$ is the matrix of cues.

One characteristic of cues in the clinical judgment ecology is that they are frequently intercorrelated. That is, there are typically a *number of cues*, rather than only one, where certain cue-values are associated with high or low criterion scores. These cues have some degree of intersubstitutability. In situations where the judge is sensitive to the correlations in the cues, judgment performance may be enhanced; intercorrelated cues may be used interchangeably.

A mathematical description of the relationships between the cues and the ecology is known as the left side of the lens model. In contrast, the right side of the model is the corresponding representation of each judge's cue utilization policy. Obtaining such an equation is known as policy capturing (Mazen, 1990; Slovic, 1966; Slovic et al., 1977).
Lens model achievement ($r_a$), is the correlation between the subject's judgments and the criterion scores.

The lens model equation is:

$$r_a = G R_e R_s + C \sqrt{(1 - R_e^2)(1 - R_s^2)}$$

(3)

This equation is a mathematical description indicating that achievement from a lens model perspective, $r_a$, is limited by the degree to which the task is predictable, $R_e$. Beyond that, such achievement is partly determined by knowledge of the task, $G$, and also by cognitive control, $R_s$. A measure of nonlinear variance-matching, $C$, has potential to increase achievement, but only if significant systematic nonlinear or unmodeled variance exists in the environmental system which can be detected and correctly used (Cooksey, 1996). Stenson (1974) defined $C$ as the linear correlation between the variance in the task system and the subject's judgment system that is unaccounted for by the linear component. Thus, if all the systematic variance in the criterion scores can be accounted for by a linear combination of cues, then $C$ will equal zero. If a subject were sensitive to nonlinearity, but used it inappropriately, $C$ would be negative (Dudycha & Naylor, 1966). The matching index, $\Sigma-d$, was described by Hammond et al. (1964), and referred to by Hammond and Summers (1965) and Hoffman (1968); this index is a measure of the extent to which a subject uses the cues, relative to the validity of the cues. Hammond et al. (1964) found that $\Sigma-d$ became smaller with experience and was highly related to lens model achievement. If matching were perfect, $\Sigma-d$ would be zero. See Figure 2.
Figure 2. Brunswik's lens model, showing mathematical relations.

\[ r_a = \text{Achievement} \]
\[ X_i = \text{Informational cues} \]
\[ Y_e = \text{Criterion state being judged} \]
\[ \hat{Y}_e = \text{Predicted criterion state} \]
\[ Y_s = \text{Subject's judgments of the criterion state} \]
\[ \hat{Y}_s = \text{Predicted judgments of the criterion state} \]
\[ G = \text{Ecology-policy matching (linear knowledge)} \]
\[ R_s = \text{Policy consistency (cognitive control)} \]
\[ R_e = \text{Predictability of ecology (task control)} \]

Critical Discussion of Linear Modeling

Separating cognitive processes from task effects. Hoffman (1968) argued that regression approaches to modeling judgments related inputs (data) and outputs (judgments) mathematically; such models were not a representation of cognitive processes. He considered these linear models to be paramorphic (Hoffman, 1960, 1968). That is, regression equations constituted a description of the judgments "as if" the cognitive processing were such.

The fact that linear models work quite well may be more because they model the task rather than the judge (Westenberg & Koele, 1994). Thus, when performance is demonstrated through regression approaches, the findings may not be informative about the actual cognitive processes used. In this study, additional tasks were used (information board and protocol analysis) to allow the nature of the cognition involved in judgment to be assessed.

Problems of interpretation of model parameters. Diamond (1989) critiqued the use of the regression method as reported in a recent medical study (Speroff et al., 1989). Schneck and Naylor (1968) also identified important limitations of the modeling process. In any sample, outlying cases have potential to significantly influence the cues and their weights. Minor differences in cue-intercorrelations can lead to large differences in the particular set of cues which enters the equation. Furthermore, results from one sample of subjects often do not cross-validate well. Relationships between the data and outcome criteria have been found to be unstable, unless the sample size is very large.

Definition and measurement of configurality. As was discussed in chapter 2, the distinction between configural and analytic processing may be an important source of difference in judgment process. Kaplan (1975) defined
configural processing as "[processing] where properties of a given stimulus are determined with reference to the other stimuli in the array" (p. 150). Stimulus importance is affected by the relationships in the configuration. In linear processing, stimulus values are simply added or subtracted, and patterning on the basis of the stimulus array is not involved.

Camerer and Johnson (1991) pointed out that experienced subjects often use configural rules, whereas novices behave more like regression models: they weight cues and add them up. One problem for experienced subjects, however, is that often even elaborate configural rules have little positive impact on performance. In the traditional lens model analyses, nonlinear cue-use is assessed by the value of C. In previous studies, comparisons of novices and experienced individuals have revealed little difference in C-scores; the reason may be that, as Green (1968) explained, configural effects are often masked by overwhelming linear effects.

N. H. Anderson (1972) argued that clinical judgment can be considered as information integration; the task facing the clinician is to integrate separate pieces of informational stimuli into some unitary judgment. Anderson claimed that "one main difficulty with previous work has been its reliance on standard multiple regression methodology. . . . The failure to detect configurality seems to have resulted . . . in part from the inherent limitations of that methodology" (p. 93). Goldberg (1968) found evidence lacking for configurality in judgment, and concluded that "the power of the linear regression model is so great that it serves to obscure the real configural processes in judgment" (p. 488).

Considerable research has been carried out in relation to linear and nonlinear models (Brehmer, 1969; L. C. Johnson & Mai, 1979; Shanteau & Anderson, 1972). Although the lens model has provision for both linear and nonlinear cue-use, the
variance explained by linear modeling has traditionally been identified first; any systematic variance remaining that matches variation in the task system is then considered nonlinear. Rozenboom (1972) stated: "The extent to which the relation between cues and focus variables is linear rather than curvilinear is very much an artifact [emphasis added] of how we choose to span the cue space" (p. 324).

To the extent that expert clinicians' differential weighting of cues for different cases is required for expertise, the lens model may not be completely adequate to reveal differences associated with expertise. Regression methods are based on the assumption that cue-weights are invariant across cases. To identify if configural processing of cues is associated with expertise, additional methods for determining configural cue-use may be necessary. In this study, the orthogonal-cue judgment task was used as an alternative method for this purpose.

**Correlational measures of performance.** Lindell (1978) expressed a warning about the interpretation of $R^2$ as an index of performance; this measure is the proportion of variance accounted for by linear regression. Lindell argued that any factor which influenced the magnitude of the response variance affected the $R^2$ index: "To the degree that subjects' responses are differentially affected by these factors, differences in levels of $R^2$ become uninterpretable" (p. 71). Lindell recommended that in comparing different subjects, the absolute amount of residual variance for each subject was more useful than an index based on $R^2$. In the lens model, multiple R ($R_m$) is used to operationalize the construct of cognitive control. O'Grady (1982) also pointed out cautions and limitations regarding measures of explained variance in psychological research. In this study, the $R^2$ index and multiple R were not used as performance measures, but as independent variables, tapping subjects' sensitivity to broad patterns of data in the ecology.
The main indicator of judgment expertise in this study was stipulated to be accuracy in judgments of healing time. Accuracy was determined by two measures. The first was the square root of the mean of squared error scores, as recommended by Lindell (1976). This measure, similar to the one employed by W. F. Wright and Anderson (1989), represents how many days, on average, each subject's judgments deviate from the actual healing time.

Because the time for incisions to heal was not precise, an interval measure of healing time was used: a 10% interval was constructed around two days preceding and two days following the reported healing time. Subjects were informed about this measure. This procedure was used so that an interval of at least five days would be assured when assessing subjects' calibration. The judgments were difficult, and a liberal interpretation of the meaning of healing time was appropriate. For each case, the "error", or deviation score, was defined as the difference between each subject's judgment of healing time and the closest limit of the interval measure of actual healing time. This measure of accuracy is residual variance, where smaller scores indicated higher accuracy. For greater interpretability, a simple linear transformation was carried out by subtracting each score from 100, so that high scores indicated higher accuracy. The equation for accuracy (transformed) assessed by mean error variance was:

\[
\text{Accuracy} = 100 - \sqrt{\frac{\sum (HT - JHT)^2}{n}}
\]

Where:

HT = Healing Time (Closest limit of interval measure)

JHT = Judgment of Healing Time

n = Number of cases (35 normal order, 35 reverse order)
The second measure of judgment accuracy was the number of correct judgments as a percent (or proportion) of total judgments. The proportion measure was used to be consistent with the literature in assessing confidence in relation to accuracy.

Error variance and the percent correct measures have an advantage over correlational measures for comparing novice and experienced clinicians on performance accuracy. Correlation methods are noted to be insensitive to differences between sets of scores that increase monotonically (Pritchard & Roth, 1991). For example, Castellan (1992) demonstrated that the use of G (task knowledge) obtained from lens model analyses had limited use as an index of performance: this author stated that "G can be large for even the most perverse or least attentive subjects" (p. 380). Evidence that G-scores quickly reached ceiling levels is found in Cooksey and Freebody (1987), where first- and second-year student teachers' G-scores were found to be extremely high (.86 to .99) in a study where reading achievement was predicted.

**The problem of the criterion.** In trying to determine expertise in clinical judgment performance, there frequently is difficulty in defining judgment accuracy in an objective manner. This point is a not a criticism that is peculiar to linear modeling; any method of investigating judgments may be subject to such a problem. Shanteau (1992) stated that external standards are seldom available and that is why expert judgment is needed; the fact that standards are defined from subjective opinion of experts, (and not the other way around) can be problematic. Salthouse (1991), in a discussion about research on expertise, and Ericsson and Smith (1991) pointed out the importance of obtaining measures of actual competence, rather than relying on consensual judgment or other index.
Situations which constitute judgment performance cannot be treated as a simple algorithm, which, if followed faithfully and systematically will lead unequivocally to accurate responses. By definition, and by their very nature, judgments are probabilistic and will contain error. How can expert judgment be distinguished from random guesses? Chan (1982) characterized good judgment as having coherence, veridicality, and reliability. In some contexts such as weather forecasting, the judgments made can be compared to the weather that actually occurs. Over time, a measure of accuracy can be calculated.

Clinical judgments, however, frequently do not have an externally verifiable criterion. Judgments about the severity of congestive heart failure (LaDuca et al., 1988), the nature of cardiac pain (Jacavone & Dostal, 1992), patients' need for blood transfusion therapy (Salem-Schatz et al., 1990) are difficult to assess for quality because there is no absolute basis for correct judgments.

In this study, the judgment task involved predicting healing times for a number of patients. Because an actual criterion existed, levels of judgment accuracy for different subjects could be compared. In some studies, expert judgment is substituted for the criterion, a practice which may be justifiable in some circumstances. For example, Carlson (1989), investigated the degree of risk of a child for abuse, and Kirwan, Chaput de Saintonge, Joyce, and Currey (1983) judged current arthritis activity. When the purpose of the study, however, is to examine aspects of expertise in judgment, an objectively verifiable criterion is essential.

Section C: Research Tasks, Conditions, and Measures

This section is a description of the methods used, including research tasks, conditions and measures.
Overview of Methods

In the literature review, it was demonstrated that judgment researchers often focused on one phase of the judgment task, and considered only a few variables. In this study, the intent was to explore the patterns of relations between experience and a number of indicators of clinical judgment performance on one task, with a group of subjects who had a wide range of experience. The patterns of relations identified for both judgment process and outcome under two different conditions were anticipated to be helpful in interpreting the results of previous judgment research.

The methods selected were intended to achieve the following goals: (1) to model the relationships between cues and a criterion in one clinical ecology; (2) to reveal aspects of conceptual structure, sensitivity to patterns in data, and judgment process; (3) to obtain data about age, education, and experience; (4) to capture subjects' judgment policies under either enhanced memory-priming or baseline memory-priming condition, and in both cue-presentation orders; and, (5) to identify variables that were most predictive of expertise in clinical judgment performance.

Research Tasks and Conditions

Association task. Benner's (1984) research on novice and expert nurses provided direction for this task. The purpose of the task was to encourage subjects to describe their experiences with surgical patients, with a particular focus on incisional healing. This was the initial research task; because the subjects had not yet been influenced by the other tasks in the study, it was of interest to pay attention to the level of specificity of the cues to which they referred, and the structure of their verbalization. In his research on expertise, Shanteau (1992) found that experts' knowledge was readily accessed through stories about past cases. Anecdotal
accounts provided a mnemonic to remember and a convenient way to organize the information.

After reading a card with information about a patient with an abdominal incision, subjects were asked to say what came to mind. They were encouraged to use their experience in any way that might be helpful. Three cards were used and responses were tape-recorded. The types of cues and their level in a conceptual hierarchy, and evidence of the extent to which a coherent theory guided subjects' thinking were assessed, as suggested by Murphy and Medin (1985).

**Information board task.** This task was based on research carried out by Payne (1976) and reviewed by Ford, Schmitt, Schechtman, Hults, and Doherty (1989), and by Harte, Westenberg, and von Someren (1994). Information search and judgment processes were assessed. The goal of the task was to have subjects rank order four simulated patients (with data based on actual case data) with respect to healing time. Cues which might provide information about the four patients were printed on overturned cards; only the labels for each cue-category were in view initially. Thirty-two pieces of information were potentially available; subjects decided which cue to examine at each point, up to a maximum of 24 pieces. Pilot testing this task showed that some data restriction was necessary in order to elicit strategic information processing.

Once sufficient data were selected and subjects integrated the cue information, they made rank order judgments of healing time. During this search and judgment task, subjects were asked to think aloud. Verbalizations were tape-recorded. Process tracings were constructed which revealed depth, variability, and pattern of information search.
Examples of strategic search processes, and compensatory and noncompensatory integration strategies were identified. Verbalizations were assessed using protocol analysis according to descriptions from Ericsson and Simon (1980, 1993), Olson and Biolsi (1991), and Payne, Bettman, and Johnson (1993a, 1993b).

**Healing time judgment task.** This task was based on research by K. R. Hammond et al. (1964). Subjects responded to two sets of 40 cases, each consisting of 35 cases for analysis, and five repeated cases to assess consistency. The cues were expressed in two brief paragraphs: one containing context cues (such as age, history, diagnosis, and surgery) and one containing the individuating or specific cues (such as blood loss, surgery time, complications, and incision data).

Cases were presented one at a time by personal computer, using a program developed for this purpose. The paragraphs were typed in white letters on a blue background for ease in reading. Once the subject signalled readiness, the initial paragraph was presented on the screen. The second paragraph appeared when the subject again indicated readiness; both paragraphs remained in view until the subject was ready to make a judgment. Subjects had control over presentation time, to a maximum of 30 seconds per paragraph; from pilot testing, such time was determined to be sufficient. Subjects read each case and made their best judgments of healing time and confidence. Latency was timed by the computer for each case, and subjects were aware of this.

In the normal order cases, context cues were presented first, followed by the individuating cues. Because this paragraph sequence would be highly familiar (particularly for experienced subjects), it was theorized that normal order presentation would encourage the use of well-learned scripts. In the reverse order
cases, the individuating cues were presented first, followed by the context cues. Reverse order presentation was theorized to require more attention and encourage more deliberate processing because the individuating cues would have little meaning until they were understood in terms of the context cues. In addition, reversing the order of the paragraphs was thought to potentially interfere with the formation of a good anchor to use as a heuristic for prediction (Friedlander & Stockman, 1983). Based on the research of K. R. Hammond and colleagues, the normal order cases were theorized to be intuition-inducing, whereas the reverse order cases were expected to be analysis-inducing; it was anticipated, therefore, that experienced subjects would demonstrate reduced performance in the reverse order cases, compared to their performance in the normal order cases. Examples of both normal order cases and reverse order cases are provided in Appendix B.

Presentation sequence for paragraph order (normal order and reverse order cases) was counterbalanced between subjects. Subjects assigned to case sequence 1 completed all normal order cases, then reverse order cases. Subjects assigned to case sequence 2 completed all reverse order cases, then normal cases. Each subject performed both sets of cases; paragraph order was a within-subject factor. Subjects were randomly assigned to case sequence, within experience levels. The presentation sequence for all tasks is shown pictorially in Appendix C.

Healing time was defined as the earliest time that incision surfaces were fully approximated, with no stitches or staples, no redness, swelling, or drainage. This definition (consistent with the definition used with surgical patients in phase 1 of the study) was explained to each of the subjects. Expertise in judgment performance was measured by accuracy, calculated as the percent (or proportion) correct, or the square root of the mean of squared errors, as stated previously. For the error variance measure, an upper limit (or "ceiling") was imposed so that a few extremely
inaccurate judgments would not have excessive influence on the score for the entire set of 35 judgments. Deviation scores of 4 weeks or more for any one case were considered irrelevant to the assessment of judgment performance.

The following assessment for the lens model measures were made. (1) Knowledge of task relations, G, (the extent to which the subject correctly detected the properties of the task), was obtained by correlating the prediction of the criterion with the prediction of each subject's judgment. (2) Cognitive control, R_s, (the predictability of the subject's responses based on the cues), was obtained from the multiple correlation between Y_s and the cues. (3) Nonlinear variance-matching, C, (an indication of configural cue-use) was the correlation between the residuals of the linear predictions of Y_e and Y_s. (4) Ecological validity coefficients, r_e (measures of the importance of the cues in the ecology) were calculated by correlating cue-values with actual criterion-values in the ecology. (5) Utilization coefficients, r_s (the importance placed on the cues by the subject) were calculated by correlating cue-values with judged criterion-values in the subject system. (6) The matching index, Σ-d, (the degree of match between the ecological validities and the utilization coefficients) was calculated in the following manner. First, two sets of differences from the lens model cases were determined: those between the model-based r_e (representing the ecology) and the subject-based r_s (representing the subject system), and the differences between the model-based regression coefficients and the corresponding regression coefficients from the subject system. The products of these differences were computed, and then added. If a subject had perfect matching ability, then Σ-d would be zero. Thus, as a matching index, Σ-d assesses the degree of mismatch.

The equation for Σ-d, based on Hammond et al. (1964), is as follows:
\[ \Sigma -d = \Sigma \left[ (r_{ei} - r_{si}) (\beta_{ei} - \beta_{si}) \right] \]  

(5)

Where

- \( r_{ei} \) = Ecological validity coefficients
- \( r_{si} \) = Utilization coefficients
- \( \beta_{ei} \) = Model-based regression weights
- \( \beta_{si} \) = Subject system-based regression weights

**Cue-importance task.** Several researchers (Birnbaum & Stegner, 1981; Brehmer & Qvarnström, 1976; Cotton, Jacobs, & Grogan, 1983; Nystedt & Magnusson, 1975; Schmitt & Levine, 1977) recommended that in terms of understanding the psychology of the judgment process, research directed at understanding the subjective weights of the cues be pursued. Therefore, in this study (in addition to the two regression policies), a third judgment policy was derived for each subject based on subjective weights of cues.

In a series of 10 lens-model type cases, for each case, subjects predicted healing time based on the cues given, and rated the importance of each of the cues to the judgment. Subjects rated cue-importance on a scale from 1 to 100. This method was used because of its simplicity. Furthermore, in a study comparing seven methods for obtaining subjective importance of cues, which included this method, Cook and Stewart (1975) reported no significant differences between methods.

A subjective policy was derived for each subject. These policies were used (with regression-based policies), to predict healing time for a new set of 35 cases as a means to assess the degree to which the subjective policy compared with the
regression-based policies in accuracy of cross-validation. The new set of cases was
developed from phase 1 data to be parallel to the other two sets.

Einhorn, Kleinmuntz, and Kleinmuntz (1979) and Surber (1985) identified that
reduction in the range of cue-values can lead to inaccurate interpretation of the lens
model equation. In developing each of the sets of cases, every attempt was made to
obtain a range of cue-values that was representative of the clinical ecology.

In Table 3-01, the means, standard deviations, and $R^2$ for normal order,
reverse order, and cross validation cases are reported, to demonstrate the extent to
which the three sets of cases were matched on these features.

Table 3-01

Means and Standard Deviations for the Dependent Variable, Healing Time, and
Regression Model R-Squared for Normal Order, Reverse Order, and Cross-Validation

<table>
<thead>
<tr>
<th>Cases</th>
<th>Summary Description for Three Sets of Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td>Types of Cases</td>
<td>Healing Time</td>
</tr>
<tr>
<td>Normal Order</td>
<td>18.40</td>
</tr>
<tr>
<td>(n=35)</td>
<td></td>
</tr>
<tr>
<td>Reverse Order</td>
<td>19.66</td>
</tr>
<tr>
<td>(n=35)</td>
<td></td>
</tr>
<tr>
<td>Cross-Validation</td>
<td>19.02</td>
</tr>
<tr>
<td>(n=35)</td>
<td></td>
</tr>
</tbody>
</table>

Note. SD = Standard Deviation
**Slide task.** Subjects viewed 15 slides of patients with abdominal incisions; each slide was a close-up view of the incision so that fine detail was possible to see. A rear-projection viewer with a 16 inch screen was used. Subjects sat about three feet from the screen. Lighting in the room was dimmed to create maximum visibility. All slides were shown in the same sequence. The first 12 slides were shown for 5 seconds; the last three slides were shown for about a minute. Each subject described and interpreted these latter three slides. Responses were tape-recorded.

This task had two purposes. The first purpose was to attempt to enhance memory-priming. The slides were anticipated to elicit memories of relevant past cases. One of the criticisms of conducting research in a laboratory context is that cues that are normally present in the natural setting are absent in the laboratory. Because clinicians make use of such cues as a guide to judgment, their absence may (in part) account for previous researchers' findings that judgment accuracy in the laboratory setting is modest. In order to reduce the performance-depressing effect that the laboratory context may have, the association task and the information board task were scheduled to precede the healing time judgment task for all subjects. The slide task, however, was theorized to instantiate an appropriate schema more effectively compared to tasks that rely on verbal cues alone. In addition, the visual detail about actual incisions available in the slides was predicted to be more helpful to *experienced* subjects, compared to novices because experienced subjects likely have many more memories. An interaction between slide condition and experience was predicted. The slide task was scheduled prior to the healing time judgment task for half the subjects at each experience level in order to assess whether this attempt at enhanced memory-priming was effective.
Two memory priming conditions were used: enhanced memory-priming, and baseline memory-priming. Subjects in condition 1 saw the slides immediately prior to engaging in the healing time judgment task. In condition 2, subjects saw the slides after completing this task. Thus, memory-priming was a between-subjects factor. Subjects at each experience level were randomly assigned to the priming condition. If evoked memories included accurate cue-criterion relationships, judgment accuracy was theorized to increase in the enhanced memory-priming condition. If memories were not available, or not in correspondence with the selected cases, such exposure to slides should not facilitate judgment performance.

The second purpose of this slide task was to obtain a measure of accessibility of particular domain-specific knowledge. This task was derived from research carried out by Myles-Worsley, Johnston, and Simons, (1988), Norman, Brooks, Allen, and Rosenthal (1990), and Norman et al. (1989). Myles-Worsley and colleagues used X-ray films as stimuli and Norman and colleagues examined diagnostic performance from slides of dermatology conditions. In the present study, verbalizations (both descriptive and evaluative comments) about the incisions seen in each slide were tape-recorded.

**Orthogonal-cue judgment task.** This task was referred to as the card-sorting task. It was based on the research conducted by Ashton (1974), Hoffman, Slovic, and Rorer (1968), Slovic (1969), and Slovic, Rorer, and Hoffman (1971). The purpose of this task was to assess the use of cues where nonlinearity was present. Case stimuli were generated by orthogonal combination of four cues, three of which were well known to influence healing: *age* (two levels: 32, 78); *diabetes* (two levels: absent, present since age 30); *weight* (four levels, expressed in pounds: 90, 150, 210, 270); *height*, expressed in feet and inches (three levels: 5' 0", 5' 5", 5' 10"). Four levels of *weight* were used to include an example of a nonlinear cue, where
both extremes could lead to problems with wound healing. Also, the weight cue and the diabetes cue were configural: the importance of the variable weight to the judgment of healing time depended on the value of height, and the importance of the diabetes depended on the value of age.

In this study, subjects made judgments about healing for 48 non-representative, hypothetical patients. Patient data (cues) were printed on cards. Subjects were informed that each case referred to a hypothetical female patient who had undergone a major abdominal operation. They were encouraged to use their knowledge and experience to help them estimate healing times for patients with particular combinations of data.

A meter-long time-line with markers at 10-day increments was placed on a desk surface. The goal was to assess each card with patient data and place the cards judged similar with respect to healing time together in a category. Subjects could construct as many categories as they could discriminate. To encourage analytical processing, subjects were informed that they could take whatever time they needed, apply any strategy they wished, and continue to sort the cards until they were satisfied with their placement.

Subjects' judgments of healing time in hypothetical patients reflected main effects of the information cues and interaction effects, or configural cue-use. Because the patients were hypothetical, there were no objectively correct responses for this task. Using the responses given by each subject as dependent variables, the data from this task were analyzed first using a single-subject Analysis of Variance (ANOVA), and then calculating omega-squared ($\omega^2$). Omega-squared provided an estimate of the proportion of the total variance in the subject's judgments that could be attributed to main effects and to interaction terms (Hayes, 1981; Slovic, 1969).
Omega-squared for main effects was a measure of linear information utilization; \( \omega^2 \) for interactions assessed configural information utilization. In this study, the latter measure was of most interest, as it provided an assessment of nonlinear cue-use that was an alternative to lens model C-scores. Omega-squared for interactions represented interactive processing under analytic conditions.

**Concept similarity judgment task.** This task was based on the research conducted by Schvaneveldt et al. (1985) in their research on the structure of expertise, and that of Roth, Gabel, Brown, and Rice (1992), who identified changes in students' cognitive structure associated with studying texts.

Each subject's conceptual structure (or cognitive structure) was revealed through multidimensional scaling (MDS). MDS is a mathematical technique in which a complex matrix of numbers representing proximities is reduced into a simpler picture. The distances between points reflected the psychological proximity of the concepts (Schiffman, Reynolds, & Young, 1981).

Similarity judgments were made of all possible unique pairs of 16 concepts related to wound healing (120 judgments in total). The particular concepts used were obtained through reading the literature on wound healing, and by consulting with a clinical nurse specialist with expertise in this field.

Each pair of concepts was displayed in sequence by personal computer, using a program developed for this purpose. An assumption was made that the proximity judgments were symmetrical; thus, the similarity rating for the concept pair *Erythema Inflammation* was assumed to be the same as that for *Inflammation Erythema*. The ordering of concept pairs was the same for all subjects. The concept names were presented on the same line in white letters on a blue
background half way down the screen. Subjects controlled the presentation rate and were encouraged to respond quickly.

The verbalizations arising from the association task (discussed previously) constituted the second source of data for conceptual structure. Verbal protocols are discussed in relation to judgment process.

**Measures of Cognitive Constructs**

**Conceptual structure.** Data pertaining to this construct were obtained from two sources. One source was the $R^2$ associated with the multidimensionally scaled ratings generated from the similarity judgment task. The other source was the measure of coherence of theory from the association task. Conceptual structure was inferred from the relationships among concepts, the structure of verbalizations, and the level of key terms subjects used (superordinate, basic, or subordinate) as subjects described experiences caring for surgical patients.

**Sensitivity to patterns in data.** Sensitivity to overall patterns of data in the ecology was assessed by the lens model measures, supplemented by the verbal protocols. Sensitivity to fine-grained patterns involving more subtle case-by-case variation in cue importance was assessed by the standard deviations of the subjective importance of cues and by the variability of information search for each alternative.

The influence of cue-presentation order (context cues followed by individuating cues, or the reverse) on the patterns of relations was assessed by comparing lens model measures and judgment accuracy separately in the two conditions.
**Judgment process.** This construct was assessed by the following measures: depth, variability, and pattern of search from the process tracing derived from the information board task, and configurality measured by the interaction terms obtained from the orthogonal-cue judgment task, and by lens model C-scores.

Judgment process was assessed also by means of protocol analysis. Protocol analysis is a technique that has been used successfully in many studies of judgment process (Elstein et al., 1978, 1990, 1993; Ericsson & Simon, 1980, 1993; Newell & Simon, 1972; Svenson, 1989). P. E. Johnson et al. (1982) used think-aloud protocols to distinguish lines of reasoning for novice and expert physicians. Nissila (1992) used this method with 5 experienced nurses and 5 less experienced nurses to identify lines of reasoning in clinical nursing decisions.

Protocol analysis was used in this study as a supplement to regression approaches, so that data could be obtained on judgment process as well as on judgment outcome. Protocols generated from the information board task were assessed for evidence of data search and cognitive processing strategies.

Making inferences about cognitive processes from verbal protocol data, however, is not without its difficulties; in fact, some authors criticize this method (Nisbett & Wilson, 1977). There are three main concerns regarding the use of think-aloud protocols as data in judgment research. The first relates to reactivity (whether the thinking aloud *influences* the judgment process). The second concern is validity; some authors question whether the protocols obtained are *accurate* as indicators of cognitive processes. The third concern is that, because experienced subjects have more automatized cognitive processes, compared to novices, both reactivity and validity may vary with experience level.
These criticisms of protocol analysis have had much debate in the literature (Ericsson & Simon, 1993; Fidler, 1983; Nisbett & Wilson, 1977; Russo, Johnson, & Stephens, 1989; P. A. White, 1988). To avoid reactivity, and to obtain maximum validity, the method must be carried out correctly; Carroll and Johnson (1990) reminded researchers that asking subjects to explain their thinking while engaged in the task represented a major threat to the validity of verbal protocols.

Henry, LeBreck, and Holzemer (1989) assessed the influence of thinking-aloud with 60 pediatric nurses and found no apparent effect on cognitive processing. Biggs, Rosman, and Sergusonian (1993), in a study of judgments of investment quality, reported that verbalizations did not affect the amount or pattern of data acquisition, or accuracy of judgments; earlier findings indicating that verbalization increased task latency were replicated.

Results from previous studies have revealed that judgment processes are adaptive and constructive (Payne, Bettman, Coupey, & Johnson, 1992; Payne, Bettman, & Johnson, 1993a). A qualitative analysis was carried out to determine whether the subjects in this study demonstrated these characteristics. For example, the process tracings were examined together with the transcribed verbal protocols to identify examples of adaptive strategies. In addition, the process tracings were divided into three parts: the first eight cues, the second eight cues, and the remainder. Each section was classified in terms of strategy use. Outcome performance in terms of accuracy was determined.

Indicators of Expertise

Accuracy of judgments. This variable was assessed using accuracy based on the error variance measure and on the percent correct measure, using the judgments of healing time from the lens model task. Judgment accuracy was
determined by comparing each judgment with the actual clinical state in the ecology. Even though nurses caring for surgical patients are expected to predict which patients have highest risk for delayed healing, and they frequently make judgments about treatment based on the level of perceived risk, these judgments were anticipated to be difficult. As a means to increase the ability of the subjects to make the required judgment, two formats for response were used in this study: one was an analogue scale anchored with extremely shorter than average on the left, and extremely longer than average on the right. To indicate their response regarding healing time, for each case, subjects moved the cursor to the desired point (which was converted to a number by the computer program). The second response format was numeric. Subjects entered their estimate of the healing time, expressed in number of days. Based on research carried out by Tversky, Sattath, and Slovic (1988), the use of two types of response formats was theorized to help subjects provide the most precise estimate of healing time of which they were capable.

**Judgment of confidence.** Subjects' confidence in judgment was determined as a component of the healing time judgment task. Following the research tradition of G. Wright (1982), Keren (1991), and Paese and Sniezek (1991), the assumption was made that confidence expressed as a proportion can be treated as a subjective probability. Reported confidence reflected subjects' degree of certainty that their judgments were within the interval measure of healing time.

Data related to confidence in judgment were obtained by two methods. The first was a concurrent measure, based on research carried out by Fischhoff et al. (1977) and Lichtenstein, Fischhoff, and Phillips (1982). Each subject moved the cursor to a point on a scale which best represented his or her confidence that the preceding judgment was correct (within the interval range of the actual healing time).
The second method was retrospective, and was motivated by research conducted by Gigerenzer et al. (1991). Each subject was asked to make two retrospective estimates of the number of judgments of healing time that were within range. Subjects made these overall estimates at the conclusion of each of the two sets of 40 judgments.

Björkman (1992, 1994) and Juslin (1994) made assumptions based on a Brunswikian perspective. In the present study, these authors' research methods have been drawn upon because they are consistent with the lens model approach. Björkman (1994) proposed that beliefs in the form of internal cue validities mediated the processing of ecological cue validities in the assessment of confidence.

**Judgment consistency.** The judgments of healing time and the judgments of confidence were assessed for consistency. Five cases in each set of normal order and reverse order cases were repeated. For each subject (and for each set of cases), consistency for healing time was obtained by correlating the repeated judgments of healing time. Similarly, consistency in judgments of confidence was obtained by correlating the repeated judgments of confidence.

**Calibration of confidence.** This indicator of judgment expertise was based on research carried out by Garb (1986, 1989) and Oskamp (1962, 1965). Judgments of confidence were aggregated over confidence intervals and compared with proportion of accurate judgments to see how well individuals were able to judge the accuracy of their predictions of healing time. Calibration measures the extent to which the subjective probabilities (confidence judgments) are realized in terms of relative frequencies. Probabilistic judgments are said to be well-calibrated when the corresponding judged probabilities match the relative frequencies of events (Ronis & Yates, 1987).
Chan (1982) argued that for a judge who is well-calibrated, the realized frequency distribution should not deviate too much from the predicted probability distribution. For example, on those days when a weather forecaster predicts a 30% chance of rain, he or she should be correct about 30% of the time. That is, in retrospect, the historical record should show that on 30% of these days it did in fact rain.

In the present study, to be able to make this type of assessment, a proportion measure of accuracy was used. Such a measure was obtained by dichotomous scoring: all judgments within the interval measure of healing time (as described earlier) were scored as correct, and all judgments outside this interval were scored as incorrect. Subjects were aware of this interval measure. All 40 judgments in each set of cases (including repeated cases) were used in the calculation of proportion correct.

**Judgment latency.** Latency was determined for the first paragraph and for the two paragraphs presented together. The two time periods for each cases were added to determine the number of seconds from the time subjects indicated readiness to read a case to the time they were ready to enter their numeric judgment. For each subject, latency was averaged over cases.

**Accessibility of knowledge.** The Knowledge Accessibility scale was developed from subjects' responses to the three slides showing close-up views of abdominal incisions; subjects provided both descriptive and evaluative data. Verbalizations were analyzed for particular knowledge, structure, and style.

In Table 3-02 an outline of the tasks, measures, and related constructs used in this study is presented.
### Table 3-02

**Outline of Tasks, Measures, and Associated Constructs**

<table>
<thead>
<tr>
<th>TASK</th>
<th>MEASURES</th>
<th>ASSOCIATED CONSTRUCTS &amp; PURPOSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Association Task</td>
<td>- Assessment of Conceptual Coherence</td>
<td>- Conceptual Structure</td>
</tr>
<tr>
<td>2. Information Board Task</td>
<td>- Depth of search and search pattern</td>
<td>- Judgment Process</td>
</tr>
<tr>
<td></td>
<td>- Variability of information search as an index of configural cue-use</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Assessment of Judgment Process in terms of metacognition, input generation, and cognitive operations</td>
<td></td>
</tr>
<tr>
<td>3. Healing Time Judgment Task</td>
<td>- Absolute accuracy (Error variance and Percent correct)</td>
<td>- Expertise in Clinical Judgment under analytic and intuitive conditions</td>
</tr>
<tr>
<td></td>
<td>- Consistency of Healing Time Judgments</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Consistency of Confidence Judgments</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Calibration of Confidence Judgments</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Judgment Latency</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Lens Model Measures ($r_a$, $R_g$, $G$, $\Sigma$-d)</td>
<td>- Sensitivity to overall patterns in data</td>
</tr>
<tr>
<td></td>
<td>- Lens Model C scores</td>
<td>- Judgment Process (Lens Model Index of Configural Processing)</td>
</tr>
</tbody>
</table>
**Table 3-02 (Continued)**

**Outline of Tasks, Measures, and Associated Constructs**

<table>
<thead>
<tr>
<th>TASK</th>
<th>MEASURES</th>
<th>ASSOCIATED CONSTRUCTS &amp; PURPOSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>4. Side Task</td>
<td>- Assessment of experience-performance interaction under memory-enhanced conditions</td>
<td>- Judgment expertise (Used to attempt to enhance memory for previous cases)</td>
</tr>
<tr>
<td></td>
<td>- Assessment of Knowledge Accessibility</td>
<td>- Judgment Expertise</td>
</tr>
<tr>
<td></td>
<td>- Assessment of Judgment Process (Omega Squared as an index of configural cue-use)</td>
<td>- Judgment Process under analytic conditions</td>
</tr>
<tr>
<td></td>
<td>- Calibration of Healing Time</td>
<td>- Judgment Expertise</td>
</tr>
<tr>
<td>5. Orthogonal Judgment Task</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Distance between relevant conconcept clusters as indices of differences in the organization of conceptual structure</td>
<td>- Judgment Expertise</td>
</tr>
<tr>
<td>7. Cue Importance Task</td>
<td>- Subjective weights of cues from which to construct subjective policies</td>
<td>- Used to compare the quality of subjective policies to regression-based policies in terms of accuracy of cross-validation to a new set of cases</td>
</tr>
<tr>
<td></td>
<td>- Variability of Importance of Information</td>
<td>- Judgment Process</td>
</tr>
</tbody>
</table>
Measures of Experience

Experience was assessed by the number of months caring for surgical patients, by the estimates of the number of surgical incisions seen during their nursing career, and by the number of references to experience from the protocol data.

Section D: Anticipated Patterns of Relations

Based on the literature, with a focus on Dreyfus and Dreyfus' (1986) theoretical views pertaining to novice to expert changes in cognition, the following patterns of relations were anticipated:

1. Experienced subjects were anticipated to have clinical judgment performance characterized by higher accuracy and consistency, greater confidence, better calibration, and shorter latency, compared to less experienced subjects.

2. Experienced subjects were expected to have more sophisticated conceptual structure, greater sensitivity to patterns in data, and more complex judgment process compared to less experienced subjects.

3. Individual differences in age and education were expected to attenuate the patterns of relations between experience and accuracy in judgment performance.

4. Cue-presentation condition was anticipated to reveal a decrement in performance for experienced subjects, compared to their performance in the normal order cases.
5. Memory-priming condition was expected to reveal higher performance for experienced subjects in the enhanced memory-priming condition, compared to that in the baseline memory-priming condition.

6. Cognitive variables were anticipated to be most predictive of expertise in clinical judgment performance, compared to other variables.

**Section E: Summary**

In this chapter, a description of the method has been provided. The methods designed to address the research questions raised in chapter 1 have been described. Each of the tasks and measures have been selected to reveal different aspects of clinical judgment performance, or to address a problem with revealing expertise in the laboratory context.

The results from one group of subjects on these tasks were anticipated to demonstrate various patterns of relations. When examined together, these patterns were anticipated to provide greater understanding of expertise in clinical judgment, and possibly to add to knowledge about changes in cognition with expertise.

The next chapter is a report of the analyses and results.
IV. ANALYSES AND RESULTS

This chapter is organized in four major sections and a summary. In Section A, the analyses and results of phase 1 of the study where the ecology was modeled (the left side of the lens model) are presented. In Section B, the sample of subjects who made judgments are described, an overview of the data analysis is provided, and the first research question is addressed. The analyses and results relating to indicators of expertise (judgment accuracy, confidence, consistency, calibration, latency, and knowledge accessibility) are reported. In Section C, the analyses results related to the cognitive constructs (conceptual structure, sensitivity to patterns in data, and judgment process), and individual differences in age, education, and experience are presented. In Section D, the results related to paragraph order condition and memory-priming condition (research questions four and five) are given. The last section is a summary of the main findings: the variables which best predict expertise in clinical judgment are identified.

Section A: Lens Model Analyses [Phase 1 of the Study]

To obtain phase 1 data (the left side of the lens model), a sample of 281 patients with abdominal incisions was investigated as described in chapter 3. The dependent variable Healing Time was missing for 13 subjects; the data from these cases, therefore, were excluded from analysis. An additional 10 cases with extreme healing times were also omitted because these scores on the dependent variable could have excessive influence on the magnitude of the regression coefficients (Barnett & Lewis, 1978; Hatch & Prihoda, 1992). Data from 99 male patients and 159 female patients (258 cases) were analyzed.

The data were screened and the degree to which the data met the assumptions underlying regression was evaluated, as recommended by Tabachnick
and Fidell (1989). The variables which predicted healing time and their relative importance were then determined.

**Data Screening**

Thirty-five variables were assessed for accuracy of data entry. Six variables were dropped from analysis because of missing data. The variable Occupation was converted to a numeric socioeconomic index as described by Blishen, Carroll, and Moore (1987). The variables Weight and Height were combined to form a single variable, Body Mass Index, following examples in medical research (Revicki & Israel, 1986; Shetty & James, 1994). The use of this index eliminated variation in weight associated with differences in height.

The frequency distributions for all variables were assessed. The distributions for Age, Hemoglobin, and Occupation were close to normal, whereas the distribution for Surgery Time was found to have a high kurtosis value of 27.60 (standard error of 0.30). An examination of the histogram revealed that the distribution for Surgery Time included one particularly deviant (but accurate) score of 933 minutes; in all other 257 cases, surgery times were between 45 and 514 minutes.

To address problems associated with departure from normality, Hawkins (1980) recommended "Winsorization", which involved truncating the deviant score to one unit beyond its nearest neighbour. Hawkins stated that, providing the outlier represented a valid observation, an advantage of Winsorization was that this technique allowed the researcher to make partial use of the data and yet avoid excessive influence. The kurtosis was reduced to 5.49 (standard error of 0.30). Two other variables with high kurtosis values were: Blood Loss (9.73, standard error of 0.30) and Body Mass Index (6.85, standard error of 0.30). Various transformations
were attempted, with some degree of success, but they all increased the complexity of interpretation. These independent variables, therefore, were not transformed.

The following seven variables were dichotomous: Gender (men/women), Contamination (clean/contaminated), Diabetes (no diabetes/diabetes), Dressing (dry/packing), Infection (no infection/infection), Incision Length (short/long) and Incision Approximation (approximated/open). Ten variables were converted to dichotomies because of uneven distributions in the original categories: Race (Caucasian/other), Blood Transfusion (none/some), Reason For Surgery (other/cancer), Prior Operations (none/some), History Of Illnesses (no/yes), Number Of Drains (none/some), Nutritional Status (good/poor), Recent Weight Loss (none/some), Severity Of Surgery (minor/major) and Complications (none/some). For all of these variables, the categories were coded as 0 and 1, respectively.

The variable Hospital Service was converted to three categorical variables, using dummy coding as illustrated by Draper and Smith (1981). One cross-product term included was that between Gender and Severity of Surgery.

The dependent variable, Healing Time, had a positive skewness of 2.07 (standard error of 0.15) and a kurtosis value of 5.11 (standard error of 0.30). To improve distributional properties this variable was logarithmically transformed. The values after transformation were 0.37 and -0.57 (standard errors of 0.15 and 0.30), respectively. Subsequent analysis was based on 258 cases, 30 independent variables, and one dependent variable.
Evaluation of Assumptions

Regression analysis was conducted to obtain the lens model equation. The dependent variable, Healing Time, was a continuous variable measured in days; the independent variables consisted of categorical and continuous variables.

Fox (1984) identified the assumptions underlying the regression model as linearity, independence of residuals, normality of the distribution of residuals, and homoscedasticity (equality of residual variance at each level of the dependent variable). These assumptions were evaluated.

Variables were assessed for nonlinear relationships with Healing Time. Nonlinearity in the data would not invalidate the analysis, but such a relationship would not be captured by a linear model. Plots of the independent variables and Healing Time were inspected, and possible deviations from linearity were assessed by eta-squared ($\eta^2$). Howell (1992) defined $\eta$ as the correlation coefficient associated with curvilinear regression, and illustrated that $\eta^2$ represented the proportion of variance accounted for in the dependent variable by an independent variable (including non-linear relationship). The variable, Admission-To-OR-Time was found to have a non-linear relationship with Healing Time. Patients who had emergency surgery, as well as patients who had surgery several days following admission, tended to experience delayed healing. This variable was re-coded to a dichotomy. Stenson (1974) pointed out that it is theoretically possible to make the nonlinear variance vanish by applying appropriate transformations to the cue-measures. Because nonlinearity is an important measure in phase 2 of the study, excessive variable transformation was viewed as unwise. In this ecology, the level of predictability was only moderate, and the re-coding of this one variable was thought to be advisable in order to better meet the assumptions of the linear model.
There is evidence that the relationship between body-mass index and healing time might be nonlinear: emaciated patients heal relatively slowly, patients of optimal weight tend to heal quickly, and obese patients heal much more slowly, other factors being equal. A comparison of $R^2$ and $\eta^2$ revealed that the distribution for Body Mass Index for this sample did not show significant nonlinearity. A possible reason for this finding is that patients with extremely low Body Mass Indices were too ill to be asked to volunteer for this study.

The second assumption underlying regression that was evaluated was the independence of residuals. All patients were independently (but not randomly) sampled from four surgical units in a large teaching hospital. Some of the patients likely had the same surgeon, which could give rise to correlated residuals. Data regarding surgeons were not obtained. Cooksey (1996) described a graphical method and a statistical method for examining the independence of residuals assumption. An assessment of the residual plot revealed no patterns which might suggest correlations. Cooksey claimed that the Durbin-Watson statistic is useful as a test of the extent to which successive residuals are correlated. In this data set, this statistic was 2.16; compared to the upper and lower critical levels extrapolated from Durbin and Watson (1951), this value of 2.16 indicated that there was no significant autocorrelation present. The assumption of independence was considered to be satisfied.

The third assumption assessed was normality. Normality of the distribution of residuals was assessed in several ways. First, the studentized residuals were compared to critical values from Lund (1975), and all were within range. Studentized residuals are a more sensitive index compared to standardized residuals (Stevens, 1984). Second, the normal probability plot of observed and expected standardized residuals was inspected. With Healing Time transformed, this plot revealed the
desired linear pattern, with a slope of one. Finally, the Mahalanobis' distance scores were examined. This measure is the distance between two populations: one containing a possible outlier, and one with this particular case deleted (Kleinbaum, Kupper, & Muller, 1988). As the model fits outlying scores, excessively large residuals may result, some of which may violate the normality assumption (Rousseeuw & van Zomeren, 1990). Two multivariate outliers were revealed by Mahalanobis' distance.

Not all outliers, however, are influential, and the automatic rejection of outliers is not recommended (Andrews & Pregibon, 1978). Cook's distance ($C^2$) is a measure used to detect cases which have excessive influence (Cook, 1977, 1979; Cook & Weisberg, 1980; Hadi, 1993). In this data set, $C^2$ was within range. Because the two outlying cases met the inclusion criteria, and because of the danger of over-fitting a regression model if all outlying cases were removed, the two cases revealed as multivariate outliers were retained.

The final assumption related to regression that was assessed was homoscedasticity. The plot of standardized residuals and predicted Healing Time demonstrated moderate heteroscedasticity. A logarithmic transformation of Healing Time led to good improvement.

**Variables Predictive of Healing Time**

To identify variables that best predicted Healing Time, a stepwise multiple regression was computed using the Regression program from the Statistical Package for the Social Sciences (SPSS). With an inclusion criterion of $p = .05$, the equation included seven variables; the $R^2$ was .59, and adjusted $R^2$ was .58. The seven variables were: Severity Of Surgery, Dressing Type, Complications, Body Mass Index, Admission-To-OR-Time, Drains, and Occupational Index.
In Table 4-01, the intercorrelations between these seven independent variables and Healing Time (logarithmically transformed) are reported.

The two variables with the highest correlation with Healing Time were Severity of Surgery and Dressing Type. These variables have substantial shared variance as well. The sign of the variable Occupation was negative, indicating that people with low scores on the Blishen Index tended to have somewhat higher healing times, compared to people with high scores. Low scores for Occupation reflect low socioeconomic status, and all that is entailed: the likelihood for poorer nutrition, less education, less desirable occupation, and possibly a less healthy lifestyle, compared to those with high scores.
<table>
<thead>
<tr>
<th>Variables</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Severity</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>2 Dressing</td>
<td>.33*</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>3 Complications</td>
<td>.23*</td>
<td>.21*</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>4 BMIndex</td>
<td>.01</td>
<td>.11</td>
<td>.07</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>5 Admission</td>
<td>.26*</td>
<td>.13*</td>
<td>.15*</td>
<td>.13*</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>6 Drains</td>
<td>.29*</td>
<td>.01</td>
<td>.07</td>
<td>-.07</td>
<td>.04</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>7 Occupation</td>
<td>.09</td>
<td>.03</td>
<td>-.04</td>
<td>-.04</td>
<td>-.15*</td>
<td>.16*</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>8 H-Time</td>
<td>.55*</td>
<td>.54*</td>
<td>.43*</td>
<td>.29*</td>
<td>.31*</td>
<td>.21*</td>
<td>-.06</td>
<td>----</td>
</tr>
</tbody>
</table>

Note.

1 = Severity of Surgery (Severity)
2 = Dressing Type (Dressing)
3 = Complications
4 = Body Mass Index (BMIndex)
5 = Admission-to-OR Time (Admission)
6 = Drains
7 = Occupational Index (Occupation)
8 = Healing Time, in Log Transformation of Days (H-Time)

Correlations were based on data from 258 surgical patients

*p < .05.
The product term Gender-Severity was not significant. The correlation between gender and severity of surgery \((r = -.40)\) was interpreted as an artifact of the particular sample of patients selected. One setting was a gynecology ward where a large number of women (coded as 1) had had minor surgery.

**Relative Importance of Variables**

Lane, Murphy, and Marques (1982) identified three indices of cue importance between a cue and a criterion: the unstandardized and standardized regression coefficients, and the squared semi-partial correlation coefficients. These indices of variable importance were derived from a stepwise regression using the logarithm of Healing Time in days as the dependent variable. See Table 4-02.

**Table 4-02**

<table>
<thead>
<tr>
<th>Significant Predictors</th>
<th>Indices of Variable Importancea</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(b)</td>
</tr>
<tr>
<td>1 Severity of Surgery</td>
<td>.21</td>
</tr>
<tr>
<td>2 Dressing Type</td>
<td>.35</td>
</tr>
<tr>
<td>3 Complications</td>
<td>.15</td>
</tr>
<tr>
<td>4 Body Mass Index</td>
<td>.01</td>
</tr>
<tr>
<td>5 Admission-To-OR-Time</td>
<td>.08</td>
</tr>
<tr>
<td>6 Drains</td>
<td>.08</td>
</tr>
<tr>
<td>7 Occupational Index</td>
<td>-.00</td>
</tr>
</tbody>
</table>

**Note.**

\(b\) = Unstandardized Regression Coefficient

\(\beta\) = Standardized Regression Coefficient

\(sr^2\) = Squared Semi-partial Correlation Coefficient

a Indices were based on Stepwise Regression using data from 258 patients.
The variable Dressing Type, followed closely by Severity of Surgery were the best predictors, using both the standardized beta weights and the squared semi-partial correlation coefficient as criteria. The last three variables (Admission-to-OR-Time, Drains, and Occupational Index), although statistically significant, accounted for negligible variance in healing time.

In summary of this section pertaining to the left side of the lens model, the results are consistent with the literature (Cooper, 1990a, 1990b; Cruse & Foord, 1973, 1980; Irvin, 1981; P. L. Jones & Millman, 1990; Viljanto, 1991). Variables such as Severity of Surgery and Complications were found to be important predictors of Healing Time. Approximately 59% of the variance in healing time for one sample of patients was accounted for by seven variables. A task requiring judgments of healing time was judged to have sufficient predictability to be useful to reveal differences in nurses' sensitivity to overall patterns in data. In addition, there was considerable unexplained variance which was expected to give subjects with expertise possibility to attain scores exceeding that obtained from a linear model.

Section B: Results Related to Indicators of Expertise

[Phase 2 of the Study]

Description of Nurse Subjects

Thirty-six subjects volunteered for this study. Of these subjects, all but 1 were women; gender was not a factor being investigated. Six subjects were senior nursing students, 6 were recent graduates with work experience of 2 months or less, and 24 were Registered Nurses who had a range of work experience from 1 year to 25 years. In terms of general education, 15 subjects had high school as the highest level attained, 3 had college diplomas, 2 had baccalaureate degrees, and 1 had a master's degree (not in nursing); a further 15 subjects had some college or university
courses. For analysis purposes, subjects were classified into one of two categories: those with basic high school preparation only, and those with courses or degrees beyond high school.

With reference to the highest level of nursing education attained, 6 subjects were students from nursing programs who were within one month of graduation, 27 had diplomas, 1 had a baccalaureate degree in nursing, and 2 had a master's degree in nursing. With nursing education, subjects were classified into two groups: the students and nurses with a diploma in one group, and the nurses with more education in the other.

Experience assessed by the number of years since nursing graduation ranged from 0 to 25. The number of months that subjects cared for surgical patients ranged from 2 to 300. The use of subjects' estimates of the total number of surgical incisions viewed during their nursing career as an indicator of experience resulted in a variable with a range from 20 to 20,000, with a median of 2640.

The indicator of experience used for subsequent analysis was the number of months of surgical nursing experience reported. Subjects with 2 months experience or less, were categorized as level 1; those with greater than 2 months to 60 months experience, comprised level 2; and subjects with more than 60 months of surgical experience were classified as level 3. This categorization resulted in three groups of 12 subjects each. In Table 4-03, the number of subjects categorized in three different ways (by age category, general education, and nursing education) is reported, by experience level.
Table 4-03

Classifications of Subjects According to Experience Level, Age Category, and Education

<table>
<thead>
<tr>
<th>Experience Level</th>
<th>&lt;25 years</th>
<th>25-34 years</th>
<th>35-44 years</th>
<th>45+ years</th>
<th>General</th>
<th>Nursing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>n=12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>4</td>
<td>6</td>
<td>1</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>n=12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>n=12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8</td>
<td>4</td>
</tr>
</tbody>
</table>

Note. HS = High School

As anticipated, inexperienced subjects tended to be younger, compared to experienced subjects; although these novices had basic nursing education only, about two thirds of them had extra general education. In contrast, only one third of the experienced group had extra general education. The middle group was the most heterogeneous, with a greater number of subjects with extra general and nursing education. These subjects also had the largest range in age.

Overview of Data Analysis

The data were initially analyzed using ANOVA or MANOVA to assess for a significant effect of experience on the various indicators of judgment performance. As anticipated, because of the low power associated with the sample size, very few of the results were statistically significant. The goal of this study was to assess for patterns of relations. Chow (1988) stated that "the statistical significance of a set of data is not informative about the practical importance (or substantive significance) of
the findings" (p. 106). Schmidt (1992, 1996) and Cohen (1990) recommended the reporting of effect sizes. Snyder and Lawson (1993) and Tatsuko (1992) described several measures of magnitude-of-effect, including eta-squared ($\eta^2$) which is a measures of association strength. Eta-squared is the proportion of variance in the dependent variable accounted for by the particular variable in question. Fowler (1985) claimed that $\eta^2$ was the least biased estimate of the population correlation coefficient and was of greatest interest to behavioral researchers. Thompson (1987) also recommended this measure. Eta-squared, therefore, was selected because the central question in this study was to assess the strength of relationship between experience and judgment performance.

Cohen (1973) suggested $\eta^2$ for use with ANOVA designs; the formula he provided is as follows, where $SS_A$ is the between sums of squares for factor A, [which, in this study is experience] and $SS_T$ is the total sums of squares:

$$\eta^2 = \frac{SS_A}{SS_T}$$

Many researchers recommended the reporting of confidence intervals in addition to some measure of effect (Cohen, 1994; Fowler, 1985; Schmidt, 1992, 1996). Frick (1995), however, identified some difficulties in the interpretation of .95% confidence levels and questioned their use. Because the effect sizes in this study were very small (which gave rise to large confidence intervals having extremely limited usefulness), confidence intervals were not reported.
Research Question 1 is:

What are the patterns of relationships among various measures of judgment performance (indicators of expertise) and experience for a group of subjects in a clinical judgment task?

The indicators of expertise considered in this study were judgment accuracy, confidence, consistency, calibration of confidence, judgment latency, and knowledge accessibility.

**Judgment Accuracy**

Judgment accuracy was determined from the healing time judgment task. In each set of cases judgments accuracy was assessed in two ways: by the square root of the mean error variance, and by the percent of correct judgments. The interval measure of healing time as described in chapter 3 was used for both measures.

For normal order cases, (cases presented in natural order with a paragraph containing context cues first, followed by specific cues) accuracy assessed by error variance ranged from 77.00 to 91.70; accuracy assessed by percent correct ranged from 10.00% to 57.50%. For reverse order cases, (cases presented with a paragraph of specific cues first, followed by context cues) accuracy assessed by error variance ranged from 77.00 to 93.70; accuracy assessed by percent correct ranged from 10.00% to 62.50%. In Table 4-04 the means and standard deviations for measures of accuracy are presented, by experience level.
Table 4-04

Means and Standard Deviations for Accuracy Assessed by Error Variance and Percent Correct for Normal Order and Reverse Order Cases, by Experience Level

<table>
<thead>
<tr>
<th>Experience Level</th>
<th>Normal Order Cases (n=35)</th>
<th>Reverse Order Cases (n=35)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean Error Variance (SD)</td>
<td>Mean % Correct (SD)</td>
</tr>
<tr>
<td>1 n=12</td>
<td>87.59 (3.65)</td>
<td>40.17 (12.55)</td>
</tr>
<tr>
<td>2 n=12</td>
<td>86.93 (3.30)</td>
<td>38.75 (9.80)</td>
</tr>
<tr>
<td>3 n=12</td>
<td>87.99 (2.03)</td>
<td>45.08 (9.92)</td>
</tr>
</tbody>
</table>

Note. SD = Standard Deviation.

Error Variance Measure has been transformed; high scores indicate high accuracy.

Experience accounted for slightly more variance in performance using the percent correct measure, compared to the error variance measure ($\eta^2 = .07$ and $\eta^2 = .02$, respectively). The proportion of variance associated with the within-subject factor (normal and reverse order cases) was much smaller using the proportion correct measure compared to the error variance measure ($\eta^2 = .01$ and $\eta^2 = .18$, respectively). Although a definite pattern in the scores could not be identified, the middle group had the lowest scores in both normal and reverse order cases. The standard deviations for the novices showed greater variability for this group, compared to the other two groups.
Confidence

As described in chapter 3, judgment confidence was assessed in two ways: by a concurrent method, and by a retrospective method. In the concurrent method, immediately after each judgment of healing time, each subject indicated his or her confidence (in percent) that the preceding judgment of healing time was within the interval measure of actual healing time. In the retrospective method, following the set of 40 judgments, each subject reported the number of judgments that he or she estimated to be within the interval measure of healing time. Means and standard deviations for concurrent confidence and retrospective confidence are reported in Table 4-05, by experience level.

Table 4-05

Means and Standard Deviations for Average Confidence (Concurrent Method) and Overall Confidence (Retrospective Method) for Normal Order and Reverse Order Cases, by Experience Level

<table>
<thead>
<tr>
<th>Experience Level</th>
<th>Normal Order Cases (n=40)*</th>
<th>Reverse Order Cases (n=40)*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Concurrent Method</td>
<td>Retrospective Method</td>
</tr>
<tr>
<td></td>
<td>Means (SD)</td>
<td>Means (SD)</td>
</tr>
<tr>
<td>1 n=12</td>
<td>60.83 (16.43)</td>
<td>47.92 (18.64)</td>
</tr>
<tr>
<td>2 n=12</td>
<td>68.33 (16.05)</td>
<td>60.50 (20.24)</td>
</tr>
<tr>
<td>3 n=12</td>
<td>75.00 (11.86)</td>
<td>57.55 (17.10)</td>
</tr>
</tbody>
</table>

Note. SD = Standard Deviation

* Repeated cases were included in the assessment of confidence.
Data from Table 4-05 revealed that experience accounted for a small proportion of the variance in confidence ($\eta^2 = .12$ with the concurrent method; $\eta^2 = .07$ with the retrospective method). There was a pattern for confidence to be somewhat higher in the normal order cases, compared to the reverse order cases.

**Judgment Consistency**

As described in chapter 3, for each subject, consistency was assessed for both judgments of healing time and judgments of confidence in normal order and reverse order cases. Consistency measures indicated by the correlations for the corresponding sets of five repeated judgments are presented. Nonparametric summaries of central tendency and variability are reported in Table 4-06.

**Table 4-06**

Nonparametric Summaries of Consistency for Repeated Judgments of Healing Time and Repeated Judgments of Confidence (Concurrent) for Normal Order Cases and Reverse Order Cases, by Experience Level

<table>
<thead>
<tr>
<th>Experience Level</th>
<th>Healing Time: Mdn (IQR)</th>
<th>Confidence Mdn (IQR)</th>
<th>Healing Time: Mdn (IQR)</th>
<th>Confidence Mdn (IQR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 n=12</td>
<td>.83 (.21)</td>
<td>.55 (.60)</td>
<td>.88 (.17)</td>
<td>.53 (.62)</td>
</tr>
<tr>
<td>2 n=12</td>
<td>.92 (.22)</td>
<td>.26 (.79)</td>
<td>.95 (.12)</td>
<td>.22 (.78)</td>
</tr>
<tr>
<td>3 n=12</td>
<td>.93 (.15)</td>
<td>.80 (.41)</td>
<td>.97 (.05)</td>
<td>.73 (.45)</td>
</tr>
</tbody>
</table>

**Note.** Mdn = Median; IRQ = Interquartile Range
Overall, there was good consistency for the judgments of Healing Time. There was a pattern for consistency to increase slightly, and for the variability in consistency to decrease, with experience. The medians for consistency of judgments of Healing Time tended to be slightly higher in the reverse order than in the normal order cases.

Consistency in judgments of confidence, however, was extremely variable. As shown by the large interquartile ranges, some of the correlations were negative. The middle group had the most diverse correlations. The most experienced subjects had the best consistency in their confidence judgments, compared to other subjects, but this level of consistency was not nearly as high as that obtained for their judgments of healing time. Medians for consistency of confidence tended to be slightly higher in the normal order cases than in the reverse order cases.

**Calibration of Confidence**

There are many ways to analyze the quality of confidence judgments: Brier (1950) proposed what is now known as the "Brier score"; Oskamp (1962) developed the "appropriateness of confidence"; G. Wright (1982) referred to "realism of confidence". These measures are all related to calibration. As discussed in chapter 3, calibration is a measure of how well the level of confidence in one's judgment corresponds to the proportion of correct judgments, within a particular range of confidence. For example, considering the range of confidence between .7 and .8, for a perfectly calibrated judge, about 75% of them should, in fact, be correct.

Confidence calibration was assessed by comparing confidence and the proportion measure of accuracy in three ways. The first method was the traditional
method, used by Lichtenstein and Fischhoff (1977), G. Wright (1982) and others to allow comparison with previous research.

The equation for concurrent judgments of confidence was:

$$ C = \frac{\sum_{t=1}^{T} n_t (x_t - c_t)^2}{N} $$

Where:

- $C$ = Calibration
- $N$ = Total number of judgments made
- $T$ = Total number of different response categories
- $n_t$ = Number of judgments in response category $t$
- $x_t$ = Reported confidence of each judgment in category $t$
- $c_t$ = Proportion correct for items with confidence $x_t$.

With this method, calibration can be expressed as the sum of the squared differences between subjects' confidence (expressed as a probability) and the proportion correct within each confidence category, divided by the number of judgments.

The second method of obtaining calibration [Björkman's (1994) method] also used concurrent ratings of confidence that had a one-to-one correspondence to each judgment made. Björkman decomposed calibration ($C$) into three additive parts: bias, the squared difference between mean confidence and proportion correct ($D^2$); resolution, the squared difference between the standard deviations of confidence judgments and proportion correct ($R^2$); and deviation from linearity, the departure of the calibration curve from a linear function ($L$).
The equation used was:

\[ C = (\bar{X} - \bar{C})^2 + (s_x - s_c)^2 + 2 s_x s_c (1 - r_{xc}) \]  

(8)

Where:

\( C \) = Calibration  
\( \bar{X} \) = Overall mean confidence  
\( x \) = Reported confidence of each judgment  
\( \bar{C} \) = Overall proportion of correct responses  
\( c \) = Judgment accuracy, where judgments were scored dichotomously.  
\( s_x \) = Standard deviation of mean confidence scores over confidence intervals.  
\( s_c \) = Standard deviation of proportion correct within confidence intervals, over confidence intervals  
\( r_{xc} \) = Correlation of reported confidence and judgment accuracy where the latter was scored dichotomously.

The simplified equation for calibration from Björkman (1994) is:

\[ C = D^2 + R^2 + L \]  

(9)

The first term in equations 6 and 7 is bias (D). The most common bias that has been reported in the literature is that of overconfidence. Overconfidence is demonstrated when mean confidence exceeds the mean proportion of correct judgments, whereas the reverse shows under-confidence. Overconfidence has been a robust finding (Fischhoff et al., 1977; Koehler, 1994; Zakay & Glickson, 1992). Whether overconfidence represents a cognitive bias, however, or is to some
extant an artifact of the method used for items selected for study is currently being debated (Björkman, 1992, 1994; Juslin, 1994). Test items which have not been selected representatively may give rise to overconfidence. In this study, the use of representative sampling may provide a more accurate assessment of bias.

The second term in equations 6 and 7 is resolution, \( R \), which reflects the subject's ability to discriminate between correct and incorrect cases. The third term is deviation from linearity (L), which describes how closely confidence judgments would fit a line drawn to represent expected proportion correct, if calibration were perfect. Ideally, this deviation is zero. In Table 4-07, the means and standard deviations for judgment calibration assessed by the traditional method and by Björkman's method are presented, for normal order and reverse order cases.

Table 4-07

Means and Standard Deviations for Two Methods of Assessing Judgment Calibration for Normal and Reverse Order Cases, by Experience Level

<table>
<thead>
<tr>
<th>Experience Level</th>
<th>Normal Order Cases (n=40)*</th>
<th>Reverse Order Cases (n=40)*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Traditional Method</td>
<td>Björkman Method</td>
</tr>
<tr>
<td></td>
<td>Means (SD)</td>
<td>Means (SD)</td>
</tr>
<tr>
<td>1 n=12</td>
<td>.09 (.07)</td>
<td>.12 (.08)</td>
</tr>
<tr>
<td>2 n=12</td>
<td>.12 (.10)</td>
<td>.14 (.10)</td>
</tr>
<tr>
<td>3 n=12</td>
<td>.14 (.09)</td>
<td>.14 (.05)</td>
</tr>
</tbody>
</table>

Note. For both methods, higher scores indicate lower calibration. SD = Standard Deviation.

* Repeated cases were included in the assessment of Judgment Calibration.
These methods both employ concurrent assessments of confidence. Responses from both methods were similar. There was a very weak relationship between experience and calibration ($\eta^2 = .06$ for the traditional method and $\eta^2 = .02$ for Björkman's method). An examination of the means suggested that calibration became somewhat worse with experience.

A third method of examining the quality of confidence judgments was based on the work of Gigerenzer et al. (1991). These authors proposed the hypothesis that the robust finding of overconfidence in the literature arises not only from the item selection procedure (items that are not representative of a reference class) but also from the conversion of confidence estimates into probabilities. Gigerenzer and colleagues argued that there is a psychological distinction between assessing confidence concurrently and making a retrospective estimate of the percent of correct judgments.

The method that Gigerenzer et al. (1991) recommended for assessing calibration consisted of computing the difference between the retrospective estimate (in percent) and the actual percent of correct judgments for each subject. In Table 4-08, the means and standard deviations for judgment quality assessed by two different methods were reported. Method 1 was the difference between the mean of the concurrent confidence judgments and the percent of correct judgments. Method 2 was the difference between the retrospective estimate of the percent of correct judgments and the percent of correct judgments, as proposed by Gigerenzer and colleagues.
Table 4-08
Means and Standard Deviations for Two Methods of Assessing Judgment Quality for Normal and Reverse Order Cases, by Experience Level

<table>
<thead>
<tr>
<th>Experience Level</th>
<th>Normal Order Cases (n=40)</th>
<th>Reverse Order Cases (n=40)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Method 1&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Method 2&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>Means (SD)</td>
<td>Means (SD)</td>
</tr>
<tr>
<td>1 n=12</td>
<td>20.67 (19.89)</td>
<td>7.75 (21.55)</td>
</tr>
<tr>
<td>2 n=12</td>
<td>29.58 (14.49)</td>
<td>21.75 (16.89)</td>
</tr>
<tr>
<td>3 n=12</td>
<td>29.92 (15.61)</td>
<td>12.27 (18.99)</td>
</tr>
</tbody>
</table>

Note. SD = Standard Deviation. For both methods, high scores indicate overconfidence.

<sup>a</sup> Repeated cases were included in the assessment of judgment quality.

<sup>b</sup> Method 1 = Average of 40 concurrent confidence judgments (in percent) minus percent of correct judgments.

<sup>c</sup> Method 2 = Retrospective assessment of percent of correct judgments minus percent of correct judgments.

The pattern of subjects' scores revealed that there was a weak relationship between experience and judgment quality (\( \eta^2 = .10 \) and \( \eta^2 = .08 \) for method 1 and method 2, respectively). The retrospective method revealed a similar pattern with experience, but showed somewhat less overconfidence.

To demonstrate subjects' ability to discriminate their level of confidence, subjects' mean confidence was calculated separately for correct and incorrect judgments, and their confidence-accuracy correlations were computed. In Table 4-09, these means and standard deviations for confidence for correct and incorrect judgments and for confidence-accuracy correlations for normal and for reverse order cases are reported, by experience level.
Table 4-09
Means and Standard Deviations for Confidence Averaged Over Correct and Incorrect Judgments and for Confidence-Accuracy

Correlations for Normal Order and Reverse Order Cases, by Experience Level

| Experience Level | Normal Order Cases (n=40) | | | | Reverse Order Cases (n=40) |
|------------------|----------------------------|-----------------|----------------------------|----------------------------|
|                  | Correct Judgments (Means (SD)) | Incorrect Judgments (Means (SD)) | Confidence-Accuracy (Means (SD)) | Correct Judgments (Means (SD)) | Incorrect Judgments (Means (SD)) | Confidence-Accuracy (Means (SD)) |
| 1                | 68.78 (14.69) | 61.82 (14.37) | .21 (.19) | 62.27 (15.75) | 57.69 (18.04) | .11 (.14) |
| 2                | 70.36 (15.60) | 66.48 (15.63) | .21 (.19) | 60.16 (23.30) | 59.65 (22.72) | .10 (.24) |
| 3                | 77.53 (10.77) | 72.56 (13.93) | .20 (.22) | 71.16 (19.99) | 68.25 (19.88) | .11 (.26) |

Note. SD = Standard Deviation

*a* Repeated cases were included in the assessment of confidence.

*b* Confidence-Accuracy Correlations underwent a Fisher r to r' transformation prior to addition, then was transformed back to the correlational scale.
As described by Howell (1992), prior to addition of the correlations, a Fisher transformation ($r$ to $r'$) was carried out; the means were then transformed back to the correlational scale. The proportion of variance associated with the within-subject factor (correct and incorrect cases) was substantial ($\eta^2 = .51$ in the normal order cases and $\eta^2 = .25$ in the reverse order cases). It was anticipated that experienced subjects would be better able to differentiate (in terms of confidence) between correct and incorrect cases. The proportion of variance associated with the interaction between experience and the within-subject factor (confidence in correct and incorrect cases), however, was small ($\eta^2 = .06$ for normal order cases, and $\eta^2 = .12$ for reverse order cases), and in the opposite direction to what was anticipated. The mean judgment-accuracy correlations were low, particularly in the reverse order cases. The standard deviations revealed that variation in confidence-accuracy correlations was not well-captured by experience level.

**Judgment Latency**

For each subject, a preliminary average judgment latency was obtained for each set of cases by omitting the first five judgments from the average. Subjects were learning the task during these initial judgments, and the latency was elevated. Each subject's mean latency for that set was substituted for the latency as timed for these five judgments. The latency used in the analysis was an average for the 35 judgments. Latencies for repeated judgments were not included. The sequence in which the two sets of cases were presented made a difference to subjects' latency. Case sequence 1 was the presentation of the normal order cases first, followed by the reverse order cases. Case sequence 2 was the reverse. In Table 4-10, the means and standard deviations for average latency are reported for normal order and reverse order cases, by sequence condition and by experience level.
Table 4-10
Means and Standard Deviations for Latency (in seconds) Averaged Over Cases for Normal Order and Reverse Order Cases, by Case Sequence and by Experience Level

<table>
<thead>
<tr>
<th>Experience Level</th>
<th>Latency for Normal Order Cases (n=35)</th>
<th>Latency for Reverse Order Cases (n=35)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Case Sequence 1&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Case Sequence 2&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>Means (SD) n&lt;sup&gt;c&lt;/sup&gt;</td>
<td>Means (SD) n&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>1</td>
<td>28.14 (8.30) 7</td>
<td>24.00 (6.28) 5</td>
</tr>
<tr>
<td></td>
<td>n=12</td>
<td>n=5</td>
</tr>
<tr>
<td>2</td>
<td>22.83 (4.17) 6</td>
<td>21.83 (5.04) 6</td>
</tr>
<tr>
<td></td>
<td>n=12</td>
<td>n=6</td>
</tr>
<tr>
<td>3</td>
<td>22.60 (6.54) 5</td>
<td>19.00 (4.24) 7</td>
</tr>
<tr>
<td></td>
<td>n=12</td>
<td>n=5</td>
</tr>
</tbody>
</table>

Note. SD = Standard Deviation

<sup>a</sup> Case Sequence 1: Subjects completed normal order cases, then reverse order cases.

<sup>b</sup> Case Sequence 2: Subjects completed reverse order cases, then normal order cases.

<sup>c</sup> Number of Subjects in each Case Sequence.

Experience accounted for a small proportion of variance in latency ($\eta^2 = .16$).

The effect of sequence condition was complex: the direct effect was negligible, but the interaction between sequence condition and the within-subject factor (normal and reverse order cases) was considerable ($\eta^2 = .48$). Average judgment latency decreased somewhat for the set of cases presented second, particularly for subjects assigned to case sequence 1; the variance associated with the effect of the within-subject factor (latency for normal order and reverse order cases) was notable ($\eta^2 = .30$).
Knowledge Accessibility

The purpose of the slide task related to the Knowledge Accessibility scale is reported here; the second purpose (memory priming) is reported later in section D.

Subjects' tape recordings of verbalizations for three slides showing close-up views of abdominal incisions were transcribed and analyzed according to the framework shown in Appendix D. The categories included in this scale were perceptual knowledge, specific item knowledge, relational knowledge, and contextual knowledge. In Table 4-11, the means and standard deviations for the number of references from each sub-scale are reported, by experience level.

<table>
<thead>
<tr>
<th>Experience Level</th>
<th>Perceptual Knowledge (Means, SD)</th>
<th>Item Knowledge (Means, SD)</th>
<th>Relational Knowledge (Means, SD)</th>
<th>Contextual Knowledge (Means, SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.75 (1.22)</td>
<td>3.17 (1.64)</td>
<td>12.17 (4.57)</td>
<td>2.17 (1.80)</td>
</tr>
<tr>
<td>2</td>
<td>2.67 (1.50)</td>
<td>3.75 (1.48)</td>
<td>12.67 (3.47)</td>
<td>1.83 (1.59)</td>
</tr>
<tr>
<td>3</td>
<td>3.50 (1.93)</td>
<td>5.08 (2.07)</td>
<td>12.25 (4.14)</td>
<td>3.58 (2.47)</td>
</tr>
</tbody>
</table>

Note: SD = Standard Deviation

There was a pattern for knowledge accessibility to increase with experience, with the exception of relational knowledge. The proportion of variance accounted for by experience in the total scores was smaller than what was anticipated ($\eta^2 = .07$).
In summary of this section, intercorrelations for experience and performance measures that were considered as possible indicators of judgment expertise are reported in Tables 4-12(a) and 4-12(b).

Table 4-12(a)

**Intercorrelation Matrix for Experience and Indicators of Expertise Based on Performance on Normal Order Cases**

<table>
<thead>
<tr>
<th>Variables</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>--</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>.03</td>
<td>--</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.32</td>
<td>.53**</td>
<td>--</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.23</td>
<td>.14</td>
<td>.23</td>
<td>--</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>.08</td>
<td>.15</td>
<td>-.21</td>
<td>.72**</td>
<td>--</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>-.12</td>
<td>-.16</td>
<td>-.05</td>
<td>.03</td>
<td>-.09</td>
<td>--</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>.21</td>
<td>.25</td>
<td>.25</td>
<td>.01</td>
<td>.10</td>
<td>-.29</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>.16</td>
<td>.06</td>
<td>.13</td>
<td>.44</td>
<td>.30</td>
<td>-.10</td>
<td>.00</td>
<td>--</td>
</tr>
</tbody>
</table>

**Note.**

1 = Experience, in Months
2 = Accuracy, Error Variance Measure
3 = Accuracy, Percent Correct Measure
4 = Mean Confidence
5 = Calibration (Traditional Method)
6 = Mean Latency
7 = Knowledge Accessibility
8 = Consistency of Healing Time Judgments

Correlations were based on 36 subjects

**p = < .01**
Table 4-12(b)

**Intercorrelation Matrix for Experience and Indicators of Expertise Based on Performance on Reverse Order Case**

<table>
<thead>
<tr>
<th>Variables</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>2</td>
<td>-.03</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>3</td>
<td>.09</td>
<td>.63**</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>4</td>
<td>.24</td>
<td>-.11</td>
<td>-.01</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>5</td>
<td>.05</td>
<td>-.42*</td>
<td>-.53**</td>
<td>.68**</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>6</td>
<td>.01</td>
<td>.13</td>
<td>.11</td>
<td>.13</td>
<td>.04</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>7</td>
<td>.21</td>
<td>.12</td>
<td>.31</td>
<td>-.02</td>
<td>-.22</td>
<td>-.15</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>8</td>
<td>.20</td>
<td>-.01</td>
<td>.12</td>
<td>.27</td>
<td>.23</td>
<td>.12</td>
<td>.33</td>
<td>--</td>
</tr>
</tbody>
</table>

**Note.**

1 = Experience, in Months
2 = Accuracy, Error Variance Measure
3 = Accuracy, Percent Correct Measure
4 = Mean Confidence
5 = Calibration (Traditional Method)
6 = Mean Latency
7 = Knowledge Accessibility
8 = Consistency of Healing Time Judgments
Correlations were based on 36 subjects
* p < .05; ** p < .01.

Several patterns of relations, all of very small magnitude, were evident. In both sets of cases, confidence and experience were positively related. The correlation between experience and accuracy was slightly higher in the normal order cases than in the reverse order cases. Knowledge accessibility demonstrated a
weak relationship with judgment accuracy in both sets of cases. As anticipated, knowledge accessibility had a negative relationship with latency. Contrary to expectation, consistency of judgments of healing time was not related to judgment accuracy, and had little association with experience.

Section C: Results Related to Cognitive Constructs and Individual Differences

Research question 2 is:

What are the patterns of relationships among measures of conceptual structure, sensitivity to patterns in data, judgment process, and performance in a clinical judgment task?

Conceptual Structure

Two tasks were used to obtain data about conceptual structure. The initial one was an association task where subjects were encouraged to talk about their experiences with surgical patients; the purpose was to identify the way in which subjects verbalized their experiences, prior to being influenced by the other tasks included in the study.

In the second task, subjects made similarity judgments for pairs of concepts. As a means to reveal information about the organization of concepts, multidimensional scaling (MDS) was carried out on the similarity judgments.

Association task. The link between conceptual coherence and conceptual structure has been documented by Medin (1989), Medin and Edelson (1988), and Murphy and Medin (1985). Subjects were tape-recorded as they responded to three brief case situations. The verbalizations were transcribed and coded according to
the framework entitled Conceptual Coherence scale available in Appendix E. This scale has four sub-scales that were labeled as follows: conceptual familiarity, conceptual boundaries, conceptual links, and conceptual structure. For three of these sub-scales, high scores represented greater coherence, and for one sub-scale (subjects' references to lack of knowledge or experience), high scores reflected less coherence. For ease in interpretation and to achieve consistency with the other sub-scales, the polarity of this latter sub-scale was reversed.

In Table 4-13, the means and standard deviations for the number of references coded in each category, by experience level, are displayed.

<table>
<thead>
<tr>
<th>Experience Level</th>
<th>Conceptual Coherence Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Conceptual Familiarity</td>
</tr>
<tr>
<td></td>
<td>Means (SD)</td>
</tr>
<tr>
<td>1 n=12</td>
<td>9.58 (3.78)</td>
</tr>
<tr>
<td>2 n=12</td>
<td>11.67 (6.05)</td>
</tr>
<tr>
<td>3 n=12</td>
<td>16.50 (6.59)</td>
</tr>
</tbody>
</table>

Note.

SD = Standard Deviation.

* Scores on this subscale have been transformed; high scores indicate greater coherence.
For three of the measures of conceptual coherence derived from the protocol data (conceptual familiarity, conceptual connections, and conceptual structure), the effect of experience was notable. The second sub-scale (conceptual boundaries) showed a only a slight increase with experience. Using subjects' total scores for this scale, experience level accounted for a substantial proportion of variance in conceptual coherence ($\eta^2 = .46$).

**Concept similarity judgment task.** To determine how subjects structured particular concepts, MDS was carried out using similarity judgments for all unique pairs of 16 concepts selected for study. Gonzalvo, Cañas, and Bajo (1994) found that MDS captured the global changes in structural representations of knowledge with learning. Wilkinson, Gimbel, and Koepke (1982) used MDS to graph symptoms of illness. Studies carried out by these groups of researchers provided rationales and guidelines for this method. Each subject's conceptual structure (as revealed through MDS of concepts relevant to healing) was examined.

The pairs of concepts for this similarity judgment task were displayed by personal computer. Subjects entered numbers from 0 to 9 expressing judgments of similarity. Using the SPSS Alscal program, data were analyzed at the ordinal level, with the program allowed to "untie" ties. Basically, the aim of MDS is to optimize the fit between any configuration and the data; the stress of the best-fitting configuration is a measure of goodness of fit (Kruskal, 1964). Mathematically, the stress is the square root of the normalized residual sum of squares. Small values of stress indicate a good fit.

The Alscal algorithm was used to compute a fit measure called S-Stress, which has two variations, S-Stress formula 1, and S-Stress formula 2. These formulae are available in Davison (1983). As recommended by Kruskal and Wish
(1978) and Davison (1983), S-Stress formula 1 was used because the data were similarities rather than preferences.

Each subject's similarity matrix was multidimensionally scaled using the unweighted Euclidian model. The unweighted model is the most commonly used spatial distance model, and was selected because the concepts were considered to be unitary; Weinberg (1991) recommended a Euclidian metric for such stimuli. Other models are more suited to stimuli that can be readily decomposed into constituent attributes, for example, for separable features such as size and shape.

To make the decision of dimensionality, several steps were followed. First, solutions were derived in one, two, three, and four dimensions. One dimension was included because Shepard (1974) pointed out that there was a tendency for investigators using MDS to extract more dimensions than were necessary. In this study, the assessment of four dimensions was considered the upper limit; this decision was based on the criterion proposed by Kruskal and Wish (1978) that the ratio of the number of stimuli to the number of dimensions should be at least 4:1.

For each subject, the stress was plotted as a function of dimensionality. If error were minimal, plots should show a bend (or "elbow") at a dimensionality that is considered to be the maximum (Schiffman et al., 1981). A slight bend either at two or three dimensions was observed in the plots for 10 subjects; no bend was noted for the other subjects.

The next step was to assess the $R^2$ values (proportion of variance of the scaled data which was accounted for by the MDS Model distances). $R^2$ was plotted as a function of dimension, as recommended by Young and Hamer (1987). A sudden rise in $R^2$ suggests a good fit of the MDS model to the data. In this study,
only a few subjects' plots showed such a rise. In Appendix F, the stress and $R^2$ for one to four dimensions are reported for all subjects.

The third method used to determine dimensionality was to compare the stress values obtained at each dimension to figures computed through Monte Carlo studies from Spence and Ogilvie (1973) and Spence (1979). All subjects' MDS solutions in two and three dimensions had stress values low enough to meet these criteria. Kruskal and Wish (1978) cautioned against accepting solutions with stress values above .1, unless the solution were in one-dimension. Using this strict criterion, only 3 subjects had two-dimensional solutions with sufficiently low stress values; when three dimensions were used, this number increased to 14.

Davison (1983), however, advised that if data contained considerable error, stress values over .1 may be accepted. Departure from monotonicity could arise from the fact that the stimuli used in this study were mental concepts pertaining to a particular field of knowledge; there is no definitively correct way to express the interrelationships among the concepts. Many MDS studies use objects with varying colour, hue and intensity, or physical stimuli such as Morse Code symbols, where objective, measurable differences among the features exist. It is reasonable to consider that the error would be less in these latter situations. In addition, Shepard (1974) expressed the opinion that investigators often place too much emphasis on numerical indicators of departure from monotonicity (stress) to the exclusion of more important considerations such as substantive interpretability of the derived configurations.

The final approach to the question of dimensionality was to assess whether the configurations were interpretable. Because two-dimensional solutions are much more easily interpreted, and because some subjects were identified as very likely to
have solutions in only two dimensions (based on stress measures and $R^2$), two dimensional representation of the concepts were obtained for each subject, and compared. They demonstrated a clear dimension that represented a continuum that could be labelled *healthy-unhealthy differentiation*. A second dimension appeared to be *phase in the healing process*.

The plots were analyzed for patterns that were predicted from the literature. One cluster of three concept-pairs was expected to become more differentiated with experience. Another cluster was anticipated to become closer together as subjects attained greater associative connection. A final pair was conceptualized to be related by the phase of the healing process (early to late). In Table 4-14, the means and standard deviations for the average distances for the concept-pairs comprising the clusters are shown, by experience level. The measures of the distance between each of the constituent concept-pairs were taken from the two-dimensional plots.

**Table 4-14**

*Means and Standard Deviations for Average Distance Between Concept Pairs in Selected Clusters*, by Experience Level

<table>
<thead>
<tr>
<th>Experience Level</th>
<th>Selected Concept Clusters</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Differentiation</td>
<td>Phase of Healing</td>
<td>Association</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Means (SD)</td>
<td>Means (SD)</td>
<td>Means (SD)</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>8.33 (1.86)</td>
<td>6.50 (3.11)</td>
<td>6.17 (2.55)</td>
<td></td>
</tr>
<tr>
<td>n=12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>8.34 (2.73)</td>
<td>4.71 (2.62)</td>
<td>5.23 (2.91)</td>
<td></td>
</tr>
<tr>
<td>n=12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>9.44 (2.06)</td>
<td>6.17 (2.48)</td>
<td>4.38 (1.64)</td>
<td></td>
</tr>
<tr>
<td>n=12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* SD = Standard Deviations. Thirty-six plots (one per subject) were analyzed. *The measures, in centimeters, were taken from two-dimensional plots derived from multidimensional scaling of similarity judgments.*
With Differentiation and Association concepts, there were very little difference in mean scores, although the differences were in the predicted direction. With Phase of Healing, the middle group had the smallest distances which did not conform to any pattern.

Schvaneveldt et al. (1985), in their research using MDS to reveal the conceptual structure of pilots, found that an analysis of all subjects' weighting of conceptual dimensions demonstrated an expertise dimension. Such an analysis was carried out in the present study, using the Indscale procedure available in SPSS. The axes were the two dimensions found relevant in the individual analyses. Coordinate points for all subjects were identified. The resulting plot based on derived subject weights showed a negative linear trend, with a wider dispersion to the left of center. Points representing subjects with varying experience levels appeared to be intermixed. Several points representing subjects with high accuracy scores were located to the lower right side of the plot. No clear expertise dimension, however, emerged.

**Sensitivity to Overall Patterns in Data**

To assess subjects' differential sensitivity to data patterns, Brunswikian lens model methodology was applied. In phase 1 of the study, the broad patterns related to incisional healing in the ecology were identified. In phase 2, the extent to which subjects of varying experience levels matched these patterns was revealed.

**Lens model task.** As described in chapter 3, subjects made judgments of healing time for two sets of 40 cases. Analysis was based on 35 cases; repeated cases were not used.
Traditionally in the lens model paradigm, judgment has been considered an *intra-individual* phenomenon captured in isolation from other individuals (K. R. Hammond et al., 1975). To be consistent with tradition, the examples of Carlström (1989) and Cooksey and Freebody (1987) have been followed: non-parametric indices are reported to provide summarization, and yet preserve the idiographic character of the data. Medians and interquartile ranges for performance on normal order and reverse order cases are presented in Table 4-15.

Considering the level of task predictability, subjects' performance was good in both the normal order and reverse order cases. For normal order cases, 13 subjects attained (or exceeded) the model $R^2$ of .70: 5 subjects from experience level 1, 3 from experience level 2, and 5 from experience level 3.

For reverse order cases, 12 subjects attained (or exceeded) the model $R^2$ of .68: 5 subjects from experience level 1, 2 from experience level 2, and 5 from experience level 3. Using the lens model measures as indicators of sensitivity to broad patterns in the ecology, no patterns emerged for increased sensitivity with experience. For both lens model achievement and G-Scores, the medians for the middle group were lower, compared to the other two groups. The median variance explained ($R^2$) and median G-Scores for the reverse order cases increased somewhat with experience, but the other lens model performance measures demonstrated no pattern.
<table>
<thead>
<tr>
<th>Experience Level</th>
<th>Performance Measures for Normal Order Cases (n=35)</th>
<th>Performance Measures for Reverse Order Cases (n=35)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Achievement (r)</td>
<td>Variance Explained ($R^2$)</td>
</tr>
<tr>
<td>Level</td>
<td>Normal Order</td>
<td>Reverse Order</td>
</tr>
<tr>
<td>n=12</td>
<td>.59 (.16)</td>
<td>.69 (.15)</td>
</tr>
<tr>
<td>n=12</td>
<td>.53 (.19)</td>
<td>.72 (.26)</td>
</tr>
<tr>
<td>n=12</td>
<td>.60 (.19)</td>
<td>.69 (.17)</td>
</tr>
</tbody>
</table>

Note. Mdn = Median; IQR = Interquartile Range
In Table 4-16, means and standard deviations for $R^2$ (variance explained) and for the matching index, $\Sigma-d$, (the degree to which subjects matched ecological validities with utilization coefficients) are reported, by experience level.

Table 4-16

Means and Standard Deviations for Measures derived from Lens Model Analyses for Normal Order and Reverse Order Cases, by Experience Level

<table>
<thead>
<tr>
<th>Experience Level</th>
<th>Normal Order Cases (n=35)</th>
<th>Reverse Order Cases (n=35)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R^2$ Means (SD)</td>
<td>$\Sigma d$ Means (SD)</td>
</tr>
<tr>
<td>1, n=12</td>
<td>.63 (.14)</td>
<td>.32 (.15)</td>
</tr>
<tr>
<td>2, n=12</td>
<td>.63 (.13)</td>
<td>.40 (.22)</td>
</tr>
<tr>
<td>3, n=12</td>
<td>.63 (.11)</td>
<td>.27 (.12)</td>
</tr>
</tbody>
</table>

Note. SD = Standard Deviation

The means for $R^2$ increased with experience in the reverse order cases, but did not increase in the normal order cases. The $\Sigma-d$ was slightly lower (more ideal) in the reverse order cases. Hammond et al. (1964) expressed a caution about the interpretation of this matching Index. Although it may be tempting to examine each difference product associated with each cue, and make inferences about the relative
contribution of each cue to $\Sigma - d$, this would not be wise. Hammond and colleagues explained that $\Sigma - d$ is not a simple sum of the mismatches, but is rather a result of interdependent difference products, or vectors. Subjects' scores for lens model measures ($r_a$, $R_s^2$, G, C, and $\Sigma - d$) and accuracy measures (proportion correct and error variance) for normal and reverse order cases are available in Appendix G.

Subject's beta weights for significant variables are available in Appendix H. No pattern was evident in relation to the beta weights and experience.

**Cue importance task.** This task was used to obtain subjective policies for the judgment of healing time. Ten new cases from phase 1 of this study, with a range of healing times, were selected for this task. The subjective weights of the cues were obtained by having subjects read case information, make a judgment of healing time, and rate the importance of each cue for each case, as described in chapter 3. To convert these cue-weights into a policy, the following procedure was followed.

First, for continuous variables, to make subjects' cue-weights comparable between cases, a variable called Caseweight was created by summing the raw cue-weights for each case; raw cue-weights were converted to relative cue-weights by dividing by Caseweight. Second, relative cue-weights were multiplied by the judged healing time to reflect the relationship with the subject's judgment of the criterion, then divided by the average value for that cue. Division was necessary so that the subjective weight could be multiplied by the value of the appropriate cue for a new case. The intermediate results were referred to as adjusted weights; these weights were components of the judged criterion in a proportion derived by the subjective weighting of cues. Finally, for each subject, the adjusted weights were averaged over the 10 cases to obtain contextual weights.
The categorical variables were treated in the same way, except that the adjusted weights were calculated only from cases coded as 1 for that variable.

The policy based on subjective weights consisted of integrating these contextual weights for both continuous and categorical variables in an additive manner.

When applying the policy derived from subjective weights to predict healing time, the polarity of the variable Occupation (Blishen et al., 1987) was reversed so that lower numbers reflected higher socioeconomic status, rather than the opposite. Einhorn and Hogarth (1975) pointed out that when predicting a criterion using policies based on subjective weights, cues must to be scaled so that they have a consistent relationship to the criterion.

The quality of the subjective policy was compared with the quality of the two regression policies. The three policies from each subject were used to predict healing times for the new set of 35 cross-validation cases. Experienced subjects with considerable schematic knowledge of the ecology were theorized to have greater accuracy on cross-validation with subjective policies, compared to inexperienced subjects.

The predictions from the three policies were compared to the actual healing time obtained from phase 1 data, using the interval measure of healing time, as before. In Table 4-17 data are presented for accuracy of cross-validation, using accuracy based on error variance and percent correct measure, by experience level.
Table 4-17

Means and Standard Deviations for Accuracy of Cross-Validation* of Judgment Policies (Based on Error Variance Measures**) and Percent Correct Measures, by Experience Level

<table>
<thead>
<tr>
<th>Experience Level</th>
<th>Normal Order Cases</th>
<th>Reverse Order Cases</th>
<th>Subjective Policy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Error Variance</td>
<td>% Correct</td>
<td>Error Variance</td>
</tr>
<tr>
<td></td>
<td>Means (SD)</td>
<td></td>
<td>Means (SD)</td>
</tr>
<tr>
<td>1 n=12</td>
<td>86.68 (5.75)</td>
<td>43.58 (18.19)</td>
<td>87.76 (2.05)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>89.15 (2.23)</td>
</tr>
<tr>
<td>2 n=12</td>
<td>85.29 (7.30)</td>
<td>38.80 (16.02)</td>
<td>86.97 (4.07)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>88.55 (2.97)</td>
</tr>
<tr>
<td>3 n=12</td>
<td>88.00 (2.42)</td>
<td>38.10 (10.27)</td>
<td>88.18 (2.04)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>88.62 (2.46)</td>
</tr>
</tbody>
</table>

Note. *Cross-Validation was assessed by 35 new cases from phase-one data.

**Accuracy based on error variance was transformed; high scores indicate high accuracy. SD = Standard Deviation.
All policies cross-validated on the new set of 35 cases moderately well, although there was little relationship with experience. Experience accounted for a negligible proportion of variance using the error variance measure. Using the proportion correct measure, a very small proportion of variance was explained ($\eta^2 = .05$ for the analysis of the policy derived from regression on normal order cases and the subjective policy; $\eta^2 = .02$ for the comparison of the policy derived from regression on reverse order cases and the subjective policy).

The effect of experience was in the direction predicted only with the regression-based policy derived from reverse order cases. Contrary to expectation, the means for accuracy for cross-validation were highest for novices' subjective policies and for this group's regression policies based on the normal order cases.

**Judgment Process**

Judgment process was assessed by two tasks. The first one was the information board task where subjects engaged in information seeking and thinking aloud; these activities provided information about the sequence, amount, and variability of data selected. Analysis of verbal protocols provided evidence about the ways in which subjects at each experience level processed information and used strategies.

The second task used to assess judgment process was the orthogonal judgment task. Measures of the subjects' ability to discriminate healing times, and a measure of configural cue-use that was considered alternative to the lens model C-scores were derived from this task.

**Information board task.** Analysis of the information board task was similar to that described by Payne (1976). The sequence of information search was
recorded on a process tracing map. The pattern of search associated with selecting
the first eight cards, the second eight, and the remainder, was classified as *inter-
dimensional* (search proceeded over the various cue-dimensions, within the same
case), *intra-dimensional* (search proceeded within the same cue-dimension, over the
set of cases), or mixed (elements of both strategies present). The depth of search
was defined as the total number of cards turned over.

A numeric index of search pattern was computed, using the method
developed by Payne (1976). This index was defined as the number of alternative-
wise moves minus the number of attribute-wise moves, divided by the sum of these
two numbers. The resulting index has a range from -1, indicating purely
*intradimensional*, (or between-alternative) search, to +1, indicating purely
*interdimensional* (or within-alternative) search. Shifts were defined as moves that
were neither alternative-wise or attribute-wise; such moves were counted.

The variability of search was indexed by the standard deviation of the
proportion of cards examined *within* the set of alternatives. Harte et al. (1994) and
Zakay (1990) described that searching a constant number of dimensions for each
alternative indicated a compensatory strategy, whereas searching a variable number
of dimensions indicated a noncompensatory strategy. Using this definition, almost
all subjects in this study used a compensatory strategy.

The depth of search ranged from 15 to 24. Twenty-eight subjects responded
accurately and 8 did not. About two thirds of the subjects began the task with an
*intra-dimensional* strategy. No relationship between strategy-use and accuracy was
evident. Detailed data about information search are reported in Appendix I. The
means and standard deviations for data related to performance on the information
board task are available in Table 4-18, by experience level.
### Table 4-18

**Means and Standard Deviations for Data Search Measures on the Information Board**

**Task and Number of Subjects with Correct Judgments, by Experience Level**

<table>
<thead>
<tr>
<th>Experience Level</th>
<th>Judgment Performance</th>
<th>Number of Subjects with Correct Judgments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Depth of Search</td>
<td>Number of Shifts</td>
</tr>
<tr>
<td></td>
<td>Means (SD)</td>
<td>Means (SD)</td>
</tr>
<tr>
<td>1 n=12</td>
<td>21.83 (2.44)</td>
<td>5.25 (2.22)</td>
</tr>
<tr>
<td>2 n=12</td>
<td>20.33 (2.35)</td>
<td>3.92 (1.16)</td>
</tr>
<tr>
<td>3 n=12</td>
<td>19.67 (3.45)</td>
<td>3.83 (1.47)</td>
</tr>
</tbody>
</table>

Note. SD = Standard Deviation.

<sup>a</sup> Search Pattern: Positive scores reflect more within-alternative (or interdimensional) search; negative scores reflect more between-alternative (or intradimensional) search.

There was a pattern for experience to account for a small proportion of variance in Depth of Search ($\eta^2 = .10$); the relationship was inverse, in that as experience increased, there was a small decrease in Depth of Search. The percent of subjects with correct judgments in each group increased with experience.

Subjects varied with respect to the degree to which they were systematic in searching for the data necessary to make their judgments. The most interesting subjects were those who initially applied a systematic, top-down strategy, and then became sensitive to particular data, and allowed bottom-up processing to interact in a reciprocal way. Such an approach demonstrated subjects' adaptiveness to the unique demands in the situation.
Analysis of the verbal protocols was carried out according to Payne (1976), Payne et al. (1993a, 1993b) and Ericsson and Simon (1993). The protocols and process tracing maps were assessed for the number of instances of 18 features of information processing. Features were categorized into a Judgment Process scale, with three sub-scales: metacognition, or executive functioning (including planning, identifying data as missing or not required, and indicating curiosity about particular data); input generation (including constructing hypotheses or possibilities in working memory, making inferences, and perceiving clinical data as a particular pattern); and cognitive operations (including connecting two or more pieces of data in working memory, and weighing the relative importance of data). The Judgment Process scale, including 18 coding categories and examples from the protocols, is available in Appendix J.

Information processing profiles were constructed for each subject. In Table 4-19, the means and standard deviations for the number of references associated with each sub-scale of the Judgment Process scale are presented, by experience level.

Table 4-19

<table>
<thead>
<tr>
<th>Experience Level</th>
<th>Metacognition Means (SD)</th>
<th>Input Generation Means (SD)</th>
<th>Cognitive Operations Means (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.42 (4.66)</td>
<td>4.17 (3.04)</td>
<td>5.83 (3.79)</td>
</tr>
<tr>
<td>n=12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>5.83 (4.15)</td>
<td>4.00 (2.30)</td>
<td>5.33 (2.90)</td>
</tr>
<tr>
<td>n=12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>10.17 (5.46)</td>
<td>8.42 (5.07)</td>
<td>7.42 (3.23)</td>
</tr>
<tr>
<td>n=12</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. SD = Standard Deviation.
There was a clear pattern evident for mean scores on each of the sub-scales to increase with experience. Using scores for the total Judgment Process scale, experience accounted for a notable proportion of variance in this measure ($\eta^2 = .21$). The increases in scores were associated primarily with the most experienced group.

To assess the degree of consistency in data coding, as recommended by Harte et al. (1994) and described by Fleis (1971), Cohen's Kappa was computed. The use of Cohen's Kappa is a means for quantifying inter-rater agreement that is greater than that which is likely to occur by chance. A second person re-coded three of the protocols derived from the information board task. Cohen's Kappa was .72.

Because not all aspects of the verbal protocols could be captured in a numerical form, qualitative analysis of the verbal protocol data was carried out. In this section, some of the strategies that subjects used to search for and process information have been identified. Examples of how subjects weighed and integrated information as they made judgments about the data are presented.

One of the strategies used was generalization based on category membership. The following are examples:

Subject 28: "This client here has ulcerative colitis ... just because ulcerative colitis patients always have a lot of problems, it's going to be long, ongoing problems with this guy."

Subject 20: "I always find out age first, and it helps me to judge how long they're going to take to heal because younger people heal a lot faster than older people."

Another similar strategy used was making backward inferences, based on data available. Examples include:

Subject 35: "The fellow with the ulcerative colitis ... He might have been on prednisone ... so he would have impaired healing."
Subject 28: "This patient with the lower abdominal transverse incision, with a supra-pubic catheter, . . . that's probably been a Burch repair.

A third strategy was making knowledge available by deductive logic, based on the cues already in view. Examples include the following:

Subject 25: "I can see that I have at least two women here, and so I don't need to turn over the age and gender on these bottom two". [This was said after turning over the cue cards for surgery; the two cards to which this subject referred were oophorectomy and Burch repair].

Subject 25: "I can tell by the surgeries what their diagnoses are."

The experienced subjects seemed to have a better idea of the relevance of particular data, compared to novice subjects. For example, when weighing the impact of particular problems and integrating data, an experienced subject [28] was highly aware of the relationship of the factors to healing time, whereas a novice subject [05] acknowledged the data, but did not seem able to integrate the data in relation to a goal:

Subject 28: "This patient with the Burch repair has stomach hyperacidity and hypertension -- she may have a lot of problems voiding, getting her catheter out, and may need to go home on home care. But in terms of the incision, she's going to be fine. There's no problems there."

Subject 05: "The patient in for a Burch repair who has hypertension and stomach hyperacidity. . . . Hm, . . . She is probably a little anxious. . . . It would be really hard to say why they would do this on an 80 year old. I don't know. . . ."

One difficulty subjects experienced was in ignoring data. Some subjects realized that particular data were not relevant to the judgments, and yet the influence was evident in the judgment made. For example:

Subject 12: "I'll take the vital signs, just to compare. . . . The temperature is getting up there a little. Well, this is on admission so this, Um. . . . I don't know if that would tell me too much about the way they'd heal because I've seen vital signs change so much while people are in the hospital. . . ."
This inexperienced subject subsequently did not give the correct rank order for healing times for the four patients, and gave the rationale that the one particular patient's vital signs "were a little off".

The protocol data provided a useful means to illustrate the clinical reasoning employed by subjects who have a range of experience.

**Orthogonal-cue judgment task.** There were two goals for this task: (1) To identify to what extent subjects make discriminations in healing time; and (2) to determine the proportions of variance associated with the main effects (Age, Weight, and Diabetic status) and with the interaction terms. Subjects' calibration of healing times was assessed by discrimination (the number of categories that subjects created when sorting the cards of patient data), and by estimates of rapid and delayed healing times.

The orthogonal-cue judgment task was analyzed for configural cue-utilization using ANOVA, following the procedure described by Hoffman et al. (1968) and Phelps and Shanteau (1978). The four cues used in this task were regarded as fixed categorical factors, each with a number of levels: Age had two levels, Diabetes had two levels, Height had three levels and Weight had four levels. The dependent variable was the judgment made for each case. The cases were constructed from an orthogonal combination of cues in a completely crossed factorial design. A separate Analysis of Variance (ANOVA) was carried out for each subject. The proportion of variance in the dependent measure explained by interactions represented the patterned use of cues. Omega-squared ($\omega^2$) was used to achieve the goals related to variance explained; this measure was computed as recommended by Slovic (1969). The equations for $\omega^2$ for main effects and interaction terms for a fixed effects model were taken from Howell (1992, p. 407).
In Table 4-20, the means and standard deviations for these information processing measures and for $\omega^2$ are displayed by experience level.

There was a pattern of relations evident with the information processing data. Experience accounted for a small proportion of the variance in the number of categories generated ($\eta^2 = .10$). Assessment of the mean scores revealed that this relationship was inverse: experienced subjects tended to generate fewer categories, compared to subjects with less experience. Experience accounted for a very small proportion of variance in subjects' estimates of low healing times ($\eta^2 = .05$), and a notable proportion of variance in the high healing times ($\eta^2 = .26$). Experience accounted for a negligible proportion of variance in $\omega^2$ for interaction ($\eta^2 = .02$); the slight relationship observed, was in the direction opposite to what was anticipated. There was variability in these mean scores, but only a small part of it was captured by experience.
### Table 4.20

**Means and Standard Deviations for Measures of Calibration of Healing Time and Omega-Squared from the Orthogonal Judgment Task, by Experience Level**

<table>
<thead>
<tr>
<th>Experience Level</th>
<th>Measures of Calibration</th>
<th>Omega-Squared&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Discrimination&lt;sup&gt;a&lt;/sup&gt; of H-Time&lt;sup&gt;c&lt;/sup&gt;</td>
<td>Low Estimate of H-Time&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>Means (SD)</td>
<td>Means (SD)</td>
</tr>
<tr>
<td>1 n=12</td>
<td>8.33 (3.01)</td>
<td>9.25 (4.99)</td>
</tr>
<tr>
<td>2 n=12</td>
<td>6.33 (2.39)</td>
<td>9.67 (3.17)</td>
</tr>
<tr>
<td>3 n=12</td>
<td>6.67 (2.61)</td>
<td>7.75 (2.49)</td>
</tr>
</tbody>
</table>

**Note.**

SD = Standard Deviation.

<sup>a</sup> Discrimination was assessed by the number of categories generated.

<sup>b</sup> Measures of Omega-Squared were converted to percent of variance explained.

<sup>c</sup> H-Time = Healing Time.
Three additional measures of configural cue-use available were C-scores from the lens model analyses, the variability of data search from the information board task, and variability of the measure of the importance of information from the cue importance task. The means and standard deviations for these indices of configural cue-use are available in Table 4-21. Because the C-scores were correlations, prior to addition, a Fisher's transformation (r to r') was applied; the mean scores were transformed back to the correlational scale for reporting.

Table 4-21
Means and Standard Deviations for Measures of Configural Cue-Use, by Experience Level

<table>
<thead>
<tr>
<th>Experience Level</th>
<th>C-Scores(^a)</th>
<th>Configural Index</th>
<th>Configural Index</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Normal Order Cases (n=35)</td>
<td>Reverse Order Cases (n=35)</td>
<td>Variability of Information Search</td>
</tr>
<tr>
<td>1 (n=12)</td>
<td>.02 (SD=.27)</td>
<td>.43 (SD=.28)</td>
<td>.95 (SD=.62)</td>
</tr>
<tr>
<td>2 (n=12)</td>
<td>-.04 (SD=.26)</td>
<td>.43 (SD=.40)</td>
<td>.86 (SD=.41)</td>
</tr>
<tr>
<td>3 (n=12)</td>
<td>.04 (SD=.22)</td>
<td>.38 (SD=.23)</td>
<td>.62 (SD=.76)</td>
</tr>
</tbody>
</table>

Note. SD = Standard Deviation.

\(^a\) C-Scores were transformed using a Fisher r to r' transformation prior to the calculation of means, then transformed back to correlations.

In contrast to expectations, the variability of information search decreased with experience \(\eta^2 = .05\), as did the variability of cue-importance \(\eta^2 = .08\). The means for C-Scores were considerably larger in the reverse order cases compared to that in the normal order cases.
In Table 4-22(a) the correlations among the four measures of configural cue-use, accuracy, and experience for normal order cases are displayed.

Table 4-22(a)

**Intercorrelations Among Measures of Configural Cue-Use, Accuracy Measures, and Experience, for Normal Order Cases**

<table>
<thead>
<tr>
<th>Variables</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>--</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>.03</td>
<td>--</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.17</td>
<td>-.03</td>
<td>--</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.01</td>
<td>-.04</td>
<td>-.03</td>
<td>--</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-.24</td>
<td>.16</td>
<td>.09</td>
<td>.24</td>
<td>--</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>.17</td>
<td>.15</td>
<td>.08</td>
<td>.12</td>
<td>.53**</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>-.16</td>
<td>.07</td>
<td>-.08</td>
<td>.08</td>
<td>.03</td>
<td>.32</td>
<td>--</td>
</tr>
</tbody>
</table>

**Note.**

1 = Omega-Squared for Interaction Terms

2 = C-scores from Lens Model Task

3 = Variability of Information Search

4 = Variability of Cue Importance

5 = Accuracy, Error Variance Measure

6 = Accuracy, Percent Correct Measure

7 = Experience, in Months

Correlations were based on 36 subjects.

**p < .01**
In Table 4-22(b) the correlations among the four measures of configural cue-use, accuracy, and experience for reverse order cases are reported.

Table 4-22(b)

<table>
<thead>
<tr>
<th>Variables</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>--</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-.07</td>
<td></td>
<td>-07</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.17</td>
<td>.07</td>
<td></td>
<td>-03</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.01</td>
<td>.01</td>
<td>.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-.34*</td>
<td>.27</td>
<td>.03</td>
<td>.06</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>-.07</td>
<td>-.13</td>
<td>.08</td>
<td>.28</td>
<td>.63**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>-.16</td>
<td>-.16</td>
<td>-.08</td>
<td>.08</td>
<td>-.03</td>
<td>.09</td>
<td></td>
</tr>
</tbody>
</table>

Note.
1 = Omega-Squared for Interaction Terms
2 = C-scores from Lens Model Task
3 = Variability of Information Search
4 = Variability of Cue Importance
5 = Accuracy, Error Variance Measure
6 = Accuracy, Percent Correct Measure
7 = Experience, in Months
Correlations were based on 36 subjects.
* p < .05.  ** p < .01.
With the exception of the two accuracy measures, these correlations tended to be very small, as anticipated with a sample size of 36. A negative relationship was observed between $\omega^2$ for the interaction term and accuracy assessed by error variance; this was not the direction predicted. The variability of cue importance and accuracy assessed by error variance demonstrated a positive relationship in the normal order cases, but not in the reverse order cases. The variability of cue importance and accuracy assessed by the percent correct measure was positively related in the reverse order cases.

**Age, Education and Experience**

Research question 3 is:

What are the patterns of relationships among individual differences in age, education, and experience and performance in a clinical judgment task?

The description of subjects with respect to age category and education was reported in Section B. One subject was over 55 by a month; to prevent this one score from having undue influence on correlations, this subject's age was adjusted to the immediately preceding age category. Because the subjects' precise ages were not known, each age was approximated by using the midpoint of the scale from the questionnaire. Thus, the ages used in the analysis became 20, 30, 40, and 50.

Experience was assessed by the number of months working in a surgical setting. Experience was also indexed from the estimate of the total number of incisions viewed during the subject's career. A further means of tapping experience was extracted from qualitative data (verbal references to experience lacking from the protocols). This measure was linearly transformed so that low scores indicated little experience.
The intercorrelations between these experience measures, age, education, and accuracy for normal order cases and reverse order cases are provided in Tables 4-23(a) and 4-23(b).

Table 4-23(a)

Intercorrelations Between Experience Measures, Age, Education, and Judgment Accuracy for Normal Order Cases

<table>
<thead>
<tr>
<th>Variables</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>.87**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.14</td>
<td>.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.62**</td>
<td>.51**</td>
<td>.35*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-.46**</td>
<td>-.36*</td>
<td>-.02</td>
<td>-.20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>.34*</td>
<td>.38*</td>
<td>.00</td>
<td>.34*</td>
<td>.06</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>.03</td>
<td>-.01</td>
<td>-.06</td>
<td>-.14</td>
<td>-.19</td>
<td>.18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>.32</td>
<td>.29</td>
<td>-.11</td>
<td>.17</td>
<td>-.22</td>
<td>.13</td>
<td>.53**</td>
<td></td>
</tr>
</tbody>
</table>

Note.
1 = Experience, In Months
2 = Experience Expressed as Number of Incisions Seen
3 = Experience, Protocol Measure
4 = Age
5 = General Education
6 = Nursing Education
7 = Accuracy, Error Variance
8 = Accuracy, Percent Correct
Correlations were based on 36 subjects.
* p < .05. ** p < .01.
### Table 4-23(b)

**Intercorrelations Between Experience Measures, Age, Education, and Judgment Accuracy for Reverse Order Cases**

<table>
<thead>
<tr>
<th>Variables</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>--</td>
<td>2</td>
<td>.87**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>--</td>
<td>.14</td>
<td>.03</td>
<td>--</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>.14</td>
<td>--</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>.14</td>
<td>.62**</td>
<td>.51**</td>
<td>.35*</td>
<td>--</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>.14</td>
<td>.62**</td>
<td>-.36*</td>
<td>-.02</td>
<td>-.20</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>.14</td>
<td>.62**</td>
<td>-.36*</td>
<td>.35*</td>
<td>.00</td>
<td>.34*</td>
<td>.06</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>.14</td>
<td>.62**</td>
<td>-.36*</td>
<td>-.02</td>
<td>.00</td>
<td>-.15</td>
<td>.03</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>.14</td>
<td>.62**</td>
<td>-.36*</td>
<td>-.02</td>
<td>-.23</td>
<td>-.20</td>
<td>.03</td>
</tr>
</tbody>
</table>

**Note.**

1 = Experience, In Months  
2 = Experience Expressed as Number of Incisions Seen  
3 = Experience, Protocol Measure  
4 = Age  
5 = General Education  
6 = Nursing Education  
7 = Accuracy, Error Variance  
8 = Accuracy, Percent Correct  
Correlations were based on 36 subjects.  
* *p < .05.  **p < .01.

Experience was strongly related to the number of incisions seen, nursing education, and to age. The percent correct measure of accuracy had a weak association with experience in the normal order cases.
The correlations between the cognitive measures, experience assessed in a variety of ways, and measures of accuracy are reported in Table 4-24.

Table 4-24

| Correlation Coefficients Between Cognitive Measures and Measures of Judgment Accuracy for Normal Order and Reverse Order Cases |
|---|---|---|---|---|
| Correlation Coefficients for Judgment Accuracy | Normal Order Cases (n=35) | Reverse Order Cases (n=35) |
| | Error Variance | Percent | Error Variance | Percent |
| Variables | Measure | Correct | Measure | Correct |
| 1 | -.02 | .22 | -.27 | -.08 |
| 2 | .13 | .25 | .13 | .21 |
| 3 | -.07 | .09 | .21 | .19 |
| 4 | .15 | .27 | .36 | .05 |
| 5 | -.20 | -.34* | -.20 | -.10 |
| 6 | -.08 | .14 | .14 | .34* |
| 7 | .04 | -.15 | -.19 | -.11 |
| 8 | -.15 | -.14 | -.19 | -.16 |

Note.
1 = R-Squared, from Similarity Judgment Task
2 = Conceptual Coherence Scale, from Protocols
3 = R-Squared, from Lens Model Task
4 = Achievement, from Lens Model Task
5 = Matching Index (εd) from the Lens Model Task
6 = Judgment Process Scale, from Protocols
7 = Depth of Search from Information Board Task
8 = Pattern of Search from Information Board Task
Correlations were based on 36 subjects
* p < .05.
With the normal order cases, Achievement from the lens model task, the Conceptual Coherence scale from the protocol data, and the $R^2$ from the similarity judgment task were the three variables with positive relations with percent correct measure of accuracy. The matching index, $\Sigma-d$, was negatively related to the percent correct measure of accuracy, which showed that subjects with more ideal matching also (as expected) tended to be more accurate.

With the reverse order cases, Achievement from the lens model task was related to the error variance measure. The Judgment Process scale was related to the percent correct measure. $R^2$ from the similarity judgment task was \textit{inversely} related to the error variance measure, which was surprising because this latter measure had been transformed so that high scores reflected greater accuracy.

The error variance measure and the percent correct measure appear to be tapping different aspects of judgment performance quality. Subjects with extreme errors in prediction were being penalized when the error variance measure was used, whereas with the percent correct measure, there was no such penalty.

**Section D: Results Related to Multivariate Analyses**

**Results Related to Research Conditions**

Research Questions 4 and 5 are:

To what extent does cue-presentation condition (context cues followed by individuating cues, or the reverse), reveal patterns of relationships in performance in a clinical judgment task?

To what extent does memory-priming condition (exposure to relevant domain-specific visual stimuli, versus no exposure), reveal patterns of relationships in performance in a clinical judgment task?
In the healing time judgment task, 18 of the subjects (those in case sequence 1) completed the normal order cases first, and then the reverse order cases; the other 18 subjects (those in case sequence 2) completed the cases in the opposite sequence. With 36 subjects, the inclusion of both between-subject factors (slide condition and case sequence), and experience level, made cell sizes extremely small. Multivariate analyses of variance (MANOVA), therefore, were carried out twice, using an alpha level of .05. The first time case sequence was included to determine that it was not significant; the second time, case sequence was omitted.

The mean squares, degrees of freedom, F-value and significance of F related to these analyses are provided in Table 4-25.

Table 4-25

Multivariate Analysis of Variance for the Effects of Slide Condition and Paragraph Order (Normal Order and Reverse Order) on Judgment Accuracy (Error Variance Measure)

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>Sig of F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Between Subjects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slide Condition (S)</td>
<td>1</td>
<td>.76</td>
<td>.05</td>
<td>.83</td>
</tr>
<tr>
<td>Experience (E)</td>
<td>2</td>
<td>2.78</td>
<td>.17</td>
<td>.85</td>
</tr>
<tr>
<td>S by E</td>
<td>2</td>
<td>44.08</td>
<td>2.68</td>
<td>.09</td>
</tr>
<tr>
<td>Error</td>
<td>30</td>
<td>16.47</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Within Subjects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paragraph Condition (P)</td>
<td>1</td>
<td>28.20</td>
<td>7.69</td>
<td>.01</td>
</tr>
<tr>
<td>S by P</td>
<td>1</td>
<td>.18</td>
<td>.05</td>
<td>.83</td>
</tr>
<tr>
<td>E by P</td>
<td>2</td>
<td>.26</td>
<td>.07</td>
<td>.93</td>
</tr>
<tr>
<td>S by P by E</td>
<td>2</td>
<td>1.97</td>
<td>.54</td>
<td>.59</td>
</tr>
<tr>
<td>Error</td>
<td>30</td>
<td>3.67</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
There was a significant difference between measures of performance based on error variance related to paragraph order. Paragraph order was a within-subject factor.

Subjects in slide condition 1 viewed slides of incisions immediately prior to performing the healing time judgment task. Subjects in slide condition 2 saw the same slides following this task. Because in this analysis case sequence was not significant, collapsing across sequence was carried out. A MANOVA was performed, using accuracy assessed by the percent correct measure on the two sets of cases as dependent variables. The mean squares, degrees of freedom, $F$-value, and significance of $F$ can be seen in Table 4-26.

Table 4-26

**Multivariate Analysis of Variance for the Effects of Slide Condition and Paragraph Order**

(Normal Order and Reverse Order) on Judgment Accuracy (Percent Correct Measure)

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>MS</th>
<th>$F$</th>
<th>Sig of $F$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Between Subjects</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slide Condition (S)</td>
<td>1</td>
<td>186.16</td>
<td>1.27</td>
<td>.27</td>
</tr>
<tr>
<td>Experience (E)</td>
<td>2</td>
<td>170.50</td>
<td>1.16</td>
<td>.33</td>
</tr>
<tr>
<td>S by E</td>
<td>2</td>
<td>566.45</td>
<td>3.85</td>
<td>.03</td>
</tr>
<tr>
<td>Error</td>
<td>30</td>
<td>147.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Within Subjects</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paragraph Condition (P)</td>
<td>1</td>
<td>18.03</td>
<td>.27</td>
<td>.60</td>
</tr>
<tr>
<td>S by P</td>
<td>1</td>
<td>84.12</td>
<td>1.28</td>
<td>.27</td>
</tr>
<tr>
<td>E by P</td>
<td>2</td>
<td>62.52</td>
<td>.95</td>
<td>.40</td>
</tr>
<tr>
<td>S by P by E</td>
<td>2</td>
<td>19.25</td>
<td>.29</td>
<td>.75</td>
</tr>
<tr>
<td>Error</td>
<td>30</td>
<td>65.74</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
There was a significant interaction between experience level and slide condition. Because paragraph order was not significant, the total percentage of correct judgments was computed and used in the calculation of means. In Table 4-27, the means and standard deviations for judgment accuracy are presented, by experience level. The cell means reveal that experienced subjects (as anticipated) tended to perform better with enhanced memory-priming; contrary to expectations, however, novice subjects who were not exposed to the slides also had high scores.

Table 4-27

Means and Standard Deviations for Judgment Accuracy (Percent Correct for all Cases - Normal Order and Reverse Order Taken Together), by Slide Condition and Experience

<table>
<thead>
<tr>
<th>Experience Level</th>
<th>Slide Condition 1&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Slide Condition 2&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Means (SD)</td>
<td>Means (SD)</td>
</tr>
<tr>
<td></td>
<td>No. of subjects</td>
<td>No. of subjects</td>
</tr>
<tr>
<td>1</td>
<td>34.33 (13.53)</td>
<td>48.50 (4.75)</td>
</tr>
<tr>
<td>n=12</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>36.75 (8.58)</td>
<td>38.06 (7.44)</td>
</tr>
<tr>
<td>n=12</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>45.56 (8.91)</td>
<td>40.13 (2.17)</td>
</tr>
<tr>
<td>n=12</td>
<td>8</td>
<td>4</td>
</tr>
</tbody>
</table>

<sup>Note</sup>. SD = Standard Deviation

<sup>a</sup> Subjects viewed slides prior to the judgment task.

<sup>b</sup> Subjects viewed slides after the judgment task.
Results Related to Methods for Confidence Assessment

Measures of confidence were significantly influenced by the method used to determine confidence. A comparison of the results of the concurrent and retrospective methods for assessing confidence using MANOVA revealed that (using an alpha level of .05), there were significant differences between the two methods. In normal order cases, $F(1, 29) = 15.27; \text{MSE} = 3039.87$. In the reverse order cases, $F(1, 29) = 9.43; \text{MSE} = 1310.30$.

Section E: Summary

Research question 6 is:

Of all the measures included in the study, which measures are most predictive of clinical judgment performance?

Prediction of Judgment Expertise

**Normal order cases.** Using a stepwise approach with an inclusion criterion of $p = .05$, and the error variance measures of accuracy as the dependent variable, no cognitive variables or experience variables entered the equation. None of the variability in accuracy was accounted for by the variables included in the study.

If, however, the percent correct measures of accuracy were used as the dependent variable, seven predictors entered the equation. These variables were: Phase Of Healing, Conceptual Association, Wound Closure Concepts, Conceptual Structure $R^2$ for Three Dimensions (from the multidimensionally scaled concepts from the similarity judgments task), High Estimate Of Healing Time, and Omega-Squared For Main Effects (from the orthogonal judgments task), and General Education. The $R^2$ was .67 (adjusted $R^2$ was .59).
In Table 4-28(a), the intercorrelations between these predictor variables, experience, and the percent correct measures of accuracy are reported.

<table>
<thead>
<tr>
<th>Variables</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>--</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>.01</td>
<td>--</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-.06</td>
<td>.21</td>
<td>--</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-.10</td>
<td>-.01</td>
<td>.32</td>
<td>--</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-.04</td>
<td>-.02</td>
<td>.30</td>
<td>.11</td>
<td>--</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>-.14</td>
<td>-.01</td>
<td>.03</td>
<td>-.00</td>
<td>.03</td>
<td>--</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>-.05</td>
<td>-.10</td>
<td>-.10</td>
<td>-.18</td>
<td>-.10</td>
<td>.11</td>
<td>--</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>.07</td>
<td>-.16</td>
<td>-.47**</td>
<td>.11</td>
<td>-.16</td>
<td>-.46**</td>
<td>-.01</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>.40*</td>
<td>-.36*</td>
<td>-.38*</td>
<td>-.24</td>
<td>.10</td>
<td>-.22</td>
<td>-.12</td>
<td>.32</td>
<td>--</td>
</tr>
</tbody>
</table>

**Note.**
1 = Phase of Healing from MDS of Similarity Judgments
2 = Conceptual Association from MDS of Similarity Judgments
3 = Estimate of Delayed Healing Time from Orthogonal Judgment Task
4 = Wound Closure concepts from MDS of Similarity Judgments
5 = Omega-Squared for Main Effects from Orthogonal Judgment Task
6 = General Education, Scored Dichotomously (Basic, Extra)
7 = $R^2$ for Three-dimensional Solutions for MDS of Similarity Judgments
8 = Experience, in Months
9 = Accuracy (Percent Correct Measure)

Correlations were based on 36 subjects.

* $p < .05$. ** $p < .01$. 
The variable Estimate of Delayed Healing was most highly correlated with experience; contrary to expectations, however, this correlation was negative. Phase of healing and Conceptual Association were most highly related to accuracy.

**Reverse order cases.** Using a stepwise approach with an inclusion criterion of $p = .05$, and the error variance measures of accuracy as the dependent variable, three variables entered the equation. Significant predictors were: Phase Of Healing (related to the concepts from the similarity judgments task), Omega-Squared For Interactions (from the orthogonal judgment task), and Achievement (from the lens model task). The $R^2$ was .46 and the adjusted $R^2$ was .41. The correlation between these predictor variables, experience, and accuracy are given in Table 4-28(b).

### Table 4-28(b)

**Intercorrelations for Predictor Variables for Accuracy for Reverse Order Cases (Error Variance Measure), and Experience**

<table>
<thead>
<tr>
<th>Variables</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.31</td>
<td>.14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.07</td>
<td>-.16</td>
<td>-.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>.52**</td>
<td>-.35*</td>
<td>.36*</td>
<td>-.03</td>
<td></td>
</tr>
</tbody>
</table>

**Note.**

1 = Phase of Healing from MDS of Similarity Judgments
2 = Omega-Squared for Interaction Effects from Orthogonal Judgment Task
3 = Achievement from the Lens Model Task
4 = Experience, in Months
5 = Accuracy (Error Variance Measure)

Correlations were based on 36 subjects.

* $p < .05$. ** $p < .01$. 
If the percent correct measures of accuracy were used, only one variable entered the equation, giving an \( R^2 \) of .13 (adjusted \( R^2 \) of .10). This variable was Low Estimate Of Healing Time. The intercorrelations between these predictor variables, experience, and percent correct measures of accuracy are presented in Table 4-28(c).

Table 4-28(c)

<table>
<thead>
<tr>
<th>Variables</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-.22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-.35*</td>
<td>.09</td>
<td></td>
</tr>
</tbody>
</table>

Note.  
1 = Estimate for Fast Healing Time from Orthogonal Judgment Task  
2 = Experience, in Months  
3 = Accuracy (Percent Correct Measure)  
Correlations were based on 36 subjects  
* \( p < .05 \).

Summary of Patterns of Relations

This chapter has been a report of the analyses and results related to clinical judgment performance.

A number of interesting patterns of relations were identified. In summary, differences in conceptual structure were found to predict accuracy in an intuition-inducing task, but not in an analysis-inducing task. With judgment process, experienced subjects showed greater metacognition, more reflectivity, and a higher level of adaptability, compared to novice subjects. Configural cue-utilization
assessed by lens model C-Scores was higher in the reverse order cases, compared to the normal order cases. Configural processing assessed by omega-squared for interaction terms decreased with experience in a task where cues were orthogonalized.

Given the level of task unpredictability, subjects performed well in the lens model task: subjects seemed to have good sensitivity to patterns of data related to incisional healing. Subjects' ability to match ecological validities was somewhat better in the reverse order cases, compared to the normal order cases. The relationship between $\Sigma-d$ and the percent correct measure of accuracy was slightly stronger under intuition-inducing conditions.

When scores were assessed in terms of absolute accuracy, however, performance would be described as moderately good. Although there was little association between performance and experience in the reverse order cases, these variables were positively related in the normal order cases.

A discussion of these results is presented in chapter 5.
V. DISCUSSION

This chapter has five sections. In section A, the results related to each of the tasks were interpreted, and related to the literature. Possible reasons for some of the results have been identified. Findings were organized by the research questions. Section B is a discussion of the findings related to the tasks and methods and their implications with respect to the philosophical tradition with which they are associated. An attempt was made to achieve unification. Some limitations to the study were identified. Section C is a commentary on how the findings add a number of important contributions to the research that has been conducted in clinical judgment and expertise. The implications of the findings to professional education and practice have been presented in section D. The final section includes a summary of the study and some suggestions for future research.

Section A: Discussion of Major Findings

Research Question 1

What are the patterns of relationships among various measures of judgment performance (indicators of expertise) and experience for a group of subjects in a clinical judgment task?

The following indicators of expertise are discussed: judgment accuracy, judgment confidence and calibration, consistency of judgment, judgment latency, and knowledge accessibility.

Judgment accuracy. There were very little differences in the mean accuracy of judgments of healing time with experience (Table 4-04). This finding
replicated previous research (Camerer & Johnson, 1991; Faust et al., 1988; Garb, 1989; Goldberg, 1968).

Subjects in experience level 2, at times, performed less well compared to the other two groups. Lesgold et al. (1988), in their research with physicians on the interpretation of x-ray films, investigated the development of expertise in a complex skill. These researchers found that, whereas traditional learning theory would predict performance improving steadily with practice, for certain problems, performance became worse, prior to improving. Lesgold and colleagues found that subjects' performance was not a monotone function of experience, and gave examples where third- and fourth-year residents performed less well than either experts or first- and second-year residents.

Lesgold et al. (1988) noted that a similar phenomenon has been reported by developmental psychologists; in a variety of developmental studies (cited in Lesgold, 1984), a skill is observed to be present at one age, is missing somewhat later, and is again present some time after that. The example was given of children's use of irregular past tense verbs and plural nouns. These researchers explained the nonmonotonicity effect they had observed in physicians by comparing it to these developmental irregularities. As practitioners develop increasing skill, the transition to more sophisticated approaches may decrease performance temporarily until the more advanced strategies have been practised sufficiently.

In the present study, the performance of the intermediate group was, at times, anomalous. For example, examination of Tables 4-04, 4-21, and 4-25, revealed that subjects in this middle group had scores that deviated from the
trend shown by the other two groups, a pattern that demonstrated "U-shaped" distributions.

**Judgment confidence and calibration.** Confidence was conceptualized as metacognitive knowledge: an estimate of the degree of trustworthiness of one's own judgments. A probabilistic task with a high level of uncertainty ought to induce limited confidence (Keren & Wagenaar, 1987). Even though subjects were informed about the probabilistic nature of the task, experienced subjects appeared unable to moderate their confidence ratings accordingly. This finding is consistent with the results of a study conducted by Baumann et al. (1991) with 40 intensive care nurses. In other studies, this pattern for confidence to increase with experience (Table 4-05) has been observed (Fischhoff et al., 1977; Koehler, 1994; Lichtenstein & Fischhoff, 1977; Einhorn & Hogarth, 1978).

An assessment of judgment calibration revealed deviations from perfect calibration. See Tables 4-07 and 4-08. This finding also replicated previous research related to overconfidence in judgment (Lichtenstein, et al., 1982; Oskamp, 1962, 1965). Novice subjects were more modest in their confidence ratings, and thus they demonstrated better overall calibration, compared to experienced subjects.

In relation to confidence for correct and incorrect judgments reported separately (Table 4-09), confidence was consistently higher for correctly judged cases, compared to cases incorrectly judged. Based on the differences shown between mean scores for correct and incorrect cases, however, most subjects did not make large distinctions. Confidence for reverse order cases was somewhat lower (and thus more appropriate) compared to that shown for normal
order cases. Based on these findings, it seems that analytic conditions induced less overconfidence, compared to intuitive conditions.

**Consistency of judgments.** Correlations between repeated judgments of healing time were considerably higher and less variable, compared to the correlations of repeated judgments of confidence (Table 4-07). One factor contributing to the low consistency of confidence judgments was that, for many subjects, the range for confidence judgments was low. Another factor was that the number of cases (five) may have been insufficient to assess consistency, a point made recently by Siegel-Jacobs and Yates (1996). A third factor was that confidence in one’s judgment is a subjective state that may lack stability over time. The experience of having made many judgments could have altered subjects’ confidence for subsequent judgments.

**Judgment latency.** There were small differences in mean latency associated with experience in the healing time judgment task (Table 4-10). There was also an effect of sequence: subjects responded more quickly to the set of cases in the healing time judgment task that was presented second. This tendency may have been a "practice" effect, or it may have represented subjects' motivation for not prolonging the research session any longer than necessary. The relationship between latency and experience was more complex than what can be shown by these data: short latencies were observed with subjects who were very capable, as well as with those who made careless errors; long latencies were noticed with subjects who were highly motivated to perform to the best of their ability, as well as with those who deliberated, but were less able to make accurate predictions, compared to other subjects.
**Knowledge accessibility.** The Knowledge Accessibility scale is available in Appendix D. The means increased only slightly with experience, indicating that experience accounted for a only small proportion of variance in scores (Table 4-11). The small mean differences may be due to the fact that for expert clinicians, knowledge is encoded in dynamic patterns which are difficult to articulate. Revealing them may require the instantiation of an authentic context, and close physical interaction with the phenomenon of interest. Several experienced subjects stated that they wished they could directly evaluate an area of the incision using forceps or a probe: they wanted "hands-on" assessment.

The findings related to knowledge accessibility were important not so much because of the quantitative measures of particular knowledge from each sub-scale, but rather because this task revealed qualitative aspects of the translation of perceptual knowledge into language. Differences in knowledge accessibility may be difficult to detect by an examination of the dialogue, in part because some knowledge may be so deeply internalized that it is not expressed overtly. This conception, consistent with Reber's (1989) and Dreyfus and Dreyfus' (1986) views about the use of intuition, constitutes a possible explanation for the findings.

There is much literature on knowledge access. Brown (1982) discussed access of mental processing, using Rozin's (1976) theory: In the course of evolution, cognitive programs have become more accessible and, therefore, can be used flexibly in a variety of situations. Brown viewed conscious control as being the highest level. Yet, with experts, there often is no conscious awareness of retrieval processes. It is possible that tacit knowledge of implicitly learned patterns could be regarded as higher than consciously-controlled knowledge.
Reber (1989) discussed the commonly assumed epistemic priority of consciousness. He argued:

Knowledge acquired from implicit learning procedures is knowledge that ... is always ahead of the capability of its possessor to explicate it. ... The implicitly acquired epistemic contents of mind are always richer and more sophisticated than what can be explicated (Reber, 1989, p. 229).

In order to address the issue of which type of knowledge is higher, the purpose must be considered. If the goal is teaching, then having consciously-controlled knowledge ought to be regarded as the highest order. If the goal is clinical practice, however, implicit knowing-in-action as Schön (1983) described, may be most useful.

Relationships among indicators of expertise. When accuracy was assessed by the percent correct measure in the normal order cases, the correlation between accuracy and experience was .32. In the reverse order cases the comparable correlation was .09 [Tables 4-12(a) and 4-12(b)]. It is possible that task factors may have influenced the strength of the relationship between experience and performance. The normal order cases could be performed through the use of either an intuitive or analytic strategy; in contrast, the structure of the reverse order cases would make intuitive processing very difficult, if not impossible. Dreyfus and Dreyfus (1986) claimed that as expertise is acquired, the use of intuition increases. The difference in the experience-performance relationship between normal order and reverse order cases constituted support for Dreyfus and Dreyfus' ideas about the use of intuition by experienced practitioners. It is also possible, however, that the particular cases selected for normal order and reverse order presentation evoked the observed differences.
The response to research question 1 can be summarized as follows: the findings replicated what has been found in the literature: little relationship was observed between performance accuracy and experience. When accuracy was assessed using an intuition-inducing task, however, experience accounted for a greater proportion of variance in performance compared to assessment under analysis-inducing task conditions. Judgment consistency and confidence tended to increase with experience; judgment calibration and latency tended to decrease.

Research Question 2

What are the patterns of relationships among measures of conceptual structure, sensitivity to patterns in data, judgment process, and performance in a clinical judgment task?

Conceptual structure. Two tasks were used to obtain data about conceptual structure: the association task, and the similarity judgments task.

The association task provided protocol data to reveal thinking related to clinical judgment. The Conceptual Coherence scale is available in Appendix E. The anticipated pattern of changes with experience was demonstrated (Table 4-13), particularly with two of the sub-scales (concept familiarity and conceptual structure). For some subjects, the predicted differences in the structure of verbalizations (for example, expression of clinical knowledge in the form of a story) were found. Such structural differences constituted evidence for more sophisticated conceptual coherence.

The similarity judgment task demonstrated small differences in subjects' conceptual structure. The multidimensional scaling of similarity judgments
revealed some patterns of relations (Table 4-14) in the proximities of concept clusters with experience.

**Sensitivity to overall patterns in data.** The pattern of medians for $R^2$ and G-Scores (Table 4-15) showed that experience had a stronger relationship under analysis-inducing conditions, compared to intuition-inducing conditions. If there were a significant number of subjects with high levels of expertise, then performance in the normal order cases should have revealed much differentiation. One explanation for this finding is that there were few true experts (as described by Dreyfus & Dreyfus) in this sample of subjects.

Considering the experience differences for each group of subjects, there was a remarkable consistency in the means of the $R^2$ for normal order cases (Table 4-16).

Hammond et al. (1964) proposed that the matching index, $\Sigma$-$d$, had promise as an important component of clinical inference. In this study there was good variation in this index of the degree to which subjects matched ecological validities with utilization coefficients (from .79 to .08 for normal order cases and .66 to .07 for reverse order cases). Hammond and colleagues recommended that $\Sigma$-$d$ is best interpreted within the context of other lens model parameters. This measure was strongly related to achievement ($r = -.77$ for normal order cases, and $r = -.84$ in reverse order cases). Subjects who had high achievement tended to have low (more ideal) scores on $\Sigma$-$d$. $\Sigma$-$d$ also related to C-scores ($r = -.35$ in the normal order cases and $r = -.37$ in the reverse order cases), indicating that subjects who used nonlinearity effectively also tended to have good $\Sigma$-$d$ scores. $\Sigma$-$d$ scores were found to be multicollinear (within rounding error) with the lens model G-scores. In this data set, G-scores were calculated by
correlating the model-predicted healing times with predictions based on subject's unique regression policies on the basis of the same cases. The calculation of $\Sigma$-$d$ involved the same two sources of data. Thus, both the $\Sigma$-$d$ scores and the G-scores were measures of the same linearity; the difference between the two measures was that G represented the degree of match, whereas $\Sigma$-$d$ represented the degree of mismatch.

Capturing subjects' policies by means of a regression equation demonstrated that different policies gave rise to similar outcomes (Appendix H). Brehmer (1994) claimed that in making judgments, subjects' particular weighting of cues does not matter much. The large inter-individual differences found in subjects' policies in this study constituted supporting evidence for Brehmer's claim. Brehmer attributed policy variation to the fact that conditions under which subjects acquired their policies were different. Most of the available cues were correlated, and the judges were not focusing on the weights.

A comparison of the regression weights from subjects' policies (Appendix H) with the regression weights from the left side of the lens model (Table 4-02), demonstrated an interesting difference. In the ecology, when 258 cases were analyzed, Severity of Surgery was the most significant predictor for healing time; in the subject systems, when two sets of 35 cases were analyzed, only one subject perceived Severity of Surgery as the most relevant cue.

The interpretation here is not that most subjects failed to detect the most important cue. Rather, because each set of lens model judgments contained 35 cases for analysis, subjects partitioned the cue Severity of surgery into its components with which it was correlated: length of time for surgery, blood transfusions needed, complications, size of incision, and so on. The subjects
possibly made qualitative distinctions among cases that related more to these other constituent cues available rather than to a classification of cases in terms of *minor* or *major* surgery. The particular cues which emerge as important depend, not only on the subject's weighting policy, and the intercorrelations among the cues, but also upon the ways the subjects conceptualize the cues, and the numbers of cases being analyzed. *Severity of Surgery* had status as the most important cue when the analysis included 258 cases, and when this variable was constrained to be dichotomous (minor or major surgery). Such simplification of a continuous variable into two levels reduces the ambiguity in the ecology. It appears that the subjects in the study did not apply this simplifying strategy.

The accuracy of subjective policies of novice subjects on cross-validation was similar to that of experienced subjects, using the error variance measure of accuracy (Table 4-17). The success of cross-validation was somewhat better for the inexperienced subjects using the percent correct measure of accuracy. It may be that looking for differences in the weighting of cues is taking too microscopic a view, as Brehmer (1994) claimed. It is possible that these novice subjects were knowledgeable and highly motivated, and applied their knowledge to the research task in a manner that was more consistent compared to experienced subjects.

One point about lens model achievement that warrants caution is that at the same time lens model achievement can be high, accuracy in an absolute sense can be low. For example, one subject in this study had good lens model achievement scores (.73 and .72 for normal order and reverse order cases, respectively), but had low accuracy scores, assessed by percent correct (28% and 18%). With this subject, high achievement scores were accomplished by providing responses that were in the appropriate direction, but were extreme; a
large variance resulted, which was translated into good lens model achievement. When assessed for accuracy, however, few of the predictions were within the interval measure of healing time. This subject appeared to have a good sense of the overall patterns in the ecology, but had poorly calibrated knowledge of typical healing times. Other subjects had low correlational measures, but good accuracy measures; these subjects' judgments were closer to the base rate and showed little variation in response to cases with large differences in cue-values.

What seems surprising is the small variation in lens model achievement across experience groups. One possible reason why novice subjects' performance on the healing time judgment task was as good as it was may be because, in the absence of experience, it is reasonable to propose that novices used a theory-based strategy; the cues employed in this study were all familiar in theory.

In contrast, the experienced subjects likely used a strategy based on the patients they had encountered during their careers. Several researchers have characterized expertise in physicians as "case-based", rather than generalized (Elstein, Shulman, & Sprafka, 1978; Feltovich & Barrows, 1984). Many of the nurses in this study seemed to have knowledge that could be characterized as case-specific. Scores may have been reduced for some experienced subjects because the match between the variety of surgical cases used in this study and the types of patients they have cared for may not have been optimal.

**Judgment process.** This construct was assessed by the information board task and by the orthogonal judgment task.

Data from the information board task demonstrated a small effect of experience with the Depth of Search and Number of Shifts. There was a small
increase in the number of accurate subjects with experience (Table 4-18). Subjects in the middle group had a positive search pattern, which indicated that on average, they used an interdimensional search pattern more frequently, compared to other subjects. The experienced subjects' search pattern was more often intradimensional.

The verbal protocols were examined to determine which cues were most emphasized. There seemed to be no relationship between the weight a cue was given in the protocol data and the weight derived from regression. For example, one subject emphasized on four occasions the importance of old age as a factor in healing (with reference to the 80 year old patient), yet Age did not enter in this subject's regression policy.

The Judgment Process scale is presented in Appendix I. The examples of metacognition, input generation, and cognitive operations were illuminating in terms of how nurses think as they make clinical judgments (Table 4-19). Nine of the protocols demonstrated reflectivity in relation to salient data; this type of verbalization was sometimes characterized by a particular affective quality that could be detected in the tone of voice from the audio-tapes. Examples of the types of reflective responses included surprise when expectations were not met, curiosity when subjects had uncovered partial data, perplexity when information seemed incongruous, satisfaction when expectations were fulfilled, or apprehension when cues of a particularly serious nature were revealed. Protocols from other subjects included the verbal processing of the cues, but no such reflection or affectivity could be detected.

The following excerpts are examples from the protocols:
S. 16: "I want to see if the diagnosis for patient A is ulcerative colitis. [Turns over card.] Just as I thought!" [Satisfaction].

S. 21: "I want to know the medical history. Oh! My goodness! I would have to watch him." [Apprehension].

S: 28: "This one has swollen thighs. At this point I can't see the connection between the thighs and the incision." [Perplexity].

This reflectivity and affective responsiveness may simply indicate individual differences in communication style, or perhaps differential ability to carry out the thinking-aloud component of the task. These features, however, could represent real differences in information processing. The reflectivity may be generating an internal response which the subject is using as a "secondary cue", as described by K. R. Hammond (1990). The reflective and responsive subjects tended to demonstrate motivation for the tasks, and good judgment performance. Experienced subjects provided all but one of these protocols. A similar example of responsiveness to internally-generated [secondary] cues is presented in Abernathy and Hamm (1995) where a resident physician paid attention to his own reactions to a patient with internal injuries, including a pelvic fracture, using his own responses as a gauge of the seriousness of the case and as a basis for clinical judgments.

The notion of secondary cues is not new; in fact, over 35 years ago, Sarbin et al. (1960) claimed that an important source of information in making clinical judgments was "the inferrer's own viscero-somatic reactions to other inputs" (p. 176). It appears that as technology becomes more sophisticated, however, much legitimacy is given to cues such as test results and laboratory data, and little value is placed on the experienced clinicians' internal cues which may have proven validity. Unless these internal cues are understood and valued,
clinicians may not develop and properly calibrate their internal sensors, and fail to reach their potential in terms of implementing the art of clinical judgment. Abernathy and Hamm (1995) stated that "intuitive thinking has atrophied as medicine and surgery have become more oriented towards numbers in laboratory data, numbers in critical care data, and the use of radiology reports as 'hard data'" (p. 4).

There has been considerable research into the manner by which subjects integrate cues into judgments. Based on evidence from the protocol data, subjects in this study seemed to use a compensatory function; they examined a certain number of cues, and then integrated them (seemingly) in an additive fashion. There were instances from the protocols where subjects processed two cues interactively, but no examples where subjects used extremely complex configural integration strategies. Whether this failure to apply such integration was related to subjects' information-processing capacities or constraints, or whether it ought to be attributed to the nature of the task was not possible to determine. The information board task was too easy to function as a test of the form of cue-integration employed by subjects. The four alternatives used in this task did not sufficiently challenge working memory, nor was there any time pressure. Unlike many information boards from the literature where preferences were elicited (for example, a choice of apartments among a set of possibilities), the board in the present research required judgments based on domain-specific knowledge [rank ordering of healing times]. With preferences, there is no normatively correct response, and thus subjects likely would feel no constraint in terms of the data they seek and the strategies they apply. In contrast, in this task, judgments were required. Subjects knew there was a normatively correct response, and that, as nurses, they may have felt accountable for applying
appropriate and defensible processes because their judgments were being scrutinized (Siegel-Jacobs & Yates, 1996). Such differences between preferences and judgments may help to account for the lack of clear patterns for data search and cue-integration strategies with experience.

Payne et al. (1992) described a constructive process view of judgment. These researchers claimed that in making judgments, individuals often develop strategies opportunistically; they alter their processing at the time, depending on the information they encounter and the contingencies inherent in the task. Thus, strategies were not merely revealed through a task, but were constructed on-line. Payne et al. (1993a, 1993b) described decision makers who flexibly applied both top-down and bottom-up strategies effectively. In the present study, the verbal protocols constituted evidence that the subjects varied in their constructiveness and adaptiveness. Most subjects were strategic in their information processing.

Discrimination of healing times and configural cue-use were assessed by the orthogonal judgment task (Table 4-20). This task required making judgment about healing times for hypothetical patients. Most of the nurses had worked in hospital settings, where they did not see patients with extremely delayed wound healing. Such patients are frequently discharged and cared for by community nurses. Although all the subjects were informed that the total time for healing was required (not just the time while the patient was in hospital), the experienced nurses apparently had difficulty in calibrating their judgments to incorporate this data, as shown in Table 4-20. On average, experienced subjects had less accurate estimates related to healing times, compared to novice subjects. Yet, such estimates very likely reflected their experience, from their hospital setting perspective.
The orthogonal judgment task required analytic thinking for subjects to consider the overall effects of the cues (main effects) as well as the joint impact of cues (interactive effects). The percent of variance accounted for by main effects and interaction effects were assessed. Omega-squared for interactions was used to assess configural processing.

The omega-squared for main effects and interaction effects tended to decrease with experience level, as shown in Table 4-20. This finding raises a number of questions. Does configural processing actually decrease with experience? Or, is the observed decrease in Omega-squared a reflection of the experienced subjects' preference for intuitive task conditions? If, as Dreyfus and Dreyfus (1986) claimed, experience fosters the intuitive processing of information, then experienced subjects' performance may have been reduced by the lack of congruence between the task demands and the subjects' preferred means of processing. K. R. Hammond et al. (1987) described how such incongruence can significantly influence performance.

The means and standard deviations for various measures of configural cue-use showed no increasing pattern with experience (Table 4-21). Variability of information search, and C-Scores in the reverse order cases demonstrated the opposite pattern: the scores decreased with experience level. Correlations among various indices of cue-use tended to be low and did not show the relationship with expertise predicted by Dreyfus and Dreyfus' (1986) theoretical position [Tables 4-22(a) and 4-22(b)]. It seems interesting that there is considerable anecdotal evidence of configural cue-use, an extensive body of judgment literature describing configurality in judgment, and yet when attempts are made to carry out empirical research, it is extremely elusive. It may be that the methods of measurement used to assess configurality interfere with the
configural use of cues. An alternative approach to the study of configural processing is needed; verbal protocols may have potential in this regard.

Research question 2 can be summarized as follows:

The patterns of relations between cognitive measures and judgment accuracy which are notable in the normal order cases are those between $\Sigma$-$d$ and performance and between lens model achievement and performance. In the reverse order cases, the patterns which are strongest are those between scores on the Judgment Process scale and performance, and the lens model achievement scores and performance (Table 4-24).

Research Question 3

What are the patterns of relationships among individual differences in age, education, and experience and performance in a clinical judgment task?

Age, education, and measures of experience. The sample size was too small to fully address the questions about age and education (Table 4-03). Three main measures of experience were the number of months of surgical experience, the number of incisions viewed, and the protocol measure that reflected experience. The correlations between experience in months, and number of incisions was substantial, but the protocol measure was unrelated [Tables 4-23(a) and 4-23(b)].
Research Questions 4 and 5

To what extent does cue-presentation condition (context cues followed by individuating cues, or the reverse), reveal patterns of relationships in performance in a clinical judgment task?

To what extent does memory-priming condition (exposure to relevant domain-specific visual stimuli, versus no exposure), reveal patterns of relationships in performance in a clinical judgment task?

Impact of paragraph order condition. The patterns of findings regarding paragraph order were interesting. Small differences in performance in the two sets of paragraphs were identified, using the error variance measure of accuracy (Table 4-25). The correlational pattern in the normal and reverse order cases was different [Tables 4-12(a) and 4-12(b)].

The median lens model achievement scores were lower in the normal order cases compared to comparable medians in the reverse order cases, as shown in Table 4-15. Curiously, the medians for $R^2$ were higher in the normal order cases compared to those in the reverse order cases (except for experience level 3 where scores were essentially the same). This pattern was not anticipated. It may be that lens model achievement was lower in the normal order cases because these scores are correlations with a real world criterion. Performance in an intuition-inducing task was not as accurate, compared to performance in an analysis-inducing task. On the other hand, measures of $R^2$ have no connection to the real world; these scores tap internal consistency within the task which was higher under intuitive-inducing conditions. It is also possible that it was the specific content of each set of cases (rather than the manipulation of paragraph order) which elicited the small differences in responses.
The fact that C-Scores were larger under analytic conditions (see Table 4-21) provided some supporting evidence that the reversal of paragraph order was successful in altering subjects' information processing. This difference suggested that subjects were better at non-linear variance matching under analytic conditions. Such a finding is consistent with what K. R. Hammond et al. (1987) described in their study of expert engineers making judgments about road safety. These authors argued that when a task is intuition-inducing, a linear model can often account for the data, whereas, under analysis-inducing task situations, subjects can better detect and use nonlinearity. This interpretation by Hammond is different from the theoretical view proposed by Dreyfus and Dreyfus (1986). These latter researchers argued that individuals with high levels of expertise could detect and use nonlinearity best in an intuition-inducing task, providing the task was relevant and familiar. The source of this difference in prediction based on theory appears to be, at least in part, a difference in the definition of expert. Dreyfus and Dreyfus applied performance criteria, whereas Hammond used experience and credentials.

K. R. Hammond et al. (1987) also claimed that confidence in the accuracy of responses would be higher under intuition-inducing task conditions; an examination of Table 4-09 showed that the means for average confidence illustrated such a pattern. Overall, confidence increased with experience, and was somewhat higher when assessed by an intuition-inducing task, compared to an analysis-inducing task.

**Impact of memory-priming condition.** When accuracy was assessed by the percent correct measure, there was a significant slide-experience interaction, as shown in Table 4-26. It can be seen in Table 4-27 that under enhanced memory-priming conditions, subjects' performance tended to increase
with experience as predicted; under baseline memory-priming, however, the subjects who did not view the slides unexpectedly achieved very good performance. If the slides had functioned as a memory prime, novice subjects were predicted to show little difference in performance with priming. However, this was not the case.

One [post hoc] interpretation for this anomalous result is that the slides may have been educative for inexperienced subjects. The slide set was not completely representative of abdominal incisions; there were more slides of abnormal incisions compared to normal ones. It could be that viewing such slides caused greater inaccuracies for the novice subjects. The experienced subjects may have been better able to withstand such an influence. This explanation, however, does not account for the high scores for novices who did not receive enhanced memory-priming.

There was a significant effect of priming condition on judgment performance. As Birnbaum (1982) described in relation to internal context, subjects have unique backgrounds in the form of memories of past and recent experiences that can (and presumably do) interact with particular cues, likely in ways unanticipated by the experimenter. Newell (1968) expressed long ago that "humans add to the information available in the situation in order to make their judgments, even when they have no internal information relevant to the situation" (p. 7).

The point here is not so much that memory influences information processing, (as that is already known), but to alert judgment researchers that it is not always possible to have control over all the cues utilized in a judgment task. When Brehmer (1994) stated that "linear modelling requires that a subject be
provided with all cues for all cases" (p. 150) [emphasis added], he may have been referring to all external cues, and was likely not precluding the possibility of internal cues being influential. Care needs to be taken to realize that any salient experiences immediately prior to a lens model type task, as well as memories evoked by the cues, have potential to influence judgment performance in the laboratory as well as in the clinical setting.

Research questions 4 and 5 are summarized as follows:

Differences in the observed patterns of scores were associated with paragraph order condition. In general, these patterns of scores provided modest support for K. R. Hammond's theory regarding a cognitive continuum. There was a significant slide-experience interaction in which experienced subjects demonstrated increased performance under enhanced memory-priming conditions, but the pattern of results in the baseline memory-priming condition was not interpretable.

Research Question 6

Of all the measures included in the study, which measures are most predictive of clinical judgment performance?

Predictors of clinical judgment performance. The configurations resulting from multidimensionally scaled similarity judgments constituted a positive finding. Four of these measures demonstrated significance in predicting judgment accuracy [Tables 4-28(a), 4-28(b), and 4-28(c)]. To the extent that the plots reflected the organization of cognitive structure, this finding demonstrated that cognitive structure predicted accuracy in judgment performance. This idea was proposed more than 35 years ago by Sarbin et al. (1960) in their discussion
of the sources of variation in clinicians: "Variations in inference . . . are primarily due to individual differences in the dimensions in the cognitive organization of the inferrer" (p. 224).

The similarity judgment task was an intuition-inducing task; subjects rated the similarity of pairs of concepts very quickly, without analysis. It may not be coincidental that these measures predicted accuracy only in the normal order cases where an intuitive strategy was likely used.

In the normal order cases, experience in months had a stronger relationship with judgment accuracy (when accuracy was assessed by the percent correct measure) compared to the comparable relationship in the reverse order cases. The question that initiated this study was whether experience was related to judgment accuracy. The conclusion on the basis of this judgment task is that there is a weak relationship between experience and judgment accuracy that was revealed by an intuition-inducing task but was not revealed by an analysis-inducing task. Experience accounted for only a small proportion of variance in performance. Experience did not predict judgment accuracy in this task.

Section B: Discussion Related to Methods and Analysis

Findings Related to Methods

Assessment of confidence and calibration. Subjects demonstrated significantly better calibration from Gigerenzer' and colleagues' retrospective method for assessing confidence, compared to the traditional method as shown in Table 4-05. There were no differences, however, for this group of subjects
between the traditional method and Björkman's (1994) method for determining calibration (Table 4-07).

The method used to determine calibration differed significantly in determining the extent of overconfidence demonstrated. Scores reflecting the use of the two confidence measures to calculate judgment quality can be seen in Table 4-08. Subjects were most accurate in estimating the number of correctly judged cases when the retrospective method was used in the reverse order cases. Inspection of the means for judgment calibration revealed that, for both methods, subjects in experience level 1 had somewhat better calibration scores; this finding was attributed to these subjects' more modest confidence, compared to experienced subjects.

Whether or not the representative selection of cases according to a natural reference class, as recommended by Björkman (1994) and Justlin (1994), had any beneficial effect on calibration could not be determined. The healing time judgment task did have representative sampling of cases, but it also had a high difficulty level; most subjects did not seem to consider this latter feature.

**Correlational and absolute measures of accuracy.** It is important that the differences between accuracy assessed by absolute measures (error variance and percent correct) and correlational measures derived from the lens model (achievement and variance explained) be understood. The subjects made judgments that were analyzed in two different ways. When absolute measures were examined, the judgments ranged in accuracy from 77 to 94 using the average error variance measure, and 10% to 63% when using the percent correct measure. When these same judgments were assessed by the lens model approach, however, 13 out of 36 subjects attained $R^2$ measures as high
as, or higher than, the regression model in the normal order cases. In the reverse order cases 12 out of the 36 subjects attained $R^2$ measures as high as, or higher than, the model. In each set of cases, 5 out of the 12 subjects from experience level 1 had scores on $R^2$ that met or exceeded the lens model $R^2$. Lens model achievement scores, particularly for novices, were impressive.

These high $R^2$ scores indicated that the subjects' cognitive control was higher than the predictability of the ecology. Compared to an external standard, performance appears fair to moderate, but when compared to a regression model, performance seems more favorable. This difference arises because the lens model yields correlational measures that contain no element of calibration.

**Value of a multi-method approach.** Green (1968) stated:

> Whether one accepts the first approximations as good descriptions of reality or as fictions contributed by the method of analysis depends partly on one's purpose. If the goal is prediction in some practical situation, an adequate description will serve. But if the goal is to understand . . . then we must beware of analyses that mask complexities" (p. 98). [Emphasis added].

Any one set of methods and analyses highlights certain aspects and obscures other aspects. In this study, the practical goal was to assess a number of cognitive variables and experience to determine which, if any, would predict accuracy in clinical judgment. Prediction, of course, would not mean causation, only an association; nevertheless, an associative link could be useful as a basis for further research. The most important goal of this study, however, was to understand judgment at a deep level: to comprehend how a person seeks, weights, and integrates data into a unitary response. This goal involved understanding of both judgment process and outcome, and required the collection and analysis of both quantitative and qualitative data. With such a comprehensive goal, a multi-method approach was necessary.
Using the lens model methodology took considerable time and effort. A major benefit was being able to compare the same subjects' lens model scores, with scores derived from other methods, as well as with scores from other lens model studies. Another benefit was obtaining healing times from authentic patients which allowed for assessing the accuracy of the subjects' judgments.

The analysis of the verbal protocols also took a great deal of time. The value of the protocols, particularly in understanding how these subjects integrated the various cues, goes beyond what can be detected from mean scores, or other numeric indicators. Their value is that they help to illuminate the processes by which subjects seek particular data in a probabilistic context, and construct meaning from a pattern of cues, with the goal to make clinical judgments. Protocol data available in published studies (Abernathy & Hamm, 1994; Kassirer, Kuipers, & Gorry, 1982; Moskowitz et al., 1988) were observed to be similar to the data generated by subjects in this study.

Recently, Cooksey (1996) pointed out the substantial impact of measurement in judgment research:

[A] very important implication for judgment research . . . parallels Werner Heisenberg's notion in physics that uncertainty exists at a fundamental level and that the act of observation changes what is observed. As in quantum physics, where it is not possible . . . to know simultaneously where a particle is in space and how fast that particle is moving (precisely measuring one characteristic alters the second characteristic), one cannot be simultaneously certain, in judgment and decision making research, of the what, where, how, and why of observed judgment behavior. In judgment research, one tends to lose sight of the fact that methodological procedures may alter the very phenomenon being observed (Cooksey, p. 331).
One example of the difficulty in measurement that may be attributed to the influence of method may be the assessment of configurality which has been discussed previously.

In chapter 4, the two measures of accuracy (error variance and percent correct) were found to be tapping different aspects of expertise. This difference illustrates how the researchers' choices of methods and measures has a significant influence on the conclusions made. For example, if only the error variance measure had been used the conclusions for this study would have been that the variance in the normal order cases was not explainable, paragraph order was significantly associated with accuracy of performance, and there was no slide-experience interaction. In contrast, if only the percent correct measure had been used, the conclusion would have been that the variance in the normal order cases was much more explainable, compared to that in the reverse order cases, the paragraph order condition was not related to performance accuracy, and there was a significant slide-experience interaction.

These contradictory results could be viewed as evidence that the methods were inadequate; another interpretation is that because all methods and measures have some effect on results, the fact that these different methods and measures generated inconsistent results provides evidence for the need for higher-order research that will help to account for such variation. Conflicting results have potential to act as an incentive for further study. A major value in using a multi-method approach is that by revealing the extent to which findings vary with method, researchers are in a better position to understand the many aspects of a question or problem, and to avoid unwarranted generalizations.
Findings Related to Research Traditions

Implications of data analysis methods. In this study, the decision was made to use less traditional approaches to the analysis of data rather than implement strict significance testing methods. The implications of this decision were far-reaching. If a criterion of statistical significance had been adopted and maintained, there would be no effects to report, nothing to discuss, and no opportunity to construct and comment upon possible connections between the findings in the present study and results of past research. In fact, following guidelines based on power analysis would mean this study would not have been conducted. In contrast, by paying attention to effect sizes (whether they are statistically significant or not), as Carver (1993) recommended, allowed the reporting of small differences. Patterns were reported not with the aim to make a strong knowledge claim, but to reflect on their possible meaning within the context of a theoretical conception, and to motivate others to discover and report interesting associated regularities so that growth in knowledge may occur.

Breaking with tradition is difficult. Schmidt (1996) claimed that "we have continued to emphasize significance testing in the training of graduate students despite clear demonstrations of the deficiencies of this approach to data analysis" (p. 116). Traditionally, statistical testing has been the predominant approach for data analysis in research in psychology. Shaver (1993) stated that a perusal of research journals, educational and psychological statistics textbooks, and doctoral dissertations will confirm that tests of statistical significance continue to dominate the interpretation of quantitative data in social science research. Yet, criticism of statistical significance testing has been reported for more than twenty years (Carver, 1978; Spielman, 1974). Huberty (1993) examined this issue from a historical perspective, and illustrated how entrenched a method can become.
Thompson (1989) noted that paradigms, including the accepted ways of thinking about research, come to be taken for granted as natural thought, and they carry normative implications about what is appropriate thinking. Thompson characterized paradigms as highly resistant to change. This study has demonstrated that although change is difficult to bring about, small changes are not impossible.

**Integration of research traditions.** Christensen and Elstein (1991) compared and contrasted two of the main research traditions where expert judgment has been investigated: information processing and judgment and decision making (JDM). Researchers from both traditions have attempted to capture the cognitive activity of experts as they make judgments under uncertainty, and they both evaluate judgment expertise. Several critical differences have been noted: whereas researchers from the information processing approach tend to focus on information search, and use verbal expressions of rules, researchers from the JDM tradition focus on the integration of weighted information.

In the information processing paradigm, the expert is given ample time to analyze and use a deliberative cognitive strategy; in the JDM approach, the expert is given a set of tasks to perform intuitively without time for analysis. The most consequential difference between these two views is that the information processing researchers have demonstrated the superiority of expert judgment, whereas the JDM researchers have found expert judgment to be poor, usually no better than a simple linear model, and often not this accurate.

K. R. Hammond (1987, 1990) proposed that the two groups examined expert judgment from different perspectives; they each present a version of
"truth" from a different point of view. Adherents from the information processing perspective use a theory that bases truth on coherence relations of similarity among features. JDM researchers use a theory in which truth is based on correspondence with some external standard, such as functional relations between proximal and distal variables.

K. R. Hammond (1987) argued that it is important to understand both views, and see the place of each in relation to the other. In seeking to construct the complementarity of these disparate perspectives, Hammond (1990) argued for a unified approach. Hammond is not alone in suggesting integration: Einhorn et al. (1979), as well as Billings and Marcus (1983), have applied both process tracing and policy capturing methodologies, illustrating advantages of a multi-method approach. Furthermore, Hammond believed that each approach used separately would provide only partial knowledge: "Unification might well provide new and important information about cognitive activity that would not have been produced by either approach alone" (Hammond, 1987, p. 14).

Research findings depend on the research tradition from which they arise. What research questions are considered important to investigate, what ontology is used to conceptualize and name the relevant variables, and what type of methods are advocated, may determine research findings and conclusions to a larger extent than what is recognized. In this study, K. R. Hammond's (1990) recommendation to conduct an investigation of clinical judgment performance using a number of methods from both the JDM and the information-processing tradition was followed. What was attempted was an investigation from two different research paradigms to see to what extent findings transcended the tradition or were determined by it.
From the JDM perspective the performance of the subjects was comparable to that reported in the literature. The conclusion is that experience accounted for only a small proportion of variance in judgment accuracy. From an information processing perspective, the subjects searched for information in adaptive ways, and were constructive in terms of their use of strategies. Results from this paradigm suggested a much more positive conclusion. Thus, it may seem that the evaluation of processes are more favorable than that of outcomes. Yet, it may be that some methods are more suited to revealing certain aspects or phases of the total judgment process than others.

In educational psychology what is needed is a metatheoretical perspective in which both research paradigms are integrated. As the twenty-first century approaches, and the amount of available information continues to accelerate, learning to process data optimally in probabilistic contexts and to make effective use of relevant information in judgment is critical. Emphasizing either process or outcome is limiting. What is needed is to understand the connections between certain characteristics of judgment process and accurate outcomes in particular contexts. Without understanding such connections, it is difficult to make a case that one process is better than another. Without examining context, variations in process and outcome that are associated with context are not clear, and may be attributed elsewhere. Such a broad theoretical perspective is necessary in order to capture the dynamic and responsive qualities of the human mind as judgments are made.
Limitations of the Study

Three limitations of this study were related to the following aspects: The sample of surgical patients who volunteered for phase 1 of the study; the artificial nature of the judgment tasks; and the narrow focus of expertise.

Limitations related to the sample of surgical patients. During phase 1 of the study the focus was the left side of the lens model; data were obtained from a large, representative sample of patients who had had abdominal surgery. The goal was to identify cues that predicted wound healing time, and to determine the relative importance of the cues as predictors. Patients could not be sampled randomly. The surgical patients who volunteered for this phase of the study deviated from that which would have been truly representative for a number of reasons. First, because of ethical considerations and the eligibility criteria, extremely ill patients were not approached, and thus not admitted to the study.

Second, because this sample was comprised of volunteers, the fact that certain patient-groups consented in greater proportions than other groups led to some degree of non-representativeness. Examples of under-represented groups included patients with drug and alcohol problems, and patients whose first language was not English. Third, departure from representativeness occurred also because follow-up was more successful with particular groups, such as highly educated, Caucasian subjects. These factors likely attenuated the relationships between predictor variables and healing time.

Limitations related to the judgment tasks. The tasks used in this study were qualitatively different from authentic judgment tasks that nurses generally perform. Schwartz, Griffin, and Fox (1989) claimed that the clinical judgments
physicians made were most typically categorical; similarly, most nursing judgments are categorical in nature (Radwin, 1995). They are often judgments related to one unique individual. Nurses are not accustomed to making quantitative judgments in relation a series of 30 or 40 cases, as lens model judgments require. Because performance was assessed as subjects engaged in these artificial tasks rather than in authentic clinical judgments, no conclusions can be made about how nurses would respond in the clinical setting. The purpose of the study, however, was to add to the body of knowledge about how people search for, weight, and integrate information when making judgments. These differences associated with tasks and setting were acknowledged as qualifications to the study.

Limitations related to the narrow focus of expertise. In the practice setting, the concept expertise in clinical judgment conveys a sense of high quality pertaining globally to a wide variety of nurses' judgments. In this study, only one judgment context was investigated. By tapping into implicit learning from experience, and demonstrating the rich networks of knowledge that experts possess, the healing time judgment task was anticipated to reveal the global nature of clinical expertise. The moderate range of accuracy scores constituted evidence that these measures did not capture as much of this global nature of expertise as was anticipated. The findings nevertheless are still useful as evidence to link particular cognitive concepts to demonstrated judgment quality.

Based on Dreyfus and Dreyfus' (1986), and Benner's (1984) framework of novice to expert, subjects included in this study seemed best categorized at the advanced beginner, competent, and proficient levels. Both true experts as well as true novices were not well represented. No generalizations can be made based on one judgment context with this small sample of subjects. Mook (1983)
argued, however, that generalization can be accomplished through replication, and that external validity must be considered in relation to the purpose of the research. In the present investigation, description and understanding, not generalization, were primary goals.

**Section C: Contribution to Theory Development**

Castellan (1993) argued that what is needed are unified theories and models that allow explanation as well as prediction in judgment and decision making. Good theories and models can lead researchers to new insights. The present study has potential to make a contribution in relation to several theoretical points.

**Linear modeling of judgment.** Brehmer (1994), in discussing the psychology of linear judgment models, stated "despite the wealth of studies showing that such models fit judgement data quite well, there has, however, been little progress in our theoretical understanding of the psychological processes that produce these data" (p. 137).

For decades, K. R. Hammond and colleagues, and Hoffman (1960, 1968) have both proposed linear models as a way to address the difficult methodological problem in the study of cognitive processes. To Hoffman, the choice was pragmatic; he asserted that the model was *paramorphic*, that is, it was not intended that cognition functioned like a linear model.

K. R. Hammond (1955) proposed a linear model for the study of clinical judgment based on theoretical principles. He was attempting to apply the general framework of probabilistic functionalism to the study of clinical judgment, as proposed by Brunswik (1955, 1956). The lens model framework was not an
arbitrary choice. It was motivated by the fact that the cues upon which judgments are based have, to some extent, *intersubstitutability*. The reason that linear models fit so well is that such models allow for such cues, and capture the capacity for vicarious functioning, a fundamental property of human cognition. To explain vicarious functioning, Hammond (1955) stated:

> Higher organisms may substitute one form of behavior for another in order to achieve a goal. . . . In the biological literature, this phenomenon has been termed *equifinality*. . . . Concerning the perception of the environment, . . . cues may substitute for one another. This phenomenon has been termed *equipotentiality* of cues. Thus, the concept of vicarious functioning refers to the variability in behavioral "output" (equifinality) and "input" (equipotentiality) of cues for an organism (p. 258).

Brehmer (1994) explained that the main reason why linear models are appropriate is that judgment tasks demand vicarious functioning and linear models capture this form of cognition.

> Brehmer (1994) provided some advice for developing a theoretical understanding of clinical judgment. First, he suggested that the use of verbal protocols in addition to linear modeling would provide a useful method for studying judgment, especially for capturing *the experience* of making judgments. Second, he recommended that psychologists adopt the Brunswikian approach of studying the interaction between the judge and the judgment task. A third recommendation was to avoid focusing analysis at the level of cue-utilization coefficients. Analysis of cue weights is too microscopic a level to establish stable empirical relations for theory development. Where the focus ought to be, Brehmer claimed, is on achievement or accuracy. "The clinician's focus is the extent to which his or her judgments agree or disagree with the distal state" (Brehmer, p. 148).
The first two suggestions have been taken in this study. An important caveat can be added to his third suggestion. The focus should be on achievement or accuracy, but, whenever possible, this ought to include comparison to an absolute measure of a criterion. Correlational measures of accuracy as demonstrated by the lens model achievement, although important, are insufficient to achieving a full understanding of the subjects' performance or the demands of the task.

**Novice-expert differences.** Patel and Groen (1991) described the progression of expertise in diagnostic reasoning in medical students and physicians. These researchers distinguished three different kinds of medical expertise: generic, specific, and domain-independent. The results related to task performance in the present study can be examined in light of these expertise categories. As practitioners become experienced, their knowledge becomes more specialized, and more difficult to employ in a generic way.

Patel and Groen (1991) concluded their paper by stating:

> Although we have a performance theory of expertise, we have no adequate description of the learning mechanism. It seems clear that theories based purely on knowledge expansion or on the development of better and better representations are inadequate (p. 122).

From an educational psychology perspective, a description of the learning mechanism is of paramount importance in all programs for professional education. In this study, evidence is presented about the significant role of differences in conceptual structure associated with expertise that has potential to add to the development of these learning theories.
Comparison of clinical and statistical prediction. For several decades a controversy in the literature has existed as to the relative accuracy of clinical methods for judgment and mechanical methods (expert systems, linear models, or other computerized approaches). Kleinmuntz (1990) framed the issue as using one's head (intuition) or using a formula (statistical or mechanical procedure) for clinical judgment.

Many authors have written about this issue over the years (Dawes, 1988, 1988; Einhorn, 1972; 1986; Goldberg, 1970; Holt, 1958; Keinmuntz, 1968; 1984; K. L. Lee et al., 1986; Meehl, 1954; J. Sawyer, 1966; Wiggins, 1981). It is obvious that human judges are required to select which variables are relevant to study in a particular situation. As early as 1972, Einhorn proposed that computers were more accurate at combining or integrating data, compared to clinicians. Even though there seems to be considerable research evidence on this point, many expert clinicians believe that their clinical predictions are superior to those made by computer.

The fact that the controversy has existed for so long indicates that there are some major philosophical foundations to the debate. Holt (1986) argued that the issue is related to the mechanist metaphysics of behaviorism. The latest resolution is to advise the use of computerized approaches where they are available and where they work best, and use human experts otherwise. Levi (1989) pointed out that expert systems ought to be more accurate because they are developed on the basis of the best experts in the field. Kleinmuntz (1990) and Holt both encouraged integration of the two approaches. The data from this study revealed that one third of the subjects made predictions where the variance explained was higher than that derived from the regression model.
In tasks where there is little predictive potential, human judges likely can out-perform a linear model, because there may be additional idiosyncratic cues available that could not be entered into any model. On the other hand, on tasks with highly valid cues, the computer will most likely perform more accurately compared to human judges because the computer applies cue-weights in a perfectly consistent manner.

**K. R. Hammond's cognitive continuum.** One example where this study has potential to add to understanding of the theoretical basis for judgment is in relation to K. R. Hammond's ideas about a cognitive continuum. Basically, Hammond (1987), Hammond et al. (1987), and Hamm (1988a) demonstrated that tasks can be located on a task continuum that ranges from analysis-inducing to intuition-inducing; these researchers recommended that, in order to better understand judgment performance, subjects should perform several tasks at a variety of locations on this continuum. The research carried out with expert engineers who judged road safety using different formats, illustrated these ideas.

The identification of tasks for their relative analysis- and intuition-inducing characteristics was carried out in this study. To add to theory development, particularly from an educational psychology perspective, the *expertise level of the subjects* also must be considered. There is considerable evidence (for example, Dreyfus and Dreyfus, 1986) that changes in cognition occur with expertise. In K. R. Hammond and colleagues' (1987) study of engineers, all of their subjects were experts. It is likely that the point on the task continuum along which a task is located is contingent on the level of expertise of the subject. Hammond found that performance was better when there was a match between the preferred means of cognition of the judge and the intuition-inducing or analysis-inducing properties of the task. By including subjects with a range of experience, there is...
potential to learn more about the cognitive changes that occur as expertise is developed.

Dreyfus and Dreyfus' (1986) claim that in making judgments, an experienced person with good performance often uses intuition where an inexperienced person employs analysis has had some support in this study. This does not mean, however, that the use of intuition will bring about expertise in judgment. Hamm (1988a) argued that without experience based on an analytic foundation, intuitive performance will be poor. "Not using rules is a privilege of the expert, not a route to becoming expert more quickly" (p. 95) [Emphasis added].

**Role of experience.** Rotter (1967) raised the question of whether it is possible for clinicians to learn from experience. He believed it was possible, providing motivation was appropriate and feedback was used properly. He claimed that some clinicians did not seem interested in learning.

It is rare that clinicians make valiant attempts to obtain systematic feedback so that they can change or improve. . . . More often they are concerned in demonstrating to others and perhaps to themselves that their clinical judgments . . . are valid (Rotter, 1967, p. 13).

Rotter maintained that motivation to prove that one is right is hardly the most appropriate incentive for discovering what one is doing wrong. This author suggested that educators ought not to try to impress students with how knowledgeable experts are, but rather with how much everyone has to learn in order to achieve reasonable prediction.

Not all the questions that could be asked about the role of experience in judgment have been answered. For example, why is it that some people require
minimal experience and others need much more experience to reach a particular level of performance? When individuals have experience with particular events and phenomena, why is it that sometimes the knowledge gained becomes generalized and abstractions are constructed, whereas, other times, the knowledge is wedded to only the particular events and phenomena encountered? More questions have been raised than have been addressed.

In this study, there is some evidence from the protocols that subjects varied in the degree of reflectiveness as they process information. Encouraging clinicians to reflect on their clinical experiences may enhance novices' ability to carry out judgments in a way that demonstrates what Schön (1983) called "reflection-on-action". Experiences that are conducive to developing reflective judgment ought to be encouraged. Considerable research has been carried out in the area of reflective judgment (King & Kitchner, 1994). Further research could be done not only linking the judgment process with the outcome, but also investigating the stage of reflective judgment as assessed by King and Kitchner, or the stage of development in terms of postformal reasoning, as studied by Arlin and Fung (1995) and Yan and Arlin (in press). It may be that increased reflectiveness has a causal role in bringing about changes to conceptual structure. It is only in reflection that people have awareness of the organization and structure of their own concepts, and the connection of such organization to successful reasoning. It may be the case that people who reach postformal reasoning are capable of reconstructing internal conceptual organization upon reflection, and that such conceptual change has potential to promote more sophisticated reasoning and judgment. Research that demonstrates causal links is of particular importance in educational psychology in order to be able to subsequently apply findings in teaching-learning contexts.
**Confidence in judgment.** Einhorn and Hogarth (1978) demonstrated that confidence in judgment persisted, despite judgment performance that is often shown to be lacking in quality. Experience does not necessarily lead to increased accuracy. These authors demonstrated how the concept "my judgment is accurate" is both *learned* and *maintained* even though judgments may be invalid. An additional point to Einhorn and Hogarth's very thorough argument can be added.

This point is based on research by Koehler (1994) and by Nelson (1996) as well as the results of this study. Judgment confidence can be considered metacognitive knowledge (knowledge about the trustworthiness of one's knowledge). One of the functions of this knowledge is to act as a cue (or signal) when individuals are at risk of making an erroneous judgment, for example when facing complex, unfamiliar tasks where the data exceed the limits of information processing.

In the course of a day, people make numerous judgments intuitively, without the need for careful attention or deliberation. Such automaticity saves cognitive effort. For people whose metacognitive knowledge includes well-calibrated judgment confidence, when a drop in confidence reaches a certain threshold, a shift from automatic intuitive processing to deliberate processing is initiated. The lowered confidence can function as a warning to interrupt intuitive processing, to pay close attention, and to analyze the situation carefully. By so doing, judgments have potential to improve. Confidence, however, is useful in this way only if it is properly calibrated to the level of one's competence.

Not many people use confidence metacognitively. Often people fail to perceive the potential value of low confidence as a signal; they learn to avoid the
negative feelings associated with low confidence by never acknowledging that it occurs. The consequence is that the signal for when to reflect and deliberately process information becomes tuned out. Without such a signal, people tend to process data intuitively and are not as open to environmental feedback: any activity that threatens the comfortable feeling of high confidence is avoided. For example, attempting to determine if past judgments were good, seeking disconfirming data, and being open to alternative perspectives, are all practices that might suggest low confidence. Engaging in these cognitive activities would mean admitting (at least to oneself) that one's judgment confidence was less than complete. It is ironic that the cognitive activities with potential to improve judgment quality (and thus increase confidence indirectly), are the very ones which are less likely to be carried out in the efforts to maintain high confidence directly.

**Section D: Implications for Education and Practice**

Several implications for professional education are derived from the literature that has been reviewed, and from the results of this study.

**Increase knowledge of the judgment process.** Faust (1986) suggested that improper judgment habits and cognitive limitations which restrict clinicians' ability to use feedback productively make learning from experience difficult. It is, therefore, important to develop approaches to increase students' and practitioners' knowledge of the judgment process, including biases and corrective measures.

Students and practitioners need to learn ways to minimize biases in judgment. By recognizing the consequences of human tendencies for seeking confirming data, failing to consider base rates, failing to ignore irrelevant data,
being overconfident, and using salience of data as an indicator of the importance of data, there is potential to increase the quality of the clinical judgments that are made. If judgments are being made intuitively, there is still merit in examining their basis and consequences in order to learn as much as possible about the factors which contribute to performance.

**Recognize trends in confidence with experience.** Practitioners may not realize that confidence in judgment tends to increase with experience, at times without any necessary relationship to increased competence. Such increasing confidence may be a consequence of the professional socialization process for students and practicing professionals; the impact on practice is that overconfidence can result which could lead to inappropriate judgments.

**Encourage judgment calibration.** Vast quantities of data are available to practitioners. Given the rapid advances in biological, psychological, and social sciences, and technology, the goal of being completely knowledgeable about all potentially relevant aspects of professional practice is clearly an impossibility. Therefore, it makes sense to advocate that practitioners aim for good discrimination between when they are capable of making quality judgments independently, and when to seek consultation and collaboration. At no previous time in history has it been so imperative that practitioners learn to differentiate between what they know and what they do not know.

**Section E: Summary**

**Summary Related to the Research Questions**

The overall research question being addressed is:
What are the patterns of relationships among measures of selected cognitive constructs (conceptual structure, sensitivity to patterns in data, and judgment process), individual difference variables (age, education, and experience), task conditions and performance in a clinical judgment task?

This study has been an examination of clinical judgment performance by 36 nurses in a probabilistic clinical judgment task. The relevant experience of the subjects ranged from 2 months to 25 years. In phase 1 of the study, 258 patients with abdominal incisions were assessed and variables that were predictive of healing time were identified. A lens model approach was employed, using representative design.

In phase 2 of the study, nurses made a variety of judgments about healing in tasks which varied in terms of their analysis- or intuition-inducing properties. The judgments were based on data from abdominal surgery patients obtained from phase 1 of the study. Policy capturing was used, and subjects' judgments were compared to the lens model equation. Subjects also made judgments about the similarity of concepts, and viewed slides of incisions, describing and evaluating what they perceived. With the use of an information board, data search strategies were revealed, and thinking that accompanied judgments about incisional healing in surgical patients was captured as subjects thought aloud.

Half of the subjects observed slides of incisions immediately prior to the lens model task. The purpose of this enhanced memory-priming condition was to attempt to reveal differences in the patterns of relations between experience and judgment expertise in novice and experienced subjects. The other subjects viewed the slides after the lens model task. When accuracy was measured by
the percent correct measure, the expected pattern of performance under enhanced memory-priming conditions was observed, but the pattern of results in the baseline memory-priming condition was not interpretable.

Based on subjects' task performance, measures of cognitive structure, sensitivity to data patterns, and judgment process were obtained. The cognitive variable which was predictive of accuracy in judgment performance was conceptual structure, assessed by applying MDS to the similarity judgments.

The influence of experience was found to be complex. Qualitative aspects of experience (experience with kinds or types of patients) likely had greater impact on performance, compared to quantity of experience, providing the quantity is beyond some minimum. Kolodner (1983) pointed out that it is not the amount of experience per se that is important, but what is learned from particular experience. Under intuitive task conditions experience had a stronger relationship with clinical judgment performance, compared to analytic task conditions.

This finding lends some support to Dreyfus and Dreyfus' theoretical ideas regarding the changes in cognition with experience. If, in this study, there had been a greater number of experts as defined by Dreyfus and Dreyfus (1986), experience may have emerged as a stronger predictor of expertise.

Performance in the lens model task was influenced by the ordering of the paragraphs. Tasks were selected on the basis of their ability to induce intuition or analysis. Performance was interpreted in light of K. R. Hammond's theory regarding a cognitive continuum, and Dreyfus and Dreyfus' (1986) theory of progression of novice to expert based on changes in cognition. In general, the differences that were observed were in the direction predicted and thus provided
support to these theories. Replication to other subjects and a greater variety of clinical judgments would be required in order to make any generalizations.

Judgment process was shown to be adaptive and constructive. Subjects with minimal experience tended to use an analytic, theory-based approach, whereas those with experience employed an intuitive, experience-based approach. Confidence in judgment tended to increase with experience, even though accuracy remained stable. Considering the high task uncertainty, lens model measures were good, but judgment accuracy assessed in absolute terms was fair to moderate.

**Recommendations for Future Research**

Based on the literature and the findings of this study, recommendations for future research can be made in the following areas:

**Issues related to confidence.** Further research is needed to determine how to best measure confidence in judgment performance, and to interpret the degree of overconfidence that actually exists. The relationship of representative design to measurement of confidence appears to be an interesting topic, not only because of practical importance, but also because of the connections to theory related to information processing and cognition generally.

**Issues related to judgment process.** Research is needed to understand the process of making judgments. Currently, this aspect of cognition, particularly with intuitive judgments, is sometimes imbued with a certain mysticism associated with expertise. Practitioners and students should discuss their thinking in relation to their judgments. Such discussion could include the data that were sought and its importance, the interpretations made and the alternative
interpretations considered, and the ways in which the information was (or could have been) combined, as well as the judgment outcome, and consequence (Arkes, 1981). If this were done, cognitive activity would become less of a "black box".

It may not be possible to recapture from the expert in explicit, formal steps the mental processes or all the elements that characterize expert clinical judgment performance. Much of the basis for intuitive judgment is implicit knowledge. Expertise often results from synthesis, not by decomposing whole situations by analytic means. It is still possible, however, to communicate some aspects of intuitive judgment. Benner (1984) recommended that expert nurses be encouraged to capture in narrative form the accomplishments and interpretative reflections on their practice. Listening to stories about paradigm cases has potential to help novice nurses gain insight into aspects of judgments, and to illuminate issues of fundamental importance in terms of professional practice.

**Information processing and connectionism.** When Brunswik initially proposed his ideas about the environment-subject interaction, and vicarious functioning, computers were extremely primitive. K. R. Hammond and colleagues used the statistical procedure of regression as a means of implementing Brunswik's ideas, not only because they believed the integrative mechanism was additive, as Brehmer (1994) claimed, but also because the choices of statistical programs were limited.

Now, with much more sophisticated computer hardware and software, researchers are using connectionist methodology for modeling judgments. For example, Brickley and Shepherd (1996) described how they trained a neural
network to make particular treatment-planning decisions that would provide reliable decision support for clinicians. There is potential for carrying out Brunswik's ideas using more modern technology. Features such as secondary cues are possible to model in a connectionist system. Research in the area of novice-expert differences in judgment performance using a combination of symbolic and connectionist approaches could prove to be more sensitive than regression to reveal changes in cognition with experience.

If, as Dreyfus and Dreyfus (1986) and Benner (1984) claim, experts excel in knowledge that is tacit, encoded in patterns, it would seem appropriate that researchers learn to incorporate new methodologies in attempting to capture the complex changes that occur in cognition.

**Future directions for judgment research.** Research needs to be conducted to tap implicit knowledge imbedded in practice. It may be that the researcher must move from the laboratory to the practice setting, so that judgments of practitioners can be studied in a more appropriate context. K. R. Hammond (1993) provided encouragement to move to a naturalistic setting, which is consistent with representative design proposed by Brunswik.

Such a change to high fidelity tasks could also be accompanied by a shift from an emphasis on quantitative designs, and statistical significance of results, to a focus on qualitative approaches and significance based on criteria that may prove more suitable for clinical judgments made by nurses, such as authenticity (Guba & Lincoln, 1989). Such a change would seem to afford excellent possibilities for enhancing understanding of clinical judgment in particular contexts. K. R. Hammond and Adelson (1976) understood judgment research in the context of human values as well as scientific aims. Referring to the prevailing
paradigms in nursing and how they are changing, Newman (1993) encouraged clarification of scientific values and methods that shape the discipline.

A move to *shift emphasis* from one perspective to another, however, should not be interpreted to mean that a quantitative approach to judgment is not useful; it seems that it is through understanding the quantitative aspects (including advantages and limitations), that one can more fully understand and appreciate the qualitative aspects. Both approaches have potential to be useful in relation to each other.

Another shift in emphasis is to move from obtaining a snapshot view of discrete, static judgments to monitoring judgments in continuous, dynamic environments over periods of time. Lusk and K. R. Hammond (1991) provided a good example where a lens model approach has been taken with weather persons' forecasts of microbursts in a dynamic environment; these researchers demonstrated how subjects changed their judgments in response to changing conditions. In addition, Hamm (1988b) provided evidence regarding dynamic variation in cognition (analysis and intuition) during task performance.

**Debate over configurality.** The subject of configurality in judgment has been a topic of debate for decades (Brannick & Brannick, 1989; Brannick & Darling, 1991; Brehmer, 1969; Edgell, 1993; Einhorn, 1970, 1971; Meehl, 1950; J. E. Sawyer, 1993). The reason that previous attempts to detect nonlinearity in clinical judgment have not succeeded, according to Ganzach (1995), is that good nonlinear models were not available. This author reanalyzed Meehl's data using a model that Ganzach referred to as a scatter model; he claimed this model provided a better fit to the data, compared to a linear model. He found patterns of nonlinearity for which he could provide psychological interpretation.
The results from the present study were not helpful in resolving the extent to which clinicians use configurality in making judgments. Further research could be carried out, reanalyzing these data, using Ganzach's (1995) scatter model.

**Teaching of judgment.** Once more research of a basic nature is carried out with respect to the theoretical aspects of clinical judgment performance, research is needed into ways to more effectively teach people to make judgments. More practical research needs to be done using methods that have validity in the settings in which they would be applied. In addition, educators need to make use of the judgment research that has already been carried out. Faust (1986) pointed out that there is considerable under-use of research on human judgment. Greater application of such research is needed to prepare students and practitioners for the cognitive challenges that are inherent in professional practice. In any professional education program, helping students to learn to make good judgments is an important and demanding goal. Faust maintained that "there is far more to know than what is known about methods for improving judgment. Continuing research on human judgment . . . can produce knowledge that helps clinicians better serve their clients" (p. 428).

Hammond and Adelman (1976) stated that "the key element . . . in the process of integrating social values and scientific facts is human judgment" (p. 389). There are distinct advantages in being able to make wise judgments. Arlin (1993) discussed the importance of wisdom and expertise in teaching; she described wisdom as entailing good judgment, manifesting itself in planning, managing, and reflecting on teaching and learning activities. Her points are relevant to nursing and other professions. In no other time in history has the teaching of judgment been of greater importance. From an evolutionary perspective, survival of the species involves groups of people (communities)
making wise judgments in relation to all important aspects of life such as health and safety, education, food production, and issues related to the environment. Hogarth (1980) stated "in the not so distant past human survival and progress depended on physical skills. There can be little doubt that the need today is for conceptual skills, that is, the ability to process information and make judgments" (p. 3).

Conclusion

The problem identified at the beginning was the finding that many studies in the literature showed no relationship between experience and expertise in judgment. To some extent, this lack of relationship may be accounted for by claiming that there actually is a relationship, but it is difficult to demonstrate in the laboratory. Such factors as the use of laboratory studies with nonrepresentative designs, failure to consider changes in cognition with expertise, and the use of tasks that have little ecological validity constitute reasons for this difficulty. On the other hand, Einhorn and Hogarth (1978), Dawes (1989), and Brehmer (1980b) explained the lack of relationship found in many studies by claiming that there is truly little relationship to be found. These authors demonstrated that factors such as the difficulty of learning in probabilistic contexts and the increase in confidence in one's judgment with experience tend to create the illusion of validity. Based on the findings in this study, a synthesis of these two explanations is proposed: Although people tend to increase in their ability to make judgments as they gain experience, standard research conditions usually do not reveal this, and because confidence increases considerably, the self-perceived level of expertise in a particular judgment context (revealed by confidence) may be greater than what is valid when compared to absolute indicators of judgment quality.
The initial question raised was to determine the extent to which experience and clinical judgment accuracy were related; this question has provided good direction to this study. A number of interesting patterns of relations were found between experience and various indicators of expertise in clinical judgment performance.

Experience was related to judgment performance under intuitive condition; in a comparable task under analytic conditions, experience was not related to judgment performance. Performance differences associated with task condition may contribute to the explanation for previous results of a lack of a relationship between experience and judgment performance. As well, such differences in performance associated with task characteristics lend support to Dreyfus and Dreyfus' (1986) model for the development of expertise.
REFERENCES


Huberty, C. J. (1993). Historical origins of statistical testing practices: The
treatment of Fisher versus Neyman-Pearson views in textbooks. Journal of
Experimental Education, 61, 317-333.


Book Medical Publishers.

experienced registered nurses and student nurses. Journal of Nursing Education,
28, 120-126.

the assessment and management of cardiac pain. Advances in Nursing Science, 15,
54-63.

Jacoby, L. L., & Brooks, L. R. (1984). Nonanalytic cognition: Memory,
perception, and concept learning. The Psychology of Learning and Motivation:
Advances in Research and Theory, 18, 1-47.


assessments: Data-based versus theory-based judgments. In D. Kahneman, P.
Slovic, & A. Tversky (Eds.), Judgments under uncertainty: Heuristics and biases

and process. In M. T. H. Chi, R. Glaser, & M. J. Farr (Eds.), The nature of expertise

nonlinear models. Organizational Behavior and Human Performance, 24, 60-66.

Multimethod study of clinical judgment. Organizational Behavior and Human
Performance, 30, 201-230.


Justlin, P. (1994). The overconfidence phenomenon as a consequence of
informal experimenter-guided selection of almanac items. Organizational Behavior
and Human Decision Processes, 57, 226-246.


Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? Organizational Behavior and Human Performance, 20, 159-183.


APPENDIX A: GLOSSARY OF TERMS

- **Additive processing**: An information processing strategy for combining data into a judgment where cues are integrated additive manner; the value of one cue does not influence the clinician's perception of importance of other cues.

- **Bias in judgment**: Systematic error in judgment performance attributed to perceptual processes, or information processing strategies.

- **Calibration of confidence**: The degree to which one's confidence in judgment corresponds to one's accuracy in judgment.

- **Clinical ecology**: The particular clinical environment or context in which a clinician makes judgments.

- **Confidence in judgment**: The strength of belief (or sense of degree of certainty) that one's judgments are accurate.

- **Configural processing**: A information processing strategy for combining data into a judgment where cues are integrated interactively; the value of one cue (or set of cues) influences a clinician's perception of the importance of other cues.

- **Expertise in clinical judgment**: Clinical judgment that is demonstrated (within the context of the research task) to have a high level of accuracy or
validity. Expertise is not assumed simply on the basis of experience or credentials.

- **Heuristics**: Information-processing strategies often used to reduce cognitive strain in judgment contexts; these efficient "shortcuts" can be functional in some situations, but they may also be associated with systematic biases.

- **Ill-structured problems**: Problems in which there is little consensus regarding the appropriate solution; they include open constraints that are resolved in the course of solution.

- **Judgment calibration**: The extent to which one's judgment performance corresponds to externally verifiable criteria.

- **Policy capturing**: A technique aimed to obtain a mathematical representation of the clinician's judgment rule or strategy for weighting and combining cues. Each individual policy, expressed in equation form, is a summary of outcome data based on responses the clinician makes to the cues.

- **Task uncertainty**: The range of possibilities of outcomes, or other parameters, that are perceived as likely in a given task or judgment context.
APPENDIX B: EXAMPLES OF LENS MODEL CASES

Examples of Normal order cases:

**EFGH** is a 19 year old waitress who has had an emergency appendectomy. She weighs 68.2 Kg (150 lbs) and is 157 cm tall (5' 2''). She is otherwise healthy, and takes no medications, except birth control pills.

**EFGH** has surgery that lasted 52 min, with minimal blood loss. She had no drains. The operation resulted a small incision in her lower right abdomen, which was stapled. She had no complications.

* * * * * *

**YZAB** is a 58 year old restaurant owner; he has had a right hemicolectomy for colon cancer. He weighs 108 Kg (238 lbs) and is 182 cm tall (6' 1''). He has had diabetes for 20 years, renal failure, UTIs, anemia, and heart disease for which he takes insulin, gentamycin, iron, and digoxin.

**YZAB** had surgery that lasted 133 minutes, with 1000 cc blood loss, requiring 2 Units of PRBC. He had no drains. The operation resulted a long midline incision which was stapled. He experienced a large wound dehiscence.
Examples of Reverse order cases:

LNPR has surgery that lasted 151 minutes with 200 cc blood loss. She has no drains. She has a small midline incision that is stapled. She has no complications.

LNPR is a 31 year old technician who has had a partial gastrectomy because of trauma. She weighs 74.8 Kg (165 lbs) and is 168 cm (5' 8'') tall; she is healthy.

* * * * * *

ADGJ has surgery that lasts 65 minutes with minimal blood loss. He has 1 drain. He has a small, transverse incision in the upper right quadrant. He had no complications.

ADGJ is an 82 year old retired construction worker. He has had an open cholecystectomy for chronic cholecystitis and stones. He has had by-pass surgery and has a pace maker. He also has stomach ulcers and has had prostate surgery.
APPENDIX C: TASK PRESENTATION SEQUENCE

<table>
<thead>
<tr>
<th>Task:</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
<th>6th</th>
<th>7th</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experience Level 1 n=12</td>
<td>A n=12</td>
<td>I n=12</td>
<td>S n=6</td>
<td>N n=3</td>
<td>R n=3</td>
<td>O n=12</td>
<td>CS n=12</td>
</tr>
<tr>
<td>Experience Level 2 n=12</td>
<td>A n=12</td>
<td>I n=12</td>
<td>S n=4</td>
<td>N n=2</td>
<td>R n=2</td>
<td>O n=12</td>
<td>CS n=12</td>
</tr>
<tr>
<td>Experience Level 3 n=12</td>
<td>A n=12</td>
<td>I n=12</td>
<td>S n=8</td>
<td>N n=3</td>
<td>R n=3</td>
<td>O n=12</td>
<td>CS n=12</td>
</tr>
</tbody>
</table>

Code: Association Task = A; Information Board Task = I; Slide Task = S; Lens Model Task (Normal Order) = N; Lens Model Task (Reverse Order) = R; Orthogonal Judgment Task = O; Conceptual Structure Task = CS; Cue Importance Task = CI
APPENDIX D

KNOWLEDGE ACCESSIBILITY SCALE

1) PERCEPTUAL KNOWLEDGE  Give 1 point each (maximum of 5)

- Montgomery tie tapes
- Adipose tissue
- Dark area on stoma
- Curve in incision around umbilicus
- no staples in open incision
- yellow skin tone for nephrectomy patient
- granulation tissue
- flange
- colostomy appliance

2) SPECIFIC ITEM KNOWLEDGE  Give 0 - 3 points each

(i) Bruising

- No mention of bruising
- Describes a colour difference, but can’t interpret
- Can identify a bruise/hematoma
- Comments on implications of bruising

(ii) Drain

- No mention of drain
- Identifies a drain (generic)
- Refers to a Penrose drain
- Comments on implication of the drain.

(iii) Absorbable sutures

- No mention of these sutures.
- Detects a texture difference at the periphery, but does not interpret this.
- Identifies absorbable sutures, but does not evaluate.
- Identifies and evaluate these sutures.

(iv) Rod supporting loop colostomy

- No mention of the rod.
- Points out rod, without naming it.
- Identifies the rod.
- Identifies the implication of the rod.
3) RELATIONAL KNOWLEDGE  Give 1 point for each aspect

(i) Knowledge of spatial relations
- Identifies relative length of incisions.
- Identifies width and depth of incisions.
- Identifies location in relation to the abdomen.
- Identifies spatial relations of particular features.

(ii) Knowledge of functional relations (Drain and suture)
- Mentions a drain.
- Mentions a suture.
- Identifies the relation between the drain and suture.
- Comments about this relation.

(iii) Knowledge of physiological relations (Open incision)
- Describes features of wound bed
- Describes drainage
- Describes skin surfaces
- Evaluates healing by secondary intention.

(iv) Knowledge of causal relations (Midline incision with staples)
- Describes incision
- Describes staples
- Identifies relation between staples and incision.
- Comments about this relation.

(v) Knowledge of relations between data and evaluation (Ileostomy stoma, or any example)
- Identifies positive evidence
- Identifies absence of negative evidence
- Indicates what additional data would be relevant
- Makes evaluation

(iv) Knowledge of practical or nursing management relations (Colostomy and open incision)
- Describes stoma
- Describes incision
- Identifies relation between stoma and incision
- Comments about this relation

4) CONTEXTUAL KNOWLEDGE  Give 1 point for each aspect

Knowledge of the healing trajectory (Open incision)
- Refers to healing as a trajectory
- Qualifies trajectory for context
- Qualifies trajectory for interaction of context and risk factors
- Extrapolates trajectory into the future
APPENDIX E

CONCEPTUAL COHERENCE SCALE

1. The protocol contains evidence that the subject is familiar with a variety of clinical concepts relevant to wound healing:

A) The subject refers to cues from the cards used for this task, suggesting what is considered salient and being attend to.

S 28: "So, this is a patient with a red incision and thick drainage at the lower end".

S 28:: "This is a big incision that's packed with saline compresses, and there's no staples or sutures".

B) The subject refers to additional relevant assessments [not included on the stimulus card].

S 13: "I'd want to know their temperature right away, and how they're generally feeling. I'd want to know how it looked yesterday".

S 09: "What I'd do is check his lab values".

C) The subject refers to risk factors.

S 06: "He has a dehiscence where his incision line opens up... . He could be diabetic, too".

S 25: "The people at risk are usually... . big beer-belly type guys".

D) The subject use terms from the Similarity Judgment Task (either by word, or synonym, or description). All verbal protocols were included in this assessment.

The concept terms are: primary intention, secondary intention, approximated, granulation tissue, pus, abscess, adipose, retention suture, steri-strips, hematoma, erythema, wound bed, oozing, inflammation, and necrotic.

S 08: "I remember one patient who had a transverse incision. She was healing okay and then, you know, one of the ends started to come open... . She had an abscess underneath... . She was in for more than a month with that".
2. The protocol contains evidence that the subject has sensitivity to knowledge limitations associated with the type of prior experiences he or she has had:

A) The subject indicates no experience in a particular situation.

S22: "I've had no experience with this type of situation".

S08: "I have never had a patient come back from the OR with an incision that was packed".

B) The subject expresses a lack of understanding in relation to a situation:

S17: "I'm not sure what I have here. . . . There's not really enough to tell me". [Said in response to the patient returning from the OR with no staples or sutures].

C) The subject indicates what help would be needed to interpret the situation.

S05: "I would find out . . . whether the staples need to be taken out". [Said in response to the patient with drainage].

S15: "I'd probably call the surgeon or a nurse to clarify" [said about the patient with the incision healing by secondary intention].

3. The protocol contains evidence that the subject uses a variety conceptual links:

A) The subject relates cues to a problem or potential problem.

S25: "I would guess that it was an infection just by the odorous drainage".

B) The subject makes reasonable interpretations of situations.

S15: "I'd question if there's possibly a hematoma that has broken underneath".

S29: "This incision has been left open for a reason. . . . its been left open to make sure that it heals from the inside out".

C) The subject identifies possible implications.

S14: "You'd have to take time to constantly be there putting dry dressings on or find some other creative way to catch that drainage before it does something to the skin".

S19: "This can be quite scary for the patient".

D) The subject states reasonable possibilities of what could (or should) be done.

S35: "I'd want to send a C & S [culture and sensitivity] of the drainage". [Said about the patient with the wound drainage].
S 22: "They take a long time to work with too because you have to be so careful and you have to be gentle and you have to talk to the patient a lot because they're scared". [Said about the patient with an incision healing by secondary intention].

S 21: "Usually just have them sit, lean back, and call a doctor to come in because it's really quite a serious problem. . . . I would always stay with them, because emotionally this is very traumatic". [Said about the patient with wound dehiscence].

4. The protocols contain evidence that the subject's concepts have structure: (All protocols are included in this assessment).

A) The subject uses a variety of super-ordinate terms.

[Examples of such terms include: Complications, risk factors, chronic health problems, immunosuppressants]

B) The subject uses a variety of highly specific idea units.

S 26: "I'd look for increased white blood cell count, and particularly for a shift in differential."

S 32: "Fifteen milligrams of prednisone is not a huge amount."

C) The subject's knowledge is structured in terms of practical knowledge or rules.

S 13: "When the incision is left open, that means lots of work doing wet dressings. They're complex dressings, you know, they take a lot of time to do".  

S 05: "Sometimes the packing has to be wet with more saline in order to take it out because it becomes dry and it would tear the new cells that are forming on the sides".

D) The subject's knowledge is structured in terms of particular clinical experiences.

S 20: "Just a couple of weeks age, I took someone's staples out and steri-stripped it and the steri-strips strips actually wouldn't stick to him and they came off and the incision opened up".

E) The subject has knowledge that is structured in terms of a clinical story:

S 35: "This happened just two weeks ago. The patient originally had cardiac surgery and then subsequently was sick and septic, went back to the OR and had a colostomy and bowel resection as well. [At first], the wound was doing well. I was helping with the patient and she eviscerated. I had never seen a true evisceration before that wasn't trauma induced and yeah, there was bowel all over the place! . . . None of us knew immediately what to do other than to throw wet saline packs on it, and get someone in there stat [immediately]. She had to go back to the OR. It was awful".
## Appendix F

### Stress and R-Squared for Multidimensional Scaling of Similarity Judgments

<table>
<thead>
<tr>
<th>Subject</th>
<th>Dimension 1</th>
<th>Dimension 2</th>
<th>Dimension 3</th>
<th>Dimension 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Stress</td>
<td>$R^2$</td>
<td>Stress</td>
<td>$R^2$</td>
</tr>
<tr>
<td>1</td>
<td>.21</td>
<td>.82</td>
<td>.13</td>
<td>.91</td>
</tr>
<tr>
<td>2</td>
<td>.24</td>
<td>.88</td>
<td>.16</td>
<td>.91</td>
</tr>
<tr>
<td>3</td>
<td>.37</td>
<td>.66</td>
<td>.22</td>
<td>.75</td>
</tr>
<tr>
<td>4</td>
<td>.42</td>
<td>.55</td>
<td>.22</td>
<td>.70</td>
</tr>
<tr>
<td>5</td>
<td>.32</td>
<td>.72</td>
<td>.19</td>
<td>.83</td>
</tr>
<tr>
<td>6</td>
<td>.15</td>
<td>.94</td>
<td>.02</td>
<td>.99</td>
</tr>
<tr>
<td>7</td>
<td>.47</td>
<td>.51</td>
<td>.21</td>
<td>.77</td>
</tr>
<tr>
<td>8</td>
<td>.35</td>
<td>.64</td>
<td>.21</td>
<td>.77</td>
</tr>
<tr>
<td>9</td>
<td>.36</td>
<td>.62</td>
<td>.19</td>
<td>.77</td>
</tr>
<tr>
<td>10</td>
<td>.26</td>
<td>.80</td>
<td>.15</td>
<td>.89</td>
</tr>
<tr>
<td>11</td>
<td>.40</td>
<td>.56</td>
<td>.21</td>
<td>.76</td>
</tr>
<tr>
<td>12</td>
<td>.33</td>
<td>.68</td>
<td>.17</td>
<td>.85</td>
</tr>
<tr>
<td>13</td>
<td>.41</td>
<td>.46</td>
<td>.23</td>
<td>.67</td>
</tr>
<tr>
<td>14</td>
<td>.35</td>
<td>.64</td>
<td>.19</td>
<td>.78</td>
</tr>
<tr>
<td>15</td>
<td>.25</td>
<td>.81</td>
<td>.16</td>
<td>.87</td>
</tr>
<tr>
<td>16</td>
<td>.47</td>
<td>.31</td>
<td>.24</td>
<td>.62</td>
</tr>
<tr>
<td>17</td>
<td>.33</td>
<td>.70</td>
<td>.18</td>
<td>.82</td>
</tr>
<tr>
<td>18</td>
<td>.39</td>
<td>.52</td>
<td>.24</td>
<td>.64</td>
</tr>
<tr>
<td>19</td>
<td>.27</td>
<td>.76</td>
<td>.14</td>
<td>.89</td>
</tr>
<tr>
<td>20</td>
<td>.30</td>
<td>.74</td>
<td>.13</td>
<td>.91</td>
</tr>
<tr>
<td>21</td>
<td>.38</td>
<td>.56</td>
<td>.22</td>
<td>.70</td>
</tr>
<tr>
<td>22</td>
<td>.43</td>
<td>.54</td>
<td>.24</td>
<td>.68</td>
</tr>
<tr>
<td>23</td>
<td>.34</td>
<td>.65</td>
<td>.18</td>
<td>.81</td>
</tr>
</tbody>
</table>
Appendix F (Continued)

Stress and R-Squared for Multidimensional Scaling of Similarity Judgments

<table>
<thead>
<tr>
<th>Subject</th>
<th>1 Stress</th>
<th>1 R²</th>
<th>2 Stress</th>
<th>2 R²</th>
<th>3 Stress</th>
<th>3 R²</th>
<th>4 Stress</th>
<th>4 R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>24</td>
<td>.25</td>
<td>.82</td>
<td>.09</td>
<td>.95</td>
<td>.03</td>
<td>.99</td>
<td>.01</td>
<td>1.00</td>
</tr>
<tr>
<td>25</td>
<td>.39</td>
<td>.56</td>
<td>.24</td>
<td>.67</td>
<td>.14</td>
<td>.82</td>
<td>.08</td>
<td>.90</td>
</tr>
<tr>
<td>26</td>
<td>.43</td>
<td>.46</td>
<td>.23</td>
<td>.67</td>
<td>.14</td>
<td>.80</td>
<td>.09</td>
<td>.88</td>
</tr>
<tr>
<td>27</td>
<td>.29</td>
<td>.74</td>
<td>.17</td>
<td>.84</td>
<td>.10</td>
<td>.90</td>
<td>.05</td>
<td>.97</td>
</tr>
<tr>
<td>28</td>
<td>.37</td>
<td>.58</td>
<td>.21</td>
<td>.73</td>
<td>.12</td>
<td>.86</td>
<td>.09</td>
<td>.91</td>
</tr>
<tr>
<td>29</td>
<td>.30</td>
<td>.74</td>
<td>.18</td>
<td>.83</td>
<td>.06</td>
<td>.96</td>
<td>.03</td>
<td>.99</td>
</tr>
<tr>
<td>30</td>
<td>.42</td>
<td>.45</td>
<td>.26</td>
<td>.60</td>
<td>.14</td>
<td>.77</td>
<td>.10</td>
<td>.82</td>
</tr>
<tr>
<td>31</td>
<td>.31</td>
<td>.72</td>
<td>.19</td>
<td>.79</td>
<td>.08</td>
<td>.94</td>
<td>.04</td>
<td>.97</td>
</tr>
<tr>
<td>32</td>
<td>.28</td>
<td>.78</td>
<td>.11</td>
<td>.93</td>
<td>.04</td>
<td>.99</td>
<td>.01</td>
<td>1.00</td>
</tr>
<tr>
<td>33</td>
<td>.43</td>
<td>.59</td>
<td>.24</td>
<td>.68</td>
<td>.13</td>
<td>.82</td>
<td>.09</td>
<td>.90</td>
</tr>
<tr>
<td>34</td>
<td>.14</td>
<td>.94</td>
<td>.09</td>
<td>.95</td>
<td>.05</td>
<td>.98</td>
<td>.03</td>
<td>.99</td>
</tr>
<tr>
<td>35</td>
<td>.39</td>
<td>.52</td>
<td>.22</td>
<td>.71</td>
<td>.12</td>
<td>.88</td>
<td>.07</td>
<td>.93</td>
</tr>
<tr>
<td>36</td>
<td>.23</td>
<td>.84</td>
<td>.17</td>
<td>.86</td>
<td>.10</td>
<td>.92</td>
<td>.06</td>
<td>.97</td>
</tr>
</tbody>
</table>

Note. The first 12 subjects were from experience level one; Subjects 13 to 24 were from experience level two; Subjects 25 to 36 were from experience level three.
APPENDIX G

LENS MODEL AND ACCURACY MEASURES FOR NORMAL ORDER AND REVERSE ORDER CASES

<table>
<thead>
<tr>
<th>Experience Level 1</th>
<th>$r_a$</th>
<th>$r_a^2$</th>
<th>$G$</th>
<th>$C$</th>
<th>$\Sigma-d$</th>
<th>% Correct</th>
<th>$\bar{x}$ Error Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Subject</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>01</td>
<td>57/56</td>
<td>62/40</td>
<td>71/66</td>
<td>13/43</td>
<td>44/43</td>
<td>38/22</td>
<td>88/85</td>
</tr>
<tr>
<td>02</td>
<td>35/58</td>
<td>37/58</td>
<td>74/82</td>
<td>-34/11</td>
<td>35/25</td>
<td>52/48</td>
<td>88/89</td>
</tr>
<tr>
<td>03</td>
<td>68/83</td>
<td>74/80</td>
<td>88/92</td>
<td>06/55</td>
<td>18/13</td>
<td>48/50</td>
<td>90/92</td>
</tr>
<tr>
<td>04</td>
<td>64/65</td>
<td>76/44</td>
<td>81/80</td>
<td>05/44</td>
<td>30/27</td>
<td>10/10</td>
<td>77/84</td>
</tr>
<tr>
<td>05</td>
<td>68/64</td>
<td>43/22</td>
<td>86/77</td>
<td>31/53</td>
<td>20/30</td>
<td>28/43</td>
<td>87/91</td>
</tr>
<tr>
<td>06</td>
<td>75/63</td>
<td>78/68</td>
<td>85/74</td>
<td>37/40</td>
<td>23/36</td>
<td>40/38</td>
<td>91/90</td>
</tr>
<tr>
<td>07</td>
<td>33/37</td>
<td>50/13</td>
<td>63/53</td>
<td>-35/31</td>
<td>52/54</td>
<td>30/55</td>
<td>88/87</td>
</tr>
<tr>
<td>08</td>
<td>56/76</td>
<td>68/41</td>
<td>83/88</td>
<td>-28/48</td>
<td>25/17</td>
<td>50/50</td>
<td>89/93</td>
</tr>
<tr>
<td>09</td>
<td>50/73</td>
<td>72/75</td>
<td>60/85</td>
<td>14/33</td>
<td>63/23</td>
<td>43/35</td>
<td>87/86</td>
</tr>
<tr>
<td>10</td>
<td>55/75</td>
<td>60/40</td>
<td>86/80</td>
<td>-25/70</td>
<td>21/27</td>
<td>53/45</td>
<td>89/89</td>
</tr>
<tr>
<td>11</td>
<td>77/74</td>
<td>79/73</td>
<td>90/95</td>
<td>29/08</td>
<td>15/07</td>
<td>50/53</td>
<td>92/90</td>
</tr>
<tr>
<td>12</td>
<td>60/76</td>
<td>73/75</td>
<td>78/85</td>
<td>09/54</td>
<td>33/21</td>
<td>40/63</td>
<td>87/93</td>
</tr>
</tbody>
</table>

Note.

The first number in each column refers to normal order cases; following the slash, the second number refers to reverse order cases. All decimals have been removed. Correlational measures have not been transformed.
APPENDIX G (CONTINUED)

LENS MODEL AND ACCURACY MEASURES FOR NORMAL ORDER AND REVERSE ORDER CASES

<table>
<thead>
<tr>
<th>Experience Level 2</th>
<th>ra</th>
<th>Ra²</th>
<th>G</th>
<th>C</th>
<th>Σ-d</th>
<th>% Correct</th>
<th>Error Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>23/62</td>
<td>40/70</td>
<td>42/67</td>
<td>-17/47</td>
<td>79/50</td>
<td>23/30</td>
<td>88/90</td>
</tr>
<tr>
<td>14</td>
<td>41/54</td>
<td>66/51</td>
<td>62/83</td>
<td>-21/-04</td>
<td>57/24</td>
<td>45/35</td>
<td>87/88</td>
</tr>
<tr>
<td>15</td>
<td>50/72</td>
<td>53/46</td>
<td>77/87</td>
<td>-22/42</td>
<td>34/18</td>
<td>45/35</td>
<td>89/90</td>
</tr>
<tr>
<td>16</td>
<td>44/40</td>
<td>67/56</td>
<td>68/53</td>
<td>-20/09</td>
<td>47/66</td>
<td>35/40</td>
<td>89/88</td>
</tr>
<tr>
<td>17</td>
<td>42/53</td>
<td>60/18</td>
<td>54/85</td>
<td>02/10</td>
<td>69/21</td>
<td>33/40</td>
<td>85/88</td>
</tr>
<tr>
<td>18</td>
<td>59/50</td>
<td>63/56</td>
<td>87/76</td>
<td>-12/-06</td>
<td>19/33</td>
<td>35/43</td>
<td>90/89</td>
</tr>
<tr>
<td>19</td>
<td>55/72</td>
<td>81/79</td>
<td>74/89</td>
<td>-30/42</td>
<td>43/15</td>
<td>35/38</td>
<td>82/84</td>
</tr>
<tr>
<td>20</td>
<td>85/94</td>
<td>78/63</td>
<td>94/96</td>
<td>52/87</td>
<td>10/06</td>
<td>55/40</td>
<td>91/93</td>
</tr>
<tr>
<td>21</td>
<td>61/83</td>
<td>68/54</td>
<td>77/93</td>
<td>17/59</td>
<td>35/10</td>
<td>45/48</td>
<td>89/94</td>
</tr>
<tr>
<td>22</td>
<td>57/89</td>
<td>60/54</td>
<td>69/95</td>
<td>15/78</td>
<td>48/07</td>
<td>33/33</td>
<td>88/92</td>
</tr>
<tr>
<td>23</td>
<td>41/75</td>
<td>42/59</td>
<td>83/88</td>
<td>-28/44</td>
<td>24/17</td>
<td>53/38</td>
<td>87/88</td>
</tr>
<tr>
<td>24</td>
<td>73/72</td>
<td>76/46</td>
<td>92/88</td>
<td>14/44</td>
<td>12/17</td>
<td>28/18</td>
<td>79/77</td>
</tr>
</tbody>
</table>

Note.
The first number in each column refers to normal order cases; following the slash, the second number refers to reverse order cases. All decimals have been removed. Correlational measures have not been transformed.
APPENDIX G (CONTINUED)
LENS MODEL AND ACCURACY MEASURES FOR NORMAL ORDER AND REVERSE ORDER CASES

<table>
<thead>
<tr>
<th>Experience Level 3</th>
<th>( r_a )</th>
<th>( R_a^2 )</th>
<th>( G )</th>
<th>( C )</th>
<th>( \Sigma-d )</th>
<th>% Correct</th>
<th>( \bar{X} ) Error Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject 25</td>
<td>45/84</td>
<td>51/78</td>
<td>75/92</td>
<td>-19/57</td>
<td>35/13</td>
<td>43/60</td>
<td>85/93</td>
</tr>
<tr>
<td>26</td>
<td>54/65</td>
<td>65/66</td>
<td>67/84</td>
<td>17/28</td>
<td>50/22</td>
<td>53/48</td>
<td>88/89</td>
</tr>
<tr>
<td>27</td>
<td>56/79</td>
<td>63/68</td>
<td>80/88</td>
<td>-14/60</td>
<td>31/17</td>
<td>43/33</td>
<td>85/85</td>
</tr>
<tr>
<td>28</td>
<td>69/69</td>
<td>74/58</td>
<td>87/85</td>
<td>-04/34</td>
<td>21/21</td>
<td>55/45</td>
<td>90/91</td>
</tr>
<tr>
<td>29</td>
<td>65/52</td>
<td>71/44</td>
<td>78/67</td>
<td>08/30</td>
<td>37/43</td>
<td>48/33</td>
<td>86/87</td>
</tr>
<tr>
<td>30</td>
<td>55/54</td>
<td>72/65</td>
<td>76/68</td>
<td>-20/15</td>
<td>37/48</td>
<td>30/38</td>
<td>89/90</td>
</tr>
<tr>
<td>31</td>
<td>64/69</td>
<td>50/56</td>
<td>80/90</td>
<td>24/20</td>
<td>29/14</td>
<td>43/35</td>
<td>85/86</td>
</tr>
<tr>
<td>32</td>
<td>77/56</td>
<td>64/66</td>
<td>95/75</td>
<td>17/13</td>
<td>08/35</td>
<td>55/58</td>
<td>91/88</td>
</tr>
<tr>
<td>33</td>
<td>78/69</td>
<td>77/74</td>
<td>92/89</td>
<td>27/18</td>
<td>13/16</td>
<td>58/43</td>
<td>89/88</td>
</tr>
<tr>
<td>34</td>
<td>56/74</td>
<td>56/74</td>
<td>89/82</td>
<td>-16/50</td>
<td>16/24</td>
<td>25/43</td>
<td>90/92</td>
</tr>
<tr>
<td>35</td>
<td>47/75</td>
<td>47/62</td>
<td>78/88</td>
<td>-19/48</td>
<td>30/17</td>
<td>43/43</td>
<td>88/92</td>
</tr>
<tr>
<td>36</td>
<td>81/82</td>
<td>72/68</td>
<td>92/88</td>
<td>42/67</td>
<td>13/17</td>
<td>45/30</td>
<td>89/89</td>
</tr>
</tbody>
</table>

Note.
The first number in each column refers to normal order cases; following the slash, the second number refers to reverse order cases. All decimals have been removed. Correlational measures have not been transformed.
APPENDIX H

REGRESSION COEFFICIENTS (STANDARDIZED) FOR NORMAL AND REVERSE ORDER CASES

<table>
<thead>
<tr>
<th>Name of Cue</th>
<th>01</th>
<th>02</th>
<th>03</th>
<th>04</th>
<th>05</th>
<th>06</th>
<th>07</th>
<th>08</th>
<th>09</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incision Approx</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incision Length</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complications</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Body Mass Index</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Urgency</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>History</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drains</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Occupation Index</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blood loss</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Severity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Surgery time</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transfusions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Decimals have been removed. Numbers before the slash refer to normal order cases; numbers after the slash refer to reverse order cases. Dots indicate no coefficients.
APPENDIX H (CONTINUED)
REGRESSION COEFFICIENTS (STANDARDIZED) FOR NORMAL AND REVERSE ORDER CASES

<table>
<thead>
<tr>
<th>Name of Cue</th>
<th>Subject ID</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
</tr>
<tr>
<td>Incision Approx</td>
<td>25/52</td>
</tr>
<tr>
<td>Incision Length</td>
<td>55/</td>
</tr>
<tr>
<td>Complications</td>
<td>...</td>
</tr>
<tr>
<td>Body Mass Index</td>
<td>42/</td>
</tr>
<tr>
<td>Urgency</td>
<td>/30</td>
</tr>
<tr>
<td>History</td>
<td>...</td>
</tr>
<tr>
<td>Drains</td>
<td>-32/</td>
</tr>
<tr>
<td>Occupation Index</td>
<td>...</td>
</tr>
<tr>
<td>Blood loss</td>
<td>...</td>
</tr>
<tr>
<td>Severity</td>
<td>...</td>
</tr>
<tr>
<td>Surgery time</td>
<td>...</td>
</tr>
<tr>
<td>Age</td>
<td>...</td>
</tr>
<tr>
<td>Transfusions</td>
<td>...</td>
</tr>
</tbody>
</table>

Note. Decimals have been removed. Numbers before the slash refer to normal order cases; numbers after the slash refer to reverse order cases. Dots indicate no coefficients.
APPENDIX H (CONTINUED)

REGRESSION COEFFICIENTS (STANDARDIZED) FOR NORMAL AND REVERSE ORDER CASES

<table>
<thead>
<tr>
<th>Name of Cue</th>
<th>SUBJECT ID</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>19</td>
</tr>
<tr>
<td>Incision Approx</td>
<td>/49</td>
</tr>
<tr>
<td>Incision Length</td>
<td>/49</td>
</tr>
<tr>
<td>Complications</td>
<td>64/36</td>
</tr>
<tr>
<td>Body Mass Index</td>
<td>20/23</td>
</tr>
<tr>
<td>Urgency</td>
<td>/31</td>
</tr>
<tr>
<td>History</td>
<td>51/</td>
</tr>
<tr>
<td>Drains</td>
<td>/49</td>
</tr>
<tr>
<td>Occupation Index</td>
<td>...</td>
</tr>
<tr>
<td>Blood loss</td>
<td>...</td>
</tr>
<tr>
<td>Severity</td>
<td>...</td>
</tr>
<tr>
<td>Surgery time</td>
<td>29/</td>
</tr>
<tr>
<td>Age</td>
<td>-24/</td>
</tr>
<tr>
<td>Transfusions</td>
<td>...</td>
</tr>
</tbody>
</table>

Note. Decimals have been removed. Numbers before the slash refer to normal order cases; numbers after the slash refer to reverse order cases. Dots indicate no coefficients.
APPENDIX H (CONTINUED)

REGRESSION COEFFICIENTS (STANDARDIZED) FOR NORMAL AND REVERSE ORDER CASES

<table>
<thead>
<tr>
<th>Name of Cue</th>
<th>28</th>
<th>29</th>
<th>30</th>
<th>31</th>
<th>32</th>
<th>33</th>
<th>34</th>
<th>35</th>
<th>36</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incision Approx</td>
<td>25/</td>
<td>24/</td>
<td></td>
<td>/52</td>
<td>46/43</td>
<td>59/65</td>
<td>25/41</td>
<td></td>
<td>/37</td>
</tr>
<tr>
<td>Incision Length</td>
<td>44/38</td>
<td>58/</td>
<td></td>
<td></td>
<td></td>
<td>48/30</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complications</td>
<td>32/</td>
<td>/44</td>
<td>42/67</td>
<td></td>
<td>45/50</td>
<td></td>
<td>57/55</td>
<td>60/28</td>
<td>49/34</td>
</tr>
<tr>
<td>Body Mass Index</td>
<td>37/69</td>
<td></td>
<td>51/44</td>
<td></td>
<td>43/40</td>
<td></td>
<td>/44</td>
<td>27/48</td>
<td></td>
</tr>
<tr>
<td>Urgency</td>
<td></td>
<td></td>
<td></td>
<td>/25</td>
<td></td>
<td></td>
<td>/25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>History</td>
<td></td>
<td></td>
<td>51/</td>
<td></td>
<td>37/</td>
<td></td>
<td>35/</td>
<td>37/</td>
<td>47/</td>
</tr>
<tr>
<td>Drains</td>
<td></td>
<td></td>
<td>51/</td>
<td></td>
<td>37/</td>
<td></td>
<td>35/</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Occupation Index</td>
<td></td>
<td></td>
<td></td>
<td>/52</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blood loss</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>/29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Severity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>47/</td>
<td></td>
</tr>
<tr>
<td>Surgery time</td>
<td></td>
<td>/46</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-/37</td>
<td>-25/</td>
</tr>
<tr>
<td>Transfusions</td>
<td></td>
<td>34/</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note.** Decimals have been removed. Numbers before the slash refer to normal order cases; numbers after the slash refer to reverse order cases. Dots indicate no coefficients.
APPENDIX I

RESULTS OF INFORMATION BOARD TASK:
DATA SEARCH STRATEGY FOR 1ST, 2ND, & 3RD SET OF 8 CARDS,
TOTAL # OF CARDS USED, AND ACCURACY

<table>
<thead>
<tr>
<th>ID</th>
<th>Search Strategy for Cards</th>
<th>Total # of cards used</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>Intra* Intra Mixed</td>
<td>24</td>
<td>Correct</td>
</tr>
<tr>
<td>02</td>
<td>Inter** Mixed Mixed</td>
<td>24</td>
<td>Incorrect</td>
</tr>
<tr>
<td>03</td>
<td>Intra Mixed</td>
<td>19</td>
<td>Correct</td>
</tr>
<tr>
<td>04</td>
<td>Mixed Mixed</td>
<td>19</td>
<td>Correct</td>
</tr>
<tr>
<td>05</td>
<td>Intra Intra Mixed</td>
<td>20</td>
<td>Correct</td>
</tr>
<tr>
<td>06</td>
<td>Intra Inter Mixed</td>
<td>23</td>
<td>Correct</td>
</tr>
<tr>
<td>07</td>
<td>Inter Inter Mixed</td>
<td>24</td>
<td>Correct</td>
</tr>
<tr>
<td>08</td>
<td>Intra Intra Intra</td>
<td>24</td>
<td>Incorrect</td>
</tr>
<tr>
<td>09</td>
<td>Intra Intra Intra</td>
<td>20</td>
<td>Correct</td>
</tr>
<tr>
<td>10</td>
<td>Mixed Inter Mixed</td>
<td>23</td>
<td>Incorrect</td>
</tr>
<tr>
<td>11</td>
<td>Intra Intra Mixed</td>
<td>24</td>
<td>Correct</td>
</tr>
<tr>
<td>12</td>
<td>Intra Intra</td>
<td>18</td>
<td>Incorrect</td>
</tr>
<tr>
<td>13</td>
<td>Mixed Inter</td>
<td>21</td>
<td>Incorrect</td>
</tr>
<tr>
<td>14</td>
<td>Intra Intra Mixed</td>
<td>23</td>
<td>Correct</td>
</tr>
<tr>
<td>15</td>
<td>Inter Inter</td>
<td>19</td>
<td>Correct</td>
</tr>
<tr>
<td>16</td>
<td>Intra Mixed Mixed</td>
<td>22</td>
<td>Incorrect</td>
</tr>
<tr>
<td>17</td>
<td>Inter Mixed Mixed</td>
<td>23</td>
<td>Correct</td>
</tr>
<tr>
<td>18</td>
<td>Intra Mixed Mixed</td>
<td>21</td>
<td>Correct</td>
</tr>
<tr>
<td>19</td>
<td>Inter Inter</td>
<td>20</td>
<td>Incorrect</td>
</tr>
</tbody>
</table>
RESULTS OF INFORMATION BOARD TASK:
DATA SEARCH STRATEGY FOR 1ST, 2ND, & 3RD SET OF 8 CARDS,
TOTAL # OF CARDS USED, AND ACCURACY

<table>
<thead>
<tr>
<th>ID</th>
<th>Search Strategy for Cards</th>
<th>Total # of cards used</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>Intra* Intra</td>
<td>17</td>
<td>Correct</td>
</tr>
<tr>
<td>21</td>
<td>Intra Intra</td>
<td>16</td>
<td>Correct</td>
</tr>
<tr>
<td>22</td>
<td>Inter** Inter Inter</td>
<td>23</td>
<td>Correct</td>
</tr>
<tr>
<td>23</td>
<td>Inter Inter</td>
<td>17</td>
<td>Correct</td>
</tr>
<tr>
<td>24</td>
<td>Inter Inter Mixed</td>
<td>22</td>
<td>Correct</td>
</tr>
<tr>
<td>25</td>
<td>Intra Mixed</td>
<td>15</td>
<td>Correct</td>
</tr>
<tr>
<td>26</td>
<td>Intra Intra Intra</td>
<td>20</td>
<td>Correct</td>
</tr>
<tr>
<td>27</td>
<td>Intra Intra Intra</td>
<td>24</td>
<td>Incorrect</td>
</tr>
<tr>
<td>28</td>
<td>Intra Intra</td>
<td>20</td>
<td>Correct</td>
</tr>
<tr>
<td>29</td>
<td>Intra Mixed</td>
<td>15</td>
<td>Correct</td>
</tr>
<tr>
<td>30</td>
<td>Intra Intra</td>
<td>20</td>
<td>Correct</td>
</tr>
<tr>
<td>31</td>
<td>Intra Mixed Intra</td>
<td>24</td>
<td>Correct</td>
</tr>
<tr>
<td>32</td>
<td>Intra Intra</td>
<td>16</td>
<td>Correct</td>
</tr>
<tr>
<td>33</td>
<td>Inter Inter</td>
<td>20</td>
<td>Correct</td>
</tr>
<tr>
<td>34</td>
<td>Inter Inter Mixed</td>
<td>23</td>
<td>Correct</td>
</tr>
<tr>
<td>35</td>
<td>Intra Mixed Mixed</td>
<td>22</td>
<td>Correct</td>
</tr>
<tr>
<td>36</td>
<td>Inter Mixed</td>
<td>16</td>
<td>Correct</td>
</tr>
</tbody>
</table>

* Intra = Intradimensional Search Strategy
** Inter = Interdimensional Search Strategy
APPENDIX J: CODING CATEGORIES AND EXAMPLES
TO USE WITH PROTOCOLS DERIVED FROM INFORMATION BOARD TASK

A) METACOGNITIVE AND OTHER ORGANIZING/EXECUTIVE FUNCTIONS

CODE 1: Planning

S 25: "The first thing I want to know about my patient is what surgery they've had, so I am going to turn over all four surgical cards ..."

S 16: "Let's see what diagnoses they have."

CODE 12: Using some type of strategy

S 12: "Vertical mid-line incision. OK, I'm just trying to visualize." [Visualization]

S 28: "Two of these people have not had any complications, so I am not going to be worrying about them right now." [Elimination by aspects]

S 25: "I want to know their relevant medical history so I can get an overall picture." [Constructing a context]

CODE 14: Recognizing missing data, data not needed, missing or unclear connections

S 26: "One hundred eight Kg. It doesn't say how tall."

S 07: "I don't care about occupation."

S 22: "I don't really want to know the incision data because I have an idea what that looks like."

S 28: "I am not sure why this patient has swollen thighs; at this point I can't see the connection between the thighs and the incision."

CODE 16: Expressing interest in/curiosity about obtaining data that may have salience

S 26: "I can't stand it. I need to know what she has had done, too. ... I like to know everything, like what did they do, what type of personality do they have, ..."

S 19: "Well, now I'm kind of curious ... age and gender. Okay. So, it's a 27 year old female."

CODE 13: Attending to, or otherwise being responsive to particular data that are available, generating a sense of salience

S 21: [In response to reading the Medical history card for patient B] "Oh! My goodness. I've got to watch him for awhile."

S 19: "OK, that's interesting ... swollen thighs ... Pretty serious."
CODE 4: Using additional aspects of metacognition

S 04: "I am trying to figure out which of these cards will give the most information."

S 17: "I didn't check that she didn't have any complications."

S 17: "I'm not surprised." [Said in response to revealing that the 27 year old patient who had the oophorectomy had no complications. By inference, these data suggest that this subject had expectations which were matched by the information obtained.]

B) INPUT GENERATION: MAKING DATA AVAILABLE IN WORKING MEMORY

CODE 2: Making correct inferences/evaluative statements

S 04: "Maybe he is on some anticoagulant like heparin." [referring to a patient with deep vein thrombosis]

S 26: "The 80 year old has hypertension; well, she's obviously on antihypertensives."

CODE 6: Generating hypotheses or possibilities

S 10: "Stomach hyperacidity. ... It could be ulcers."

CODE 9: Having relevant declarative knowledge become activated

S 09: "Maybe they have diabetes or something that would [make healing] take longer."

CODE 11: Having knowledge of clinical patterns derived from experience become activated

S 09: "I'm going to check the surgery [card], because sometimes it [healing] depends on what they had [done]."

S 33: "She just had a Burch repair, which isn't that uncommon for that age."

CODE 7B: Obtaining data as a result of operations to rule in or rule out hypotheses

S 29: Obtains relevant medical history after generating the following hypothesis: "It's sort of not usual for a 58 year old to have infection, so there must be something else going on."

CODE 3: Making incompletely-connected or seemingly irrelevant inferences/evaluative statements

S 06: "This patient will have dumping syndrome I think because the bowel was removed."

S 16: "Now this lady here, ... maybe is living in a nursing home."
C) CARRYING OUT COGNITIVE OPERATIONS ON DATA IN WORKING MEMORY

CODE 5: Connecting two or more pieces of data, or understanding connections that previously were unclear

S 20: "Younger people heal up a lot faster ..."

S 21: "Circulation problems, renal failure, diabetes ... that all slows down healing."

S 28: "Patient A is on prednisone and has a DVT [deep vein thrombosis] -- the DVT takes care of the swollen thighs."

CODE 7A: Seeking data to rule in or rule out hypotheses

S 28: "Just for fun I would say the patient with ulcerative colitis is the one that will take the longest to heal, but I can't predict that yet; I've got some more information to take a look at first. For both, I need the medical history. Okay, Oh! That makes a big difference."

CODE 8: Referring to the goal of the task, or relating data to the goal

S 26: "If the question is healing time, then to me ..."

S 28: "This woman with the Burch repair, may have a lot of problems with voiding, getting her catheter out, ... but in terms of the incision, she's going to be fine."

CODE 11: Summarizing data

S 26: "Patient A -- 108 Kg, that's heavy ... and he is diabetic and has renal failure ..."

CODE 15: Recognizing the dynamic and probabilistic nature of the data

S 12: "I'll look at the vital signs ... These are on admission; I've seen vital signs change so much."

S 25: "He is on prednisone. That isn't necessarily going to cause infection ..."

S 27: "Also, of course, it would depend on how long he's been on it [prednisone]."

CODE 17: Weighing the relative importance of data

S 36: "This 80 year old female with hyperacidity, stress incontinence, hypertension -- age is against her, but she's relatively healthy given her age."

S 07: "The 80 year old has no infection at all, but she is old."

CODE 18: Making conclusions or judgments relative to the goal of the task [One point for all]

S 12: "I'd say the 27 year old female would heal the fastest."

S 17: "Lastly, I would choose patient B, because of his heart and kidney problems, as well as diabetes, and he has complications which could be quite serious."