INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.
In search of optimal human-expert system explanations: Empirical studies of human-human and human-expert system interactions

Halgren, Shannon Lee, Ph.D.
Rice University, 1993
RICE UNIVERSITY

IN SEARCH OF OPTIMAL HUMAN-EXPERT SYSTEM EXPLANATIONS:
EMPIRICAL STUDIES OF HUMAN-HUMAN AND HUMAN-EXPERT SYSTEM
INTERACTIONS

by

SHANNON LEE HALGREN

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE
DOCTOR OF PHILOSOPHY

APPROVED, THESIS COMMITTEE

Nancy J. Cooke
Nancy J. Cooke, Ph.D., Director
Assistant Professor, Department of Psychology

David Lane, Ph.D., Chair
Associate Professor, Department of Psychology

Ken Laughery, Ph.D.
Professor, Department of Psychology

Henry Roediger, Ph.D.
Professor, Department of Psychology

Ian Walker, Ph.D.
Assistant Professor,
Department of Electrical Engineering

Houston, Texas
April, 1993
IN SEARCH OF OPTIMAL HUMAN-EXPERT SYSTEM EXPLANATIONS:
EMPIRICAL STUDIES OF HUMAN-HUMAN AND HUMAN-EXPERT
SYSTEM INTERACTIONS

by

Shannon Lee Halgren

Abstract

In this project explanations were studied along a continuum ranging from
human-human interactions to human-expert system interactions with the goal of
identifying features of successful expert system explanations. The project
consisted of five distinct phases or steps: a) defining what a successful
explanation entails, b) observing human-human explanation and formulating
hypotheses about the features of successful explanations, c) testing hypotheses
formulated in step b, d) extending results to an expert system domain and
testing again, and e) from this empirical data, formulating recommendations for
expert system explanation designers. The progressive nature of this study
allowed conclusions to be drawn about both human-human and human-expert
system interactions and the role explanations play in these exchanges. The
most salient conclusion drawn from these studies was that explanatory
interactions are complex and explanation success is dependent on more than
just features of the explanations involved. Individual differences such as an
explanation recipient's initial abilities and their participation level in the
interaction influence their understanding and performance as much, if not more
so, than explanation features. Consistently subjects' participation level
interacted with explanation content level. Individuals who are active participants in interactions with an expert perform better when given explanations with low levels of content, whereas passive participants benefit from explanations with high levels of content. Overall, an active participation level increases performance and understanding in human-human interactions, but this result does not generalize to human-expert system interactions where an active participation style is detrimental to performance. This and other inconsistencies between human-human and human-expert system interactions are discussed as well as the advantages of the research approach employed in this project. Finally, recommendations based on the results of these studies are provided for expert system explanation designers.
Acknowledgments

A mere mention in the acknowledgments section of this dissertation does not describe my gratefulness to my committee director, Nancy Cooke. Throughout this project, and my graduate school career, Nancy has become a respected advisor, mentor, and friend. Her input on this project was invaluable. A special thanks must also go to David Lane who willingly took over the role of committee chair upon Nancy’s leave from Rice. Our discussions about statistics and methodology strengthened this dissertation a great deal. I would also like to thank my committee members, Henry Roediger, Ken Laughery, and Ian Walker for the contributions and support. Each was able to add valuable comments and criticisms. I am also extremely grateful to Roger Schvaneveldt and the psychology department at New Mexico State University for allowing me to temporarily join their department while I finished this project. They generously allowed the use of their lab space, equipment, and subject pool, and made me feel welcome. A special thanks must go to members of the cognitive engineering lab for allowing me to tie up their lab computer and lab space for several weeks while I ran subjects. Love and thanks must also go to my parents who were always available for support and encouragement even during the worst of times. Finally, I'd like to thank Tracy, my best friend, colleague, and sister, for making my last few months of graduate school a very special time.
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Tables</td>
<td>iii</td>
</tr>
<tr>
<td>List of Figures</td>
<td>iv</td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Expert Systems</td>
<td>1</td>
</tr>
<tr>
<td>Expert System Explanations</td>
<td>6</td>
</tr>
<tr>
<td>Human-Human Explanation</td>
<td>15</td>
</tr>
<tr>
<td>Research Project</td>
<td>23</td>
</tr>
<tr>
<td>Definition of a Good Explanation</td>
<td>24</td>
</tr>
<tr>
<td>Experiment 1</td>
<td>66</td>
</tr>
<tr>
<td>Method</td>
<td>27</td>
</tr>
<tr>
<td>Results and Discussion</td>
<td>34</td>
</tr>
<tr>
<td>Summary</td>
<td>41</td>
</tr>
<tr>
<td>Experiment 2</td>
<td>41</td>
</tr>
<tr>
<td>Method</td>
<td>44</td>
</tr>
<tr>
<td>Results and Discussion</td>
<td>44</td>
</tr>
<tr>
<td>Experiment 3</td>
<td>47</td>
</tr>
<tr>
<td>Method</td>
<td>49</td>
</tr>
<tr>
<td>Results and Discussion</td>
<td>53</td>
</tr>
<tr>
<td>General Discussion</td>
<td>57</td>
</tr>
<tr>
<td>Recommendations for Expert System Explanations</td>
<td>63</td>
</tr>
<tr>
<td>References</td>
<td>65</td>
</tr>
</tbody>
</table>
List of Tables

<table>
<thead>
<tr>
<th>Tables</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1</td>
<td>133</td>
</tr>
<tr>
<td>Table 2</td>
<td>134</td>
</tr>
<tr>
<td>Table 3</td>
<td>135</td>
</tr>
<tr>
<td>Table 4</td>
<td>136</td>
</tr>
<tr>
<td>Table 5</td>
<td>138</td>
</tr>
<tr>
<td>Table 6</td>
<td>139</td>
</tr>
<tr>
<td>Table 7</td>
<td>141</td>
</tr>
<tr>
<td>Table 8</td>
<td>142</td>
</tr>
<tr>
<td>Table 9</td>
<td>143</td>
</tr>
<tr>
<td>Table 10</td>
<td>145</td>
</tr>
</tbody>
</table>
# List of Figures

<table>
<thead>
<tr>
<th>Figures</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>147</td>
</tr>
<tr>
<td>Figure 2</td>
<td>148</td>
</tr>
<tr>
<td>Figure 3</td>
<td>149</td>
</tr>
<tr>
<td>Figure 4</td>
<td>150</td>
</tr>
<tr>
<td>Figure 5</td>
<td>152</td>
</tr>
<tr>
<td>Figure 6</td>
<td>153</td>
</tr>
<tr>
<td>Figure 7</td>
<td>155</td>
</tr>
<tr>
<td>Figure 8</td>
<td>158</td>
</tr>
<tr>
<td>Figure 9</td>
<td>159</td>
</tr>
<tr>
<td>Figure 10</td>
<td>160</td>
</tr>
<tr>
<td>Figure 11</td>
<td>161</td>
</tr>
</tbody>
</table>
INTRODUCTION

Expert computer systems that aid humans in decision making and problem solving have the potential to be extremely valuable across a wide range of domains, however, their success is dependent in part on their ability to explain their actions to the user. There seems to be wide agreement in the literature on several facts: a) few, if any, expert systems have succeeded in providing "successful" explanations to all their users across all their tasks, b) a major reason that the explanation components of expert systems have not been entirely successful is because what makes a "successful" system explanation has yet to be empirically determined, c) researchers from the domains of philosophy, social psychology, cognitive science, and education who study explanation (either human-human explanation or human-expert system explanation) can offer some useful insights and data that address the issue of successful explanations, but d) more empirical work needs to be done.

This paper discusses each of these ideas in detail and presents research directed toward the further study of human explanation (specifically, determining the components of a successful human explanation). An introduction to expert systems and their explanation components is presented below.

Expert Systems

Several types of expert systems exist that incorporate numerous purposes and applications. Madni (1988) has classified existing expert systems into several groups based on their type, purpose, application, and intended user. This classification is depicted in Table 1. According to Madni, two broad
types of expert systems exist: autonomous expert systems and expert consultation systems.

Autonomous expert systems are stand-alone systems that entirely replace the human in the performance of a task (e.g., robots). Modesitt (1987) has argued that these are the only type of system that should be called "expert systems" because these systems, contrary to others, "attain or surpass human expertise." Only 60 to 120 of these "real" expert systems exist, most of which are unavailable to the general public.

Table 1
Expert System Classification from the Human Factors Prospective (modified from Madni, 1988).

<table>
<thead>
<tr>
<th>Type</th>
<th>Purpose</th>
<th>Application</th>
<th>User</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autonomous</td>
<td>Perform Task</td>
<td>Diagnosis</td>
<td>-</td>
</tr>
<tr>
<td>Expert Systems</td>
<td>Planning</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Expert Consultation</td>
<td>Assist in Task</td>
<td>Diagnosis</td>
<td>Novice through</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Expert</td>
</tr>
<tr>
<td>Teach Task</td>
<td>Diagnosis</td>
<td>Novice or</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intermediate</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Planning</td>
<td>Novice or</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intermediate</td>
<td></td>
</tr>
</tbody>
</table>
The second, more popular, type of expert system is called an expert consultation system. This type constitutes a much larger, broader group of systems. Expert consultation systems either provide assistance with certain tasks, teach the task to users, or they may do both. Systems providing task assistance help the user make decisions by one of two methods. When the user's task is to incorporate several pieces of information into a decision, the system can help by making the important information more salient by adjusting its format and display (Hollnagel, 1990). Systems that provide this type of assistance are called decision support systems (DSS). Systems also provide task assistance by providing advice about the decision or task. For example, consultation or second opinion systems critique the users' conclusions or answers questions similar to a human colleague.

Teaching systems are designed to teach domain knowledge and to test users' skill against some preset criterion of performance. This grouping of systems also includes "what-if" systems, which answer questions about hypothetical situations, a valuable teaching technique. Intelligent tutoring systems (ITT) or intelligent computer-assisted instruction (ICAI) systems should also be included in the teaching systems group. These systems normally provide some sort of on-line workbook and have the ability to pinpoint user misconceptions and adapt the lesson to correct these "bugs" (Dede, 1986).

The three expert system purposes (task performance, task assistance, and teaching the task) can be broken down further by considering their potential applications. Regardless of the system's purpose, it might be applicable to a diagnostic task involving analytical problem solving, or to a planning task
involving synthetic problem solving. The consultation systems can also be calculated by considering the intended user of the system. Task assistance systems (for diagnosis or planning) can be designed for any level of user expertise, novice through expert. Systems built to teach tasks, on the other hand, are normally not intended for experts who already have a firm grasp of the task domain, but for domain novices or intermediates.

The remainder of this paper will use the term "expert system" loosely to describe all the intelligent systems discussed above. Modesitt (1987) has argued for the term knowledge-based system (KBS) to be used for any system that uses heuristic knowledge captured from intelligent sources, including expert systems and all the systems listed above. While KBS is more descriptive for reference to general intelligent systems because it is more broad, convention will be followed in this paper and the term "expert system" will be used instead.

Although expert systems' architecture changes somewhat with the system type, purpose and application, most expert systems have similar architecture. A common expert system architecture discussed by Lachman (1989) is presented in Figure 1.

The user interface gathers information from the user and presents system output. Most developers strive to use natural language in their interfaces as opposed to a programming language or computer-ease foreign to most end users. However, as discussed by Lachman, there is a great deal of work to be done before technology can support a truly natural language interface that does not depend on unnatural canned text.
The knowledge base is the storage area for the system's productions or rules, structured objects (e.g., frames and semantic networks), and propositional logic. This storage area is different from the database which contains the domain information, inferred domain facts and system status information. The inference engine comprises a set of algorithms responsible for determining the information needed to solve a problem and initiating the decision process.

Finally, the support software is responsible for utility and support functions such as rule editing, graphics support, and file location and reformatting (Lachman, 1989). Some systems concerned with user acceptance also include an explanation generator to create explanations which justify the system's advice to the user (Neches, Swartout & Moore, 1985).

Advantages of Expert Systems
Advances in artificial intelligence have given expert systems the ability to out-perform humans on several tasks and to carry out some tasks that humans are incapable of performing (Modesitt, 1987). For example, tasks requiring rapid computation of higher-order math problems or the immediate detection of an abnormal system state from a complex display are much easier for artificial intelligence to handle. In addition to the superior performance expert systems can provide, there are numerous other advantages of using artificial over human expertise as well. Waterman (1986) has defined five of the most significant advantages of machine expertise:

1) **Permanence:** A machine's knowledge is permanent and not dependent on its use. Human knowledge can be easily forgotten when it is not rehearsed or practiced for prolonged periods of time. Additionally, human experts have a limited mental (and physical!) life.

2) **Transferability:** Transferring machine knowledge to another machine is trivial and quick. Transferring human knowledge to another human is a lengthy, difficult (if not impossible in some cases), and expensive process.

3) **Documentation:** Machine knowledge is easy to document. It merely requires determining the mapping between the expertise representation in the system and the natural language description of that representation. It is difficult and time-consuming to document human knowledge.

4) **Consistency:** A machine given the same problem will consistently provide the same solution, whereas human performance can be affected by current emotions, stress level, the weather, or by several other unpredictable internal or external factors.
5) **Cost**: Top notch human experts are scarce and can demand high salaries. By contrast expert systems are relatively inexpensive. The high cost of system development is offset by the low operational costs and the ability to make additional copies of the program cheaply.

**Limitations of Expert Systems**

Whereas expert systems have several advantages over human experts, they also have limitations. These limitations can be broken down into three broad areas: a) expert systems have a limited application, b) their success depends on user acceptance, and c) poor interfaces and explanation components limit expert systems' appeal. The first of these limitations is probably the most severe. According to Modesitt (1988), intelligent systems should only be used when the problem domain supports the use of an intelligent system. Examples of these domains include situations where the task is primarily cognitive, the task requires no common sense, human experts are scarce, or the task is routinely taught to humans (this would ensure that the domain is simple enough to be broken down into if-then rules). It can be a dangerous temptation to overuse and therefore misuse expert system technology. Significant technological advances will have to be made before expert systems can move into a wider range of domains.

The second significant limitation of expert systems, user acceptance, can be affected by several factors such as a lack of understanding about how the system works, a distrust in the system's ability, or poor usability (these factors are probably causally related (Muir, 1987)). Regardless of the reason for low
user acceptance, the result is the same: rejection of the expert system
(Shortliffe, 1982; Teach & Shortliffe, 1981).

It is easy for a user to wonder, "If expert systems are so intelligent, why
am I needed to operate them?" The answer to this question should point out
that artificial intelligence still has a long way to go before it can replace human
intelligence. Humans have several abilities and skills machines have not yet
been able to master. These shortcomings play a significant role in the low
acceptance of and trust in artificial advice. Waterman (1986) has also
summarized the main limitations of artificial intelligence:

1) **Creativity**: Humans are creative, innovative, imaginative and have the
ability to synthesize new information by merely reorganizing old
information. Even the most intelligent systems are uninspired and routine.

2) **Adaptability**: Humans can adapt to changing situations by quickly shifting
their strategies in accordance to the changing environment. Machines
have difficulty learning new rules and concepts without human
intervention.

3) **Sensory Input**: The complex sensory experience humans can gather from
their environment can be directly assimilated and used in problem solving,
whereas this information must be translated into a symbolic format in order
for a computer to use it. A significant amount of information and time might
be lost in this translation.

4) **Broad Focus**: Humans can focus on "the big picture" of a problem,
considering all the related concepts and influences that might relate to the
central issue. Machines, however, tend to only concentrate on the problem at hand, ignoring issues not central to the problem.

5) **Commonsense**: The ability to realize what can be known and what cannot, as well as realistic limitations, gives humans a significant advantage over machines that may waste a large amount of time looking for a nonexistent answer.

Most of these shortcomings are readily noticed in interactions with expert systems, as is the necessity of the human component in the problem-solving loop. Attempts to totally eliminate humans from these tasks would eventually lead to an inefficient and brittle system (Roth & Woods, 1989).

User acceptance is also affected by the third limitation of expert systems; their generally poor interface and user support (Kidd & Cooper, 1985). Only a few systems reach practical utility because the user is considered only half-heartedly in system design and evaluation. Usually expert system literature has been more concerned with the internal function of the machine than with the user (Gordon, 1987; Hendler & Lewis, 1988). This trend is not surprising when one considers the origin of expert systems.

Expert system designers are typically software engineers who are less likely to be aware of human factors issues than others. Fortunately, interface problems are less severe today than they were when expert systems were first developed. This trend can be attributed to the increased awareness of usability testing benefits (Berry & Hart, 1990a).

Probably the most important area of system usability is the presence of ample explanation facilities. Expert system explanations help the user
understand the system and its advice. A growing number of researchers in artificial intelligence agree that system explanations need to be more than just a statement of the system rules that fired during the decision-making process, or a solution with justification (Berry & Broadbent, 1987; Hendler & Lewis, 1988; Langoltz & Shortliffe, 1989; Woods, Roth, & Bennett, 1990). Rather, explanations should be structured around the problem-solving process. They should involve close cooperation with the user in formulating the problem and in identifying and evaluating the solution paths (Roth & Woods, 1989). Expert system explanation are also of considerable importance to system acceptance (Diederich, 1992) and therefore, deserve an in depth discussion.

Expert System Explanations

An explanation component is provided in expert systems primarily to facilitate users' understanding by educating them about aspects of the domain or the system that are most relevant to their goals. Tasks and goals that users bring to the expert system are varied and have changed drastically throughout the history of expert systems.

The explanation components available in early expert systems were included to assist the developers of the system. At that time, developers were the primary system users due to the complexity of expert systems. Explanations were used by these individuals to perform two important tasks: system testing and program debugging. Developers needed to confirm that the mechanisms employed in the derivation of the system advice were correct (Jackson, 1986). When a problem or failure occurred, they needed the explanation to make the system's actions transparent, so bugs could be quickly identified (Buchanan &
Shortliffe, 1984; Wallis & Shortliffe, 1984). Both of these purposes could be satisfied by simply listing a rule trace. The rule trace displayed each rule that the system fired during its problem solving activity (generally listed by the rule number) and the order of rule firing. Rule-trace explanations provided developers with the information they needed to perform their tasks efficiently, but the usefulness of this type of coded explanation was limited to these specific users.

When advances in technology and artificial intelligence moved expert systems into several new domains involving non-programmer users, these limited explanations were no longer sufficient. Expert systems were being built to be used for a variety of purposes such as for advice-seeking and consultation. Users of advice-giving systems referenced explanations for completely different reasons than did the developers. Most importantly, these users needed explanations to provide assurance that the reasoning behind the system advice was correct, logical, and appropriate for the particular problem that the user was solving (Jackson, 1986; Wallis & Shortliffe, 1984). This might sometimes mean that the explanation needed to persuade the user that unexpected advice was appropriate and should be taken (Buchanan & Shortliffe, 1984; Scott, Clancey, Davis, & Shortliffe, 1984; Wallis & Shortliffe, 1984). In short, the explanation was needed to help users understand the system's functioning so they could accept or reject its advice, whichever was appropriate (Erdman, 1987).

While the implicit goal of the developers and advice-seeking users was to learn about the system, some systems were developed solely to teach
(Buchanan & Shortliffe, 1984; Wallis & Shortliffe, 1984). Explanations in these systems focused on providing the users with the strategies that the expert system was using to solve problems. It was assumed that if users could understand the system’s strategies, they should be able to go off and apply these rules on their own. Providing explicit explanations would allow users to incorporate information into their own knowledge base, develop a maximum understanding of the problem (Berry & Broadbent, 1987), and help them remember the information longer (Clancey, 1983).

The Role of Explanations in User Acceptance

After considering the important purposes explanations serve in human-expert system interaction, it is not surprising that physicians (typical expert system users) rate this as the most important feature of an expert system (Teach & Shortliffe, 1981). Providing users with adequate information about the domain and the system via an explanation has several positive side effects such as facilitating effective system use and creating positive user attitudes towards the system. A hypothesized causal path connecting the expert system’s explanation and users’ attitudes towards the system is shown in Figure 2. Several steps along this path have been suggested by others, but no empirical work has been done to support these ideas. The pathway takes the following course:

1. A good explanation promotes an accurate conceptual model.

Regardless of the user’s task, the explanation component should have a common goal- to educate the user about the system and its domain. This information forms the user’s conceptual model of the system and the task at
Figure 2. Hypothesized Causal Pathway Between an Expert System Explanation and Effective System Use.
hand. To achieve this end, the system in some cases needs to show users how it makes its decisions (Erdman, 1987) and to give them access to as much system knowledge (e.g., the rules in the knowledge base or the line of reasoning used) as necessary (Buchanan & Shortliffe, 1984; Scott, Clancey, Davis, & Shortliffe, 1984). Providing users with the reasoning behind system advice and information about how the system reached its decision should facilitate their understanding of the system (Buchanan & Shortliffe, 1984; Kidd & Cooper, 1985). This understanding will be enhanced if the system also communicates its boundaries, limits, and side effects (Hayes & Reddy, 1983; Kidd & Cooper, 1985; Roth, Bennett, & Woods, 1987).

2. An accurate conceptual model is a prerequisite for accurate calibration of trust. If users can establish an accurate conceptual model of the system through the explanation, they are more likely to appropriately calibrate their confidence or trust in the system (Buchanan & Shortliffe, 1984; Berry & Hart, 1990b). According to Muir (1987), appropriate understanding of the system allows users to assess the machine expert's responsibility. They then use this system information to calibrate their trust in the system. Several other factors probably enter into a user's trust calibration such as past experiences with the expert system or attitudes towards computers and artificial intelligence (labeled "Other Knowledge" in Figure 2). All else being equal, however, users with an appropriate conceptual model should allocate trust to the system more accurately than those with inappropriate conceptual models.
3. **Accurate trust calibration leads to effective interaction.** By calibrating their trust appropriately, users are more likely to use the system effectively (Moray & Lee, 1991). Effective system use means knowing when to allow the system to make the decisions, thereby freeing up time for the user to concentrate on other tasks, and when to distrust the system and override its advice or take over the decision making task (Muir, 1987). Often, inappropriate levels of trust cause users to misuse or misinterpret the advice. For example, some users who lack an understanding about how the system operates can become "mystified" by the system and put too much trust or "blind faith" into the system (Muir, 1987; Sheridan, 1980). Trust in the system can also increase after using a system for an extended time. In this situation, users lose their own skills and rely on the system to do all the decision making tasks (Hollnagel, 1990). These situations may result in incorrect decision to accept the system's advice. Ineffective system use can also cause users to incorrectly reject the system's advice, or to put more of their own cognitive effort into the task than necessary, or to make less than optimal decisions. Again, other factors such as the system's usability and the user's experience with the system probably also affect efficient system use. But, when these other factors are held constant, it is assumed that those who have accurate levels of trust in the system will have more effective interactions.

According to Zuboff (1988), if a sufficient level of trust is established, allowing efficient system use, user acceptance of the system will normally follow (Moray & Lee, 1991). A low level of trust, however, will usually lead to system rejection (Hendler & Lewis, 1988; Muir, 1987). For example, experts, such as
medical doctors typically distrust artificial advice and are very cautious about using expert systems to assist in patient diagnosis. Doctors are keenly aware that they are held accountable for their treatment advice whether it was derived from their own thought processes or from artificial thought processes performed by an expert system (Jagodziniki & Holmes, 1989). Doctors in a position in which they must demand that expert system's provide ample, explicit explanations describing how the advice is formulated and providing the knowledge upon which the advice is based.

The challenge of providing an explanation that promotes understanding, and therefore efficient system use and user acceptance, is made difficult by the fact that the same explanation does not always lead separate users to form the same level of trust in the system. Factors such as the user's prior experience with expert systems (Ellis, 1989) and level of domain expertise (Erdman, 1987) seem to affect how well an explanation helps the user make efficient use of the system. In a study by Erdman (1987), subject matter experts (psychiatrists) and non-experts (physicians) were asked to use a clinical-disorder diagnosis/consultation system to help them diagnose and treat eight fictional cases. After subjects had identified which treatment they thought was appropriate for each case, the consultation system provided its treatment advice. Half of the subjects received an explanation for this advice, and half did not. The presence of explanations significantly affected the experts' treatment decisions and increased the non-experts' attitudes toward the computer system. Other significant interactions caused Erdman to conclude that the non-expert group may have been inappropriately influenced by system explanations (i.e.,
explanations made this group less dubious of the system advice even when it was mistaken), but the expert group was able to use the explanations in appropriate ways. Thus, system explanations influenced both groups' judgments, but not always in the appropriate ways.

Clearly, the explanation component of expert systems is critical. Explanations have the ability to assist user judgments appropriately, as well as the ability to hinder accurate judgments. Results like these along with findings that users do not always want to see or use explanations (Berry & Broadbent, 1987; Berry & Hart, 1990b) are causing several researchers to realize the complexity of providing useful explanations. Most researchers agree that good or effective explanations can be valuable to the user, but they also realize the mere presence of an explanation does not guarantee that users will incorporate them into their decisions or will be helped rather then hindered by them. Critical thought must be put into the development of system explanation facilities in order to generate useful and effective explanations.

**Explanation Guidelines**

The unanswered questions about the characteristics of useful explanations are not the result of a lack of armchair intuition. The literature is full of ideas about what makes a good explanation and other explanation do's and don'ts. Because few of these ideas have been supported by data, however, it is difficult to distinguish the true characteristics of good explanations from the merely intuitive characteristics. Nonetheless, reviewing researchers recommendations is an excellent way to overview what characteristics might influence an explanation's usefulness.
A common theme running through researcher's recommendations about expert system explanations is that they should mimic human-human explanations as closely as possible (Goguen, Weiner, & Linde, 1983). For instance, some feel systems should allow for iterative interactions with the explanation component (Cawsey, 1988; Putnam & Duffy, 1984; Teach & Shortliffe, 1981). It has also been suggested that system explanations should be goal driven as they are when a person generates an explanation to rationalize an event in their environment (Leake, 1991). This requires that the system determines the user's goals. Once done, the system could anticipate the user's questions and canned text could be generated to answer these questions (Jagodzindki & Holmes, 1989).

Several other researchers (e.g., Clancey, 1983; Gordon, 1987; Kidd & Cooper, 1985; Stenton, 1987) have also made suggestions about system explanations. Berry & Hart (1990b) provided a useful way to summarize these suggestions by saying that explanations need to provide "the right information, in right format, at the right time."

1. **Provide the "right information"**. There seem to be as many ideas about what constitutes the "right information" as there are researchers who have thought about this question. Some feel that a good explanation component needs to be able to answer specific questions about its own abilities, its domain knowledge, and hypothetical situations (Chandrasekaran, Tanner, & Josephson, 1988; Hayes & Reddy, 1983; Shortliffe, 1976). The "right" level of abstraction needs also to be provided for each particular user. An explanation should be possible at lowest level of interest (Hasling, Clancey, & Rennels,
1984) and the highest, most abstract level, in order to provide the causal rules behind the system's reasoning (Clancey, 1983). A good explanation should inform the user about the problem features and their relations (Berry & Broadbent, 1987), why the recommended advice or action is desirable (Johns, 1990), and the appropriateness of the system knowledge and reasoning process (Berry & Hart, 1990b), and focus on the problem-solving process over a content centered approach that merely makes the system's body of knowledge explicit (Jagodzinski and Holmes, 1989; Woods et al., 1990). It should also keep track of past failures and use this information to help the user decide about the correctness of the system advice (Johns, 1990). A useful system should be able to structure explanations so any incorrect beliefs held by users are first dismissed so users will be open to the correct facts (Weiner, 1989). Similarly, explanations should rule out any competing explanations (Josephson, Chandrasekaran, Smith, & Tanner, 1987). Finally, explanations should provide complete and comprehensive answers to most users’ questions (Scott et al., 1984).

2. Use the "right" format. Ideas about explanation format are not as prevalent as explanation content ideas. Only a few have made suggestions about format or structure. Goguen et al. (1983) suggest that when explanations are complex, they should list reasons for the statement first followed by the actual statement or advice. This format as opposed to the opposite order (statement then reasoning) requires less memory for the recipient of the explanation. When the explanation is simple, either ordering is fine.
3. Present at the "right" time. Most expert systems automatically provide explanation with their advice and provide more detailed information when prompted by user questioning. Some have suggested that follow-up explanations should be available if the first explanation fails to promote user understanding (Cawsey, 1993; Wick, 1989). On the other hand, some feel that automatically providing advice and explanation can lead to user frustration when the advice is not wanted (Langoltz & Shortliffe, 1984; Wick, 1989). This scenario is most common with critiquing systems, which monitor and critique user performance. Users of a cancer treatment consultation system, ONCOCIN, quickly became frustrated by the system's constant evaluation that pointed out minor, insignificant differences between its model of the correct user action and the actual user action (Langoltz and Shortliffe, 1984). ONCOCIN was later restructured to only give advice and explanation when the user's actions deviated significantly from the system's model. Users found this new format more enjoyable to use.

**Difficulties Associated with Providing Expert System Explanations**

Clearly, the literature is not lacking in ideas about what explanation components need to provide for their users, yet, as mentioned above, few of these ideas have been successfully implemented or studied empirically. Why is it so difficult for an expert system to produce explanations that provide users with everything they need to know for their specific task? Hasling et al. (1984) have broken down the difficulties of providing useful explanations into three problem areas: epistemological, user modeling, and rhetoric. Others have
derived similar breakdowns (e.g., Chandrasekaran et al., 1988; Hasling et al., 1984; Lehner & Kralj, 1988; Weiner, 1989).

1. Epistemological. This problem area is concerned with determining the knowledge required to solve a problem and the aspects of a problem solver's behavior that need to be explained. An explainer, human or machine, must determine what to say (should it be simple or complex?) and how to say it (what amount of detail should be provided?). The explainer must also identify and speak to any additional goals of the recipient that need addressing (Weiner, 1989).

One reason that it is so difficult to develop concrete rules for determining explanation content is that the term "explanation" is used to represent several different ideas. "Explanation" has been used to describe ANY request for further information such as requests for further instruction, more data feedback, explication of terms, justification for advice or the reasoning process involved. All of these requests require different content (Berry & Hart, 1990b). Clearly, when discussing appropriate explanation content, one must be specific about which type of user request is being discussed.

Another challenge is to agree upon the goal of the explanation at any given time. "It is still an open issue whether to strive for ways to make the expert system's reasoning clear to operators, or to provide alternative ways for operators to validate the expert system's conclusions" (Shafto & Remington, 1990, pg 2). An assumption is often made that providing users with information about the system activity in the inference component is sufficient basis for an
understandable explanation. This information is important, but not sufficient. The user model must also be considered (Mescheder, 1990).

2. User modeling. This problematic area is concerned with generating an explanation that takes into account user knowledge and preferences. User modeling has been attempted by several system designers and is widely regarded as being important in expert system design (Weiner, 1989).

Whereas few would argue with the utility of determining the user's model for tailored explanations, the problems of constructing and incorporating the user model into an explanation is still largely unsolved. How should the system construct the user model? Some systems have simply asked users to input their skill level on a scale from one to ten. However, users are normally bad at determining their own skill level, which could lead to inappropriate system explanations. For this reason it has been suggested that systems should try to determine user expertise through their actions (Carroll & McKendree, 1987).

Some researchers have identified ways to accomplish this difficult task by monitoring users' discourse (Cawsey, 1993; McKeown, Wish, & Mathews, 1985) or users' command frequencies and requests for online help (Vaubel & Gettys, 1990). Determining expertise through users' actions may or may not be a more accurate method for determining user models than direct user input of their own skill level, but it certainly requires more complex programming and decision making by the system.

Assuming an accurate user model can be derived, then what? The system still needs to determine what kind of information, amount, and timing is appropriate for each user model, all of which have proven difficult for artificial
intelligence (Carroll & McKendree, 1983; Clancey, 1983; Roth & Woods, 1989). Answering these questions require empirical studies. It is quite possible that determining the appropriate information for each user is impossible due to the wide range of user types and preferences. Ellis (1989) points out that knowing a user's expertise level may not provide enough information to determine the amount of detail to present. Some experts may prefer high levels of detail, whereas some novices might ignore detailed information.

3. Rhetoric. Rhetoric is concerned with stating the explanation so it will be understandable to the receiver, that is, the semantics of the explanation. It is particularly difficult for most expert systems to provide the reasoning behind the problem in natural language since this information is not stored in English, but rather in the system's programming language. In order for this information to be presented in an explanation, it must first be translated into English. This step requires additional complex programming (Chandrasekaran et al., 1988).

New Developments in Expert System Explanations

Despite the numerous difficulties involved in developing expert systems that include a useful explanation generator, significant advances have been made. According to Chandrasekaran, Tanner and Josephson, (1988) the advancements made in providing artificial explanations can be described in terms of three significant breakthroughs or "good ideas." A brief discussion of these breakthroughs will clarify the gains made by explanation technology and where more work is still needed.

The first breakthrough in expert system explanations was made by Shortliffe and his colleagues while developing MYCIN, an expert system for
diagnosing infectious blood diseases (1976). MYCIN was the first expert system to incorporate the idea that a rule trace of the problem-solving activity at the low architectural level could provide explanations about the system actions (Chandrasekaran et al., 1988). Shortliffe felt that explanations were necessary to provide system transparency and encourage user acceptance (Buchanan & Shortliffe, 1984). When it was completed, MYCIN was capable of answering how and why questions about the steps used to reach its advice (Chandrasekaran et al., 1988). This was a new idea for the time and the start of a new research domain for those in artificial intelligence, computer science, and psychology.

Originally, the rule trace explanation was provided in MYCIN's programming language, LISP, and then was later translated into English so the intended users (physicians with a little programming knowledge) could make use of them (Buchanan & Shortliffe, 1984). An example of an explanation written in both LISP and English provided in Figure 3 should foster a sincere appreciation for MYCIN's additional translation step.

Another important area was also being worked on at this time: user modeling. In an expert system interaction, a user model is the information the system has about its user (e.g., the user's expertise level, strategy preferences, tasks, goals, etc.). This information can be used to tailor explanations to be congruent with the user's expertise, goals, and preferences by altering, among other things, vocabulary and level of detail (Gordon, 1987; Pew, 1988; Wick, 1989). Systems were built that were able to tailor the level of detail provided in their explanations to the user's wishes (e.g., TEIRESIAS built by Davis in 1976
Rule trace explanation in LISP:

RULE 009

Premise:  \((\text{AND}(\text{SAME CONTEXT GRAM GRAMNEG})\\(\text{SAME CONTEXT MORPH COCCUS}))\)

Action:  \((\text{CONCLUDE CONTEXT IDENTITY NEISSERIA TALLY 800})\)

Same rule trace explanation in English:

IF:  1) the gram stain of the organism is gramneg, and
   2) the morphology of the organism is coccus

THEN:  There is strong suggestive evidence (.8) that the identity of
        the organism is Neisseria.

Figure 3. Example of MYCIN explanation written in LISP, MYCIN's
programming language, and then translated into English (example taken
from Scott, Clancey, Davis, & Shortliffe, 1984).
(Weiner, 1989)) and build models of the users' knowledge in order to identify misconceptions (e.g., GUIDON developed by Clancey (Millet & Giloux, 1989)).

The second breakthrough according to Chandrasekaran et al. (1988) was NEOMYCIN's (Clancey & Letsinger, 1981) ability to provide strategic information to the user. NEOMYCIN was a teaching tool built to represent MYCIN's diagnosis task more explicitly by providing strategic explanations. To accomplish this goal, MYCIN's database was restructured so that explanations were generated from a goal-subgoal level of the rule base rather than a more abstract task level as had been used in previous systems. This allowed access to strategic information used by human experts but not explicitly programmed into expert systems (Chandrasekaran et al., 1988).

The advantages of providing strategic explanations for users are numerous. This type of explanation makes the plans and methods that the system used in attaining a goal available to the user (Jackson, 1986). Strategic explanations are thought to emulate the way that human experts think (Clancey & Letsinger, 1981; Ellis, 1989) because they provide more information about the logic behind the system expertise.

Chandrasekaran et al. (1988) identified XPLAIN's (Swartout, 1983) ability to justify its advice as the third breakthrough in expert system explanation. XPLAIN is programmed with both a domain model composed of descriptive facts about the domain (called deep knowledge) and strategic information describing how to diagnose and treat a patient. Deep knowledge of this sort and strategic information are assumed to be necessary for truly explainable expert systems (Neches, Swartout & Moore, 1985). For the first
time, a system could make strategic information explicit and justify these strategies by pulling out information from its domain model or deep knowledge (Chandrasekaran et al., 1988; Leake, 1991). According to Neches et al. (1985), XPLAIN’s explanations are still somewhat limited in the types of user questions they can directly address. However, the team is working on extending these capabilities in their work on explainable expert systems.

Human-Human Explanation

Despite the significant breakthroughs in providing artificial explanations, there is little disagreement that more empirical work is needed to justify some of the decisions made (e.g., Berry & Broadbent, 1987; Erdman, 1987). Researchers interested in numerous facets of explanations such as explanation logic (Jackson, 1986), explanations in graphical interfaces (Johns, 1990), and expert system trust (Muir, 1987), have petitioned for more empirical research on system explanation components. Several important questions have yet to answered such as "how can the user still benefit from expert systems when they know they’re not 100% reliable?" (Pew, 1988), and "how do system failures affect trust in the system?" (Muir, 1987). Future empirical work on the explanation component is also needed to create guidance for interface designers (Hollnagel, 1990; Potter & Woods, 1990), to identify the benefits vs. potential dangers of providing explanations (Erdman, 1987), and to identify techniques for creating user friendly interfaces and user acceptance (Roesner, 1988).

Unfortunately, the researcher interested in solving some of these problems by gathering empirical data is faced with the difficult problem of
deciding where to begin. The seemingly endless supply of recommendations for improving the explanation component (the discussion above only included a sampling of the numerous suggestions found in the literature) and the long list of difficulties in implementing those recommendations provides a wide-open domain for research.

The most logical starting point is to take a step back and review what we know about explanation in general and then build upon the existing knowledge, advice advocated by Suchmund (1990) and Potter and Woods (1990). Because several (e.g., Goguen et al., 1983; Leake, 1991) have suggested that artificial explanation should be modeled after human explanation, it seems logical to review not only what we know about expert system explanation, but also what is known about human explanation.

The study of human-human explanation has an interesting past ranging across several domains and covering several centuries of research. The first documented interest in explanation was held by the philosophers, Aristotle and Plato. The study of explanation is still a popular topic in philosophy today (Achinstein, 1983; Ruimschotel, 1989). Explanation as a topic for research has also branched out into the domains of social psychology (Berry & Broadbent, 1987; Lalljee & Abelson, 1983), education (Bush, Kennedy, & Cruickshank, 1977; Putnam & Duffy, 1984; Roehler, Duffy, Book, & Wesselman, 1983), and recently, cognitive science (Cawsey, 1993; Goguen et al., 1983; Chi, Bassok, Lewis, Reimann, & Glaser, 1989). Information on explanations that maybe useful in expert system explanation design will be reviewed in each of these domains. Specifically, focus will be given to literature that identifies the
characteristics of an expert's explanation that make the explanation successful or useful to the recipient.

**Philosophy**

Philosophers are usually interested in scientific explanations or the explanations we give each other about natural occurrences. Explanation, like knowledge, is viewed as an epistemic concept to philosophers, and therefore has a place within their study of knowledge (Rubin, 1990). Philosophical theories of explanation are mostly focused on the formal requirements for explanations, that is, determining what part of discourse should be called an explanation and what should not (Leake, 1991; Ruimschotel, 1989). Philosophical theories also focus on the explanation product that can be characterized solely by the kind of information it conveys; no reference to the act of explaining is required (Rubin, 1990). Whereas these points of focus are somewhat different from those of expert system designers and the applicability of the knowledge to system design is therefore, limited, some useful advice about what makes explanations "successful" can be found in this literature.

According to Achinstein (1983), explanations are only understandable to the recipient if the explanation is a "complete content-giving proposition with respect to the question" and if the recipient knows the explanation is correct. This implies that expert system users need to believe that the expert's knowledge incorporated into a system is correct and that the system created the correct explanation for the information given. A successful explanation should also contain hypotheses that are plausible and do not widely contradict user models of the domain (Smart, 1990). This is explained further: "an explanation
may be bad if it fits only into a bad web of belief. It can also be bad if it fits into a (possibly good) web of belief in a bad sort of way." (Smart, 1990, p. 9). In order for an expert system explanation to fit into the user's "web of beliefs," the system must determine these beliefs. This would support systems that develop a user model and then use the model to tailor explanations accordingly.

Finally, some philosophers argue that an explanation for an event is chosen based, in part, on the context of the event (Leake, 1991). An event will be explained differently by different explainers according to their expertise and interests (Hanson, 1961) or by differences in the "causal field" of situations to be distinguished by the explanation (a context-dependent approach (Mackie, 1965)). These findings also support the employment of user modeling in expert systems, as well as the ability to track the user-system interaction in order to develop the appropriate explanation for the particular event context.

Most philosophical accounts of explanation are only moderately helpful to expert system designers. For example, Ruimschot’s (1989) dissection of explanations into seven components (form, content, inference, alternative prediction, alternative explanation, causality, and truth) would be more helpful to designers if each dimension were not only described, but also associated with information about how to improve explanation along that dimension. Philosophy does provide additional support for the importance of explanations in a users’ trust or acceptance of the system, as well as the system’s ability to develop an accurate model of the user and the context of the event, yet little help is given beyond these points. Questions about how to increase user trust,
how to develop an accurate user model, and what constitutes a good explanation are still left unanswered.

**Social Psychology**

Explanation is discussed in a wide variety of areas in psychology, especially in social psychology, yet it has received relatively little systematic empirical study (Berry & Broadbent, 1987; Leake, 1991). Heider's attribution theory is probably the best known work on explanation in this domain. Attribution theory is concerned with how one explains the actions of another. Specifically, it makes predictions about when attributes of the environment as opposed to attributes of the persons involved in the event are used as the basis of the explanation (Leake, 1991; Zechmeister & Johnson, 1992).

Heider's work led to several other theories about explanation such as the covariation principle developed by Kelly in 1967 (Leake, 1991). This principle states that when deciding what to attribute a specific behavior to (the environment or the actor) we sample across various situations, people, and time. On the surface this work does not appear to be directly applicable to expert system explanation, but Berry & Broadbent (1987) point out that a study done by Jaspars, Lalljee & Jaspars (1986) has tied these two areas together. From Jaspars et al.'s observation of students using computer systems in the classroom, they concluded that if users are repeatedly given explicit explanations containing sufficient justification for expert system advice, then they will tend to attribute their success to the system's ability. Insufficient justifications would cause users to distrust the system and attribute any successes to other factors (Berry & Broadbent, 1987). This idea is echoed by
Muir (1987) who states that the degree of trust a user will have in a system depends on the amount of sampling the user does in order to confirm the system's correctness. If the system appears to be giving correct advice and a reasonable explanation for the advice, users will start trusting the system and reduce the degree of critical checking they do before acting on the advice.

The work of Kelly was later challenged by Hilton (1990) who has developed a model of causal explanation. This model directly addresses interpersonal and functional aspects of causal attribution. The idea behind Hilton's model is that explanations should identify and make salient the factors that distinguish between the recipient's belief about the cause of an event and the causes of the event which are be provided by the explainer (Hilton, 1990). This agrees with Hesslow's (1983) thought that explanations describing why a puzzling event occurred in the target case, but not in some alternative explanation, are chosen over explanations that merely explain the puzzling event (Hilton, 1990). Finally, Hilton states that a causal explanation must be relevant to a question as well as true of the target event and should follow the four maxims of conversation proposed by Grice (1975): quality (be truthful), quantity (provide enough information, but not too much), relevance (stick to the question at hand), and manner (be precise and orderly).

Like philosophers, social psychologists have several ideas about what good explanations entail: they explain the "why" behind an event not just "what", should eliminate alternative explanations, and should consider the task and recipient of the explanation. Unfortunately, like philosophers, social
psychologists have little empirical evidence or detailed suggestions to offer about the characteristics of a successful explanation.

**Education/Educational Psychology**

Of the areas discussed thus far, the educational domain appears to be the most directly applicable to the design of expert system explanations. If the characteristics of a successful human explanation can be identified in the classroom in which the goal is to teach or promote understanding, it seems likely that this knowledge could be usefully incorporated into expert system design.

Although several studies discuss the content of teachers' instructional explanations (Bush et al., 1977; Leinhardt; 1987; Putnam & Duffy, 1984; Roehler et al., 1983), few distinguish between successful and unsuccessful teacher explanations. Whereas one may assume that all teachers are expert explainers, it seems that more can be learned by comparing good and poor teacher explanations and describing the properties that distinguish the two. This is what Duffy, Roehler, Meloth, and Vavrus (1986) have done.

Duffy et al. (1986), studied second-, third- and fifth-grade teachers responsible for teaching students with below average reading abilities. These researchers were interested in determining whether explanations containing explicit reading strategies help students improve their awareness of what was taught as well as their performance on reading comprehension measures. Half of the teachers participating in the study were taught to incorporate explicit explanations into their lectures, the other half were not instructed how to teach these skills. For a period of several years, the teachers' lessons were
periodically observed and students were interviewed about the lesson to gain a measure of awareness of lesson content. Students' test achievement was also measured. For purposes of this research, effective explanations were defined as ones that increased students' lesson content awareness and achievement. The result of this lengthy study: four distinguishing features of effective verbal explanation were identified. Effective explanations 1) are responsive to the student's changing knowledge structure, 2) try to put students in conscious control, 3) present declarative, conditional, and procedural information, which is conceptually accurate, explicit and meaningful, and 4) attempt to assist student efforts to build understandings by sequencing and providing restructuring "hooks."

These "hooks" are provided for novices to "hang onto" until they develop their own expertise. The hooks identified are listed below. More effective explainers:

1) present information from the perspective of a novice
2) avoid cognitive overload by simplifying or "chunking" the information
3) mark informational chunks with clear verbal signals
4) direct (and re-direct) attention to the crucial features of the task
5) question students to remind them of previously communicated information
6) help students assume responsibility for doing the necessary thinking
7) help students re-structure their understanding by spontaneously creating analogies that students can use to build accurate interpretations of what the teacher wants them to learn
8) help students understand the intended outcome by narrowing the gap between the skill and its use in real reading

These classifications easily lend themselves to guidelines for human teachers and those trying to learn how to be effective explainers. The authors warn, however, that explanations will differ with each new context, and therefore, rigid rules or guidelines should not be developed. Rather, resourceful teachers will combine and re-combine these methods in an "infinite number of ways." This warning makes the knowledge about successful explanations offered by this study more difficult to apply to an expert system, which would benefit from rigid rules of explanation. Another limitation of these classroom studies is that the explanations studied are not short statements or paragraphs as required by expert systems, but rather long interactive dialogues with students that may continue over several lessons and several days (e.g., Leinhardt, 1987). It is difficult to generalize suggestions about successful instructional explanations derived from these situations to expert systems, which are limited in their ability to participate in full natural language conversations with their users.

Cognitive Science

The domain of cognitive science which combines, among other disciplines, cognitive psychology and artificial intelligence, is rich with researchers interested in both human and machine explanations. The discussion of this research is divided into two sections which classify cognitive science research in terms of its relevance to expert system explanations.
General cognitive research. Research in areas such as problem solving, decision making, and thinking has indirect applications to the generation and use of explanations. For example, Miyake's (1986) conclusion that when learning a complex system, subjects' go in and out of phases of understanding suggests an explanation's timing during the learning process might be important. When subjects are in an understanding phase, detailed explanations may be viewed as unnecessary and even annoying. However, detailed explanations should be more welcomed when subjects are in a phase characterized by lack of understanding.

In addition to a user's level of understanding, the effect of explanations on a user's confidence level should also be kept in mind when building explanations. When faced with a problem solving task, subjects gain more confidence in their decision with higher levels of information gained. Unfortunately, subjects' decision correctness does not linearly increase with their confidence level; it often trails behind (Oskamp, 1983). Explanations that provide information that causes user confidence to go up erroneously might be more harmful than helpful.

Probabilistic information often is used as the basis for decision making (Tversky & Kahneman, 1983). Unfortunately, a person's ability to use this type of information correctly is often quite poor. Consider an informal poll taken by Eddy (1983). Doctors were given the results of a patient's mammograph and asked to estimate the probability that the patient had a malignant lesion. Of the 100 doctors polled, approximately 5 were able to give the correct probability of 8%. The other 95 doctors guessed the probability to be around 75%. The
seriousness of these types of errors in judgment are obvious. Perhaps a goal of expert diagnosis systems which provide probabilities along with their diagnosis should be to also provide explanations about how to correctly use the probabilistic information.

These are just a few examples of the cognitive science research that have an indirect application to expert system explanation design. This research cannot offer recommendations for system design, but can provide information on which to base hypotheses about successful expert system explanations.

**Explanation research.** Research focusing explicitly on explanations can also be found in the cognitive science literature. There is no shortage of papers that attempt to define the purposes and components of human explanation (e.g., Brown & Van Kleeck, 1989; Goguen et. al, 1883; & Turnbull, 1986), but research that compares successful and unsuccessful explanations is rare. Nonetheless, one study was found that is relevant to this discussion.

Chi, Bassok, Lewis, Reimann, and Glaser (1989) studied the variance between explanations generated by good and poor physics problem solvers. After an extensive knowledge acquisition phase lasting several hours over several weeks, students gave talk-aloud protocols while studying worked-out examples and solving related problems. It was concluded that explanations created by superior problem solvers were complete and integrated with other problems solved previously or with concepts in the text. Explanations from this group also tended to be goal-oriented. Quite possibly, the explanations generated by the better students were at least partially responsible for their
superior performance and therefore, the explanations generated by this group could be viewed as more successful.

There is some explanation research that is even more directly relevant to expert system explanations because it was done with the goal of generalizing conclusions about human explanations to intelligent systems. Maybury (1992) studied human explanation and developed a taxonomy of explanation utterances based on content, form, and communicative function of the explanation. During this process he discovered one difference between human explanation and current artificial explanation is that humans are very good at recovering from failed communication. If one explanation is misunderstood, it is reworded or another explanation is given. Intelligent systems are rarely able to detect failed communication much less recover from it. Maybury suggests that multiple explanations should be programmed into intelligent systems to allow for user misunderstandings.

Cawsey has done similar research focusing on the interactive nature of human explanation (1988, 1993). Through thorough discourse analyses, she has concluded that the explanation process is a bi-directional as opposed to a uni-directional process assumed by some expert systems. The EDGE system has been developed based on Cawsey's conclusions that artificial experts should allow for misunderstandings, interruptions, and clarification questions often present in human explanation.

Finally, Goguen et al. (1983) studied human explanation and reasoning with the goal of applying what was learned to the development of BLAH, an expert system developed by Weiner (1980). In this investigation several pre-
existing transcripts of human discourse were studied in detail to develop a comprehensive model of natural explanation. The model is able to suggest helpful orderings of explanations (e.g., for long explanation chains, stating the reason before the statement requires less memory for the recipient of the explanation than explanations in the statement/reason form) and recognizes that explanations are a social process and depend on the particular situations and on the assumptions of both the explainer and audience.

This model is helpful because it suggests optimal explanation organization and supports user modeling by the explainer, but it has relatively little to say about the benefits of various explanation content. Three ways to justify a statement are covered by the model (provide a reason, provide an example, or eliminate the alternatives), but there is no evidence or advice which states when each type of justification is most useful to the recipient. The models developed by Goguen et al. (1983) and Maybury (1992) have provided a useful way to represent natural, human explanations, but they seem to assume that if explanations are natural, they are useful and worth emulating in an expert system. Although expert system explanations should be modeled after human explanations, those explanations selected to be models should be based on their success.

The above research is helpful in that they all offer methods for analyzing explanations and provide ideas about how to apply rules used by human explanation to artificial explanation. Nonetheless, they fail to provide data or discussion about whether the incorporation of their ideas will lead to more successful artificial explanations. Unfortunately, none of the above research
was able to offer information about the components of a successful explanation and provide advice about how to apply these findings to an expert system. Recall that the philosophy and social psychology literature have the same shortcomings.

**Summary.** In summary, researchers in the fields of philosophy, social psychology, education, and cognitive science have suggested several features of good explanations. Successful explanations should 1) adapt to the user by taking into account the recipient’s expertise and interests (Hanson, 1961), 2) establish the explainer’s expertise so the recipient is able to determine if the explanation is correct (Achinstein, 1983), 3) sufficiently show why alternative explanations are not correct (Hilton, 1990), 4) follow the four maxims of conversation (Hilton, 1990), 5) provide information explicitly (Duffy et al., 1986), 6) provide conceptual "hooks" for novices to "hang onto" until they develop their own expertise (Duffy et al., 1986), 7) explanations should integrate previously learned material (Chi et al., 1989), 8) be able to recover from failed communication by rewording or providing an alternative explanation (Maybury, 1992), 9) allow for interactive dialogues (i.e., user interruptions, clarification questions, etc. (Cawsey, 1993)), and 10) state the reason before the statement to decrease the memory load required by the recipient (Goguen et al., 1983).

These features of a good explanations are numerous, however, only features 5, 6, 7, and 10 are based on empirical data measuring explanation success. The other features are merely recommendations or speculations. The fact that only four features of successful explanations could be found across the broad domains of philosophy, social psychology, education, and cognitive
science suggests more empirical research is desperately needed before expert system programmers can confidently build explanations into the system that will help rather than hinder system interactions.

Given this review of what is known about explanations, human-human or otherwise, it is not surprising that expert system explanations are currently unsuccessful; users are misinterpreting them, ignoring them, or being annoyed by them. The next section of this paper will present a research project developed to identify features of successful human explanations in order to incorporate them into expert systems.

RESEARCH PROJECT

One path to good expert system explanations can be stated in terms of several subgoals: 1) Define what is meant by a "good explanation" in the expert system domain and in human interactions, and determine how this can be empirically measured; 2) Observe human explanations to identify possible elements of good and poor explanations; 3) Test in a controlled study the hypothesized elements of good and poor human explanations; 4) Incorporate the elements of good and poor human explanations into an expert system environment and test them empirically; 5) Take the results from these studies and incorporate them into recommendations on expert system explanation.

The following project was designed to meet all of these subgoals in hopes of being able to define good from poor expert system explanations at its completion. After a definition was developed for a "good explanation", an observational study of human-human explanation (Experiment 1) was
conducted and hypotheses were developed about the elements of good and poor explanations. These hypotheses were tested in two empirical studies which tested various explanations when given between two humans orally (Experiment 2) and when given between two humans or a human and a computer electronically, via the computer (Experiment 3). Recommendations about expert system explanations and expert system interactions in general were then made based on the results of these studies. Each of these subgoals will be discussed in turn, beginning with the process of defining a "good explanation".

Definition of a Good Explanation

To define what is meant by a good explanation a decision must be made about the purpose of the explanation. As previously discussed, expert system explanations seem to be tied to effective interactions with the expert system. Effective system use depends on, among other things, users' trust in the system and the accuracy of their conceptual model. Because effective interactions also lead ultimately to appropriate system acceptance or rejection, a major concern for system designers (Langoltz & Shortliffe, 1989), a good explanation should be one that promotes an accurate conceptual model and therefore, facilitates effective interactions.

This definition of a good explanation should generalize to human explanation as well. Explanations that teach recipients about the explainer and domain should allow an accurate calibration of trust in the explainer. If trust is accurately calibrated, the explanation recipient should appropriately accept or reject the explainer as an authoritative source for the domain.
Most likely, the type of explanation that promotes an accurate conceptual model will vary depending on several factors such as the recipient’s goals, expertise, and the specific task being performed (Leake, 1991). These factors should, therefore, be held constant when studying explanations. To identify additional factors of good and poor human-human explanations, an observational study will be performed to observe explanations of subjects with the same level of expertise performing the same task. A good explanation, which is defined here as one that results in the formation of an accurate conceptual model, can be measured most directly by comprehension tests covering domain facts and processes. It is critical that indirect measures of user understanding, such as task performance, should be taken to measure other aspects of explanation success.

Experiment 1

This section describes an exploratory study in which human domain experts were observed as they explained their advice or problem solutions to others. The goal of this study was to investigate features or dimensions that distinguish good from poor explanations. Because the literature offers few clues about the features of good vs. poor explanations and there are any number of features that could hypothetically distinguish good and poor explanations (see Halgren, Flowers & Cooke, 1991), this study was by necessity exploratory.

Because one of the goals of this project is to apply what is learned about explanations among humans to expert systems, an effort was made to observe human-human interactions in the lab that mimicked human-expert system
interactions as closely as possible. To accomplish this objective, the task constructed for this study as well as the domain knowledge required by subjects were carefully planned.

The task domain chosen for this study was medical diagnosis, a task widely assisted by expert systems (e.g., MYCIN, Buchanan & Shortliffe, 1984; XPLAIN, Swartout, 1983). Subject pairs worked together on a computerized monitoring and diagnosis task. The task involved monitoring an alien life form's health, and diagnosing and treating any problems that occurred. Each subject pair was made up of one subject playing the role of the expert system (this subject will subsequently be referred to as the "expert") and the other subject playing the role of the expert system user (to be subsequently referred to as the "user").

Decisions about the amount of information, training, and instructions given to each subject were based on the features of a typical expert system user and expert system. Thus, the expert had access to all the procedural, conceptual and anecdotal information available about the alien. Expert systems have a perfect memory and so should always have access to all of the available domain information. Although expert system users may have learned all of this information at some time, it is unlikely they will be able to retrieve all of it at any particular moment (if they could, they probably would not have a need for an expert system). For this reason, subjects assigned to the user role were only provided with a portion of the procedural, conceptual, and anecdotal information.
One of the limitations of expert systems is a lack of common sense knowledge (Waterman, 1986). To simulate this limitation and to create instances in which the user could justify ignoring the expert's advice, users were given current environmental information on their computer screen. Because experts were not given this information, instances should occur where obvious diagnosis or treatment solutions involving manipulation of the environment are missed by the expert, but caught by the user. There are real life situations in which a physician will be able to make better judgments using common sense than an expert system which is limited to the analytic rules that are built into it. For example, imagine the following scenario: An expert system taking readings from several monitors connected to a critically ill patient detects a problem—one of the vital signs being measured has gone far beyond its normal range. The expert system offers a diagnosis and a treatment for the problem. The doctor finds the extreme vital sign very unusual, especially since it is not accompanied by any other abnormal symptoms. The doctor takes a closer look at the monitor showing the extreme reading and notices smoke emerging from the back compartment. The doctor concludes that the monitor has malfunctioned and orders it to be serviced.

This scenario points out the need to keep humans in the decision loop for diagnosing and treating failures in complex systems. Had the expert system alone made the decision, the patient would have been treated for a nonexistent
medical problem. The complexities behind the common sense knowledge that humans have is, for the time being, beyond the capabilities of expert systems.

Several measures were taken of users' diagnosis and treatment performance. It was predicted that these measures would be positively correlated so that users who performed well on one performance measure would perform well on the others. It was also assumed that users who performed well on the performance measures would have acquired more knowledge about alien medicine through the explanations they had received from the expert. Another prediction was that over time subjects playing the expert role would become more knowledgeable about the problem domain and would, therefore, become better explainers. Thus, experts were paired with three different users. The explanations given by the expert to the first user would probably not be as good as those given to later users by the same expert.

The ability to provide superior explanations might also be linked to practice explaining, regardless of the domain. It was predicted that users who were paired with experts who had experience teaching or tutoring, should receive superior explanations, learn more domain information, and subsequently perform well on the performance measures.

Because of the exploratory nature of this study, the goal is the identification of specific characteristics of good explanations as opposed to testing for features judged to be critical a priori. However, it was expected that some features such as content (e.g., the use of examples, analogies, statistics, redundancy, anecdotes, etc.), form (e.g., the use of mechanistic form, explanation-advice order, sentence structure, etc.), and timing (was the expert's
advice and explanations prompted by the user or volunteered without a prompt?) would be relevant.

**Method**

**Subjects**

Subjects were 12 undergraduate and 8 graduate students from Rice University. The 9 undergraduates playing the user role participated in the study in exchange for course credit and the 3 undergraduates playing the expert role were given course credit and paid $30. Graduate students were paid either $30 (expert role) or $8 (user role) for their participation. Five subjects played the expert role and each interacted with three subjects playing the user role for a total of 15 unique subject pairs.

Because it was assumed that being a good explainer is a skill that takes more than a few hours of practice to master, subjects playing the expert role were chosen who had varying levels of explaining or teaching experience. Two experts had no prior teaching or tutoring experience (experts 1 and 2). Expert 3 had experience tutoring for one year's time and had taught a small class in an informal situation (swimming lessons). Experts 4 and 5 were more experienced teachers, especially expert 5 who had taught 2 college level courses. Expert 4 had taught one college level course and had some experience tutoring a non-academic subject (piano lessons). Experts 1 and 2 were considered novice explainers, expert 3 was considered an intermediate, and experts 4 and 5 were considered experienced explainers.

Data from only 14 of the 15 subject pairs are reported here because of a tape recorder malfunction for one of the pairs.
Materials

The subject pairs worked together on a computerized monitoring and diagnosis task. Because it was important to control subjects' expertise with the domain, the task domain, an alien and its anatomy, were created by the experimenter. In this way, amount of domain information provided could be manipulated according to the subject's role in the study.

Laboratory setup. Two Macintosh Classics were connected via a local network so that the expert and the user saw the same information. Both subjects were located in the same room, but separated by a partition through which they could speak to each other. The partition was placed between the two subjects to minimize some of the advantages human explainers have over artificial explainers. Humans use gestures, eye contact, and body language to enhance meaning, behaviors a computer are incapable of performing. The presence of the partition between the two subjects forced all interactions to be strictly spoken, thereby limiting the advantage of unspoken communication. A tape recorder was used to record the subjects' spoken interactions.

The diagnostic program. The diagnostic program was written in HyperCard 2.0. Both subjects (the expert and user) saw identical screens throughout the experimental session with the exception that the user's screen had an extra window labeled "Environmental Information". Only the subject playing the user was allowed to interact with the computer program. The subject playing the expert role could not see the user's cursor, but could see when an option had been selected.
A diagram of the alien connected to eight monitors was pictured on the main screen (see Figure 4a). The information on the monitors updated about every three seconds. When one or more of the monitors showed an abnormal reading, the computer sounded a beep and a message stating "Failure" was flashed on the screen for approximately 2 seconds. The monitor(s) displaying abnormal readings were then highlighted, specifying which part or system of the alien was failing. The user was able to treat the alien by clicking on a button labeled "Treatment" and then selecting the desired treatment option and the "OK" button from the treatment window that was displayed (see Figure 4b).

After a treatment was selected, feedback about treatment success was displayed on both subjects' screens (see Figure 5). Feedback was either positive (the alien recovered) or negative (the alien got worse). If the feedback was negative, no details about the correct treatment were given. Treatment success was, to some extent, based on chance. Because medical diagnosis is not an exact science, even the most skilled diagnosticians cannot always be accurate in their diagnoses. For this reason, an element of chance was built into the computer program. For each medical failure that the alien suffered, two possible diagnoses and treatments were built into the program and given to the expert in her materials (these materials will be described shortly). The expert was also given the probability associated with each diagnosis for each problem. One diagnosis was always much more probable than the other (one always had an 80% chance or higher of being correct). This made the decision between the two diagnosis easy to make, but still allowed for the possibility of error.
Figure 4. A) Main screen from the Diagnosis Program as seen by the subject playing the user role (the expert subjects saw the same screen without the "Environmental Information" window)
Click on the circle to the left of the treatment you wish to administer to the alien.

- add 50% nickel sludge
- add 75% nickel sludge
- add 50% aluminum sludge
- add 75% aluminum sludge
- inject vitamin B12
- inject dopamine
- increase light intensity
- decrease light intensity
- increase heat intensity
- decrease heat intensity
- add carbon monoxide
- add lithium
- add oil
- add microorganisms
- add hydrogen
- send a random signal
- clear path between lamp and tank
- move lead sheet away from monitors
- remove large foreign elements from tank
- service monitors
- clean tank water
- extract aluminum
- other

**Figure 4.** B) Treatment window seen by both subject types.
A.

Congratulations!

Gus appears to have recovered completely. Job well done.

Continue

B.

We regret to inform you that Gus has gotten much worse after your recommended treatment was administered. You have been taken off the case until Gus has recovered fully.

Continue

Figure 5. A) Positive feedback screen shown when treatment administered was successful, B) Negative feedback screen shown when treatment was unsuccessful.
After the feedback was presented to the subjects, selecting the "Continue" button started the next trial. Regardless of treatment success and feedback given, the alien's health was set back to nominal. If an incorrect diagnosis had been made, subjects did not see the consequences of the unsuccessful treatment that they had administered.

Although this was an exploratory investigation of natural human explanation, a limit needed to be placed on the explanation so it could have some potential for being useful to expert system designers. Contrary to expert system advice, human explanation is not limited by length. Human explainers normally have the luxury of talking for an undefined amount of time to get their point across. On the other hand, screen real-estate limitations in expert systems impose constraints on explanation length. To eliminate this advantage, a time limit was going to be imposed on subjects playing the expert role to limit the amount of explanation given at any one request by the user. However, during pilot testing no limit was imposed and under these task conditions, long explanations were rarely given. Therefore, in order to allow as natural explanations as possible, no time limit was imposed during the actual experimental trials.

**Instructional support and materials.** The information describing the alien anatomy was presented to the subjects through several formats. The amount and type of information that the subject was given depended on the subject's role in the study, expert or user. The information was different for the expert and the user in an attempt to be representative of typical user-expert system interactions (see Appendix A for both the user and expert versions of these
materials). The four types of information that were provided for treatment decisions are as follows:

1) **procedural information**: This information was presented in a diagnosis and treatment chart. This chart provided procedural rules describing the symptom combinations associated with each medical failure and the treatment associated with each failure. As previously discussed, each symptom combination on the chart was associated with two possible failures. For each possible failure, the probability (in percent) that it was the cause of the symptom combination was also listed. One failure was always far more likely than the other. Providing procedural rules to subjects was an attempt to mimic the mechanistic knowledge of a domain.

2) **conceptual information**: Conceptual information was taught through a diagram of the alien's body and through an alien medical training booklet. The anatomical diagram pictured the alien's seven anatomical systems, the parts that comprised these systems, and their connections. Each system was colored-coded for easy identification. Conceptual information underlying the anatomy was taught through the training booklet. Each system was described in turn and the rules behind its functionality were given. The possible failures or problems that might occur, how they could be recognized, and their recommended treatment were explained. This information provided not just the "how" behind a diagnosis or treatment, but also the "why."

3) **anecdotal information**: To simulate the knowledge gained from experience within the domain, a listing of past case histories of specific
alien diseases and treatments was provided in a treatment log. For each past problem the alien had, the hypothesized diagnosis was listed in addition to all the successful and unsuccessful treatments administered.

4) **environmental information**: This information described the environment around the observation lab containing the alien. A sentence describing the actions of others, unexpected noises, or the position of objects in the lab was provided during some problems. This information was included to simulate information a human would automatically have and use if they were in the lab, but a machine would not.

Subjects playing the expert role were responsible for providing diagnosis and treatment assistance and had access to all the procedural, conceptual and anecdotal information available about the alien. Subjects assigned to the user role were only provided with a portion of the procedural, conceptual, and anecdotal information, however, they did have access to the environmental information which the experts did not.

The user's version of the diagnosis and treatment chart contained diagnoses and treatments for 11 of the 22 problems given to the expert. Only eight of these were actually encountered during the experimental session. The user's version of the anatomical diagram included the color-coded alien systems and their parts, but most of the connections between the parts were removed. The users' training booklet described all the alien's systems including the various anatomical parts and their function. However, no details about how the systems functioned were given and only a few possible failures were discussed. Only rarely were failure treatments given or explained. Finally,
the users' version of the treatment log containing anecdotal information listed only 6 of the 14 cases included in the experts' version of the log.

**Problems.** All subject pairs were given the same 22 alien problems to diagnose and treat. The order of these problems was randomized across sessions. Users were given enough information to solve 8 of the problems on their own, 3 of these 8 problems were caused by an environmental influence, but had symptoms that were identical to one of the failures given to both subjects caused by non-environmental sources. The remaining 14 problems required information that had been given only to the expert. The expert could provide the correct diagnosis and treatments for these 14 problems by referring only to the procedural information in the diagnosis and treatment chart they were given. However, it was assumed that in order to provide useful explanations, experts needed to develop a deep, conceptual understanding of these problems which could only be done if the other forms of conceptual, anecdotal, and procedural information in addition to the treatment chart were consulted and understood.

**Procedure**

Subjects were tested in pairs consisting of an expert and a user. The expert first participated in one hour of training.

**Expert Training.** During training the expert was introduced to the study, followed by an explanation of the anatomical diagram and the treatment log. At this point the expert was given the training booklet to read and study. When the expert had completed reading the booklet, the diagnosis and treatment chart was explained. During the explanation of this chart, the experimenter went over
each possible problem or symptom combination on the chart and reviewed the rules and some strategies for accurate diagnosis.

After all the expert's questions were answered, a test covering the procedural and conceptual domain knowledge was administered. This test is presented in Appendix B. The first seven questions that required the expert to identify each anatomical system on the alien diagram by color were given orally and the subject was allowed to look at the diagram while answering. After these problems were completed, the diagram was taken away and the remainder of the exam was given to the expert to complete. The test was scored in front of the expert and any questions answered incorrectly were discussed with the expert until the experimenter was satisfied that the expert had a firm understanding of the correct answer.

The computer program and task was then explained. Experts were told that their task was to help the user administer the correct treatment when the alien had a problem because they, contrary to the user, had been given all the available information about the alien anatomy. It was also added that users had access to environmental information that they did not and that the users would be told not to communicate this information to the expert. Experts were told that their goal was to teach the user as much as possible about the alien anatomy through their explanations so that the user could provide the best care for the alien. The ultimate goal was to teach enough information to the user so that he or she could make correct diagnosis and treatment decisions without the expert's help and do well on a conceptual test to be given after the experimental session. The expert was then left alone to work through the same problems as
the user would in the experimental session (the window containing
environmental information was not present during this training session and the
three failures caused by environmental causes were not seen at this time).
Experts were told that during the practice session they should think about how
they would explain their diagnosis decision and treatment actions to the user if
asked to do so. This entire training period lasted approximately 90 minutes.

**User Training.** Subjects playing the user role went through a shorter
training session. The study was introduced and the diagram of the alien and
the treatment log were explained. The users were then asked to read through
his version of the training booklet. Afterwards, the treatment and diagnosis
chart was explained and the task and goals were described.

Users were told they had three goals to focus on during the study: 1) to
attempt to treat the alien successfully, 2) to learn as much as possible about
alien medicine so that they could guess the correct treatment before the expert
was consulted, and 3) to perform well on a test covering the rules behind alien
medicine which was to be administered after the experimental session. Users
were also told what pieces of information they were given that the expert was
not, and what information the expert was given that they were missing. Users
were told not to share any environmental information with the expert during the
experimental session.

**Experimental Trials.** A trial consisted of the following sequence of
events: The monitors on the screen periodically updated to convey the alien’s
current medical condition. When a problem occurred, a failure message
appeared on both screens and the monitor(s) which had detected the failure
was highlighted. At this point, users were required to verbally suggest a treatment for the failure before they were allowed to ask the expert for any information or advice. After users' suggestion or guess was made, they could ask the expert any questions they had, or ask the expert for diagnosis and treatment advice. Users were not required to ask the expert for advice if they were confident that they already knew the correct treatment to administer. No constraints were put on users about the number or the types questions they could ask the expert and experts were free to answer prompts for help however they wanted. Users were required to make the final decision about treatment because they were the only ones allowed to interact with the computer. Users did not have to follow the expert's advice if they felt it was incorrect.

After the user administered a treatment, a feedback message was displayed on the computer screen describing whether the treatment administered caused the alien to recover or get worse. At this point the user was also given feedback from the experimenter describing the accuracy of his or her guess made before the expert was consulted. The experimenter then told users the number of correct guesses they had made thus far along with the number of problems completed. Regardless of the treatment administered, the trial was completed and the monitors reset to normal ranges for the next trial.

**Post-session test and questionnaire.** After the 22 trials were completed, the user was given the same test that had been administered to the expert during training. Unlike the experts, the users were allowed to use their anatomical diagram and treatment chart during the test. This decision was made because pilot testing revealed that the test was very difficult. The
anatomical diagram and treatment chart served as useful aids on the test, but test questions could not be answered correctly by merely referencing these items. Instead, a deeper, conceptual understanding was usually necessary (an exception to this were the first seven test questions which could be answered by looking at the anatomical diagram; they were removed from test version given to the user). Both experts and users were also given a final questionnaire asking about their relationship with the other subject with which they interacted, and any past conversations they might have had. The questionnaires given to both subject types are listed in Appendix C.

At this point the user was debriefed and sent home and the expert remained to play the expert role for two more users. The expert was used for multiple sessions for three reasons: 1) the training involved for the subject playing the expert role was long and involved, 2) it was predicted that the formation of good explanations may take more practice than can be gained from completing only 22 problems, and 3) large individual differences might exist between the types of explanations that are successful for different people. One expert might provide good explanations for one user, but the same explanations could be poor for another user. Testing the same expert with several users would help separate explanation characteristics that are successful for most recipients from characteristics that help only a few or no recipients.

Dependent measures. Several dependent measures were taken. All verbal interactions were recorded. This included the users' guesses made prior to the experts' advice, the experts' advice and explanations, users' questions, and experts' answers. To test for the success of the experts' explanations,
several performance measures were taken. Correctness of the users' guesses should provide an indication of knowledge held prior to any explanation. Given that there should be no knowledge of this fictional domain prior to this study, differences in users' scores on the comprehension test should be partly dependent on the quality of the explanations they were given. Explanation quality should also affect treatment performance. Explanations that provide a superior conceptual model should eventually result in users' superior treatment performance. Explanations given to those who score highest on the comprehension test and provide superior guesses and treatments can be contrasted to those explanations given to users who perform worse on these measures.

Results and Discussion

Performance Measures

Analyses were first performed on the three main dependent measures: users' guess, treatment, and test performance. Guesses were the initial suggestion that the user made about the required treatment which may or may not have been identical to the ultimate treatment administered. Guesses were considered correct if the most probable of the two possible treatments was given or if the diagnosis that would lead to the most probable treatment was given. For each diagnosis, there was only one correct treatment, so it could be assumed that if the subject had guessed the correct diagnosis, the correct treatment would have been administered. Treatments were considered correct if the most probable treatment was administered. The treatment score was not affected by the feedback provided at the end of the trial describing whether the
The data describing the 14 subject pairs and their performance on the major dependent variables are listed in Table 2. The fact that the experts' pre-session test scores are higher than the users' post-session test scores ($t(17) = 2.69, p < .05$) and the experts' practice session performance is greater than the
<table>
<thead>
<tr>
<th>Subject Pair</th>
<th>Expert's Class Experience</th>
<th>Expert's Tutoring Experience</th>
<th>Expert's Test Score (%)</th>
<th>Expert's Practice Session Score (%)</th>
<th>User's Guess Score (%)</th>
<th>User's Treatment Score (%)</th>
<th>User's Test Score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>87.3</td>
<td>89.5</td>
<td>50.0</td>
<td>85.7</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td>50.0</td>
<td></td>
<td>86.3</td>
<td>80.4</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
<td>77.3*</td>
<td></td>
<td>90.9*</td>
<td>76.8*</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>80.3</td>
<td>100.0</td>
<td>43.8</td>
<td>81.3</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td></td>
<td></td>
<td>40.9</td>
<td></td>
<td>86.4</td>
<td>48.2</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td></td>
<td></td>
<td>26.7</td>
<td></td>
<td>86.7</td>
<td>68.9</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>91.8</td>
<td>61.1</td>
<td>63.6*</td>
<td>100.0*</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td></td>
<td></td>
<td>63.6*</td>
<td></td>
<td>90.9*</td>
<td>71.4*</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>73.82</td>
<td>100.0</td>
<td>27.3</td>
<td>100.0*</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td></td>
<td></td>
<td>54.5*</td>
<td></td>
<td>59.1</td>
<td>78.6*</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td></td>
<td></td>
<td>44.4</td>
<td></td>
<td>88.9*</td>
<td>70.7</td>
</tr>
<tr>
<td>12</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>87.5</td>
<td>87.5</td>
<td>63.6*</td>
<td>95.5*</td>
</tr>
<tr>
<td>13</td>
<td>5</td>
<td></td>
<td></td>
<td>45.5</td>
<td></td>
<td>81.8</td>
<td>76.8*</td>
</tr>
<tr>
<td>14</td>
<td>5</td>
<td></td>
<td></td>
<td>54.5*</td>
<td></td>
<td>95.5*</td>
<td>87.5*</td>
</tr>
</tbody>
</table>

\(^a\) Refers to teaching experience prior to the study rated on a 5 point scale (1= no prior experience, 5= has taught more than one college level course)

\(^b\) Refers to tutoring experience prior to the study rated on a 5 point scale (1= no prior experience, 5= has tutored an academic subject from more than one year)

\(^*\) Score was above the mean
users' guess performance ($t (17) = 4.93, p < .0001$) indicates that the experts had a greater understanding of alien medicine from their training than did the users after they had completed the 22 trials. It appears that the one-on-one structured tutoring the experts received from the experimenter during the training session led to better performance compared to the training that the users' received. In other words, foregoing extensive structured training for a learning-by-doing training method was not beneficial in this situation, even when an expert was available to lend advice and answer questions.

It should also be mentioned that another explanation exists for these results. The subjects playing the expert role could be more capable of performing the task (due to overall aptitude, intelligence or motivation) than subjects playing the user role were. Without pre-training measures of ability, however, it is not clear which of these explanations is most plausible.

It was predicted that the users' performance measures would be significantly positively correlated with each other. Subjects who performed well on one of the performance measures should perform well on the other two measures. Surprisingly, this was not the case. As seen in Table 3, none of the correlations between the three performances measures were statistically significant (significance level for correlations was set a priori at $\alpha = .05$, this level will remain constant throughout this paper unless otherwise noted).

The fact that treatment performance is not correlated with the other two measures could be explained by the fact that treatment performance is the only measure directly influenced by the expert's help. Only rarely was the expert's treatment advice not followed by the user (users chose a non-recommended
treatment only eight times (.03 %) across all subjects and trials; four of these times (.015%) the problem was caused by an environmental problem so the expert's advice was incorrect). So, in a sense, this variable is indicative of the experts' ability to provide the correct advice and the users' ability to follow that advice. It was assumed that guess and test performance depended more heavily on the users' conceptual understanding of alien medicine. These two variables do have the highest correlation between the three performance measure pairs ($r = .338$); however, this correlation is not significant ($p = .24$). It would appear that these measures require different types of conceptual knowledge, or perhaps guess performance does not require conceptual knowledge at all. It is also possible that the test was not an accurate measure of subjects' conceptual knowledge. A split-half reliability analyses was performed on the test and the correlation between halves corrected using the Spearman Brown adjustment formula. The reliability of the conceptual test was $r = .73$ suggesting that the test was at least somewhat reliable.

Table 3

<table>
<thead>
<tr>
<th></th>
<th>Guess Performance</th>
<th>Treatment Performance</th>
<th>Test Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guess Performance</td>
<td>1.000</td>
<td>.068</td>
<td>.338</td>
</tr>
<tr>
<td>Treatment Performance</td>
<td>1.000</td>
<td>.008</td>
<td></td>
</tr>
<tr>
<td>Test Performance</td>
<td></td>
<td></td>
<td>1.000</td>
</tr>
</tbody>
</table>
The different experts used in this study probably also contributed to this result. Explanations across experts varied a great deal and therefore, so did the training that the users received. Most likely, the training received by some users included more information necessary for superior test performance, whereas other users' training provided more information necessary for accurate guess performance. Differences in the type of information users received most likely affected their guess and test performance differently, thereby causing low correlations between these two measures.

**Verbal Report Scoring**

The next phase of the analyses was to transcribe the verbal protocols of the 14 sessions. For each protocol, all comments made by the two subjects and the experimenter were recorded even when the comments were not directly relevant to the problem being performed at the time. An effort was made to accurately capture every event that happened during the experimental session in the transcription log. The transcribed logs were then studied and the frequency of numerous events and processes counted.

Various aspects of the explanations' presentation (was the advice given first then explanation or vice versa?, etc.), content (did the expert state if-then rules? use metaphors? refer to the probabilities? give procedural or conceptual information?, etc.), and timing (was the explanation volunteered by the expert or prompted by the user?, etc.) were noted. Characteristics of the expert (did she joke around, what types of questions did she ask of the user?, etc.) and the user (what types of questions did he ask? how often did he prompt for advice or an explanation?, etc.) were also measured. A total of 69 different variables were
coded. Most variables were measured by counting the frequency of their occurrence, whereas a few others such as the amount of control the user had over the interaction with the expert, were given subjective rankings by the experimenter.

**Correlational analyses - User performance**

Correlational analyses were run between each of the verbal protocol variables and the three user performance variables. Also included in this analysis were measures describing expert performance on the training test, expert performance during the practice session, and both subject's responses to the post experiment questionnaire. Because the goal of this study was to identify some elements of good vs. poor explanations, and it was assumed that good explanations lead to superior performance, the variables that correlate significantly with the performance measures are of most interest here. Various factors relevant to expert characteristics, user characteristics, and the explanations themselves correlated significantly with users' performance. Table 4 describes the correlations that reached statistical significance.

**Expert Characteristics.** Four variables describing characteristics of the expert subject were found to correlate significantly with user performance. As predicted, past teaching experience was significantly correlated with users' test performance ($r = .56, p < .05$). Apparently, teachers are able to teach test knowledge better than those without teaching experience.

Experts' performance during training correlated with users' performance as well. Experts' test scores were positively correlated with users guess performance ($r = .62, p < .05$), whereas experts' performance on the practice
Table 4
Significant Correlations between User Performance and Expert, User, and Explanation Characteristics from Experiment 1

<table>
<thead>
<tr>
<th>Expert Characteristics</th>
<th>Performance Measure</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test performance during the practice session</td>
<td>Guess</td>
<td>.62</td>
</tr>
<tr>
<td>Performance on practice problems</td>
<td>Guess</td>
<td>-.60</td>
</tr>
<tr>
<td>Prompts for users' understanding of the explanation</td>
<td>Guess</td>
<td>-.59</td>
</tr>
<tr>
<td>Teaching experience gained prior to the study</td>
<td>Test</td>
<td>.56</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>User Characteristics</th>
<th>Performance Measure</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question asking if expert agrees with the guess</td>
<td>Guess</td>
<td>.54</td>
</tr>
<tr>
<td>Number of times the expert is prompted for advice</td>
<td>Treat</td>
<td>.61</td>
</tr>
<tr>
<td>Question asking about the information displayed on one of the monitors</td>
<td>Treat</td>
<td>.56</td>
</tr>
<tr>
<td>Number of times the user said they didn't understand something</td>
<td>Test</td>
<td>-.64</td>
</tr>
<tr>
<td>Question asking the expert to clarify something they had said</td>
<td>Test</td>
<td>.56</td>
</tr>
<tr>
<td>Question about the treatment required</td>
<td>Test</td>
<td>.52</td>
</tr>
<tr>
<td>Number of times an explanation was received after an incorrect guess was made</td>
<td>Test</td>
<td>.62</td>
</tr>
</tbody>
</table>
(Table 4, Continued)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Performance Measure</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explanation of a symptom-diagnosis rule</td>
<td>Guess</td>
<td>-.61</td>
</tr>
<tr>
<td>Explanation of a diagnosis-treatment rule</td>
<td>Guess</td>
<td>-.54</td>
</tr>
<tr>
<td>Alien symptoms restated by the expert</td>
<td>Guess</td>
<td>-.60</td>
</tr>
<tr>
<td>Rules provided describing where required elements can be found in the alien's environment</td>
<td>Guess</td>
<td>-.62</td>
</tr>
<tr>
<td>Expert responses that provided the advice before the explanation</td>
<td>Guess</td>
<td>.58</td>
</tr>
<tr>
<td>Expert responses that provided the explanation before the advice</td>
<td>Guess</td>
<td>-.62</td>
</tr>
</tbody>
</table>
problems was negatively correlated with users' guess performance ($r = -.60$, $p < .05$). This result could be explained by differentiating "knowing that" versus "knowing how" (Anderson, 1983). That is, experts with conceptual domain knowledge as measured by a test are able to explain this domain to others. Experts who are able to perform a task well, however, are not necessarily able to explain the knowledge they are using in task performance. This is a conclusion knowledge engineers have often reached after attempting the difficult task of eliciting an expert's knowledge for use in an expert system (e.g., Mrózek, 1992). These correlations also support the general conclusion that the test and task performance require different types of domain knowledge.

Finally, expert prompts for feedback from the user about their understanding were negatively correlated with guess performance ($r = -.59$, $p < .05$). Intuition suggests that this correlation would be in the opposite direction: experts who prompt the user for feedback should be able to identify users' misunderstandings and provide tailored explanations to address the misunderstandings, resulting in increased user understanding. Perhaps the correlation makes more sense when viewed in the other direction. When experts noticed users' guess performance decreasing, they prompted the user for feedback about their understanding level more often.

**User characteristics.** The six user characteristics significantly correlated with performance are all measures of users' participation level with the expert. Users who prompted the expert for advice (especially after they had made an incorrect guess) and asked questions of the expert had higher performance scores. Lower performance scores were associated with users who acted more
passively during the study, that is, they asked the expert for help less often. Perhaps those who were performing poorly did not have enough domain knowledge to identify the holes in their understanding and to formulate a useful question. This conclusion supports previous findings that only those that know enough about a domain, ask questions (Miyake & Norman, 1979). Extreme novices do not understand the material well enough to formulate a question. It is only after some knowledge has been gained that subjects have the knowledge to ask questions.

**Explanation features.** Two of the explanation features significantly correlated with users' guess performance described the presentation order of the explanation. Superior guess performance is associated with explanations that state the advice first followed by an explanation for the advice. The direction of this correlation does not support the previous recommendations made by Goguen et al. (1983) that when complex explanations are given, explanations should be given in the form of reasoning then statement (i.e., explanation then advice). This research team suggested that order does not matter when explanations are not complex which might explain the contradictory results. However, the explanations provided in this study were quite integrated and complex.

Four additional explanation features that are significantly correlated with guess performance describe the content of the explanation (explanation of a symptom-diagnosis rule, explanation of a diagnosis-treatment rule, abnormal symptoms restated, and rules provided describing where required elements can be found in the alien's environment). These were all negatively correlated with
guess performance. It would appear that users were better at making correct guesses when they were provided with short, concise explanations by the expert. Yet, it seems most plausible that superior guess performance should require a deep, more conceptual understanding of alien medicine and therefore be associated with higher content explanations. Nonetheless, a few explanations exist for this result.

The example user-expert interaction shown in Figure 6 provide some clues about the negative correlation between performance and explanation content. The first interaction is between expert 5 and user 13, one of the bottom performers, and the second interaction is between expert 5 and user 14, one of the top performers. Each interaction lists the entire conversation between the expert and the user discussing the same problem. Each user makes the wrong guess, gets advice from the expert and administers the correct treatment. The explanation given to user 13 would appear to have more conceptual substance and be longer than the one given to user 14. However, user 14 clearly gained more from the interaction because she, contrary to the expert, was the one adding the conceptual knowledge underlying the advice to the interaction. The expert then only had to agree or disagree with her comments and questions. The expert provided less content when interacting with user 14 than with user 13 (66 words vs. 98 words), but the information shared was probably more meaningful to user 14. From this example it would appear that the content of the expert’s advice is not useful unless the user wants the information or realizes how it can be helpful. User 13 was provided with ample conceptual information, but it is not clear if she wanted it or felt she needed it. Through
**Figure 6.** Two interactions between Expert 5 and User 13, a bottom performer, and Expert 5 and User 14, a top performer.

### Interaction Between Expert 5 and User 13

**User:** I think that um, the failure is ink production, low heat, and I should increase heat. What do you think, expert?

**Expert:** Well, the temperature gauge doesn't seem to be highlighted so the temperature is probably normal.

**User:** Oh, that's right.

**Expert:** Um, but, there is some obvious, something with the ink.

**User:** Yeah, I suppose.

**Expert:** Now, one of the first things to look at is the communication, uh, box, the out signal?

**User:** Right.

**Expert:** It's sending in a, a short tall tall tall which indicates that there's something wrong with the water. And, oil is found in the water and oil is used to make ink, ... so since the ink system is failing, there's probably a shortage of oil. So you probably want to add some oil to the water.

**User:** OK.

(Figure Continues)
Interaction Between Expert 5 and User 14

User: OK. I'm going to guess that there is ah, ... I don't know, and ink production problem due to, ... I don't know, due to low lithium. OK, expert ...

Expert: M-hm.

User: Um, let's see, I've got, it would be, ... can the fact that it can't produce ink tell itself that it's in trouble and it's sending out a danger signal? What's the, what do you think the problem is?

Expert: Well, that's not a danger signal coming out.

User: Well, I mean, but there's something wrong with it.

Expert: Right. That signal, short tall tall tall?

User: M-hm.

Expert: Means that there's something wrong with the water environment and ...

User: Oh, um, not the air though, right?

Expert: No, the, do you remember what the air was?

User: Air has something ... OK, but it's something with the water?

Expert: Right.

User: OK, ... oil? Does it make ink from oil?

Expert: Yes.

User: Do you think it's the oil?

Expert: Yes, I think there's low oil.

User: Are there any previous cases for that? ... new cases that I don't ...

Expert: No. This is um, ... this is the first time. But there's a 95% chance that it's wrong.

User: So, if it's the oil, what would you do?

Expert: I'd add oil to the tank.

User: Add oil? OK.
actively communicating with the expert (i.e., by asking questions), however, user 14 received the exact information she needed and directed the expert's explanation to fill the gaps in her understanding.

The negative correlation between explanation content and guess performance could also be explained in another way. Rather than assuming that a high content level leads to poor performance, it could be that poor performance triggers high content explanations. As the expert notices the user frequently making incorrect guesses, they might try to give more detailed explanations to improve the user's understanding.

**Correlational analyses - Expert practice effects**

There were several interesting effects that practice teaching or tutoring or practice explaining the alien domain during the experimental sessions had on the types of explanations experts' provided. Whereas this data is not the main focus of this study, it is nonetheless worthy of a brief review.

The five experts in this study were recruited for their varying levels of teaching or tutoring experience prior to this study. It was hypothesized that those who had experience explaining in a classroom or tutoring situation would provide different explanations from those who had no formal explanation practice. The frequency of occurrence of several explanation features were significantly correlated with experts' teaching and tutoring experience.

When the user guessed the wrong treatment, experienced teachers were more likely to explain why the guess was incorrect than were subjects with no teaching experience ($r = .68, p < .01$). Experienced teachers were also more likely to provide information about the rules behind the diagnosis and its
corresponding treatment ($r = .65, p < .05$). They were more likely to make references to information or problems discussed earlier ($r = .61, p < .05$), and to provide help in the form of an explanation, followed by the advice, followed by more explanation ($r = .62, p < .05$). When users suggested the correct treatment, experienced teachers usually waited for the user to prompt them before they offered an explanation rather than always offering advice whether the user wanted or needed it ($r = .59, p < .05$).

Subjects with experience tutoring were more likely to make anthropomorphic statements about the alien such as calling it a "him" or saying it was "thinking" or "feeling" something ($r = .75, p < .005$). Experienced tutors also used more words to explain their treatment recommendations ($r = .62, p < .05$), and asked users to clarify their comments or questions more frequently ($r = .57, p < .05$) than those with no tutoring experience.

Another hypothesis made was that explanations would change with practice explaining the domain as experts interacted with more users. Expert practice of this type was significantly correlated with the number of words used to give diagnosis or treatment advice ($r = -.59, p < .05$) and with the number of words used to explain the treatment advice ($r = -.61, p < .05$). As experts became more experienced, their advice and explanation became less wordy.

**Summary**

This study was an observational study of human explanations that was performed to generate hypotheses about some explanation features that affect user performance. Several characteristics of the experts, users, and explanations were significantly correlated with user performance. The goal of
the next study was to confirm hypotheses about explanation features found in this study. The explanation variables that were significantly correlated with user performance describe explanation presentation order and content. These two features were manipulated and tested in Experiment 2. Explanations were either presented in advice/explanation order or vice versa and content level was varied. To keep the number of between-subject conditions to a minimum, the four features corresponding to explanation content were combined to produce three levels of content (none, low, and high).

Experiment 2

The goal of this study is to test the explanation features that were significantly correlated with performance in Experiment 1. Presentation order of the advice and justification and amount of conceptual information provided were both manipulated in this experiment.

Based on the data from Experiment 1, it was predicted that the advice-explanation order presentation and the low content condition would lead to superior guess performance over the other conditions. It was also hypothesized that a curvilinear relationship would be found between content level and user performance. Experiment 1 showed performance was best with a low rather than high amount of conceptual information. It seems logical, however, that subjects given no conceptual information would perform even worse than those given a small amount.
Method

Subjects

A total of 50 subjects participated in this study. All subjects were Rice undergraduates voluntarily participating to partially fulfill course requirements. The data from only 48 of the 50 subjects tested are presented here. A system error occurred during another subject's experimental session and one subject was hearing-impaired and could not accurately hear the experimenter's advice and explanation when given.

All subjects played the user role in this study. Because the explanation features were manipulated in this study, the expert role used in Experiment 1 was played by the experimenter.

Design

A 2 x 3 factorial between-subjects design was used in this study. Two levels of explanation presentation were tested (advice followed by an explanation or the explanation followed by the advice), as well as three levels of conceptual information (no, low, or high content).

Because the underlying purpose of this study was similar to the first study's goal (to determine which of the explanation features are helpful in the sense that they lead to the most accurate performance by the user), the same dependent measures of the user's performance and conceptual information were observed. Guess and treatment performance were measured as in Experiment 1 and the same comprehension test covering the conceptual knowledge behind alien medicine was given to the users after the session had ended. In addition, a final questionnaire asking for the subjects' opinions about
the expert and her advice and explanations was administered after completion of the monitoring portion of the study.

Materials

The same materials were used in this study as in the last. The experimenter (rather than an "expert" subject as in Experiment 1) and the user interacted with the same HyperCard program and communicated with each other through a partition. Subjects (all playing the user role) received the same procedural, conceptual, anecdotal, and environmental information as the users in Experiment 1. Recall that this is a subset of the information that the subjects playing the expert role were given in the first study (Appendix A lists the materials given to the subjects playing the user and the expert roles in Experiment 1).

The experimenter used pre-written explanations and, when possible, pre-written answers to subjects’ questions. Because each subject worked through the same problems, the correct advice and explanations could be anticipated. The features of the explanations were different depending on the condition to which the subject was assigned prior to the experimental session. Therefore, six explanations were generated for every problem, each representative of a different condition.

Each explanation contained one or two sentences stating the specific problem the alien was suffering (the diagnosis) and what treatment should be administered to treat the problem. These sentences were either placed at the beginning or the end of the experts’ response depending on the presentation order condition (advice/explanation order or explanation/advice order).
For each presentation order, the explanations were written to contain either high, low, or no conceptual information. Conceptual information was defined as a combination of explaining why the diagnosis decision is correct, explaining why the recommended treatment will cure the problem, stating or pointing out a symptom, and describing where in the alien's environment an important element can be found. All four of these explanation elements were significantly correlated with users' guess performance in Experiment 1 (recall, however, the correlation was negative).

Decisions about how much conceptual information should be placed in the high, low and no content explanations were made by examining the explanations from Experiment 1. The amount of conceptual information given to each subject was calculated. This measure was plotted against users' guess performance. The natural breaks in this plot served as the boundaries for the high, low and no conceptual information groups to be used in Experiment 2. Explanations provided by the experts in Experiment 1 in the high, low, and no content ranges, were used as a basis for the explanations created for Experiment 2. Summed across all 22 problems, Experiment 2 explanations in the high content group contained 97 "pieces" (separate ideas or facts) of conceptual information, explanations in the low content group contained 29 pieces, and explanations in the no content group contained 0 pieces of conceptual information. Appendix D lists the explanations created for use in each of the six conditions tested in Experiment 2.

To keep answers to users' questions consistent within and across conditions, answers to anticipated questions and additional explanation of
major concepts in alien medicine were prepared. When the subject asked about a specific point, or asked for some information to be repeated, the relevant information was read. The content condition the user was assigned to influenced the amount of additional information provided. When unanticipated questions were asked, an effort was made to answer within the limitations of the subject's condition.

A questionnaire was also developed to inquire about the subjects' perceptions of the expert, her advice and her explanations. The questionnaire consisted of 15 7-point Likert scales and 3 short answer questions (this questionnaire is listed in Appendix E).

**Procedure**

Subjects were randomly assigned to one of the six conditions. All procedures were identical to those of Experiment 1. Contrary to the Experiment 1, however, subjects in this study saw the 22 problems in the same order. A consistent presentation order was used so that performance on the first few problems could be used as a pretest of users' initial problem solving abilities. After the subjects completed the 22 problems, they were given the same comprehensive test as in Experiment 1 as well as a questionnaire asking about their satisfaction with the expert and her advice.

**Results and Discussion**

**Content Manipulation**

One thing that became evident during this study was that the number of times users prompted the expert for advice varied a great deal across subjects. Additionally, users did not ask for the expert's advice as much as anticipated
and, therefore, did not receive much explanation. On average, subjects asked for expert input on only 48% of the problems. This meant that the subjects' a priori assignment to an explanation content condition may not have accurately reflected the amount of explanation content they actually received. It is possible that a subject in the low content group who asked for advice on nearly every problem could have received more conceptual information than a subject in the high content condition that rarely asked for help. For this reason a manipulation check on content was performed.

The amount of conceptual information the subjects received through the expert's explanations and answers to questions was calculated. Using the original criterion for content level assignment, the subjects were reassigned to content groups based on these new content scores. Five subjects were reassigned to a different group. Further examination of the conceptual information scores created some concern that the high, low and no content groups were not distinct from each other. Several conceptual information scores fell close to group boundaries. For example, several subjects assigned to the high content group received only one or two more pieces of conceptual information than some subjects assigned to the low content group. Therefore, subjects whose conceptual information score was greater or less than 2 standard deviations away from the content group mean were eliminated from the analysis. Under this criterion, 5 subjects were eliminated from the data set and the new distinct content groups were defined as follows: high content- 31 to 73 pieces of conceptual information across all trials (n = 14), low content- 11 to 26 pieces (n = 15), and no content- 0 to 1 piece (n = 14).
Performance Measures

Subjects' guess, treatment and the test performance were all scored following the same criteria as described in Experiment 1. Since everyone in this study completed all 22 problems or trials, there was no need to create individual criteria for scoring the test questions as was done in the first study.

It was noticed during this study that as in Experiment 1, there seemed to be wide individual differences in how the users approached the task which affected their performance. Although these differences are interesting in their own right, it is informative to evaluate explanation quality independent of as many of these factors as possible. By analyzing percent improvement scores rather than raw performance scores, differences in subjects' initial task abilities can be equated.

Subjects' guess performance on the first four problems was used as a measure of initial (pre-explanation) performance. Guess, treatment, and test performance scores were all converted to percentages from which percent initial performance was subtracted to create gain scores. These gain scores were then converted to percent improvement scores by dividing each by the original performance scores. Converting the raw performance scores to percent improvement scores gives high performance scores earned by subjects with low initial abilities more weight than high performance scores earned by subjects with high initial abilities. This data transformation technique was a way to decreased the variance in the data without changing the direction of the underlying trends. The raw data (before transformation) from this study is listed Appendix F.
Contrary to the first experiment, the three performance measures were significantly correlated with each other in this study (See Table 5), that is, high scores on one performance measure are associated with high scores on the other two measures. Perhaps the careful planning that went into the explanations used in this study ensured that they all contained information necessary for guess, treatment, and test performance. As was discussed in Experiment 1, some of the experts used in that study may have unconsciously provided explanations containing only one type of knowledge which could explain the low correlations between performance measures found in that study.

Table 5

Correlation Matrix for Percent Improvement Scores from Experiment 2

<table>
<thead>
<tr>
<th></th>
<th>Guess Performance</th>
<th>Treatment Performance</th>
<th>Test Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guess Performance</td>
<td>1.000</td>
<td>.898***</td>
<td>.829***</td>
</tr>
<tr>
<td>Treatment Performance</td>
<td>1.000</td>
<td>1.000</td>
<td>.855***</td>
</tr>
<tr>
<td>Test Performance</td>
<td></td>
<td></td>
<td>1.000</td>
</tr>
</tbody>
</table>

*** p < .0001

The means and standard deviations for each of the three measures in the six conditions are listed in Table 6. Scores on the three performance
Table 6

Cell Means and Standard Deviations (in parentheses) of Percent Improvement Scores from Experiment 2

<table>
<thead>
<tr>
<th>Dependent Measure</th>
<th>Presentation Order</th>
<th>Content</th>
<th>None</th>
<th>Low</th>
<th>High</th>
<th>Row Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guess Performance</td>
<td>Advice First</td>
<td>41.48</td>
<td>39.92</td>
<td>43.75</td>
<td></td>
<td>41.53</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(17.83)</td>
<td>(24.04)</td>
<td>(4.59)</td>
<td></td>
<td>(17.48)</td>
</tr>
<tr>
<td></td>
<td>Explanation First</td>
<td>40.12</td>
<td>45.97</td>
<td>48.93</td>
<td></td>
<td>45.19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(16.14)</td>
<td>(24.13)</td>
<td>(14.49)</td>
<td></td>
<td>(18.03)</td>
</tr>
<tr>
<td>Column Mean</td>
<td></td>
<td>40.80</td>
<td>42.74</td>
<td>46.71</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(16.36)</td>
<td>(23.42)</td>
<td>(11.32)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Table Continues)
<table>
<thead>
<tr>
<th>Dependent Measure</th>
<th>Presentation Order</th>
<th>Content</th>
<th>Row Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>None</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Treatment</td>
<td>Advice First</td>
<td>11.79</td>
<td>1.42</td>
</tr>
<tr>
<td>Performance</td>
<td></td>
<td>(30.67)</td>
<td>(45.50)</td>
</tr>
<tr>
<td></td>
<td>Explanation First</td>
<td>-0.36</td>
<td>8.32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(14.10)</td>
<td>(41.60)</td>
</tr>
<tr>
<td></td>
<td>Column Mean</td>
<td>5.71</td>
<td>4.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(23.79)</td>
<td>(42.30)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dependent Measure</th>
<th>Presentation Order</th>
<th>Content</th>
<th>Row Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>None</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Test</td>
<td>Advice First</td>
<td>39.70</td>
<td>35.92</td>
</tr>
<tr>
<td>Performance</td>
<td></td>
<td>(21.53)</td>
<td>(24.48)</td>
</tr>
<tr>
<td></td>
<td>Explanation First</td>
<td>31.23</td>
<td>41.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(17.85)</td>
<td>(25.98)</td>
</tr>
<tr>
<td></td>
<td>Column Mean</td>
<td>35.47</td>
<td>38.74</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(19.50)</td>
<td>(24.47)</td>
</tr>
</tbody>
</table>
measures were submitted to a 2 x 3 between subjects ANOVA. There were no significant effects found for any of the performance measures, although all three measures showed a slight, though nonsignificant, advantage of high content explanations presented in the explanation/advice order in the guess data.

Overall, subjects' raw data scores on the performance measures were below expected performance levels, especially on the guess measure (guess performance mean: 43.6%, treatment performance mean: 84.1%, test performance mean: 79.0%). The wide ranges of raw scores (guess: 22.7% - 63.6%, treatment: 54.5% - 100%, test: 49.2% - 95.2%) supports the absence of a floor effect and merely suggests the task was quite difficult. In fact, it is likely that the task was too complex for performance to be affected by any differences in explanations. Perhaps, explanation effects, if present, are evident in the later trials. To test this hypothesis, guess and treatment percent improvement scores for the last 11 trials were analyzed. Again no significant effects were found.

**Subjective Ratings**

The post-task questionnaire was scored and each question was submitted to a 2 x 3 between subjects ANOVA. The means and standard deviations for the questions associated with significant effects are presented in Appendix G. The more content the subjects were given, the more reliable and useful they felt the expert's advice was ($F(2,42)=2.98$, $p=.06$; $F(2,42)=3.28$, $p=.05$, respectively). Apparently, even though the amount of content did not affect performance, it did affect subjects' perception of the advice. Providing subjects with high content explanations increased their perception of the explanation's reliability and usefulness. Although not significant, users'
preference for explanations with high content agrees with their performance which improves with higher levels of content. This effect is in the opposite direction from content effect from the first study that supported the use of explanations with lower levels of content.

When asked how difficult it was to decide whether or not to take the expert's advice, those in the advice/explanation presentation order condition felt it was more difficult than those in the explanation/advice presentation order condition ($F(2,42)=4.57$, $p<.05$). This was the only question that was significantly affected by the presentation order manipulation.

**Individual Differences**

On the whole, the results of this study suggest that explanation content level and presentation order have little affect on domain performance which would indicate that correlations found in Experiment 1 may not imply causal relations. However, observations of the subjects in both studies suggest that subject performance is heavily dependent on user characteristics. Even though initial task ability was taken into account in this study, other individual differences were also apparent. It is plausible that the possibly large effects of user characteristics were overwhelming any of the more subtle effects of the explanations themselves.

As mentioned before, several measures of user characteristics were significantly correlated with user performance in Experiment 1. Specifically, users who were more active (i.e., they prompted the expert for advice and asked the expert questions), tended to have higher performance scores than the more passive users. To determine whether these effects were robust enough to be
replicated in Experiment 2, correlations between the user characteristics measured in Experiment 2 and user performance were performed. The significant correlations are listed in Table 7. As found in Experiment 1, active users tended to have higher performance scores. Prompting the expert for advice, asking questions, and taking notes all led to superior performance in this study as well.

Table 7

**Significant Correlations between User Performance and Characteristics of the User from Experiment 2**

<table>
<thead>
<tr>
<th>User Characteristics</th>
<th>Performance Measure</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert's advice prompted</td>
<td>Treat</td>
<td>.52</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>.35</td>
</tr>
<tr>
<td>Expert asked to clarify something they just said</td>
<td>Treat</td>
<td>.46</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>.38</td>
</tr>
<tr>
<td>Total number of questions asked of expert</td>
<td>Test</td>
<td>.47</td>
</tr>
<tr>
<td>Amount of notes taken during the monitoring task</td>
<td>Test</td>
<td>.46</td>
</tr>
</tbody>
</table>
Experiment 3

The results of correlational analyses of Experiments 1 and 2 support the conclusion that a high user participation level is associated with improvements in user performance. For this reason, the degree of subject participation in the task was manipulated in Experiment 3. Subjects were either forced to interact actively or were allowed to interact passively with the expert. Subjects in the active condition always received the expert's advice and explanation and were forced to ask at least one question on each trial. Subjects in the passive condition were given the option to ask for advice, an explanation, or other questions.

Ideally, both explanation content level and presentation order should be tested along with subject-participation level in the third experiment. In efforts to keep the number of conditions to a manageable number, however, the presentation order variable was not tested. There was little support in Experiment 2 that this explanation element had any affect on performance or subjects' subjective opinion. Explanation content level was instead chosen for manipulation in this study.

In addition to testing subject participation level, the task was altered slightly to increase the experiment's external validity. That is, it would be premature to assume that the conclusions drawn from this human-human explanation task could generalize to expert system explanations. A leap like this must first be tested. Although several feel expert system explanations should be based on human explanation (e.g., Goguen, et al., 1983; Leake, 1991; Shortliffe, 1982), some have warned that human explanation should not
be incorporated verbatim into an expert system environment (Richards & Underwood, 1984). Of course, there is still a great deal to be learned about human-human explanation. To determine the effect of explanation content level and subject participation in human-expert system interactions (the ultimate goal of this project), the spoken interaction between expert and user used in the first two studies was replaced by electronic communication.

The purpose of Experiment 3 was to test, in an ecologically valid paradigm, hypotheses regarding the effects of explanation features and user characteristics on user performance. The amount of conceptual information was again manipulated (a low amount versus a high amount), as well as subjects' participation level with the expert (passive versus active). Subjects in this study performed the same alien monitoring task as those in Experiments 1 and 2, but the oral explanation given by the expert was replaced with a typed explanation displayed on the computer screen. This manipulation made the explanation environment more similar to interactions with an expert system.

Several studies have investigated the differences between the system interactions of subjects who think they are interacting with another human through a computer and subjects who think they are interacting with an intelligent system (e.g., Falzon; 1990; Richards & Underwood, 1984). Usually subjects communicate quite differently in these two situations. For example, subjects who think they are communicating with a machine tend to simplify their expression and limit their dialogue (Falzon, 1990). If differences exist in the way users communicate in these situations, perhaps differences in the helpfulness of some explanation features will also differ across these two
conditions. Manipulation of the user's perceived communication partner
(human or computer) should provide useful information about which
explanation features are supported at various levels of electronic
communication. This manipulation should also allow the results of this study to
be interpreted more precisely. If a human expert condition was not included in
this study, any significant results in the computer expert condition could not be
attributed to either the manipulated variables or the expert system environment
with any degree of confidence. Explanation components that affect
performance in one condition but not the other are somehow affected by user's
attitudes towards electronic communication or intelligent systems.

Effects of explanation content level and the manipulation of
communication partner type (computer or human) from previous studies are
mixed. However, based on the results of the last two studies and on
convention, it was predicted that subjects given explanations with low content
levels and told that they are interacting with a human expert would achieve
superior performance scores. Only the active subject-participation level has
been supported by data as repeatedly enhancing subject performance.
Therefore, it was also hypothesized that subjects who were forced to actively
interact with the expert should out-perform those who had no constraints on
their interactions. It was also considered that forcing subjects to interact actively
may not have the same beneficial results as if they had actively interacted on
their own initiative. It could be that active interactions cannot be forced on
subjects.
Method

Subjects

A total of 66 subjects participated in this study. All subjects were undergraduates enrolled in the Introduction to Psychology course at New Mexico State University. Subjects voluntarily participated in the study to partially fulfill course requirements.

The data from only 64 of the 66 subjects tested were analyzed. All subjects’ performance was monitored by the experimenter and it was obvious that two of the subjects were not putting effort into the task. These subjects moved from screen to screen too fast to have read any advice or to have thought about the correct treatment before administering one. The data for these two subjects were dropped from the analyses.

Design

A 2 x 2 x 2 factorial between subjects design was used in this study. Two levels of explanation content (low or high content) were tested as well as two levels of user participation (active and passive) and perceived expert type (human or computer). The explanations used in the low and high content conditions were identical to those used in the high and low content, advice-first presentation conditions in Experiment 2.

---

1 Between running the second and third studies, the author moved to New Mexico. For the purposes of this study, the student population used from New Mexico State University is similar enough to the student population of Rice University used in the first two studies to dismiss any concerns about comparing data from two different populations.
As in the previous two studies, users' guess, treatment and test performance was measured as well as subjective measures of users' attitudes towards the expert's advice and explanations.

**Materials**

One of the conclusions drawn from Experiment 2 was that the rules behind the task were very complex and they took a great deal of practice to learn. For this reason, the alien domain was modified and made slightly easier in this study. The diagnosis portion of the task was simplified by requiring users to monitor 9 symptoms on 7 monitors rather than 12 symptoms on 8 monitors as required by the previous studies. To allow for more improvement from practice, the number of unique problems was reduced from 22 to 16 and 7 problems were repeated for a total of 23 problems. Of the 16 unique problems, 7 were given to the subject on their treatment chart and 3 were solved by treating a problem in the environment. None of the problems that were repeated were on the treatment chart or caused by environmental influences. Subjects in this study all solved the same 23 problems in the same order.

Subjects performed the same alien monitoring task as the users in Experiments 1 and 2. However, rather than communicating orally with the human expert located behind a partition, subjects interacted with the perceived human or computer expert through the HyperCard program. For the subjects in the computer condition, the program appeared to perform the same actions (e.g., provide advice and answer questions) as the human expert in the first two studies. The subjects in the human condition were led to believe a human in another room was performing these actions, but communication was only
allowed via the computer. The computer program looked and operated exactly
the same in these two conditions.

In reality, both conditions used a Wizard of Oz simulation (Green & Wei-
Haas, 1985). The subjects’ computer was connected via a local network to the
experimenter’s terminal located in another room. All expert explanations and
subjects’ questions and answers were typed and sent to the other terminal. As
in the second study, explanations and answers to typical user questions were
developed prior to the study. These responses were then stored on the
experimenter’s terminal which allowed the experimenter to quickly provide
responses to subjects’ requests for advice or questions. Fast response time
was particularly necessary in the computer expert condition to enhance the
believability that the expert was a computer rather than a human. As in
Experiment 2, some questions were asked that the experimenter did not
anticipate and the answer to these questions were composed and typed on the
fly. However, an effort was made to keep these answers consistent within
conditions and across subjects.

Subjects in the active participation and the passive participation
conditions saw slightly different HyperCard programs. The program used by the
active participation group presented the expert’s advice and explanation on
each trial and forced the subject to ask at least one question for each problem.
After the users entered their guesses, a message was automatically sent to the
experimenter’s terminal requesting advice. The advice and explanation was
then sent to the subject by the experimenter. If the subject attempted to treat the
alien before they asked the expert a question, an error message appeared on
the screen.

The program used by the subjects in the passive participation condition
gave subjects the option to ask for advice and ask questions. A button labeled
"Advice" and one labeled with a question mark were presented on the screen
after the subjects entered their guesses. When selected, these buttons sent a
message to the experimenter's terminal requesting advice or brought up a
dialogue box prompting for a question to be entered, respectively. Subjects in
this condition were allowed to treat the alien without selecting either of these
buttons.

Because the domain of alien medicine changed slightly in this study, new
training materials (e.g., the training booklet, the treatment chart, etc.) were
created. A new conceptual knowledge test was also developed. Questions
from the conceptual knowledge test that were answered correctly by more than
95% of the subjects in the first two studies and that were no longer relevant to
the simplified domain were eliminated. Several more questions were added to
make the revised version roughly the same length as the original. Finally,
questions were added to the questionnaire used in Experiment 2 which asked
for subjects opinions of their activity level during the study. The paper materials
which were revised for use in this study are presented in Appendix H.

Procedure
Subjects were randomly assigned to one of the eight conditions.
Subjects were provided with the same types of information as those in the
previous studies, although all the materials were revised somewhat to be
congruent with the simplified task domain. The same training and basic instructions about the task and performance goals were given to subjects in this study as were given to subjects in Experiments 1 and 2. The only difference between this study and the past two was the additional training on how to use the computer program and a short explanation about the expert with which the subjects would be interacting. Subjects in the human expert condition were told they were interacting with a person who was helping with the study and had been thoroughly trained in the task domain. Subjects in the computer expert condition were told they were using an expert system that had been programmed with the knowledge of an expert in the task domain.

Since the experimenter was not present in the room with the subject to answer questions as in previous studies and the task involved more interaction with the program in this study, more time was spent explaining how to use the program and subjects were walked through several practice problems. The programs in this study worked similarly to the ones used in previous studies, however there were a few additional steps. After a failure occurred, subjects in both participation conditions entered their guess into the computer as users had in the first two studies. At this point if the subject was in the active condition, a window appeared containing the expert's advice and explanation as well as a button labeled with a question mark. When this button was selected, a dialogue box was displayed that prompted the subject to type in a question and select the "OK" button. The subject's question and expert's response were both placed in the scrolling explanation window after the explanation. After the subject had read the advice and asked at least one question of the expert, they
were required to treat the alien using the same methods as subjects in previous studies.

If the subject was in the passive condition, after the initial guess was made a button labeled "Advice" and the question button appeared. If subjects choose to select the "Advice" button, a window containing the expert's advice was displayed. Subjects in this condition had the option of asking the expert a question by selecting the question button, however, they were not required to do so (see Figures 7a-7c for example screens of the diagnosis program used in this study). Treatment administration was performed in the same way as those in the active condition.

Feedback was given to subjects in both conditions. Subjects were told whether or not the alien got better or worse as a result of their treatment. They were also given a running tally of the number of guesses they had gotten right, the number of treatment administrations that cured the alien, and the number of problems completed. See Figure 8 for example feedback messages used in this study.

After the subjects completed the 23 problems, they were given the comprehension test and asked to complete the final questionnaire. Subjects were then debriefed and dismissed.

**Results and Discussion**

**Manipulation Check**

As in the second study, the assignment of subjects in the passive condition to high or low content conditions was checked post hoc. Contrary to those in the active condition, subjects in the passive condition did not always
Figure 7. A) Main screen from the Diagnosis Program used in Experiment 3. In this example, the user has just guessed "Increase Heat Intensity" as the correct treatment.
To treat this problem, you should decrease the temperature in Gus's tank. The temperature is over 80 degrees, the maximum temperature Gus can operate normally under. Gus is sending out a tss signal, which is his above water warning signal. This further confirms that there is something wrong with the air temperature.

Environmental Information
The principle scientist just entered the lab.

Your Suggestion: Increase heat intensity
Figure 7. C) Users' questions and the expert's answers (listed in between the stars) are displayed beneath the explanation and advice.
CONGRATULATIONS!

Gus appears to have recovered completely.
Job well done.

Your treatment suggestion was correct.
Correct treatment suggestions: 2
Correct treatment administrations: 2
Problems completed: 3

We regret to inform you that Gus has gotten much worse after your recommended treatment was administered. You have been taken off the case until Gus has recovered fully.

Your treatment suggestion was correct.
Correct treatment suggestions: 2
Correct treatment administrations: 2
Problems completed: 3

Figure 8. A) Positive feedback screen shown when treatment administered was successful. B) Negative feedback screen shown when treatment was unsuccessful.
receive expert advice and thus, may not have received the correct amount of content for their assigned condition. Across trials, subjects asked for the experts' advice on 87% of the trials. The amount of information actually given to these subjects was calculated and unlike the second study, no subjects needed to be reassigned to a different content group. However, to ensure the high and low content groups were truly distinct from each other, subjects receiving more or less information "pieces" than 2 standard deviations from their group mean were eliminated from the data set. Four subjects were dropped for this reason.

Performance Measures

Subjects' guess, treatment, and test performance were scored using the same criteria as used in the first two studies. To account for individual differences in task performance caused by initial abilities held by subjects prior to the study, percent improvement scores were calculated for the three performance measures as in Experiment 2. The pretest in this study, however, consisted of the first three problems performed by the subjects rather than the first four as used previously. In this study nearly every subject got the fourth problem correct thereby limiting its usefulness as a predictor of inherent diagnostic ability. Pre-transformation raw data is listed in Appendix I.

As in the second study, the correlations between scores on the three performance measures are all significant and positive. The correlation matrix between these three measures is presented in Table 8. The performance means and standard deviations for each condition are listed in Table 9. The three performance scores were each submitted to a 2 x 2 x 2 way between subjects ANOVA to determine the effects of expert type, content level, and
subject participation. The expert-type by participation interaction was significant for the guess measure ($F(1, 59) = 5.04, p < .05$). When subjects are forced to participate actively, interaction with a human rather than a computer results in superior guess performance. When subjects are not restricted in their interactions (passive interaction), interacting with a computer expert produces slightly higher guess scores than interacting with a human expert. The type of expert that the subject interacted with affected test performance as well. Subjects who thought they were interacting with a human expert had higher test scores than those who thought they were interacting with a computer expert ($F(1, 59) = 5.06, p < .05$). No other effects were significant for the three performance measures.

Table 8
Correlation Matrix for Percent Improvement Scores on the Performance Measures from Experiment 3

<table>
<thead>
<tr>
<th></th>
<th>Guess Performance</th>
<th>Treatment Performance</th>
<th>Test Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guess</td>
<td>1.000</td>
<td>.942****</td>
<td>.833****</td>
</tr>
<tr>
<td>Treatment</td>
<td></td>
<td>1.000</td>
<td>.841****</td>
</tr>
<tr>
<td>Test</td>
<td></td>
<td></td>
<td>1.000</td>
</tr>
</tbody>
</table>

**** $p < .0001$
Table 9  
**Percent Improvement Score Cell Means and Standard Deviations (in parentheses) from Experiment 3**

<table>
<thead>
<tr>
<th>Dependent Measure</th>
<th>Expert Type</th>
<th>User Participation</th>
<th>Content Low</th>
<th>Content High</th>
<th>Row Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guess Performance</td>
<td>Human</td>
<td>Active</td>
<td>48.28</td>
<td>53.54</td>
<td>50.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(25.84)</td>
<td>(33.29)</td>
<td>(28.91)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Passive</td>
<td>24.97</td>
<td>28.33</td>
<td>26.41</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(45.24)</td>
<td>(16.21)</td>
<td>(34.73)</td>
</tr>
<tr>
<td></td>
<td>Computer</td>
<td>Active</td>
<td>39.63</td>
<td>1.52</td>
<td>20.57</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(11.70)</td>
<td>(55.41)</td>
<td>(43.41)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Passive</td>
<td>29.07</td>
<td>42.28</td>
<td>34.73</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(14.56)</td>
<td>(37.56)</td>
<td>(26.51)</td>
</tr>
<tr>
<td></td>
<td>Column Mean</td>
<td></td>
<td>35.49</td>
<td>30.86</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(27.88)</td>
<td>(42.79)</td>
<td></td>
</tr>
</tbody>
</table>

(Table Continues)
<table>
<thead>
<tr>
<th>Dependent Measure</th>
<th>Expert Type</th>
<th>User Participation</th>
<th>Content</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Low</td>
<td>High</td>
<td>Row Mean</td>
</tr>
<tr>
<td>Treatment</td>
<td>Human</td>
<td>Active</td>
<td>65.32</td>
<td>73.65</td>
<td>69.49</td>
</tr>
<tr>
<td>Performance</td>
<td></td>
<td></td>
<td>(17.85)</td>
<td>(16.56)</td>
<td>(17.18)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Passive</td>
<td>60.93</td>
<td>57.15</td>
<td>59.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(22.42)</td>
<td>(12.49)</td>
<td>(18.29)</td>
</tr>
<tr>
<td></td>
<td>Computer</td>
<td>Active</td>
<td>64.43</td>
<td>51.84</td>
<td>58.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(2.27)</td>
<td>(18.00)</td>
<td>(14.00)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Passive</td>
<td>60.66</td>
<td>64.80</td>
<td>62.43</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(11.27)</td>
<td>(22.13)</td>
<td>(16.16)</td>
</tr>
<tr>
<td></td>
<td>Column Mean</td>
<td></td>
<td>62.83</td>
<td>61.99</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(14.82)</td>
<td>(18.77)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dependent Measure</th>
<th>Expert Type</th>
<th>User Participation</th>
<th>Content</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Low</td>
<td>High</td>
<td>Row Mean</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test</td>
<td>Human</td>
<td>Active</td>
<td>56.42</td>
<td>65.67</td>
<td>61.05</td>
</tr>
<tr>
<td>Performance</td>
<td></td>
<td></td>
<td>(21.89)</td>
<td>(22.61)</td>
<td>(22.02)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Passive</td>
<td>43.85</td>
<td>44.92</td>
<td>44.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(32.85)</td>
<td>(14.18)</td>
<td>(25.65)</td>
</tr>
<tr>
<td></td>
<td>Computer</td>
<td>Active</td>
<td>42.95</td>
<td>21.08</td>
<td>32.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(13.72)</td>
<td>(37.00)</td>
<td>(29.23)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Passive</td>
<td>33.48</td>
<td>39.70</td>
<td>36.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(20.46)</td>
<td>(66.47)</td>
<td>(43.99)</td>
</tr>
<tr>
<td></td>
<td>Column Mean</td>
<td></td>
<td>44.17</td>
<td>42.92</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(23.61)</td>
<td>(40.50)</td>
<td></td>
</tr>
</tbody>
</table>
Even though the subjects' task was made simpler in this study, they still found it quite difficult producing low raw performance scores for the guess measure (performance mean = 49.6%). As in the second study, variance in the performance scores suggests a floor effect is not present (guess: 26.1% - 73.9%, test: 25.6% - 93.0%). It was therefore felt that elimination of early trials may reveal effects that manifested themselves later in learning. The last 10 trials were separated from the complete data set and the guess data were analyzed. In these analyses, the guess measure was the only measure significantly affected by the independent variables. Means and standard deviations for this measure are listed in Table 10. Significant interactions between participation and both content \( (E(1, 59) = 4.05, p < .05) \) and the expert type \( (E(1, 59) = 4.88, p < .05) \) were found. Subjects' guess performance in the active condition benefited from a human over a computer expert. This pattern was reversed for subjects in the passive interaction condition. Additionally, there was a benefit of active over passive participation when the expert type was human, whereas, when the expert was a computer, a passive interaction style was superior. The participation by expert type interaction is identical to the one found in the full data mentioned previously. Active participants benefit from explanations with low content levels and passive participants benefit from high levels of content. These interactions are shown in Figures 9 and 10. A marginally significant 3-way interaction between content, subject participation, and expert type \( (E(1, 59) = 3.46, p = .06) \) suggests that the content by subject participation interaction described above is only be present when subjects think they are interacting with a computer expert. In the human expert conditions,
content level had no effect on guess performance; however, active participants were superior performers to passive participants.

Table 10
Percent Improvement Score Cell Means and Standard Deviations (in parentheses) from the Last 11 Problems in Experiment 3

<table>
<thead>
<tr>
<th>Dependent Measure</th>
<th>Expert Type</th>
<th>User Participation</th>
<th>Content</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Low</td>
<td>High</td>
<td>Row Mean</td>
</tr>
<tr>
<td>Guess Performance</td>
<td>Human</td>
<td>Active</td>
<td>43.55</td>
<td>44.28</td>
<td>50.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(34.59)</td>
<td>(53.04)</td>
<td>(28.91)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Passive</td>
<td>-1.74</td>
<td>3.79</td>
<td>26.41</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(65.55)</td>
<td>(42.74)</td>
<td>(34.73)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Computer</td>
<td>Active</td>
<td>34.33</td>
<td>-49.40</td>
<td>20.57</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(12.86)</td>
<td>(120.41)</td>
<td>(43.41)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Passive</td>
<td>0.00</td>
<td>37.50</td>
<td>34.73</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(36.13)</td>
<td>(42.81)</td>
<td>(26.51)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Column Mean</td>
<td>35.49</td>
<td>30.86</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(27.88)</td>
<td>(42.79)</td>
<td></td>
</tr>
</tbody>
</table>

Subjective Ratings

The data from each question on the final questionnaire were submitted to a 2 x 2 x 2 way between subjects ANOVA. Ratings on several of the questions were affected by the subjects' assignment to explanation content, participation, and expert-type conditions. The cell means for all questions with significant effects are listed in Appendix J.
Figure 9. Effect of Explanation Content Level and Subject-Participation Level on Guess Percent Gain Scores in Experiment 3.
Figure 10. Effect of Subject-Participation Level and Perceived Expert Type on Guess Percent Gain Scores in Experiment 3.
Trends in the questionnaire data supported users' preference for explanations with low content, a human expert, and a passive interaction style. However, only the main effect of content level was statistically significant. The level of explanation content affected the way subjects viewed the expert's advice. Overall, subjects in the low content condition rated the expert more useful than subjects in the high content conditions did ($E(1, 56) = 4.21, p<.05$). If subjects were in the active condition where they were forced to ask a question on each trial, they were asked if they minded having to ask a question. Subjects in the low content condition reported that they disliked asking questions more than subjects in the high content condition did ($E(1, 31) = 8.84, p<.01$). Perhaps subjects in the low content condition were gaining everything they needed from the shorter explanations and did not feel it was necessary to ask additional questions. Subjects receiving explanations with high content, however, may have felt they were too complex to be useful, which left them confused and eager to ask questions.

Often participation level influenced content level effects. Ratings of both the usefulness and reliability of the expert's advice were influenced by a two-way interaction between content and participation ($E(1, 56) = 3.82, p=.05$; $E(1, 56) = 4.62, p<.05$, respectively). The direction of these interactions are similar to each other. Subjects in the active participation condition rated high content explanations as more useful and more reliable than subjects who received low content explanations. This pattern was reversed for users interacting passively. For this group of subjects, low content explanations were given superior ratings to high content explanations. These interactions are presented in Figure 11.
Figure 11. Effect of Explanation Content Level and Subject-Participation on Subjects' Ratings of Advice Usefulness and Reliability in Experiment 3.
Interestingly, this is completely opposite to the pattern of the content by participation interaction in the guess performance data discussed previously. In the guess data, subjects in the active condition receiving low content explanations and those in the passive condition receiving high content explanations were superior guessers. It was these superior guessers who rated the advice as less useful and reliable than the poorer guessers in the active, high content conditions and the passive, low content conditions. Evidently, poor guess performers who were having difficulty learning the task domain found the expert’s explanations more beneficial than subjects who were performing well and, therefore, rated them highly. Another explanation for this result is that explanation quality is difficult to judge, and subjects are not accurate at doing so.

A three-way interaction between content, participation and expert type was found for a question that asked subjects if they learned the domain as a result of interacting with the expert or more on their own ($E(1, 52) = 5.83, p<.05$) and for a question that asked if it was difficult to decide whether or not to follow the expert’s advice ($E(1, 52) = 5.34, p<.05$). Overall, subjects in all conditions rated the decision as being more easy than hard and learning more from the expert than themselves— it was merely a matter of degree that caused these three-way interactions.

In the human expert conditions, the content by participation interaction pattern is similar to the ones shown in Figure 11 for the advice usefulness and reliability data. Those who felt the experts advice was useful and reliable (active participation with low content and passive participation with high content
conditions) also felt they learned more from the expert than they picked up on their own, and felt the decision to take or ignore the expert's advice was easy to make.

When subjects thought they were interacting with a computer expert, the two-way interaction pattern described above was reversed. Low content levels were perceived as being beneficial (i.e., subjects learned more from the expert than on their own and felt it was easier to decide whether to follow the expert's advice) when they were forced to interact actively. Receiving high content explanations was beneficial when subjects were in the passive interaction condition. This pattern is similar to this interaction in the guess performance data discussed above- subjects who are superior guessers reported that the decision to take or ignore the advice was easy to make, and that they learned most of their knowledge from the expert.

Clearly, guess performance, subjective ratings of advice usefulness and reliability, and ratings of decision difficulty and the learning source are not related to each other in intuitive ways, if they are related at all. In the human expert condition, the content by participation interaction pattern found in the guess performance data is exactly opposite from the patterns for this interaction found in the four subjective ratings' data. In the computer expert condition, the guess data agree with the decision difficulty and learning source data, but is also opposite from the advice usefulness and reliability data. The differing effects between the two expert conditions is difficult to explain. Apparently, expert type influences subjects' ratings of decision difficulty and learning source, but not of advice usefulness and reliability. This effect is most likely
caused by the difference in subjects' knowledge and expectations about a human vs. a computer expert.

General Discussion

This project's goal was to study explanations along a continuum ranging from human-human interactions with some limitations imposed (i.e., visual cues eliminated) to human-expert system interactions. The progressive nature of these studies allowed conclusions to be drawn about both human-human and human-expert system interactions and the role explanations play in these exchanges. Perhaps the most salient finding across all three studies presented here is that simply manipulating explanation features will not affect the recipient's performance a great deal. Interactions between an instructor and a learner are complex, and numerous factors in addition to differences in explanations' features facilitate learners' understanding.

Wide differences exist across subjects such as how they approach the diagnosis task and their abilities held prior to the study. Overall, these individual differences appeared to affect performance as much, if not more than differences among explanations. This conclusion taken alone has little practical utility. The presence of individual differences is certainly not new to psychologists. However, this project was successful at identifying some of the differences that affect performance and interact with some features of explanations. Once these critical differences were either manipulated or controlled, explanation content level and expert type were both found to affect user performance. The individual difference effects observed in this series of
studies will be discussed in detail below, but a discussion of the explanation content level effects will first be presented.

**Explanation Content.** The effect of explanation content on user performance which was predicted to be simple— the more conceptual information provided, the better— turned out to be much more complex. The effect explanation content had on user performance differed across all three studies and appeared to interact with several other variables. Looking at each of the three studies independently, might lead to three different conclusions about beneficial levels of explanation content. However, when considered as a whole, the results across the studies tell a different story.

Recall that performance in Experiment 1 tended to be better for explanations with low content levels, whereas subjects' performance in Experiment 2 was not significantly affected by content level, although trends suggest that high content was superior to low. On the whole, low content levels were preferred by subjects in Experiment 3, but content effects on performance were not always in the same direction as effects on preference ratings and were mediated by participation. It is this content by participation interaction in Experiment 3 that provides an explanation for why the results from Experiment 1 were not replicated in Experiment 2.

As discussed above, subjects' guess performance in Experiment 3's active participation condition was facilitated by low content levels. Subjects' performance in the passive interaction condition was enhanced when high content explanations were provided. In general, subjects in Experiment 1 interacted more actively with the expert than the subjects did in Experiment 2.
The average subject in Experiment 1 asked for the expert's advice 76% of the time and asked the expert 11.43 questions across all 22 problems compared to the average subject in Experiment 2 who asked for advice 54% of the time and asked 2.03 questions across the 22 problems. These differences in activity level are probably due to the person playing the expert role in the two studies. In the first study subjects probably felt more at ease with another subject playing the expert role. In many cases, the two individuals knew each other. In the second study, the experimenter was playing the expert role and may have made interactions slightly more intimidating or uncomfortable. These differences in participation level by subjects in Experiments 1 and 2 apparently affect the optimal explanation content level. An advantage for explanations with low content was found in the first study where subjects were more active. This result was not replicated in Experiment 2 where subjects interacted in a more passive manner. These results agree with the content level by participation interaction in Experiment 3 which controlled for users' participation level.

User Characteristics. There is ample evidence from all three studies that user characteristics interacted with explanations and task characteristics to affect performance. It became clear early in Experiment 1 that some subjects walked into the experimental session with more initial abilities to perform the task than others. These abilities seemed to affect performance and possibly overwhelm any effect of explanation elements. Even after initial ability was somewhat equaled by converting performance scores to percent gain scores in Experiments 2 and 3, however, there were still wide differences in performance that seemed to be caused by individual differences.
Differences in users' participation with the expert was one of the most salient characteristics that was found to affect performance and explanation effects. When interacting with a human expert as in Experiments 1 and 2 and the human condition in Experiment 3, subjects who were more interactive by prompting for advice and asking questions performed better than those who were passive. The benefits of this active style were evident even when high participation levels were forced on subjects in Experiment 3. This finding has relevance for researchers in education and training domains. Trainers or instructors should strongly encourage or even force learners to request help and ask questions during training. The benefit of an active style was demonstrated in the subject interactions shown in Figure 6. Active learners tend to have more control over the direction of the conversation and therefore, are more likely to acquire the information they need to fill in gaps in their understanding.

Caution should be taken when forcing learners to participate actively however. On the whole, subjects in Experiment 3 did not mind getting expert advice on every trial, but they did dislike being forced to ask a question on each trial, especially if they were receiving explanations with high content levels. More research needs to be done to identify ways to subtly encourage learners to participate actively without sacrificing their satisfaction.

The expert by participation interaction from Experiment 3 suggests that the benefits of active participation do not generalize to interactions with expert systems. In fact, subjects who were forced to participate actively with the computer expert tended to have lose scores indicating that over the course of
the study, their guess performance got worse. Similar results have been found
by Carroll and Rosson (1987) when studying how users learn to use a word
processing program. In this situation, people are usually active learners, but
tend to focus too much on their desired end goal, rather than on the most
efficient method for achieving this goal. As a result, the most efficient methods
for performing the task are not learned. Carroll and Rosson describe the
learning process as follows: "...people are so busy trying things out, thinking
things through, and trying to relate what they already know (or believe they
know) to what is going on that they often do not notice the small voice of
structured instruction crying out to them from behind the manual and the system
interface" (pg. 82). Carroll and Rosson have named this phenomena the
"production paradox" because users who appear to be very productive by
interacting with the system more often perform less efficiently than those who
approach a new system cautiously and confirm actions with instructional
support.

Perhaps a similar phenomenon was occurring in Experiment 3. For
subjects in the active condition, who were forced to ask a question on every
problem, part of their end goal became, not only to acquire knowledge about
alien medicine, but also to successfully ask the expert a question. When the
subjects believed the expert was a human, this goal was simple and beneficial
to the learning goal. When the expert was believed to be a computer, however,
the question-asking goal became much more difficult. Because subjects did not
immediately understand what types of questions the expert system was capable
of answering, the question-asking task became an hurdle to overcome rather
than a learning aid. As a result of this added question-asking goal, effort was
taken away from reaching the learning goal and performance suffered.
Subjects in the passive condition, where question asking was optional, only
had one goal: to learn alien medicine. For these subjects question asking was
merely an optional step to help them reach this goal.

It is not clear from the data in Experiment 3 if users who are naturally
active and not forced to be so would exhibit similar performance decrements.
The subjects in Carroll and Rosson's (1987) studies were not forced to be
active, they merely preferred this style. Subjects in this study, however, did not
have a choice. This is an area worthy of future research.

Another explanation exists for the poor performance of subjects in the
active-participation, computer-expert condition. Perhaps subjects came to the
study with preconceived notions about what a computer expert was capable of
knowing and doing. This phenomenon was also observed by Carroll and
Rosson (1987) and termed the "assimilation paradox". According to these
researchers, users tend to apply prior knowledge to a new computer task even
when it does not apply, that is, prior knowledge is not always assimilated into
the new domain correctly. Users in Experiment 3 may have been slightly wary
of the computer expert because they did not bring as much prior knowledge
about this type of expert to the task or they underestimated the ability of the
computer expert to assist in task performance. Users in the human expert
condition, however, most likely entered the study with more accurate knowledge
about how the expert might be able to assist them. Due to their incomplete
knowledge and possible negative biases about computer experts, subjects in
the computer expert group might have preferred to gradually ease into using the expert system for consultation. Instead, users in the active participation condition were immediately forced to actively use the artificial expert. These subjects may have become frustrated with this approach and, therefore, not able to focus on task performance. Forcing otherwise passive users to interact actively with an expert system might have serious detrimental effects on their performance.

**Design for the User.** Carroll and Rosson (1987) suggest several design solutions to correct the production and assimilation paradoxes in user performance. According to these researchers, system design should, among other things, design for end-product focus by making the system friendly to active users. This can be achieved by decreasing the consequences of errors and increasing the potential for learning-by-doing. The assimilation paradox can be corrected by exploiting the accommodation users typically make when they do not correctly assimilate the metaphor used by the system interface. The system should be described as something truly similar to a system with which the user is familiar. The aspects of the system that do not overlap with a users’ prior knowledge must be identified and typical errors predicted and steps taken to help the user avoid making them. No doubt these suggestions can be generalized as solutions to the paradoxes noticed in expert system use as well.

Most likely subjects in the computer expert condition found the diagnosis task more difficult since they had to learn what the computer expert was capable of doing and may have had to overcome some negative biases about the abilities of artificial intelligence. This was probably especially difficult for users
who were forced to create questions they felt the expert system could handle on every problem. Perhaps if these subjects were given more information about the expert system itself (i.e., system knowledge or mental model), they would have had fewer problems using the expert effectively. Instructions prior to system use could emphasize that the information programmed into the system came from a human expert and explicit information could be provided about what types of questions the system can and cannot handle. Providing accurate system knowledge about the expert system to users should also help them calibrate their trust in the system to an accurate level (Muir, 1987).

Another design solution which the results of these studies suggest is in the form of adaptive systems that tailor their explanations to the individual user. Clearly, differences between expert system users exist and affect their interactions. Good human instructors easily adapt to these differences in knowledge, interaction style, and preference and tailor their explanations accordingly (Weiner, 1980). Expert systems should strive to do the same. This conclusion is not new. Several researchers (e.g., Carroll & McKendree, 1987; Finn, Joshi, & Weber, 1986; Gordon, 1987; Jackson, 1986) have advocated the need for intelligent systems to acquire information about the user and tailor their output to the specific goals, preferences, and needs of the user.

From this results of these studies, it can be concluded that the ideal expert system would deduce early in the users' session whether the user is a passive or active learner and adjust its output accordingly. To assess user activity level the system could measure the number of times the user asks for advice and the number of questions they ask. There are most likely several
other measures of user activity level that could be measured as well. Alternatively, there might be a way of determining activity level through a short questionnaire that users take before the session begins. However, these possibilities will have to be left for future research.

Once the system has determined the user's activity level it can adjust its explanations accordingly. Most performance and subject rating measures from these studies suggest users with passive interaction styles should be given explanations with high content levels and users with active interaction styles should be given explanations with low content levels. Logically, subjects who are receiving advice and explanations on every trial (active interaction) do not need explanations that are packed full of information. However, if users are not receiving advice on every problem and are not asking many questions, when explanations are provided, they should contain higher levels of information.

The above conclusions were based on the assumption that expert system designers will place more emphasis on user performance than on user satisfaction. As discussed in Experiment 3, performance and subjective ratings of the task and explanations were often in the opposite direction. Unfortunately, this makes it difficult to provide a system that optimizes performance and satisfaction. In the critical domain of patient diagnosis and treatment, optimizing performance (i.e., diagnosis and treatment accuracy) is probably more critical than the user's ratings of satisfaction, although user satisfaction should certainly be a goal of expert system design.

Lessons Learned about Experimental Design. Prior to the exploratory analysis there was little research available to identify the explanation features
that might affect user performance. Choosing to manipulate features from the myriad of possibilities would have been no more scientific than spinning a roulette wheel. The results of this project point out the importance of an initial exploratory analysis, especially in new and complex task domains. Psychologists, as a whole, acknowledge that people differ, but often they do so along some salient dimension, rather than haphazardly.

In the course of this project, it became clear that a lot of mileage could be gained by acknowledging the simple distinction between active and passive learners. Without the exploratory analysis, the variation resulting from this factor may have never been apparent. Data from both "types" of individuals would have been averaged together and important effects washed out.

The success of this project supports the following research approach to applied problems:

1) **Explore a real world problem to formulate hypothesis.** In this case, human explanation was observed to formulate hypotheses about the features of an explanation that affect recipient performance. The exploratory study suggested that features of explanation as well as characteristics of the user and the expert subjects, were systematically associated with variation in user performance.

2) **Test hypotheses in the lab.** Experiment 2 and the human condition in Experiment 3 tested the hypotheses about explanation features and user participation generated from the exploratory study.

3) **Extend results to the application and test again.** Conclusions drawn about human explanation were tested in an expert system domain in
Experiment 3's computer expert condition. Generalization to the applied domain (in this case the expert system domain) is not justified without empirical support. As found in Experiment 3, not all results generalize to the applied domain. Subjects' active participation which enhanced performance when interacting with a human expert was detrimental to performance when subjects thought they were interacting with an expert system.

Recommendations for Expert System Explanations

The final step required for the creation of successful expert system explanations is the translation of the results from these studies to recommendations for expert system designers. It is hoped that these recommendations can help designers determine what types of expert knowledge should be programmed into a system's explanation component and provide guidance for interface design.

1. Whenever possible, measure the user's interaction level and adjust explanation content level accordingly. Users who are active participants, that is they prompt for help often and ask a lot of questions, benefit from explanations with low levels of content, whereas users who interact passively with the system should be provided with high content explanations.

2. If it is impractical to create an adaptive system, high content explanations should be incorporated into an interface that allows the user to prompt for advice and ask questions at will. Whereas
designers struggling with screen limitations might prefer to provide low detailed explanations and force users to be active learners, forcing an active style on expert system users has detrimental effects on performance when interacting with a computer expert.

3. Strive to make system knowledge salient to the user. System instructions should explicitly state that the knowledge inherent in the system comes from a human expert. It should also give explicit instructions about what types of questions the system can and cannot answer. This type of system knowledge should increase users trust in the system, their satisfaction with the system, and their performance.

4. Design for end-goal focus by reducing error consequences and increasing the benefits of a learning-by-doing user style. Expert system users who prefer an active interaction style should not be encouraged to continue this type of interaction style.

These recommendations on expert system design should, by no means, be considered complete. Most likely, several explanation features that could enhance performance or satisfaction were missed due to their subtlety or the particular nature of the task and domain used here. Also, explanation features that are beneficial in an expert system environment and have no effect in human explanation would not have been observed in these studies. In general, it is quite likely that there are several more user characteristics and explanation features that mediate performance.
Care should be taken not to over-generalize these results. As discussed previously, wide differences between types of expert systems, their domains, purposes, and users should limit the usefulness of recommendations formulated from this project (Leake, 1991; Berry & Broadbent, 1987). Therefore, these recommendations can serve as suggestions or clues to expert system designers, but should not be used to replace usability testing of each system and should not discourage further work in this area.

Results from this project should contribute to our knowledge of explanation behavior, an area of interest to psychologists, educators, philosophers, and cognitive scientists. The complexity of this area and the fact that relatively little is known about explanation still leaves many important questions unanswered (e.g., are there other ways to measure explanation success that might affect the conclusions drawn from these data?, what explanation features are helpful in an expert system explanation, but not in a human explanation? What is the best way to encourage users to interact actively with the expert system without decreasing their performance?). For this reason, it is hoped that this work will merely be among the first attempts to improve our knowledge of human and machine explanations through empirical study.
References


Appendix A

Procedural, Conceptual, and Anecdotal Information Given to Expert and User Subjects in Experiment 1

The materials are listed in this appendix are given in the following order:
1. Expert Diagnosis and Treatment Chart (procedural information)
2. Expert Anatomical Diagram (conceptual information)
3. Expert Training Booklet (conceptual information)
4. Expert Treatment Log (anecdotal information)
5. User Diagnosis and Treatment Chart (procedural information)
6. User Anatomical Diagram (conceptual information)
7. User Training Booklet (conceptual information)
8. User Treatment Log (anecdotal information)
<table>
<thead>
<tr>
<th>Body Mov</th>
<th>Eye Mov</th>
<th>Eye Pos.</th>
<th>Ink Cover</th>
<th>Ink Test</th>
<th>Signal Out</th>
<th>Signal In</th>
<th>Valve Pos</th>
<th>in/Out Cycle</th>
<th>Temp</th>
<th>Case</th>
<th>%</th>
<th>Failure</th>
<th>Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>YES</td>
<td>None</td>
<td>Both IN</td>
<td>2</td>
<td>0</td>
<td>20</td>
<td>1</td>
<td>Communication System</td>
<td>1 add 50% Ni sludge</td>
<td>2 inject B12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 3 MPH</td>
<td>TTSS</td>
<td>&lt;75 Watts</td>
<td>15</td>
<td>15</td>
<td>Communication System</td>
<td>1 add 50% Ni sludge</td>
<td>2 inject B12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 3 MPH</td>
<td>TTSS</td>
<td></td>
<td>15</td>
<td>15</td>
<td>Food Production-low light</td>
<td>1 increase light</td>
<td>2 add 75% Ni sludge</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 3 MPH</td>
<td>TTSS</td>
<td></td>
<td>15</td>
<td>10</td>
<td>Food Production-low Al</td>
<td>1 add CO</td>
<td>2 add 50% Al sludge</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 3 MPH</td>
<td>TTSS</td>
<td></td>
<td>15</td>
<td>10</td>
<td>Food Production-low Al</td>
<td>1 add CO</td>
<td>2 add 50% Al sludge</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 3 MPH</td>
<td>STST</td>
<td></td>
<td>10</td>
<td>10</td>
<td>Muscle System-low Al</td>
<td>1 add 75% Ni sludge</td>
<td>2 add 50% Ni sludge</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 MPH</td>
<td>STST</td>
<td>1-2-12-12</td>
<td>12</td>
<td>12</td>
<td>Muscle System-low Al</td>
<td>1 add 75% Ni sludge</td>
<td>2 inject B12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FAIL</td>
<td>TTSS</td>
<td></td>
<td>10</td>
<td>10</td>
<td>Ink Production-low heat</td>
<td>1 increase heat</td>
<td>2 add Li</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FAIL</td>
<td>TTSS</td>
<td></td>
<td>10</td>
<td>10</td>
<td>Ink Production-low Li</td>
<td>1 add Li</td>
<td>2 add Oil</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fail</td>
<td>STST</td>
<td></td>
<td>10</td>
<td>10</td>
<td>Ink Production-low Oil</td>
<td>1 add Oil</td>
<td>2 add 50% Ni sludge</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>TTS</td>
<td></td>
<td>10</td>
<td>10</td>
<td>Ink Production-low Oil</td>
<td>1 add Oil</td>
<td>2 add 50% Ni sludge</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>random</td>
<td>12-12-12-12</td>
<td>12</td>
<td>12</td>
<td>Discharge-receptor</td>
<td>1 inject dopamine</td>
<td>2 inject B12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>random</td>
<td>12-12-12-12</td>
<td>12</td>
<td>12</td>
<td>Discharge-receptor</td>
<td>1 inject dopamine</td>
<td>2 inject B12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>random</td>
<td>10-10-10-10</td>
<td>10</td>
<td>10</td>
<td>Intake-receptor</td>
<td>1 inject dopamine</td>
<td>2 inject B12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>random</td>
<td>10-10-10-10</td>
<td>10</td>
<td>10</td>
<td>Intake-receptor</td>
<td>1 inject dopamine</td>
<td>2 inject B12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 3 MPH</td>
<td>TTSS</td>
<td>&gt;80 Watts</td>
<td>10</td>
<td>10</td>
<td>Ink &amp; Food Production-</td>
<td>1 send random signal</td>
<td>2 inject B12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>TTSS</td>
<td></td>
<td>10</td>
<td>10</td>
<td>Excess Heat</td>
<td>1 reduce heat</td>
<td>2 add 50% Ni sludge</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>TTSS</td>
<td></td>
<td>10</td>
<td>10</td>
<td>Eye Irritation</td>
<td>1 clean tank water</td>
<td>2 inject dopamine</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>TTSS</td>
<td></td>
<td>10</td>
<td>10</td>
<td>Excess Al</td>
<td>1 extract Al</td>
<td>2 inject dopamine</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YES</td>
<td>SS55</td>
<td>Both IN</td>
<td>15</td>
<td>15</td>
<td>Danger signal in</td>
<td>1 send random signal</td>
<td>2 add CO</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

141
Expert Alien Medical Training Booklet

Alien Background

The alien (Gus) you are about to monitor was found three months ago in its space ship which apparently crashed into a coral reef off the coast of St. Thomas. After extensive testing of the environment inside the space ship, a comparable atmosphere was simulated and the alien was moved to its new Houston home—a large tank of sludge (a mud-like substance), water, and gaseous atmosphere.

Since the discovery of this St. Thomas alien, five more similar aliens have been found around the globe. Simulated atmospheres have also been made for these aliens and we have managed to keep three alive including the one here in Houston. During the past three months, a great deal has been learned about these alien's anatomy, disease, and treatments. At this point we don't know that all of our information is correct, but we are able to give estimates about the correctness of each piece of information we have about the alien.

We will begin by giving a detailed description of the alien’s seven anatomical systems. It might help you to know that when you are actually performing the monitoring task you will be able to refer to all the materials we are about to use, and to take notes if you like. However, you will not be allowed to use these materials while you take the test.

Do you have any questions at this point?
ANATOMICAL SYSTEMS

The alien body is composed of seven separate systems. Each system has its own unique parts and functions, yet most systems are dependent on each other. Each system will be described separately below. Refer to the anatomical diagram of the alien as you read through each system description. On the diagram, each system is a different color. The descriptions below will tell which color to look for. One thing must be kept in mind while studying alien medicine: this is an alien system which means that everything might not make sense to us. For instance, this alien seems to use chemical elements in unconventional ways. This is one of the important reasons we are concerned with keeping this species alive.

**Control System (colored orange)**

The control system is perhaps the most important anatomical system because it contains the control center which manipulates all of the other systems. A control center failure can potentially affect all the other systems which means it can have a number of symptoms. However, it is quite rare the a control center problem affects all the other systems. Rather, a problem with the control center usually only affects one or two of the systems. Although control system failures are difficult to treat, we have had reasonable success with injections of vitamin B12 in the past.
In addition to the control center, the tank receptors responsible for the timing of the water and sludge cycle are also a part of the control system. You will learn more about the function of these receptors in the description of the in/out system. Tank receptors failures can be treated with an injection of dopamine.

**Water-Sludge Intake/Discharge System (In/Out System (colored purple))**

This system is responsible for gathering water and sludge from the alien's underwater environment, filtering out the useful water and sludge elements, and discharging the unusable elements. Water and sludge are sucked into the alien through two intake valves located on the right side of the alien's body. The water intake valve gathers water and floating oil from the surface of the water while the sludge intake valve gathers sludge and silt from just below the water floor. Both intake valves empty into a water and sludge filter located inside the alien body where the water and sludge are held until they separate from each other. Once the sludge has settled to the bottom of the tank and the oil has floated to the top of the water, the oil is skimmed off the top of the water and sent to the ink chamber and the sludge is sent to the food production and storage area. The ink chamber and food production and storage area will be discussed later. The contents remaining in the water and sludge filter are then pushed out of the alien body through the output valve on the left side of the alien. Both intake valves are left out under normal
conditions. When a problem occurs, however (i.e., the water elements are at abnormal concentrations or the alien feels threatened), they can be retracted into the alien body.

The muscle located below the water and sludge filter is responsible for the water and sludge in/out flow. When the muscle expands it pushes against the tank, forcing water out. After the water has been pushed from the tank, the muscle relaxes. The output valve is closed, the intake valves are opened again, and more water and sludge are sucked into the tank. This in/out cycle is continuous under normal conditions. This cycle can be monitored by keeping a constant record of the alien's body circumference. When the alien has a full tank, its circumference measures 12 inches. The water is held for two minutes and then pushed out at which point the circumference measures 10 inches. It takes 30 seconds to push the water out of the tank, and another 30 seconds to suck new water and sludge into the tank. Therefore, if the system is functioning correctly, the entire cycle takes three minutes. A diagram of a few normal cycles as measured by the cycle monitor is pictured below.
Tank sensors located on the left side of the tank are responsible for controlling the amount of water let into the tank, and the timing of the cycle. A receptor failure could result in abnormal in/out cycles. Abnormal in/out cycles include slow intakes and discharges as well as failures to intake or discharge at the correct times. Cases 004 and 005 on the Alien Treatment Log describes a instances when control system failures were successfully treated.

The cycle can also become irregular when one of the input or output valves becomes clogged. Oil, tar and other water and sludge particles can build up in the valves, reducing the water or sludge flow. When this happens, the in/out cycle is slowed. After several attempts at treating this symptom (see cases 006 and 007), it was finally found that the introduction of tar-eating microorganisms to the tank will solve this problem. The organisms are sucked into the alien with the water and sludge where they feed on the tar built up in the valve. These organisms are not harmful to the alien.

As mentioned above the two valves are normally in the out position which allows them to gather water and sludge. At times, however, one or both of the valves are retracted. It seems that several conditions can cause this action. Among them, abnormal concentrations of elements in the water (e.g., too much aluminum) and an incoming danger signal causing defensive actions. As can be seen in case 012, this situation can prove to be fatal if left untreated.
Communication System (colored blue)

The antenna located between the two leaves above the water are the alien's communication system. These antenna receive and emit signals very similar to radio waves. In fact, at times the alien seems to respond to radio waves. The signals are stronger than normal radio waves, however. Only a lead wall or barrier can cut off the alien's transmissions or receptions.

Nickel seems to be the element most closely associated with the communication system. When the nickel concentration in the sludge is low the signals sent out by the alien may not be reliable messages. If the concentration of nickel is extremely low, a food production shortage is caused in addition to the unreliable outgoing signals. As was learned in case 001, adding the right type of nickel sludge to the tank can cure transmission problems.

The control center is responsible for sending and interpreting receptions and transmissions. Control center failures may have a drastic effect on the communication system (e.g., transmissions or receptions are completely inoperable).

When all systems are operating normally, a steady signal of long waves are sent out. Deviations from this normal signal seem to indicate some sort of problem or failure. Below are several signals and our hypothesis about their meaning:
Steady tall wavelengths: "All is well" message, perhaps this also communicates the alien's location to other nearby aliens.

Two tall, two short signal (t-t-s-s): this signal seems to correlate with problems that result from having a deficiency of some above water element (e.g., light). Perhaps it is asking other nearby aliens about their above water conditions so it can determine which direction to move.

One short, three tall signal (s-t-t-t): this signal seems to correlate with problems that result from having a deficiency of some water element (e.g., oil).
One short, one tall signal (s-t-s-t): this signal seems to correlate with problems that result from having a deficiency of some sludge element (e.g., aluminum).

Random signal: a random signal seems to correlate with control center problems. See case 004.

We can also monitor the signals the alien is receiving. When the signals are random, the alien seems to disregard them. However, when a patterned signal is detected or no signal is registered, the alien might react in a variety of ways. When the alien does not receive any incoming signals it appears that it gets frightened and protects itself by squirting ink and pulling in its intake valves. The best way to treat this is to send the alien a random signal.

From observing some of the aliens' reactions to various incoming signals (see cases 008, 009, 010, and 013), we have been able to make some hypothesis about successful treatments. In most cases it seems
that interrupting the abnormal signals coming in with a random "all's well" signal will relieve any symptoms.

Food Production System (colored green)

The food production system is composed of the left leaf and the food production and storage area. The left leaf gathers carbon monoxide gas from the air. Light hitting the leaf helps assimilate this gas into a liquid substance which is transported into the food storage area. The light intensity needs to be between 75 and 100 watts to assist this process. Sludge from the water and sludge filter is also brought to the food production and storage area. The two elements aluminum and nickel found in the sludge along with the carbon monoxide from the air are the most important elements for food production.

Shortages in any of the elements necessary for food production (carbon monoxide, light, aluminum, or nickel) will cause a food production failure. The food produced by this system is distributed to every other system. For this reason, a food production failure can be quite serious and result in a number of problems. Some systems are more dependent than others on different nutrients. Therefore, depending on the cause of the food production failure (e.g., low aluminum, low carbon monoxide), some systems will show symptoms before others. Generally, the first indication of a food production failure, is a slowing down of all non-essential operations to conserve energy. The alien will
begin to move slower and, if not needed to search for the needed
element, the eyes will be housed or brought in.

Ink Production System (colored pink)

The ink production system is the alien's defense system. The
alien produces a cloudy, black poisonous ink which, when squirted into
the water, provides a protective barrier intolerable to most of the alien's
predators. The ink is produced from the oil gathered from the water and
from hydrogen and lithium gases gathered from the air by the right leaf.
Lithium gas and hydrogen (converted from a gas into a liquid) are both
sent to the ink chamber. The temperature needs to be between 60 and
80 degrees Fahrenheit to convert the hydrogen and lithium into liquid
form. In the ink chamber these elements are mixed with the oil to form
the liquid ink.

When provoked, the alien will expel the ink from four pores (two
are located on the left side of the body and two are on the right side) into
the surrounding water. Enough ink is expelled to cover the alien
thoroughly and make him virtually impossible to see through the clouds.

Because ink is constantly being produced (only fresh ink is
poisonous), shortage of any of the elements needed to produce the ink
will result in some sort of ink production failure. This type of failure is not
overtly noticeable. It is only when the alien is provoked and does not
expel ink that an ink system failure becomes noticeable. Hydrogen is the
only element that if missing will not interrupt ink production. However,
hydrogen is partially responsible for the ink’s dark color. When hydrogen is low or missing, the ink has a light cloudy appearance as opposed to a dark color (see case 003). Case 014 also shows an example of a successful ink production failure treatment.

**Vision System (colored red)**

The vision system contains two eyes connected to the body by two long stems. The stems have the ability to move in all directions allowing the alien to look in any direction. Under normal conditions, the eyes are out and moving around. If some sort of failure occurs, the eyes may stop moving or be housed in the alien body. The eyes seem to be closely tied to the control system. When a control system problem occurs, the eyes normally stop moving. Eyes can also stop moving when there is a muscle system failure as well.

Sometimes the eyes can become irritated by some foreign element in the water. If this happens, the eyes are housed until the eyes have recovered. If this should occur, the alien’s water should be cleaned (see case 011).

**Muscle System (colored yellow)**

The muscle system is made up of the feet which are nearly all muscle, the water and sludge filter muscle located below the filter, and the muscles responsible for housing the intake valves and eyes. The muscle system seems to be highly dependent on aluminum. Even the
smallest shortage of this substance in the alien's diet can cause paralysis-like symptoms. Under normal conditions, the alien is moving between 3 and 6 mph, the eyes are out and moving, and the intake valves are out. If a shortage of aluminum occurs, the alien might slow down or stop moving, or the eyes might stop moving and be left out. If the concentration of aluminum is extremely low, all of these symptoms might occur (see case 002).

Other failures such as a food production failure, might also affect the muscle system. As mentioned previously, when food is scarce, the alien will conserve energy be slowing down body movement stop moving the eyes, and housing the eyes if the element needed cannot be seen.
### Expert Alien Treatment Log

<table>
<thead>
<tr>
<th>Case:</th>
<th>001</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date:</td>
<td>1/15/92</td>
</tr>
<tr>
<td>Lab:</td>
<td>Australia</td>
</tr>
<tr>
<td>Symptoms:</td>
<td>Signal in monitor shows no transmissions</td>
</tr>
<tr>
<td>Diagnosis:</td>
<td>Communication System Failure - cause unknown</td>
</tr>
</tbody>
</table>
| Treatment: | 25% nickel sludge added to tank  
50% nickel sludge added to tank  
| Result: | No change, Symptom disappeared |

<table>
<thead>
<tr>
<th>Case:</th>
<th>002</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date:</td>
<td>1/16/92</td>
</tr>
<tr>
<td>Lab:</td>
<td>Houston</td>
</tr>
</tbody>
</table>
| Symptoms: | Alien stops moving  
No eye movement  
Circumference 12-12-12-12  
T-S-T-S signal sent |
| Diagnosis: | Muscle System Failure - cause unknown |
| Treatment: | 50% nickel sludge added to tank  
75% nickel sludge added to tank  
25% aluminum sludge added to tank  
50% aluminum sludge added to tank  
75% aluminum sludge added to tank |
| Result: | No change  
No change  
No change  
No change  
Symptoms disappeared |

<table>
<thead>
<tr>
<th>Case:</th>
<th>003</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date:</td>
<td>1/20/92</td>
</tr>
<tr>
<td>Lab:</td>
<td>Mexico City</td>
</tr>
</tbody>
</table>
| Symptoms: | Ink test produces light ink  
T-T-S-S signal sent |
| Diagnosis: | Ink Production Failure - cause unknown |
| Treatment: | Hydrogen added to tank |
| Result: | Symptoms disappeared |
### Case: 004
**Date:** 2/10/92  
**Lab:** Paris  
**Symptoms:** Circumference 12 -12-12-12  
- No eye movement  
- Random signal output  
**Diagnosis:** Water Discharge Failure - cause unknown  
**Treatment:**  
- vitamin E injected  
- carbon monoxide increased  
- vitamin B12 injected  
**Result:**  
- No change  
- No change  
- Symptoms disappeared

### Case: 005
**Date:** 1/16/92  
**Lab:** London  
**Symptoms:** Circumference 10-10-10-10  
- No eye movement  
**Diagnosis:** Water Intake Failure - cause unknown  
**Treatment:**  
- vitamin B12 injected  
- dopamine injected  
**Result:**  
- No change  
- Symptoms disappeared

### Case: 006
**Date:** 2/11/92  
**Lab:** Australia  
**Symptoms:** Circumference 10-11-11-12  
**Diagnosis:** Water Intake Failure - cause unknown  
**Treatment:**  
- add oil to tank  
- inject with vitamin B12  
- inject with dopamine  
- add 75% nickel sludge to tank  
- add 50% aluminum sludge to tank  
**Result:**  
- Symptom gets worse  
- No change  
- No change  
- Symptom gets worse  
- ALIEN DIES
Case: 007
Date: 2/17/92
Lab: Houston
Symptoms: Circumference 10-11-11-12
Diagnosis: Water Intake Failure - artery clog
Treatment: hydrogen added to tank
tar-eating microorganisms added to tank
Result: No change
Symptom disappears

Case: 010
Date: 3/2/92
Lab: Amsterdam
Symptoms: T-T-S-S signal in
Body movement slow
Ink test fails
Eyes are housed
Diagnosis: Above water condition warning in, Ink & Food Production Failure
Treatment: send random "all's well" signal
Result: symptoms disappear

Case: 011
Date: 3/9/92
Lab: Houston
Symptoms: Eyes are housed
Diagnosis: Unknown
Treatment: inject dopamine
inject vitamin B12
clean tank water
Result: no change
no change
symptom disappears
Case: 012
Date: 3/16/92
Lab: Mexico
Symptoms: Sludge valve retracted
Diagnosis: Unknown
Treatment: add 50% nickel sludge
           add 50% nickel sludge
           add 50% aluminum sludge
           add 75% aluminum sludge
Result: no change
        no change
        body movement slowed
        ALIEN DIES

Case: 013
Date: 3/19/92
Lab: London
Symptoms: S-S-S-S signal in
           Ink covering
           Both intake valves in
Diagnosis: Danger signal coming in
Treatment: send random "all’s well" signal to alien
Result: Symptoms disappear

Case: 014
Date: 3/20/92
Lab: Amsterdam
Symptoms: L-L-S-S signal sent
           Air temperature 58 degrees F
           Ink test fails
Diagnosis: Ink Production Failure - cause unknown
Treatment: carbon monoxide added to tank
           hydrogen added to tank
           lithium added to tank
Result: No change
        No change
        Symptoms disappear
<table>
<thead>
<tr>
<th>Body Mov.</th>
<th>Eye Mov.</th>
<th>Eye Pos.</th>
<th>Ink Cover</th>
<th>Ink Test</th>
<th>Signal In</th>
<th>Signal Out</th>
<th>Valve Pos.</th>
<th>In/Out Cycle</th>
<th>Temp</th>
<th>Light</th>
<th>Case</th>
<th>%</th>
<th>Failure</th>
<th>Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>YES</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>Both</td>
<td>80</td>
<td>20</td>
<td>1 Communication System 2 Control Center</td>
<td>1 add 50% Ni sludge 2 inject B12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 3 MPH</td>
<td>TTSS</td>
<td>&lt;75 Watts</td>
<td>70</td>
<td>30</td>
<td>1 Food Production-low light 2 Food Production-very low Ni</td>
<td>1 increase light 2 add 75% Ni sludge</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 3 MPH</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>TTSS</td>
<td>60</td>
<td>40</td>
<td>1 Food Production-very low Ni 2 Muscle System-low Al</td>
<td>1 add 75% Ni sludge 2 add 75% Al sludge</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FAIL</td>
<td>TTSS</td>
<td>&lt; 60°</td>
<td>014</td>
<td>55</td>
<td>45</td>
<td>1 Ink Production-low heat 2 Ink Production-low Li</td>
<td>1 increase heat 2 add Li</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>random</td>
<td>12-12-12</td>
<td>12-12</td>
<td>70</td>
<td>30</td>
<td>1 In/Out system-control center 2 In/Out system-receptor -low Ni</td>
<td>1 inject B12 2 inject dopamine add 50% Ni sludge</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>&lt;10°-0-10°</td>
<td>TSS 55</td>
<td>65</td>
<td>35</td>
<td>2 In/Out system-receptor 1 In/Out system-control center</td>
<td>1 inject dopamine 2 inject B12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>IN</td>
<td>Sludge IN</td>
<td>60</td>
<td>40</td>
<td>20</td>
<td>1 Eye irritation 2 Control Center</td>
<td>1 clean tank water 2 inject dopamine</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSSS</td>
<td>Both</td>
<td>IN</td>
<td>012</td>
<td>80</td>
<td>20</td>
<td>1 Excess Al 2 Control Center</td>
<td>1 extract Al 2 inject dopamine</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YES</td>
<td>SSSS</td>
<td>Both</td>
<td>70</td>
<td>50</td>
<td>2 low CO 1 Danger signal</td>
<td>1 send random signal 2 add CO</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
User Alien Medical Training Booklet

Alien Background
The alien (Gus) you are about to monitor was found three months ago in its space ship which apparently crashed into a coral reef off the coast of St. Thomas. After extensive testing of the environment inside the space ship, a comparable atmosphere was simulated and the alien was moved to its new Houston home—a large tank of sludge (a mud-like substance), water, and gaseous atmosphere.

Since the discovery of this St. Thomas alien, five more similar aliens have been found around the globe. Simulated atmospheres have also been made for these aliens and we have managed to keep three alive including the one here in Houston. During the past three months, a great deal has been learned about these alien’s anatomy, disease, and treatments. At this point we don’t know that all of our information is correct, but we are able to give estimates about the correctness of each piece of information we have about the alien.

We will begin by giving a detailed description of the alien’s seven anatomical systems. It might help you to know that when you are actually performing the monitoring task you will be able to refer to all the materials we are about to use, and to take notes if you like. However, you will not be allowed to use these materials while you take the test.

Do you have any questions at this point?
ANATOMICAL SYSTEMS

The alien body is composed of seven separate systems. Each system has its own unique parts and functions, yet most systems are dependent on each other. Each system will be described separately below. Refer to the anatomical diagram of the alien as you read through each system description. On the diagram, each system is a different color. The descriptions below will tell which color to look for. One thing must be kept in mind while studying alien medicine: this is an alien system which means that everything might not make sense to us. For instance, this alien seems to use chemical elements in unconventional ways. This is one of the important reasons we are concerned with keeping this species alive.

Control System (colored orange)

The control system contains the control center. This organ controls the other six systems. For this reason, the control system is probably the most important system. The tank receptors responsible for the timing of the water and sludge filter cycle are also a part of the control system. You will learn more about the function of these receptors in the description of the in/out system.

Water-Sludge Intake/Discharge System (In/Out System (colored purple))
This system is responsible for gathering water and sludge from the alien's underwater environment, filtering out the useful water and sludge elements, and discharging the unusable elements. Water and sludge are sucked into the alien through two intake valves located on the right side of the alien's body. The water intake valve gathers water and floating oil from the surface of the water while the sludge intake valve gathers sludge and silt from just below the water floor. Both intake valves empty into a water and sludge filter located inside the alien body. After the useful elements are filtered from the water and sludge, the remaining waste is pushed out of the alien body through the output valve on the left side of the alien. Both intake valves are left out under normal conditions. When a problem occurs, however, they can be retracted into the alien body.

The muscle located below the water and sludge filter is responsible for the water and sludge in/out flow. When the muscle expands it pushes against the tank, forcing water out. After the water has been pushed from the tank, the muscle relaxes. The output valve is closed, the intake valves are opened again, and more water and sludge are sucked into the tank. This in/out cycle is continuous under normal conditions. This cycle can be monitored by keeping a constant record of the alien's body circumference. When the alien has a full tank, its circumference measures 12 inches. The water is held for two minutes and then pushed out at which point the circumference measures 10 inches. It takes 30 seconds to push the water out of the tank, and another
30 seconds to suck new water and sludge into the tank. Therefore, if the system is functioning correctly, the entire cycle takes three minutes. A diagram of a few normal cycles as measured by the cycle monitor is pictured below.

Tank sensors located on the left side of the tank are responsible for controlling the amount of water let into the tank, and the timing of the cycle. When this system fails, abnormal cycles result. Abnormal cycles might be slower than 3 minutes or stop altogether. Cases 004 and 005 on the Alien Treatment Log describe instances when a receptor and control center failure were successfully treated.

As mentioned above the two valves are normally in the out position which allows them to gather water and sludge. At times, however, one or both of the valves are retracted. We think the valves are retracted when the alien feels threatened. There may also be other circumstances when the valves are retracted. As can be seen in case 012, this situation can prove to be fatal if left untreated.

**Communication System (colored blue)**
The antenna located between the two leaves above the water are the alien's communication system. These antenna receive and emit signals very similar to radio waves. The communication waves appear to be stronger than radio waves, however. Only a lead barrier can block these waves.

When all systems are operating normally, a steady signal of long waves are sent out. Deviations from this normal signal seem to indicate some sort of problem or failure. Below are a few signals and our hypothesis about their meaning:

![Wave Height Graph]

Steady tall wavelengths: "All is well" message, perhaps this also communicates the alien’s location to other nearby aliens.

![Wave Height Graph]

Two tall, two short signal (t-t-s-s): this signal seems to correlate with problems that result from having a deficiency of some above water element (e.g., light). Perhaps it is asking other nearby aliens about their above water conditions so it can determine which direction to move.

We can also monitor the signals the alien is receiving. When the signals are random, the alien seems to disregard them. However, when
a patterned signal is detected or no signal is registered, the alien might react in a variety of ways. When the alien does not receive any incoming signals it appears that it gets frightened and protects itself by squirting ink and pulling in its intake valves. The best way to treat this is to send the alien a random signal.

**Food Production System (colored green)**

The food production system is composed of the left leaf and the food production and storage area. The left leaf gathers carbon monoxide gas from the air. Light hitting the leaf helps assimilate this gas into a liquid substance which is transported into the food storage area. The light intensity needs to be between 75 and 100 watts to assist this process. Sludge from the water and sludge filter is also brought to the food production and storage area. The two elements aluminum and nickel found in the sludge along with the carbon monoxide from the air are the most important elements for food production.

**Ink Production System (colored pink)**

The ink production system is the alien's defense system. The alien produces a cloudy, black poisonous ink which, when squirted into the water, provides a protective barrier intolerable to most of the alien's predators. The ink is produced from the oil gathered from the water and from hydrogen and lithium gases gathered from the air by the right leaf. Lithium gas and hydrogen (converted from a gas into a liquid) are both
sent to the ink chamber. The temperature needs to be between 60 and
80 degrees Fahrenheit to convert the hydrogen into liquid form. In the ink
chamber these elements are mixed with the oil to form the liquid ink.

When provoked, the alien will expel the ink from four pours (two
are located on the left side of the body and two are on the right side) into
the surrounding water. Enough ink is expelled to cover the alien
thoroughly and make him virtually impossible to see through the clouds.

Problems with the ink production system are not always overtly
noticeable. You can only be certain that there is a ink system failure when
the alien feels threatened and doesn't expel ink. Case 014 shows an
example of a successful ink production failure treatment.

Vision System (colored red)

The vision system contains two eyes connected to the body by two
long stems. The stems have the ability to move in all directions allowing
the alien to look in any direction. Under normal conditions, the eyes are
out and moving around. If some sort of failure occurs, the eyes may stop
moving or be housed in the alien body.

Sometimes the eyes can become irritated by some foreign
element in the water. If this happens, the eyes are housed until the eyes
have recovered. If this should occur, the alien's water should be cleaned
(see case 011).

Muscle System (colored yellow)
The muscle system is made up of the feet which are nearly all muscle, the water and sludge filter muscle located below the water and sludge filter, and the muscles responsible for housing the intake valves and eyes. Under normal conditions, the alien is moving between 3 and 6 mph, the eyes are out and moving, and the intake valves are out.
# User Alien Treatment Log

<table>
<thead>
<tr>
<th>Case:</th>
<th>001</th>
<th>Case:</th>
<th>004</th>
<th>Case:</th>
<th>005</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date:</td>
<td>1/15/92</td>
<td>Date:</td>
<td>2/10/92</td>
<td>Date:</td>
<td>1/16/92</td>
</tr>
<tr>
<td>Lab:</td>
<td>Australia</td>
<td>Lab:</td>
<td>Paris</td>
<td>Lab:</td>
<td>London</td>
</tr>
<tr>
<td>Symptoms:</td>
<td>Signal in monitor shows no transmissions</td>
<td>Symptoms:</td>
<td>Circumference 12 -12-12-12</td>
<td>Symptoms:</td>
<td>Circumference 10-10-10-10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>No eye movement</td>
<td></td>
<td>No eye movement</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Random signal output</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diagnosis:</td>
<td>Communication System Failure- cause unknown</td>
<td>Diagnosis:</td>
<td>Water Discharge Failure - cause unknown</td>
<td>Diagnosis:</td>
<td>Water Intake Failure - cause unknown</td>
</tr>
<tr>
<td>Treatment:</td>
<td>25% nickel sludge added to tank 50% nickel sludge added to tank</td>
<td>Treatment:</td>
<td>vitamin E injected carbon monoxide increased vitamin B12 injected</td>
<td>Treatment:</td>
<td>vitamin B12 injected dopamine injected</td>
</tr>
<tr>
<td>Result:</td>
<td>No change Symptom disappeared</td>
<td>Result:</td>
<td>No change No change Symptoms disappeared</td>
<td>Result:</td>
<td>No change Symptoms disappeared</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


Case: 011
Date: 3/9/92
Lab: Houston
Symptoms: Eyes are housed
Diagnosis: Unknown
Treatment: inject dopamine
          inject vitamin B12
          clean tank water
Result: no change
        no change
        symptom disappears

Case: 012
Date: 3/16/92
Lab: Mexico
Symptoms: Sludge valve retracted
Diagnosis: Unknown
Treatment: add 50% nickel sludge
          add 50% nickel sludge
          add 50% aluminum sludge
          add 75% aluminum sludge
Result: no change
        no change
        body movement slowed
        ALIEN DIES

Case: 014
Date: 3/20/92
Lab: Amsterdam
Symptoms: L-L-S-S signal sent
          Air temperature 58 degrees F
          Ink test fails
Diagnosis: Ink Production Failure - cause unknown
Treatment: carbon monoxide added to tank
          hydrogen added to tank
          lithium added to tank
Result: No change
        No change
        Symptoms disappear
Appendix B

Alien Anatomy Test

Below are several anatomical parts, phrases describing a system, and environmental elements used by the alien. Write the letter of the system that the phrase corresponds most closely with:

A. Water & Sludge Intake/Discharge (In/Out) System
B. Ink Production System
C. Food Production System
D. Communication System
E. Vision System
F. Muscle System
G. Control System

___ colored orange on the anatomical diagram
___ colored pink on the anatomical diagram
___ colored red on the anatomical diagram
___ colored green on the anatomical diagram
___ colored blue on the anatomical diagram
___ colored purple on the anatomical diagram
___ colored yellow on the anatomical diagram
___ the ink chamber is part of this system
___ this is probably the most important system
___ this system is responsible for pushing the sludge and water out of the water and sludge filter
___ the control center is part of this system
___ monitoring the circumference of the alien body helps monitor this system’s activity
___ the eyes are part of this system
___ food production and storage area
___ the sludge intake valve is part of this system
___ this system is responsible for alien locomotion
___ the output valve is part of this system
___ light is required for this system to function normally
this system is responsible for receptions and transmissions
hydrogen is required for this system to function normally
lithium is required for this system to function normally
this system is responsible for gathering elements from the underwater environment
oil is required for this system to function normally
heat is required for this system to function normally
the antenna is part of this system
the foot is part of this system
when this system fails, most other systems are shut down to conserve energy
the tank receptors are part of this system
aluminum is required for this system to function normally
the water and sludge filter is part of this system
carbon monoxide is required for this system to function normally
the water intake valve is part of this system
this system is severely affected when there is a shortage of aluminum
the tank muscle is part of this system
this system is responsible for regulating the timing of the in/out cycle and the amount of sludge and water gathered
waves from this system are similar to radio waves
this system is responsible for the functioning of all other systems
this system is responsible for filtering out the usable elements from the sludge and water gathered from the underwater environment
failures of this system are caused by even moderately low levels of nickel
failures of this system are caused by very low levels of nickel
these two systems are connected to every other system
the eyes often stop moving when these two systems have problems
Each letter should be used only once:

A. Nickel
B. Aluminum
C. Carbon Monoxide
D. Hydrogen
E. Lithium
F. Oil

___ When this element is missing the alien experiences paralysis-like symptoms
___ If the above water warning signal is sent out, the alien is moving below 3 MPH and the eyes are retracted, this element is probably missing
___ When this element is missing, clear ink is produced
___ This element can cause the in/out cycle to be slowed if it starts building up in the valves
___ Even small shortages of this element can make the signals sent out or received unreliable
___ When this above water element is missing the ink test result will be "FAIL"
These letters may be used more than once:

A. TTTT
B. TTSS
C. TTTS
D. TSTS
E. SSSS
F. Random

___ This signal sent out means there is probably a control center problem
___ This signal sent out means there is something wrong with the water
    conditions
___ This signal sent out means there is something wrong with the above
    water conditions
___ This signal sent out means there is something wrong with the sludge
    conditions
___ This signal received probably means danger
___ This signal received will cause the alien to slow down, bring the eyes
    in, and produce a FAIL ink test result
___ This signal received will cause the water valve to be retracted, the
    in/out cycle to slow, and produce a FAIL ink test result
___ This signal received will cause the sludge valve to be retracted, In
    in/out cycle to slow, and the eyes and body to stop moving

Short Answer
If the alien feels threatened what defensive actions does he take besides
expelling ink?

When a food production shortage occurs, under what conditions will the
alien retract its eyes in order to save energy?

Under what conditions will the alien retract its intake valves?
Appendix C

Final Questionnaire Given to Subjects in Experiment 1

User Questionnaire

1. Please select the letter of the phrase that best describes your relationship with the other subject:

   a. I have never talked with this person before
   b. I have talked with this person on a few occasions (under 5), but the conversations were short and impersonal (i.e., "do you have the time?")
   c. I have talked with this person on several occasions (over 5), but the conversations were short and impersonal (i.e., "do you have the time?")
   d. I have talked with this person on only a few occasions (under 5), but at least one of the conversations was long and involved (over 5 minutes).
   e. I have talked with this person on several occasions (over 5), many of the conversations were long and involved (over 5 minutes).
   f. I talk with this person nearly every day
   g. I talk with this person nearly every day, we are roommates

2. Has the other subject ever explained a problem (i.e., homework) to you before? If yes, please describe the circumstances including the topic area and the approximate number of hours.
Expert Questionnaire

1. Have you ever taught in a classroom situation? If yes, please describe the circumstances including topic area, number in the class, and the approximate number of hours.

2. Have you ever tutored one or more individuals? If yes, please describe the circumstances including topic area, number of people, and the approximate number of hours.

3. Familiarity with the other subject:
   a. I have never talked with this person before
   b. I have talked with this person on a few occasions (under 5), but the conversations were short and impersonal (i.e., "do you have the time?")
   c. I have talked with this person on several occasions (over 5), but the conversations were short and impersonal (i.e., "do you have the time?")
   d. I have talked with this person on only a few occasions (under 5), but at least one of the conversations was long and involved (over 5 minutes).
   e. I have talked with this person on several occasions (over 5), many of the conversations were long and involved (over 5 minutes).
   f. I talk with this person nearly every day
   g. I talk with this person nearly every day, we are roommates

For each subject, please select the letter of the phrase that best describes your relationship with that person:

   Subject #1 _____
   Subject #2 _____
   Subject #3 _____
4. Have you ever explained a problem (i.e., homework) to this person? If yes, please describe the circumstances including the topic area and the approximate number of hours.

Subject #1

Subject #2

Subject #3
Appendix D

Explanations Developed for the Six Conditions Tested in Experiment 2

For each of the six conditions tested in Experiment 2, unique explanations were created for each of the 22 problems. Presented below are the explanations for two problems used in each of the six conditions. The explanation content and form differences between the conditions seen in these two problems is representative of the other 20 problems not shown here. The remaining of the 20 explanations used in the high content, advice/explanation presentation order condition follows.

Two Representative Explanations from the High Content, Advice/Explanation Presentation Order Condition

1. Gus is currently experiencing an ink production failure because he’s low on hydrogen. To treat this problem, you should override that short signal with a random signal, his normal incoming signal. There is a communication problem with Gus, but it’s not an internal problem. He’s receiving a short short short short signal from somewhere, maybe there’s a radio turned on or something. Gus had interpreted this signal as a danger signal so he has pulled his valves in and he has gone ahead and spurted ink out to form a camouflage. When Gus receives a random signal he thinks everything is o.k., and he relaxes.
2. I would recommend you add 50% nickel sludge to the tank to treat the current communication system failure. When there is no signal coming in, it means that Gus is not perceiving any signals and there's probably something wrong with his communication system. The other symptoms, the ink covering and the pulled in valves are residual effects of this problem. When Gus couldn't perceive his normal, all's well signal, he got scared and took his defensive actions. Even a small shortage of nickel can cause the communication system to fail. So adding a relatively small amount of nickel sludge should allow the communication system to operate normally again.

Two Representative Explanations from the High Content, Explanation/Advice Presentation Order Condition

1. There is a communication problem with Gus, but it's not an internal problem. He's receiving a short short short signal from somewhere, maybe there's a radio turned on or something. Gus had interpreted this signal as a danger signal so he has pulled his valves in and he has gone ahead and spurted ink out to form a camouflage. When Gus receives a random signal he thinks everything is OK, and he relaxes. So what you can do is go ahead and override that short signal with a random signal, his normal incoming signal.

2. When there is no signal coming in, it means that Gus is not perceiving any signals and there's probably something wrong with his communication system. The other symptoms, the ink covering and the pulled in valves are
residual effects of this problem. When Gus couldn't perceive his normal, all's well signal, he got scared and took his defensive actions. Even a small shortage of nickel can cause the communication system to fail. So adding a relatively small amount of nickel sludge should allow the communication system to operate normally again. So to treat the current communication system failure, you should add 50% Nickel sludge to the tank.

Two Representative Explanations from the Low Content, Advice/Explanation Presentation Order Condition

1. Gus currently thinks he's receiving a danger signal and to treat this, you should send Gus a random signal. There is a communication problem with Gus, but it's not an internal problem. Gus thinks he's receiving a danger signal and got frightened. What you can do is go ahead and override that short signal with a random signal, his normal incoming signal, so that Gus will know that everything is all right.

2. I would recommend you add 50% nickel sludge to the tank to treat the current communication system failure. In this case there is clearly something wrong with Gus's communication system. Because Gus couldn't perceive his normal, all's well signal, he got scared. The communication system is highly dependent on nickel and fails when there is a shortage of this element.

Two Representative Explanations from the Low Content, Explanation/Advice Presentation Order Condition
1. There is a communication problem with Gus, but it's not an internal problem. Gus thinks he's receiving a danger signal and becomes frightened. When Gus receives a random signal he thinks everything is OK, and relaxes. So what you can do is go ahead and override that short signal with a random signal, his normal incoming signal.

2. In this case there is clearly something wrong with Gus's communication system. Because Gus couldn't perceive his normal, all's well signal, he got scared. The communication system is highly dependent on nickel and fails when there is a shortage of this element. So to treat the current communication system failure, I would recommend you add 50% Nickel sludge to the tank.

Two Representative Explanations from the No Content, Advice/Explanation Presentation Order Condition

1. To treat this problem, you should send Gus a random signal. There is a communication problem with Gus, but it's not an internal problem. Gus thinks he's receiving a danger signal.

2. To treat this problem, add 50% nickel sludge to the tank. In this case there is clearly something wrong with Gus's communication system probably because there is a shortage of nickel.
Two Representative Explanations from the No Content, Explanation/Advice

Presentation Order Condition

1. There is a communication problem with Gus, but it's not an internal problem. Gus thinks he's receiving a danger signal. To treat this problem, you should send Gus a random signal.

2. In this case there is clearly something wrong with Gus's communication system probably because there is a shortage of nickel. To treat this problem, add 50% nickel sludge to the tank.

The remaining 20 Explanations from the High Content, Advice/Explanation

Presentation Order Condition

1. Here, the problem is probably a communication system failure and you should add 50% nickel sludge to the tank. Since the only symptom is that there are no signals being sent out, the problem probably lies with the communication system. Nickel is the only thing that affects the communication system and it tends to mess up that system when there isn't enough of it. If there is a severe shortage of nickel, then it messes up the food production system as well. But since there are no other symptoms indicating that has happened, it probably has only a moderate shortage of nickel.
2. This is most likely a in/out system failure caused by a problem with the tank receptors. To treat this problem, you should inject Gus with dopamine. The in/out cycle has stopped expelling waste, staying at a circumference of 12 inches. This could either be caused by a tank receptor or control center problem which are each part of the control system. Anytime the control system has failed, the eyes will stop moving since they are dependent on the control system. In this case, it's most likely the tank receptors have failed which are responsible for keeping the in/out cycle on a regular track. Currently, this isn't happening. If it were a problem with the control center, more symptoms would be visible since the control center regulates all the other systems.

3. Gus is currently experiencing an ink production failure because he's low on hydrogen. To treat this problem, you should add hydrogen to the tank. You can tell there is something wrong with the ink production system since Gus didn't pass his ink test. He's still making ink, but it's clear which isn't normal. The warning signal being sent out, the tall tall short short pattern, means there's a problem with an above water element. The above water elements that affect the ink system are hydrogen, lithium and heat. The temperature is within range, so it can't be a problem with the heat, leaving hydrogen or lithium. When there is a shortage of heat or lithium, the ink test will produce a "fail" result and Gus won't be able to make any type of ink. The only way he can produce clear ink, is if there's a lack of
hydrogen. In this case, he can still make ink, but it's going to come out lighter or clear because the hydrogen is what gives the ink its dark color.

4. Gus's eyes are probably just irritated so you should clean the water in the tank. Since the only symptoms are that the eyes have been pulled in and there is no eye movement, the problem is most likely that the water is dirty causing this problem.

5. Gus is experiencing a food production failure due to low light so you should increase the light intensity in the tank. When the body movement slows down it usually indicates that there's something wrong with the food production system and he's trying to conserve energy. The signal Gus is sending out the, pattern, means there's something wrong above the water. The elements required for food production found about the water are light and carbon monoxide. The light intensity monitor shows the light intensity is below the normal range which is in turn causing a food production failure.

6. Gus is receiving an above warning signal, so you should send him a random signal. The signal coming in, the is interpreted as an above water warning signal. Since no other alien is around to send it maybe there's a radio around or something else that sends off radio waves. But, what's important is that Gus is interpreting the signal as a warning. Because he thinks there's something wrong with the above water environment but doesn't know what exactly, he's stopped taking in all above water elements.
He takes above water elements in through his two leaves, the left one that's part of the food production system and the right one that's part of the ink production system. The failed ink test shows the ink production system is down and the other symptoms indicate the food production system is down. Because his food production is shut off he is trying to conserve energy by only partially using his vision and muscle systems. If Gus were sent his normal random signal, he would start taking in his above water elements again.

7. To treat this problem, a 75% nickel sludge mixture should be added to the tank to cure the food production and communication system failures currently occurring. The absence of an outgoing signal indicates that there may be a problem with the communication system. The fact that the body movement is slower than it should be is almost always an indication that there's something wrong with the food production system. The link between the communication and food production systems is that they're both dependent on nickel. Small shortages of nickel will cause a communication system failure, and more severe shortages will cause a food production failure in addition to a communications system failure. Normally, Gus is able to send a signal out describing where in his environment the problem lies. However, his failed communication system wouldn't allow it in this case.
8. Gus has clogged intake or discharge valves and you should treat this problem by adding microorganisms to the tank. Something has gone wrong with the in/out system. If the problem was caused by some failure of the control system, like of the tank receptor or control center, you would probably see more symptoms. If you look at this in/out cycle, it's just slowed down, not stopped completely like it would if it were a control system failure. So, what might be the case is that there's something wrong with either the discharge or the intake valves. On occasion, these valves can get clogged because the oil and sludge sucked in builds up in the valves and eventually clogs them, slowing down the in/out cycle. Tar-eating microorganisms can be sucked into the valves where they will eat the oil and sludge without harming the alien.

9. Gus has suffered a control center failure and you should inject vitamin B12. Your main clue about the problem is the outgoing communication signal, it's a random pattern which generally indicates a problem with the control center. The control center regulates the communications and is also closely associated with the eyes, explaining why they have stopped moving. The in/out system has also stopped as a result of the control center failure. So, all the symptoms are due to the fact that the control center has failed, affecting all these other systems. The control center requires high doses of vitamin B12 to operate normally, so the best way to treat a control center problem is to inject this vitamin.
10. Gus is experiencing an ink production system failure due to low oil so you should add oil to the tank. The ink production system relies on four elements: it takes 3 from the air, hydrogen, lithium and heat, and it takes oil from the water. If there's a shortage of any one of these elements, the ink production system will not pass its ink test. The outgoing communication pattern, the short tall tall tall, means that there's a deficiency with some water element. So that pretty much narrows the choices down to the oil which is the only element found in the water.

11. Here, there's an ink production failure due to low lithium so you should add lithium to the tank. The failed ink test tells you that there's a problem with the ink production system and signal being sent out means it has something to do with the above water environment. The ink production system relies on heat, hydrogen and lithium from the above water environment. The temperature is fine so it's not a problem with heat. If the problem was a lack of hydrogen, the ink test would not produce a fail result, but instead clear ink would be produce because the hydrogen lends the dark color to the ink. But he's currently failed the ink test, which suggests that the problem is caused by a shortage of lithium.

12. There is a tank receptor failure so you should inject dopamine. The in/out cycle indicates that Gus is not taking in water and sludge because there's no increase in circumference, it's staying at 10. Since the in/out system is regulated by the control system, the problem probably lies there.
This is also why the eyes have stopped moving. The control system is closely associated with the vision system, so when a problem occurs with the control system, the eyes are one of the first things to be affected. The control system consists of the tank receptors and the control center. If there were a control center problem then there would be more symptoms because the control center regulates all the other systems.

13. A food production failure has occurred because Gus is low on CO, so you should add carbon monoxide to the tank. Whenever, the body movement is slow, there is usually a problem with food production. The food production system basically runs on four things: carbon monoxide and light from the above water environment, and aluminum and nickel from the sludge. If you look at the signal being sent out, you have the signal which means the problem is an above water deficiency which limits it to either light or carbon monoxide. Since the light intensity monitor isn't highlighted, that means it's OK. That would leave carbon monoxide. Another way you can tell that the problem is with carbon monoxide is that whenever there's a problem with the food production system, Gus tries to save energy by shutting off all non-essential systems. For example, he will pull in his eyes when the element he is missing is not visible. The only above water element necessary for food production that is not visible is carbon monoxide, so that must be the problem.
14. Gus is experiencing a food production and muscle system failures due to low aluminum so you should add 75% aluminum sludge to the tank. The fact that Gus has slowed down his body movement, in this case it's not even moving at all, means something is wrong with the food production system. The outgoing pattern, the short tall short tall pattern means that there's some sort of problem in the sludge. This narrows the problem down to either aluminum or nickel, the two elements in the sludge which are both necessary for food production. Deficiencies of aluminum have a paralysis effect on the alien because the muscle system is highly dependent on aluminum to function normally. This explains why the other systems depending on the muscle system have failed. aluminum is required by the food production system as well.

15. There is a communication failure happening because he is low on nickel, so you should add 50% nickel sludge to the tank. He's not perceiving any communication signals in and that frightens him. There may be a signal coming in, but he can't detect, so he gets scared because even though there might be a normal random signal there, he doesn't know it. When Gus gets scared he takes his defensive actions. The communication system is very dependent on nickel, in fact, nickel is essential to it. So when there's even a mild shortage of nickel, the communication system doesn't function properly and he may not be able to detect incoming signals.
16. To treat the clogged valves Gus most likely has, you should add microorganisms to the tank. Since the in/out cycle is going really slowly in, something must be wrong with its intake or discharge valves. There are no other indications that anything else is going wrong, so there really aren't any deficiencies of elements or failures of the control system. It's probably a situation similar to our heart valves when they become clogged with cholesterol and things like that. I think the same thing has happened here.

17. The temperature in the tank should be increased because there is an ink production failure. The signal being sent out tells you that the ink production failure is caused by some abnormality in the above water environment. The ink production system requires that the temperature be between 60 and 80 degrees F. This is so the hydrogen and lithium can be converted from gases into liquid form. This mixture is then combined with oil in the ink chamber to form the liquid ink. So when the air environment isn't warm enough for him to convert the hydrogen and lithium, that sort of precludes him being able to combine them with the oil and that's why he isn't producing any ink. You can tell by the fact that the temperature monitor is highlighted, that the temp. is out of the normal range.

18. Gus has a control center failure and you should inject him with vitamin B12. In this case the alien is failing to intake any sludge or water so something is wrong with the in/out system. Because there are also symptoms showing in the vision system and the communication system, it
can be assumed that the failure is in some central area, probably the control center. The control center regulates the 3 systems showing symptoms here. So the problem is most likely in the control center.

19. Here the valves have become clogged so to treat this problem, add micro to the tank. Because the only symptom is that the in/out cycle has slowed, probably the only thing wrong here is that oil and sludge have built up inside the intake or discharge valves.

20. Gus is receiving a sludge warning signal and my advice is to send a normal random signal back to Gus. The communication pattern he’s receiving, the tall short tall short, tells him that something or at least he thinks something is wrong in the sludge. So Gus has pulled in the sludge valve which has slowed down the in/out cycle and eliminated the intake of sludge elements. The paralysis-like symptoms are caused by the lack of aluminum usually gathered from the sludge. So Gus thinks something is wrong with the sludge when there really isn’t. If the incoming warning signal would stop, Gus would correct these problems himself and start taking in sludge again.
Appendix E

Post-Session Questionnaire Given to Subjects in Experiment 2

Please answer the following questions by circling the number that most closely matches your answer:

1. How would you rate the expert's treatment advice (e.g., add 50% nickel sludge)?
   Correct 1 2 3 4 5 6 7 Incorrect
   Reliable 1 2 3 4 5 6 7 Unreliable
   Useful 1 2 3 4 5 6 7 Useless
   Confusing 1 2 3 4 5 6 7 Clear

2. How would you rate the expert's treatment explanation (any words offered to tell you why you should add 50% nickel sludge)?
   Relevant 1 2 3 4 5 6 7 Irrelevant
   Useful 1 2 3 4 5 6 7 Useless
   Confusing 1 2 3 4 5 6 7 Clear

3. How would you rate the amount of explanation provided by the expert?
   Just right
   Too much 1 2 3 4 5 6 7 Too little
4. Please rate the expert you interacted with:
   Knowledgeable 1 2 3 4 5 6 7  Clueless
   Useful 1 2 3 4 5 6 7  Useless

5. Did the expert make your decision about how to treat Gus easier or harder to make?
   Easier 1 2 3 4 5 6 7  Harder

6. How difficult was it to decide whether or not to take the expert's advice?
   Very difficult 1 2 3 4 5 6 7  Very easy

7. How often did you take the expert's advice?
   Never 1 2 3 4 5 6 7  Always
   If your answer was less than 7, why didn't you always take the expert's advice?

8. If Gus were a real alien that you were responsible for (use your imagination here), would you want the expert available for consultation?
   Definitely 1 2 3 4 5 6 7  Definitely not
9. Did your interactions with the expert help you learn the rules behind alien medicine or did you learn what you learned on your own? I learned from the expert 1 2 3 4 5 6 7 my own

10. What is your major? ______________________

11. What year are you at Rice? __________________

12. Do you feel your knowledge of "real" chemistry or medicine got in the way of learning this "pretend" chemistry and medical knowledge? If yes, please explain:

13. Is there anything else about your participation in this study you think would be helpful for the experimenter to know?

Thank you for your time!!!!
Appendix F

Pre-Transformation Cell Means and Standard Deviations (in parentheses) from Experiment 2

Scores below have been converted into percentages.

<table>
<thead>
<tr>
<th>Dependent Measure</th>
<th>Presentation Order</th>
<th>Content</th>
<th>Row Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>None</td>
<td>Low</td>
</tr>
<tr>
<td>Guess Performance</td>
<td>Advice First</td>
<td>54.54</td>
<td>56.82</td>
</tr>
<tr>
<td></td>
<td>Explanation First</td>
<td>51.95</td>
<td>58.44</td>
</tr>
<tr>
<td></td>
<td>Column Mean</td>
<td>53.25</td>
<td>57.58</td>
</tr>
</tbody>
</table>

(Table Continues)
<table>
<thead>
<tr>
<th>Dependent Measure</th>
<th>Presentation Order</th>
<th>Content</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>None</td>
<td>Low</td>
<td>High</td>
<td>Row Mean</td>
</tr>
<tr>
<td>Treatment</td>
<td>Advice First</td>
<td>76.64</td>
<td>78.57</td>
<td>86.51</td>
<td>80.20</td>
</tr>
<tr>
<td>Performance</td>
<td></td>
<td>(10.52)</td>
<td>(8.36)</td>
<td>(7.95)</td>
<td>(9.53)</td>
</tr>
<tr>
<td></td>
<td>Explanation First</td>
<td>68.71</td>
<td>82.54</td>
<td>82.34</td>
<td>78.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(11.76)</td>
<td>(9.52)</td>
<td>(5.33)</td>
<td>(10.85)</td>
</tr>
<tr>
<td></td>
<td>Column Mean</td>
<td>72.68</td>
<td>80.42</td>
<td>84.13</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(11.48)</td>
<td>(8.30)</td>
<td>(6.64)</td>
<td></td>
</tr>
<tr>
<td>Test</td>
<td>Advice First</td>
<td>51.95</td>
<td>51.70</td>
<td>56.06</td>
<td>53.03</td>
</tr>
<tr>
<td>Performance</td>
<td></td>
<td>(7.81)</td>
<td>(7.26)</td>
<td>(8.46)</td>
<td>(7.65)</td>
</tr>
<tr>
<td></td>
<td>Explanation First</td>
<td>46.75</td>
<td>51.30</td>
<td>56.25</td>
<td>51.62</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(10.73)</td>
<td>(8.59)</td>
<td>(8.39)</td>
<td>(9.67)</td>
</tr>
<tr>
<td></td>
<td>Column Mean</td>
<td>49.35</td>
<td>51.51</td>
<td>56.17</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(9.41)</td>
<td>(7.62)</td>
<td>(8.09)</td>
<td></td>
</tr>
</tbody>
</table>
Appendix G

Cell Means and Standard Deviations for Experiment 2 Post-Questionnaire Items with Significant Effects

How would you rate the expert's treatment explanation (any words offered to tell you why you should add 50% nickel sludge)?

1 = Relevant  7 = Irrelevant

Significant Content Level Effect

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Content</td>
<td>2.24</td>
<td>(0.89)</td>
</tr>
<tr>
<td>Low Content</td>
<td>1.67</td>
<td>(0.98)</td>
</tr>
<tr>
<td>High Content</td>
<td>1.43</td>
<td>(0.65)</td>
</tr>
</tbody>
</table>
How would you rate the expert's treatment explanation (any words offered to tell you why you should add 50% nickel sludge)?

1 = Useful    7 = Useless

Significant Content Level Effect

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Content</td>
<td>2.43</td>
<td>(1.22)</td>
</tr>
<tr>
<td>Low Content</td>
<td>1.67</td>
<td>(1.05)</td>
</tr>
<tr>
<td>High Content</td>
<td>1.50</td>
<td>(0.65)</td>
</tr>
</tbody>
</table>

How difficult was it to decide whether or not to take the expert system's advice?

1 = Very Difficult    7 = Very easy

Significant Presentation Order Effect

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advice-Explanation Order</td>
<td>5.19</td>
<td>(1.54)</td>
</tr>
<tr>
<td>Explanation-Advice Order</td>
<td>6.09</td>
<td>(1.06)</td>
</tr>
</tbody>
</table>
Appendix H

Revised Procedural, Conceptual, and Anecdotal Information Given to Subjects in Experiment 3

The materials are listed in this appendix are given in the following order:

1. Diagnosis and Treatment Chart (procedural information)
2. Training Booklet (conceptual information)
3. Treatment Log (conceptual information)
4. Conceptual Knowledge test
5. Post Questionnaire

The anatomical diagram given to the users in Experiments 2 and 3 was left unaltered and given to subjects in this study.
<table>
<thead>
<tr>
<th>Body Mov.</th>
<th>Eye Mov.</th>
<th>Eye Pos.</th>
<th>Ink Test</th>
<th>Signal Out</th>
<th>In/Out Cycle</th>
<th>Temp</th>
<th>Light</th>
<th>Case #</th>
<th>%</th>
<th>Failure</th>
<th>Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>None</td>
<td>None</td>
<td>TTSS</td>
<td>&lt;75 Watts</td>
<td>001</td>
<td>80</td>
<td>20</td>
<td>1</td>
<td>Communication System</td>
<td>1 add 50% Ni sludge</td>
<td></td>
</tr>
<tr>
<td>&lt; 3 MPH</td>
<td>TTSS</td>
<td>&lt;75 Watts</td>
<td>70</td>
<td>30</td>
<td>2 Food Production-low light</td>
<td>1 increase light</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 3 MPH</td>
<td>None</td>
<td>None</td>
<td>60</td>
<td>40</td>
<td>2 Food Production-very low Ni</td>
<td>2 add 75% Ni sludge</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FAIL</td>
<td>TTSS</td>
<td>&lt; 60°</td>
<td>014</td>
<td>55</td>
<td>2 Ink Production-low heat</td>
<td>2 add Li</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>45</td>
<td>2 Ink Production-low Li</td>
<td>1 increase heat</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>random</td>
<td>12-12-12</td>
<td>004</td>
<td>70</td>
<td>1 In/Out system-control center</td>
<td>1 inject B12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>30</td>
<td>2 In/Out system-receptor</td>
<td>2 inject dopamine add 50% Ni sludge</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>10-10-10</td>
<td></td>
<td>005</td>
<td>65</td>
<td>1 In/Out system-receptor</td>
<td>1 inject dopamine</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>35</td>
<td>low Ni</td>
<td>2 inject B12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TTSS</td>
<td></td>
<td>&gt; 80°</td>
<td>90</td>
<td>10</td>
<td>1 Excess Heat</td>
<td>1 reduce heat</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>IN</td>
<td></td>
<td>011</td>
<td>60</td>
<td>1 Eye Irritation</td>
<td>2 add 50% Ni sludge</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>40</td>
<td>2 Control Center</td>
<td>1 clean tank water</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Al = Aluminum  
Li = Lithium  
Ni = Nickel
Alien Medical Training Booklet

Alien Background

The alien (Gus) you are about to monitor was found three months ago in its space ship which apparently crashed into a coral reef off the coast of St. Thomas. After extensive testing of the environment inside the space ship, a comparable atmosphere was simulated and the alien was moved to its new Houston home—a large tank of sludge (a mud-like substance), water, and gaseous atmosphere.

Since the discovery of this St. Thomas alien, five more similar aliens have been found around the globe. Simulated atmospheres have also been made for these aliens and we have managed to keep three alive including the one here in Houston. During the past three months, a great deal has been learned about these aliens' anatomy, disease, and treatments. At this point we don't know that all of our information is correct, but we are able to give estimates about the correctness of each piece of information we have about the alien.

We will begin by giving a detailed description of the alien's seven anatomical systems. It might help you to know that when you are actually performing the monitoring task you will be able to refer to most of the materials we are about to use, and to take notes if you like. However, you will not be allowed to use these materials while you take the test.
ANATOMICAL SYSTEMS

The alien body is composed of seven separate systems. Each system has its own unique parts and functions, yet most systems are dependent on each other. Each system will be described separately below. Refer to the anatomical diagram of the alien as you read through each system description. On the diagram, each system is a different color. The descriptions below will tell which color to look for. One thing must be keep in mind while studying alien medicine: this is an alien system which means that everything might not make sense to us. For instance, this alien seems to use chemical elements in unconventional ways. This is one of the important reasons we are concerned with keeping this species alive.

Control System (colored orange)

The control system contains the control center. This organ controls the other six systems. For this reason, the control system is probably the most important system. The tank receptors responsible for the timing of the water and sludge filter cycle are also a part of the control system. You will learn more about the function of these receptors in the description of the in/out system.
Water-Sludge Intake/Discharge System (In/Out System (colored purple))

This system is responsible for gathering water and sludge from the alien's underwater environment, filtering out the useful water and sludge elements, and discharging the unusable elements. Water and sludge are sucked into the alien through two intake valves located on the right side of the alien's body. The water intake valve gathers water and floating oil from the surface of the water while the sludge intake valve gathers sludge and silt from just below the water floor. Both intake valves empty into a water and sludge filter located inside the alien body. After the useful elements are filtered from the water and sludge, the remaining waste is pushed out of the alien body through the output valve on the left side of the alien.

The muscle located below the water and sludge filter is responsible for the water and sludge in/out flow. When the muscle expands it pushes against the tank, forcing water out. After the water has been pushed from the tank, the muscle relaxes. The output valve is closed, the intake valves are opened again, and more water and sludge are sucked into the tank. This in/out cycle is continuous under normal conditions. This cycle can be monitored by keeping a constant record of the alien's body circumference. When the alien has a full tank, its circumference measures 12 inches. The water is held for two minutes and then pushed out at which point the
circumference measures 10 inches. It takes 30 seconds to push the water out of the tank, and another 30 seconds to suck new water and sludge into the tank. Therefore, if the system is functioning correctly, the entire cycle takes three minutes. A diagram of a few normal cycles as measured by the cycle monitor is pictured below.

![Circumference Graph]

Tank receptors located on the left side of the tank are responsible for controlling the amount of water let into the tank, and the timing of the cycle. When this system fails, abnormal cycles result. Abnormal cycles might be slower than 3 minutes or stop altogether. Cases 004 and 005 on the Alien Treatment Log describe instances when a receptor and control center failure were successfully treated.

**Communication System (colored blue)**

The antenna located between the two leaves above the water are the alien's communication system. These antenna receive and emit signals very similar to radio waves. In fact, Gus sometimes seems to respond to radio waves. The waves Gus emits appear to be
stronger than radio waves, however. Only a lead barrier can block these waves.

When all systems are operating normally, a steady signal of long waves are sent out. Deviations from this normal signal seem to indicate some sort of problem or failure. Below are a few signals and our hypothesis about their meaning:

Steady tall wavelengths: "All is well" message, perhaps this also communicates the alien's location to other nearby aliens.

Two tall, two short signal (t-t-s-s): this signal seems to correlate with problems that result from having a deficiency of some above water element (i.e., light). Perhaps it is asking other nearby aliens about their above water conditions so it can determine which direction to move to get the element it is currently missing.

Food Production System (colored green)
The food production system is composed of the left leaf and the food production and storage area. The left leaf gathers carbon monoxide gas from the air. Light hitting the leaf helps assimilate this gas into a liquid substance which is transported into the food storage area. The light intensity needs to be between 75 and 100 watts to assist this process. Sludge from the water and sludge filter is also brought to the food production and storage area. The two elements aluminum and nickel found in the sludge along with the carbon monoxide from the air are the most important elements for food production.

**Ink Production System (colored pink)**

The ink production system is the alien's defense system. The alien produces a cloudy, black poisonous ink which, when squirted into the water, provides a protective barrier assumed to be intolerable to most of the alien's predators. The ink is produced from the oil gathered from the water and from hydrogen and lithium gases gathered from the air by the right leaf. Lithium gas and hydrogen (converted from a gas into a liquid) are both sent to the ink chamber. The temperature needs to be between 60 and 80 degrees Fahrenheit to convert the gases into liquid form. In the ink chamber these elements are mixed with the oil to form the liquid ink.

When provoked, the alien will expel the ink from four pours (two are located on the left side of the body and two are on the right
side) into the surrounding water. Enough ink is expelled to cover the alien thoroughly and make him virtually impossible to see through the clouds.

Problems with the ink production system are not always overtly noticeable. You can only be certain that there is a ink system failure when the alien feels threatened and doesn't expel ink. Case 014 shows an example of a successful ink production failure treatment.

**Vision System (colored red)**

The vision system contains two eyes connected to the body by two long stems. The stems have the ability to move in all directions allowing the alien to look in any direction. Under normal conditions, the eyes are out and moving around. If some sort of failure occurs, the eyes may stop moving or be pulled into the alien body.

Sometimes the eyes can become irritated by some foreign element in the water. If this happens, the eyes are housed until the eyes have recovered. If this should occur, the alien's water should be cleaned (see case 011).

**Muscle System (colored yellow)**

The muscle system is made up of the three feet which are nearly all muscle, the water and sludge filter muscle located below the water and sludge filter, and the muscles responsible for moving
the intake valves and eyes. Under normal conditions, the alien is moving between 3 and 6 mph, the eyes are out and moving, and the intake valves are out.
# Alien Treatment Log

<table>
<thead>
<tr>
<th>Case:</th>
<th>001</th>
<th>Date:</th>
<th>1/15/92</th>
<th>Lab:</th>
<th>Australia</th>
</tr>
</thead>
</table>
| **Symptoms:** | Signal in monitor shows no transmissions | **Diagnosis:** | Communication System Failure - cause unknown | **Treatment:** | 25% nickel sludge added to tank  
|             |                   |             |           | 50% nickel sludge added to tank | **Result:** | No change |
|             |                   |             |           |               | Symptom disappeared |

<table>
<thead>
<tr>
<th>Case:</th>
<th>004</th>
<th>Date:</th>
<th>2/10/92</th>
<th>Lab:</th>
<th>Paris</th>
</tr>
</thead>
</table>
| **Symptoms:** | Circumference 12-12-12  
No eye movement  
Random signal output | **Diagnosis:** | Water Discharge Failure - cause unknown | **Treatment:** | vitamin E injected  
carbon monoxide increased  
vitamin B12 injected | **Result:** | No change  
No change  
Symptoms disappeared |

<table>
<thead>
<tr>
<th>Case:</th>
<th>005</th>
<th>Date:</th>
<th>1/16/92</th>
<th>Lab:</th>
<th>London</th>
</tr>
</thead>
</table>
| **Symptoms:** | Circumference 10-10-10-10  
No eye movement | **Diagnosis:** | Water Intake Failure - cause unknown | **Treatment:** | vitamin B12 injected  
dopamine injected | **Result:** | No change  
Symptoms disappeared |
Case: 011
Date: 3/9/92
Lab: Houston
Symptoms: Eyes are housed
Diagnosis: Unknown
Treatment: inject dopamine
inject vitamin B12
clean tank water
Result: no change
no change
symptom disappears

Case: 014
Date: 3/20/92
Lab: Mexico
Symptoms: T-T-S-S signal sent
Air temperature 58 degrees F
Ink test fails
Diagnosis: Ink Production Failure - cause unknown
Treatment: carbon monoxide added to tank
hydrogen added to tank
lithium added to tank
Result: No change
No change
Symptoms disappear
Alien Anatomy Test

Below are several phrases describing a system and environmental elements used by the alien. Write the letter of the system that the phrase corresponds most closely with:

A. Water & Sludge Intake/Discharge (In/Out) System
B. Ink Production System
C. Food Production System
D. Communication System
E. Vision System
F. Muscle System
G. Control System

___ this system is responsible for pushing the sludge and water out of the water and sludge filter
___ the control center is part of this system
___ monitoring the circumference of the alien body helps monitor this system's activity
___ this system is responsible for alien locomotion
___ light is required for this system to function normally
___ this system is responsible for receptions and transmissions
___ hydrogen is required for this system to function normally
___ lithium is required for this system to function normally
___ this system is responsible for gathering elements from the underwater environment
___ oil is required for this system to function normally
___ heat is required for this system to function normally
___ when this system fails, some other systems are shut down to conserve energy
aluminum is required for this system to function normally

carbon monoxide is required for this system to function normally

this system is severely affected when there is a shortage of aluminum

this system is responsible for regulating the timing of the in/out cycle and the amount of sludge and water gathered

this system is responsible for the functioning of all other systems

this system is responsible for filtering out the usable elements from the sludge and water gathered from the underwater environment

failures of this system are caused by even moderately low levels of nickel

failures of this system are caused by very low levels of nickel

these two systems are connected to every other system

... the eyes often stop moving when these two systems have problems
These letters may be used more than once:

A. TTTT
B. TTSS
C. TTTS
D. TSTS
F. Random

__ This signal sent out means there is probably a control center problem
__ This signal sent out means there is something wrong with the water conditions
__ This signal sent out means there is something wrong with the above water conditions
__ This signal sent out means there is something wrong with the sludge conditions
__ This signal would be sent out if there's a shortage of carbon dioxide
__ This signal would be sent out if there's a shortage of aluminum
__ This signal would be sent out if there's a shortage of oil
__ This signal would be sent out if there's a shortage of lithium
__ This signal would be sent out if there's a shortage of nickel
__ This signal would be sent out if there's a shortage of hydrogen
__ This signal would be sent out if the temperature was out of the normal range
__ This signal would be sent out if the light intensity was out of the normal range
Each letter should be used only once:

A. Nickel  D. Hydrogen
B. Aluminum  E. Lithium
C. Carbon Monoxide  F. Oil

___ When this element is missing the alien experiences paralysis-like symptoms
___ Even small shortages of this element can make the signals sent out or received unreliable
___ If the above water warning signal is sent out, the alien is moving below 3 MPH and the eyes are retracted, this element is probably missing
___ When this element is missing, clear ink is produced
___ This element can cause the in/out cycle to be slowed if it starts building up in the valves
___ When this above water element is missing the ink test result will be "FAIL"
Questionnaire Given to Subjects in the Active Participation, Computer Expert Condition

Please answer the following questions by circling the number that most closely matches your answer:

1. How would you rate the expert system's treatment advice (i.e., add 50% nickel sludge)?
   Correct 1 2 3 4 5 6 7 Incorrect
   Reliable 1 2 3 4 5 6 7 Unreliable
   Useful 1 2 3 4 5 6 7 Useless
   Confusing 1 2 3 4 5 6 7 Clear

2. How would you rate the expert system's treatment explanation (any words offered to tell you why you should add 50% nickel sludge)?
   Relevant 1 2 3 4 5 6 7 Irrelevant
   Useful 1 2 3 4 5 6 7 Useless
   Confusing 1 2 3 4 5 6 7 Clear

3. How would you rate the amount of explanation provided by the expert system?
   Just right
   Too much 1 2 3 4 5 6 7 Too little
4. Please rate the expert system you interacted with:
   Knowledgeable 1 2 3 4 5 6 7 Clueless
   Useful 1 2 3 4 5 6 7 Useless

5. Did the expert system make your decision about how to treat Gus easier or harder to make?
   Easier 1 2 3 4 5 6 7 Harder

6. How difficult was it to decide whether or not to take the expert system's advice?
   Very difficult 1 2 3 4 5 6 7 Very easy

7. How often did you take the expert system's advice?
   Never 1 2 3 4 5 6 7 Always
   If your answer was less than 7, why didn't you always take the expert system's advice?

8. If Gus were a real alien that you were responsible for (use your imagination here), would you want the expert system available for consultation?
   Definitely 1 2 3 4 5 6 7 Definitely not
9. Did your interactions with the expert system help you learn the rules behind alien medicine or did you learn what you learned on your own?

I learned from the expert system 1 2 3 4 5 6 7 my own

10. Did you have enough control over your interaction with the expert system?

I wish the expert system would have only given me advice when I asked for it 1 2 3 4 5 6 7 I didn't mind getting the expert system's advice on each problem

I wish I didn't have to ask a question on each problem 1 2 3 4 5 6 7 I didn't mind having to ask a question on each problem

comments on above two:

11. What is your major? 

12. What year are you at NMSU? 

13. Do you feel your knowledge of "real" chemistry or medicine got in the way of learning this "pretend" chemistry and medical knowledge? If yes, please explain:

14. Is there anything else about your participation in this study you think would be helpful for the experimenter to know?
Appendix I

Raw Pre-Transformation Cell Means and Standard Deviations (in parentheses) from Experiment 3

Scores listed below have been converted into percentages.

<table>
<thead>
<tr>
<th>Dependent Measure</th>
<th>Expert Type</th>
<th>User Participation</th>
<th>Content Low</th>
<th>Content High</th>
<th>Row Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guess Performance</td>
<td>Human</td>
<td>Active</td>
<td>61.96</td>
<td>50.54</td>
<td>56.25</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(9.51)</td>
<td>(15.75)</td>
<td>(13.88)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Passive</td>
<td>53.26</td>
<td>52.90</td>
<td>53.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(12.68)</td>
<td>(18.12)</td>
<td>(14.59)</td>
</tr>
<tr>
<td></td>
<td>Computer</td>
<td>Active</td>
<td>53.80</td>
<td>52.17</td>
<td>52.99</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(8.68)</td>
<td>(15.24)</td>
<td>(12.01)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Passive</td>
<td>51.09</td>
<td>57.97</td>
<td>54.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(10.33)</td>
<td>(13.09)</td>
<td>(11.66)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Column Mean</td>
<td>55.03</td>
<td>53.12</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(10.74)</td>
<td>(14.98)</td>
<td></td>
</tr>
</tbody>
</table>

(Table Continues)
<table>
<thead>
<tr>
<th>Dependent Measure</th>
<th>Expert Type</th>
<th>User Participation</th>
<th>Content</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Low</td>
<td>High</td>
<td>Row Mean</td>
</tr>
<tr>
<td>Treatment</td>
<td>Human</td>
<td>Active</td>
<td>95.65 (2.32)</td>
<td>95.11 (7.51)</td>
<td>95.38 (5.38)</td>
</tr>
<tr>
<td>Performance</td>
<td></td>
<td>Passive</td>
<td>96.20 (4.31)</td>
<td>89.86 (11.88)</td>
<td>93.48 (8.65)</td>
</tr>
<tr>
<td></td>
<td>Computer</td>
<td>Active</td>
<td>94.02 (5.66)</td>
<td>95.11 (2.79)</td>
<td>94.56 (4.34)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Passive</td>
<td>95.11 (5.90)</td>
<td>94.20 (5.27)</td>
<td>94.72 (5.44)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Column Mean</td>
<td>95.24 (7.03)</td>
<td>94.64 (4.80)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dependent Measure</th>
<th>Expert Type</th>
<th>User Participation</th>
<th>Content</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Low</td>
<td>High</td>
<td>Row Mean</td>
</tr>
<tr>
<td>Test Performance</td>
<td>Human</td>
<td>Active</td>
<td>77.91 (8.88)</td>
<td>71.51 (14.48)</td>
<td>74.71 (12.07)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Passive</td>
<td>71.51 (16.57)</td>
<td>70.54 (11.92)</td>
<td>71.10 (14.24)</td>
</tr>
<tr>
<td></td>
<td>Computer</td>
<td>Active</td>
<td>64.12 (14.76)</td>
<td>63.95 (20.54)</td>
<td>64.03 (17.44)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Passive</td>
<td>59.59 (18.68)</td>
<td>72.87 (20.20)</td>
<td>65.28 (19.79)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Column Mean</td>
<td>68.42 (16.10)</td>
<td>69.44 (16.69)</td>
<td></td>
</tr>
</tbody>
</table>
Appendix J

Cell Means and Standard Deviations for Experiment 3 Post-Questionnaire Items with Significant Effects

How would you rate the expert's treatment advice (e.g., add 50% nickel sludge)?

1 = Reliable  7 = Unreliable

Significant Content Level by User Participation Effect

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Content</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Active</td>
<td>2.31</td>
<td>(1.18)</td>
</tr>
<tr>
<td>Passive</td>
<td>1.63</td>
<td>(0.81)</td>
</tr>
<tr>
<td>High Content</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Active</td>
<td>1.88</td>
<td>(0.89)</td>
</tr>
<tr>
<td>Passive</td>
<td>2.50</td>
<td>(1.57)</td>
</tr>
</tbody>
</table>

Did you have enough control over your interaction with the expert?

1 = I wish I didn't have to ask  7 = I didn't mind having to
a question on each problem ask a question on each problem

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Content</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.69</td>
<td>(1.25)</td>
</tr>
<tr>
<td>High Content</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3.63</td>
<td>(2.36)</td>
</tr>
</tbody>
</table>
How would you rate the expert's treatment advice (e.g., add 50% nickel sludge)?

1 = Useful  7 = Useless

**Significant Content Level by User Participation Effect**

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Content</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Active</td>
<td>1.64</td>
<td>(1.01)</td>
</tr>
<tr>
<td>Passive</td>
<td>1.40</td>
<td>(0.63)</td>
</tr>
<tr>
<td>High Content</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Active</td>
<td>1.44</td>
<td>(0.73)</td>
</tr>
<tr>
<td>Passive</td>
<td>2.08</td>
<td>(1.00)</td>
</tr>
</tbody>
</table>

Please rate the expert you interacted with:

1 = Useful  7 = Useless

**Significant Content Level Effect**

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Content</td>
<td>1.62</td>
<td>(0.68)</td>
</tr>
<tr>
<td>High Content</td>
<td>2.14</td>
<td>(1.27)</td>
</tr>
</tbody>
</table>
Did the expert make your decision about how to treat Gus easier or harder to make?

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>Active</td>
<td>High</td>
<td>1.38</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low</td>
<td>1.75</td>
</tr>
<tr>
<td></td>
<td>Passive</td>
<td>High</td>
<td>2.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low</td>
<td>1.63</td>
</tr>
<tr>
<td>Computer</td>
<td>Active</td>
<td>High</td>
<td>2.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low</td>
<td>1.50</td>
</tr>
<tr>
<td></td>
<td>Passive</td>
<td>High</td>
<td>2.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low</td>
<td>2.50</td>
</tr>
</tbody>
</table>
Did your interactions with the expert help you learn the rules behind alien medicine or did you learn what you learned on your own?

1 = I learned from the expert  
7 = I learned on my own

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Human</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Active</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>2.50</td>
<td>(1.31)</td>
</tr>
<tr>
<td>Low</td>
<td>2.75</td>
<td>(1.67)</td>
</tr>
<tr>
<td>Passive</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>3.00</td>
<td>(2.19)</td>
</tr>
<tr>
<td>Low</td>
<td>1.75</td>
<td>(1.16)</td>
</tr>
<tr>
<td><strong>Computer</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Active</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>2.75</td>
<td>(1.16)</td>
</tr>
<tr>
<td>Low</td>
<td>2.25</td>
<td>(0.89)</td>
</tr>
<tr>
<td>Passive</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>1.67</td>
<td>(0.82)</td>
</tr>
<tr>
<td>Low</td>
<td>3.38</td>
<td>(2.07)</td>
</tr>
</tbody>
</table>