

# Modeling the Effect of Trend Information on Human Failure Detection and Diagnosis in Spacecraft Systems

by

Rachel L. Owen

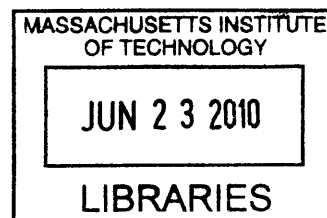
B.S. Space Systems Engineering  
United States Air Force Academy, Colorado Springs CO, 2008

Submitted to the Department of Aeronautics and Astronautics  
in partial fulfillment of the requirements for the degree of

Master of Science in Aeronautics and Astronautics  
at the  
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2010

**ARCHIVES**



© 2010 Massachusetts Institute of Technology. All rights reserved.

Author.....

Department of Aeronautics and Astronautics

May 21, 2010

Certified by .....

R. John Hansman

Professor of Aeronautics and Astronautics

Thesis Supervisor

Certified by .....

Lauren J. Kessler

Principal Technical Staff, Charles Stark Draper Laboratory

Thesis Supervisor

Accepted by .....

Eytan H. Modiano

Associate Professor of Aeronautics and Astronautics

Chair, Committee on Graduate Students

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government.

# **Modeling the Effect of Trend Information on Human Failure Detection and Diagnosis in Spacecraft Systems**

by

Rachel L. Owen

Submitted to the Department of Aeronautics and Astronautics on May 21, 2010 in partial fulfillment of the requirements for the degree of Master of Science in Aeronautics and Astronautics.

## **Abstract**

Systems are performing increasingly complicated tasks, made possible by significant advances in hardware and software technology. This task complexity is reflected in the system design, with a corresponding increased demand on comprehensive design efforts. Fundamental to the safety and mission success of these systems is the tradeoffs between human tasking and system tasking, and the resultant human interface. The research presented in this thesis was motivated by the development of an early-stage system design tool. This tool includes models of human decision making in order to evaluate system design tradeoffs with regard to human performance. An experiment was conducted to evaluate the effect of trend information displays on human decision making performance. Decision latency and accuracy were examined as performance metrics. To elicit information regarding the subjects' decision making process, the Lens model was used to gather metrics on achievement and decision consistency. The experimental results showed that both detection latency and diagnosis accuracy improved when trend information about dynamic system parameters is explicitly available to operators of spacecraft systems. The presence of this additional information also improved decision consistency. However, it made no significant difference for subjects' detection accuracy, diagnosis latency or achievement. Other predictors of latency and accuracy included the type of failure and the spacecraft trajectory. This was expected as both of these factors are important contributors to an operator's mental model of normal system behavior, which is critical to detecting and identifying failures. From these results, it can be concluded that operators of spacecraft systems could benefit from the inclusion of trend information, since it improves failure detection and diagnosis performance which can improve overall mission safety and success.

## Acknowledgments

This thesis was prepared at the Charles Stark Draper Laboratory under an internal research and development program.

Publication of this thesis does not constitute approval by Draper Laboratory of the findings or conclusions contained herein. It is published for the stimulation and exchange of ideas.

Many special thanks to Lauren Kessler for being an excellent mentor and friend in my time at Draper and for offering me a wide range of interesting opportunities over the last two years. The balance of guidance and independence and her unwavering confidence in my ability during the course of this project has been incredibly rewarding for me. I am so grateful to have had the opportunity to work under her supervision.

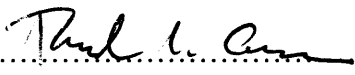
Thanks to Professor John Hansman of MIT for his consistent advice and immensely valuable outside perspective on this project. His critical mix of challenging technical and theoretical questions and suggestions for alternatives were instrumental in shaping this project and thesis.

Thanks to Kevin Duda for welcoming me onto his project team at Draper and sharing his considerable expertise in human factors engineering. He has provided me with invaluable support through all stages of the project and thesis preparation, going above and beyond to help me prepare a quality product.

Special thanks to many other members of the technical staff at Draper Laboratory, Brian Collins, Nicholas Borer, and John Irvine for their invaluable help with the complex mathematics, modeling, and statistical analysis required for this project. Their open doors and sacrifices of time and resources were invaluable to my learning.

To Joel Cohen, friend and outstanding computer programmer, who made possible the coding of my experimental interface in an incredibly short amount of time. Without his generous, faithful support, this project never would have been completed on schedule.

Finally, to my parents, brother, and friends for their constant love, encouragement, and unwavering faith in me.

  
.....  
Rachel L. Owen



# Table of Contents

<b>Abstract .....</b>	<b>3</b>
<b>Acknowledgments .....</b>	<b>4</b>
<b>List of Figures .....</b>	<b>10</b>
<b>List of Tables.....</b>	<b>13</b>
<b>1 Introduction .....</b>	<b>14</b>
1.1 Background and Motivation.....	14
1.2 Research Objectives.....	16
1.3 Thesis Organization .....	18
<b>2 Background .....</b>	<b>19</b>
2.1 Human Information Processing.....	19
2.1.1 Memory.....	20
2.1.2 Situation Awareness .....	22
2.1.3 Decision Making Heuristics.....	28
2.2 Human Decision Making .....	32
2.2.1 Decision Making Performance .....	35
<b>3 Lens Model.....</b>	<b>38</b>

3.1	History and Formulation.....	38
3.1.1	Limitations of Correlation and Skill Score.....	43
3.2	Previous Applications.....	46
3.3	Limitations.....	47
3.4	Implementation and Data Post Processing.....	48
<b>4</b>	<b>Focus Area .....</b>	<b>52</b>
4.1	Failure Types.....	52
4.2	Failure Detection .....	55
4.3	Failure Diagnosis.....	58
<b>5</b>	<b>Methodology.....</b>	<b>61</b>
5.1	Motivation .....	61
5.2	Research Questions and Hypotheses .....	62
5.2.1	Research Questions .....	62
5.2.2	Hypotheses.....	62
5.3	Experiment Overview.....	63
5.3.1	Background .....	63
5.3.2	Task .....	71

5.3.1	Subjects .....	78
5.4	Procedure.....	78
5.4.1	Apparatus and Graphical User Interface .....	79
5.4.2	Data Collection .....	80
<b>6</b>	<b>Data Analysis and Results .....</b>	<b>82</b>
6.1	Purpose for Analysis.....	82
6.2	Analysis and Results.....	84
6.2.1	Latency Data .....	84
6.2.2	Accuracy Data.....	88
6.3	Lens Model Achievement and Consistency Data.....	93
6.3.1	Detection Achievement and Consistency .....	94
6.3.2	Diagnosis Achievement and Consistency.....	95
6.4	Detection False Alarm Data .....	96
6.5	Discussion.....	97
6.5.1	Detection.....	97
6.5.2	Diagnosis .....	101
6.5.3	Lens Model Parameters .....	103

<b>7</b>	<b>Conclusions and Future Work.....</b>	<b>105</b>
7.1	Conclusions.....	105
7.2	Future Work.....	107
	<b>Appendix A: Lens Model Implementation Validation .....</b>	<b>110</b>
	Naval Aircraft Identification Example.....	110
	<b>Appendix B: Graphical User Interface Details.....</b>	<b>113</b>
B.1	User Interface Development .....	113
B.2	Information Displayed.....	115
B.3	Variable Definitions .....	116
	<b>Appendix C: Experimental Consent Forms and Training Materials .....</b>	<b>119</b>
	<b>Appendix D: Demographic Survey and Results.....</b>	<b>144</b>
	<b>Appendix E: Detailed Analysis Results Tables and Plots .....</b>	<b>146</b>
E.1	Detection Latency .....	146
E.2	Detection Accuracy .....	149
E.3	Diagnosis Latency.....	150
E.4	Diagnosis Accuracy.....	152
E.5	Detection Achievement and Judgment Consistency and False Alarms	153

E.6 Diagnosis Achievement and Judgment Consistency ..... 155

**Appendix F: Linear Multiple Regression Overview ..... 158**

**References..... 166**

## List of Figures

Figure 1: Wickens' model of the human information processing cycle[14].....	19
Figure 2: Endsley's model of human situation awareness.....	23
Figure 3: Representations of memory heuristics and where they act on the decision making process .....	30
Figure 4: A visual representation of Brunswik's Lens Model .....	40
Figure 5: Diagram of Skill Score and its components .....	45
Figure 6: Diagram of three different lunar trajectories selected for experiment and highlighted which portion of the trajectory is incorporated in the experimental task ....	65
Figure 7: Square configuration of RCS thruster pods around lunar lander (left), full lunar lander with axis definition and close up of RCS Thruster pod (right) .....	67
Figure 8: Conceptual schematic of the RCS including fuel tanks, thrusters and fuel pumps .....	68
Figure 9: Display presented to subject at the start of each scenario .....	71
Figure 10: Failure button to be selected when subjects detected a failure.....	72
Figure 11: Failure identification panel where a subject selected a button to diagnose a failure after detection.....	73
Figure 12: Trend information display shown to each subject during half the scenarios .....	74
Figure 13: Primary flight display shown to subjects for every scenario .....	75

Figure 14: Trend display at scenario completion .....	76
Figure 15: The area of a deviation is the integral between the normal trajectory signature (top curve), and the failure case (bottom curve) from the scenario start to the time of decision indicated by the white line. ....	83
Figure 16: Data suggesting that trend information has an effect on detection latency.....	86
Figure 17: Data suggesting that trend information has no significant effect on diagnosis latency .....	87
Figure 18: Detection accuracy results showing no significant effect for trend information .....	89
Figure 19: Results indicating a significant effect on decision accuracy for order. ....	90
Figure 20: Results showing the effect of failure type on detection accuracy .....	91
Figure 21: Diagnosis accuracy data showing a significant effect for trend information .....	92
Figure 22: Results showing the effect of failure type on diagnosis accuracy .....	93
Figure 23: Detection achievement data showing no effect for trend information .....	94
Figure 24: Diagnosis achievement data suggesting a possible effect for trend information .....	95
Figure 25: Results suggesting a trend towards increasing numbers of false alarms with trend information.....	97

Figure 26: Results showing the combined effects of trend information, approach, and failure type on detection latency .....	98
Figure 27: Results suggest an increase in the number of false alarms with trend information present .....	100
Figure 28: Results of the combined effects of trend information, approach, and failure type on diagnosis latency .....	102
Figure 29: Primary Flight Display shown to all subjects for each trial .....	114
Figure 30: Trend Information display with same system parameters represented, shown to subjects for half of the trials .....	114
Figure 31: Example of multiple regression plot with regression model trend line .....	158
Figure 32: Example normal histogram with distribution curve.....	163
Figure 33: Plot of residuals vs. the independent variable for a data set that shows a likely linear relationship .....	164
Figure 34: Example spread and level plot for determining equal variance .....	165



## List of Tables

Table 1: Experimental test matrix for all three trajectory types .....	77
Table 2: Factor Effects for Detection Accuracy Data (Significance: $p < 0.05$ ) .....	90
Table 3: Factor Effects for Diagnosis Accuracy Data (Significance: $p < 0.05$ ) .....	92
Table 4: Table showing the effects of trend information on Lens model output parameters for detection (Significance: $p < 0.05$ ) .....	94
Table 5: Table showing the effects of trend information on Lens model output parameters for diagnosis (Significance: $p < 0.05$ ) .....	96
Table 6: Beta-weights calculated for synthetic validation case and compared to Bisantz, et al. (2000) .....	112

# 1 Introduction

## 1.1 Background and Motivation

Advances in hardware and software have created an opportunity to build systems that are capable of increasingly complicated tasks. This opportunity is accompanied by a corresponding increase in system complexity, making the design effort particularly demanding. As part of this effort, engineers are tasked to balance the demands of four drivers: reliability, safety, cost, and performance [1]. Because these drivers have aspects that are often mutually exclusive, engineers are forced to trade one characteristic for another. Evaluating these tradeoffs becomes more difficult, as the system complexity increases. The complexity increases even further when human interaction coupled with high levels of automation are added to the design equation [2]. Even though humans have unique ways of processing information, and are influenced by factors such as stress, fatigue, and hunger, there are still tasks for which humans are better suited than automation. This, in turn, makes the design of the human-system interaction critical to both mission safety and success [3]. Exploring these design challenges was the driving motivation for this thesis, focused on the area of human-system interaction.

While systems can now be engineered to be highly reliable for many known conditions, the noisiness of the real world still introduces complications. Real operating environments introduce unplanned variations in operating conditions, not to mention

the inherent difficulty of predicting future conditions. Real systems must cope with unexpected behavior of system components or human operators, and system malfunctions. All of these inconsistencies mean that there will always be a set of conditions under which the system will fail or automation will reach an incorrect decision [3]. In order for the mission to continue, these failure conditions must be compensated for either by the human or the system[11]. The possibility of failure makes human failure detection and diagnosis critical functions of human operators in both manual control situations and monitoring situations [4, 12]. Poor failure detection and resolution can result in not only loss of productivity but also in loss of expensive equipment and ultimately human lives[10]. Ensuring good failure detection and diagnosis requires understanding and evaluating the operator's decision making process. The area of failure detection and diagnosis was further refined in this research to investigate how the addition of trend information displays for dynamic system parameters affect these decisions.

Trend information on system parameters eliminates extra cognitive workload for operators who would normally be required to store this information in memory. This should allow operators to make faster, more accurate and more consistent decisions, three key decision making performance metrics. A model of human decision making was selected from current literature to aid in examining operator decision performance in a human subject experiment [4-9]. Then, an experiment was conducted to evaluate the effect

of trend information specifically in failure detection and diagnosis tasks. Improving performance in these tasks is an important issue in the design of complex systems because good failure detection and diagnosis are prerequisites for increased reliability, safety, system availability, and also for the enhancement of system performance, and reduced cost [7, 10].

## **1.2 Research Objectives**

The goal of this thesis was to select and evaluate a model of human decision making from past literature and to apply this model in a dynamic environment to investigate the effect of trend information on failure detection and diagnosis tasks. In order to select an appropriate decision making model, several basic requirements were established. These requirements were set in the hope that this model could be useful in future human-system modeling work. These requirements included:

- The model must be applicable and expandable to a variety of different decision types and task domains.
- It must be capable of switching between different human mental models and system task models in order to allow modeling of full mission phases with different actions and tasks
- The model must highlight human decision making performance in both nominal and failure conditions

The Lens model was selected and implemented as the model that best fulfilled these requirements. The Lens model was originally developed for the study of perceptual psychology and is unique in its ability to compare the decisions of a human judge to the true condition in the corresponding environment through a variety of correlations

coefficients which are the outputs of the model[5]. Once selected and implemented, the actual implementation needed to be validated to ensure that it had been implemented accurately. Secondly, the model needed to be calibrated using actual human decision data to prove that it was a legitimate choice that satisfied the basic requirements.

The validation of the model in the context of a dynamic complex system was done via human subject experimentation. The experiment was designed to investigate the possible human performance benefits of explicitly displayed trend information of system parameters for failure detection and diagnosis. Performance metrics included the speed and accuracy for each of these decisions. The Lens model was also used to post process data and two model parameters were also selected as metrics of human decision making performance. The speed and accuracy results have implications for designers as they attempt to build safe and functional complex systems. The model parameter results are useful, both to calibrate the model for future use, and to illustrate the model's use for two different decision types in a dynamic complex system domain.

Since most realistic decision making environments are dynamic, it was important that the model be able to represent dynamic system behavior. However, the Lens model is primarily a static model that has not been used to incorporate dynamic information in its analysis of human decision making. Therefore, this research also extends the use of the Lens model in the area of dynamic decision making. While the Lens model has been previously applied in dynamic environments in literature[13],

dynamic information has not been directly inserted into the Lens model as a part of the human judgment strategy. This thesis seeks to represent dynamic trend information in the Lens model to get a more accurate representation of how this information and its dynamic characteristics affect the judgment performance of complex system operators.

### **1.3 Thesis Organization**

- Chapter 1, Introduction, outlines the motivation, research objectives, and provides an overview for this research.
- Chapter 2, Background, provides information on previous work in human performance modeling, situation awareness, judgment theory, and failure detection and diagnosis.
- Chapter 3, Lens Model, describes the selection and implementation of the Lens model, as well as its history and formulation.
- Chapter 4, Focus Area, describes the two specific decisions, failure detection and diagnosis, that this thesis is focused on.
- Chapter 5, Methodology, describes the procedures and design of a human performance experiment used to test the hypotheses of this research regarding the effect of trend information on failure detection and diagnosis.
- Chapter 6, Data Analysis, Results and Discussion, details the analysis conducted on the data from the experiment, presents the results and a discussion about interesting findings.
- Chapter 7, Conclusions and Future Work, reviews important conclusions drawn from the results section and discusses directions for future work in this area.

## 2 Background

A key issue in complex system design involves understanding the way humans perceive and process system information and how they make decisions. While this thesis is focused on the decision making activities, components of the other cognitive processes help to provide the foundation for decision making.

### 2.1 Human Information Processing

In order to examine human decision making, it is important to understand the human information processing cycle that it fits into. This cycle defines how humans interact with their environment as they perceive, process, and respond to environmental cues and stimuli. A basic framework of human system interaction is provided in Figure 1 below.

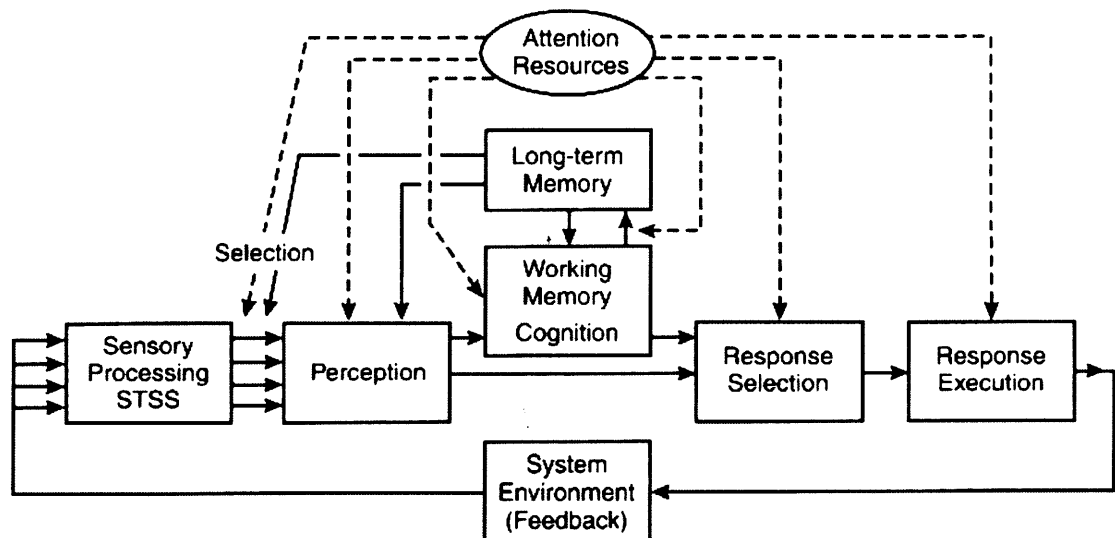


Figure 1: Wickens' model of the human information processing cycle[14]

This model of human information processing includes the following stages: sensory processing, perception, attention, memory, and response. Although this model is an abstraction of the complex nature of human sensation and cognition, it provides a useful platform from which to approach human performance.

This thesis examines the portion of the information processing cycle that begins with the outputs of perception and addresses the cognitive tasks of situation awareness and decision making. Perception is distinct from cognition in that cognition requires significantly more time, effort and attention than perception alone, though the lines between these two processes can be blurred[15]. Cognitive processes are conscious activities which transform or retain information, are resource limited, and are highly vulnerable to disruption when attentional resources are diverted to another task. Situation awareness, which is achieved through perception and augmented by these cognitive transformations, is often the basis for appropriate response selection[15].

### **2.1.1 Memory**

Memory is a foundational piece of the human information processing cycle. It is at work in virtually every stage of human cognition. Memory is both the place where information is stored long term and where information is integrated and transformed. There are generally thought to be two different types of memory: working memory and long-term memory. Working memory is the temporary, attention demanding store that humans use to retain information for a short period of time until they are ready to use it



or commit it to long-term storage. Working memory is also the “cognitive workbench” where we examine, evaluate, and compare different mental representations. It is where judgment and decision-making take place. Working memory has limitations in both capacity and duration. Research has shown that when rehearsal is prevented, humans retain information for approximately 20 seconds. Working memory also has capacity limits which interact with time. The capacity of working memory is generally limited to 7 +/- 2 pieces of unique information. Faster decay occurs the more items are held in working memory because rehearsal is not instantaneous. Therefore working memory tends to perform best if immediate rehearsal of information is possible and if this information is approximately 5-9 pieces in length. This type of memory is significantly affected by stress, and is considered a fragile resource in the context of decision making, because it can so easily be derailed by distractions[16]. Information can be encoded from working memory into the more permanent type of memory: long-term memory. Long-term memory stores facts about the world and how to do things that operators use repetitively. This is a more stable form of memory which is ideally where operators draw on their experience and pattern recognition skills to identify problems with a system or environmental conditions. However it is largely developed through individual experience and therefore not always reliable, particularly for novice operators or monitors who do not have the repetition necessary to have committed such

information to long-term memory. As a result, the success or failure of human memory can have a major impact on the success and safety of a system.

### **2.1.2 Situation Awareness**

Situation awareness (SA) is another metric for evaluating human performance and a critical component of human decision making[17]. In the information processing model, it generally falls somewhere on the border between perception and cognition. SA is described as an emergent property of the model, as opposed to a structural part of the model itself. It is the product and result of many different components and in practice is often treated as a “black box”, where the internal workings are essentially unknown[18]. Endsley defines situation awareness as "the perception of elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future"[19]. Situation awareness is tasked with three things: recognizing and interpreting environmental cues, assessing present and future risk, and assessing the time available to make decisions [16, 20]. It is closely tied to understanding the interrelationships inside the system and the system interaction with the environment or larger domain. Misunderstanding these relationships can result in adverse consequences. Therefore, situation awareness is critical to an operator’s ability to adapt and function effectively in its environment. Endsley presents the following model of situation awareness and its different levels below in Figure 2 [19].

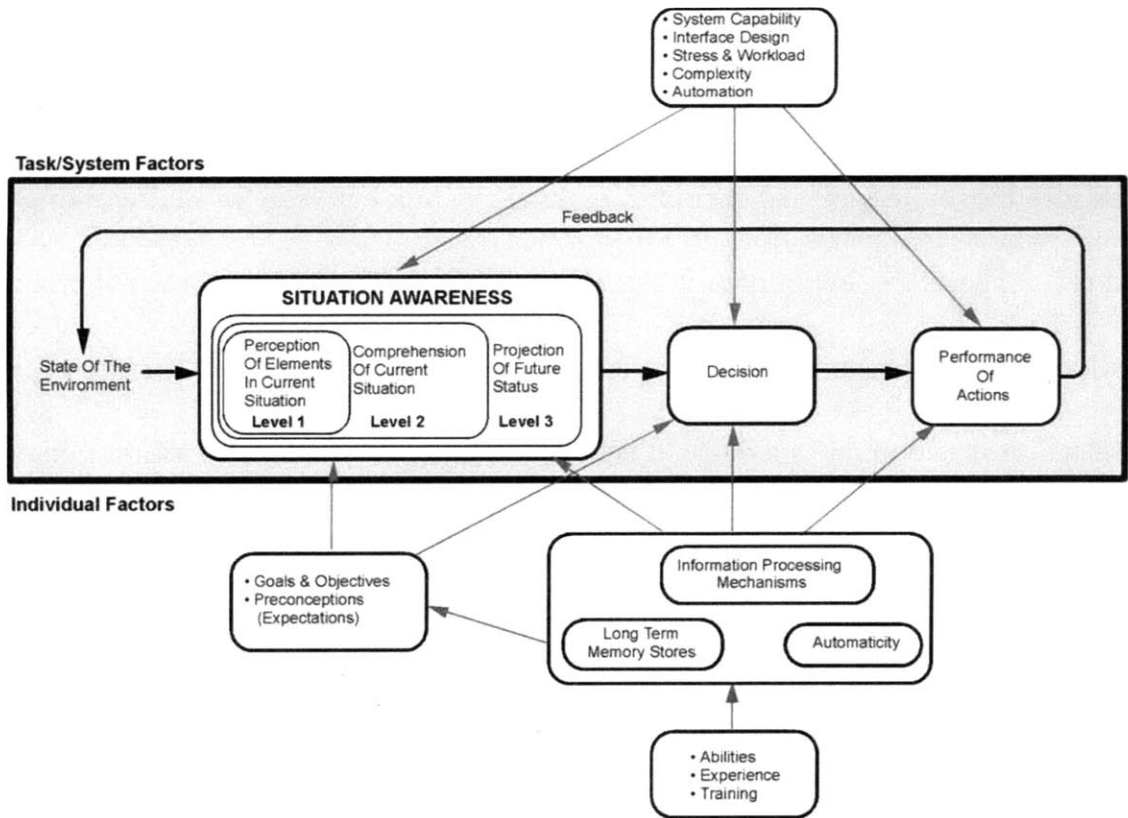


Figure 2: Endsley's model of human situation awareness

Situation awareness is a complex integration of a variety of information which begins with basic perception and ends with projection of environmental states and decision outcomes into the future. In order to plan or problem solve effectively in dynamic, changing environments operators must have a relatively accurate awareness of the evolving situation which is the role of situation awareness. It is a prerequisite for accurate decision making and response selection[15]. It is not purely a raw sense impression, but includes projections of future states and an integrated knowledge of the complete current picture[18].

Understanding which possible states of the world might be in effect, forms the foundation of effective choice[15]. This awareness of the current system condition resides mostly in working memory, making the link between SA and working memory direct. Therefore, awareness degrades as both memory and attentional resources are allocated to different competing tasks. While accurate situation awareness depends on selective attention and memory, it is not the same. The process of maintaining situation awareness, supported by attention, working memory and long term memory is not the same as the awareness itself. It is usually domain specific and is affected by the expertise of the operator. This means that good situation awareness must be developed for every operator and task individually, and is critical to operator's response in new situations. Therefore, supporting broader situation awareness has been shown to be a critical component in decision making.

Situation awareness is measured using a variety of subjective, performance-based, and physiological techniques. Memory probe measurement is a common method which seeks to evaluate the contents of the users working memory for a given task. This is generally done in the form of a questionnaire or evaluation which is designed to determine the overall picture that the operator had of a task situation at the current time. This is a widely used method and generally accepted as a viable measurement of SA[18]. The most common memory probe is known as the Situation Awareness Global Assessment Technique (SAGAT), which freezes a screen and queries

subjects on their knowledge of the situation at given times during an experimental task. This is a knowledge based measurement of situation awareness, which is more accurate at providing detailed theoretical assessment of an operator's awareness. However performance based measurements look directly at user response to realistic situations and are sometimes preferred since knowledge based measurement can only make inferences about what a subject would do given their current knowledge base. Some performance based measures of situation awareness include: workload, user confidence, latency, and accuracy [21].

#### **2.1.2.1 *Environmental Cues and Stimuli***

The information that operators require to make good decisions and maintain SA is acquired through cues in the environment. Typically cues are estimated initially through perception, then information is selected and integrated by attention and cognition using long term memory and working memory to provide background update and revise different hypotheses about the exact nature of the situation. Initially, the decision maker is confronted with a series of cues or sources of information that have some impact on the true state of the world. Some or all of these cues are attended to with the goal of using them to influence the operators current belief in one of several alternative hypotheses about what is going on within the system. Each cue can be characterized by three properties:

- Diagnosticity-Represents how much evidence a cue can offer the decision maker regarding one or the other hypothesis. Cues may be high or low in diagnosticity and also have polarity to favor one hypothesis or another. A perfectly diagnostic cue means that the presence of this cue is a perfect predictor or indicator of a particular hypothesis, and obviates the need for a decision.
- Reliability-Refers to the likelihood that the physical cue can be believed or that the information is accurate. The product of reliability and diagnosticity reflects the information value of a cue. The higher the information value of a cue, the more useful that cue is in terms of evaluating decision alternatives.
- Physical features – Refer to the actual physical representation of the cue such as its salience. These characteristics have bearing on the attention and subsequent processing that any given cue receives.

Selective attention must be utilized in order to process a variety of different cues and give them a subjective weight associated with their predicted information value. First, all of the raw perceived information from any available cues must be integrated together to form some sort of understanding of the current situation. Then, the operator incorporates any expectancies or prior beliefs from long-term memory and mental models, which bias one hypothesis over any others. Finally, they iteratively test and retest that hypothesis in order to attain the final belief which forms the basis for choice[15].

This can be interrupted by three vulnerabilities of human attention and cue integration that can prevent cues from effectively informing decision making:

- Informational cues are missing- The operator doesn't have all the necessary information at hand. Good decision makers are often aware of what they don't know and seek out these missing cues.
- Information overload-Available cues are numerous. Beyond three distinct cues, people do not use the information to make more accurate decisions. This is

because human operators deploy a selective filtering strategy on available information. This filtering can be time consuming and affect decision quality.

- Cue salience- How noticeable a piece of information is to the operator. This affects the attention and processing devoted to any particular cue. If one cue is highly salient it may end up being paid too much attention and over processed. For example, alerting cues (high salience) are not necessarily compatible with effective fault diagnosis since salience should be related to the information value of a piece of information for full failure resolution, not just in detecting that a failure has occurred.

Once cues are processed and integrated, the operator must assign a weight or importance to each one to use it in their judgment strategy. Processed cues are often not differentially weighted effectively based on their information value. More valuable cues would theoretically be weighted higher; however this requires estimating the information value for all cues. It requires less cognitive effort to simply weight all cues equally. Operators subjective weighting tends to vary in an “all or none” manner, instead of as a linear function of the cue’s correlation to the actual environment. When people are asked to estimate differences in the reliability of a set of cues, they can often do so accurately. However, when these estimates are to be used as part of a larger aggregate, the values become distorted. This is more commonly known as the “as-if” heuristic. Values are typically interpreted “as-if” they were all equal in their predictive value. The relative insensitivity of humans to differences in the predictive validity or reliability of a cue should infer that we are poorly equipped for performing tasks where diagnosis or prediction involves multiple cues with different information values. It has been suggested that humans should only be involved in identifying the relevant pieces

of information for a task, and machines should be tasked with the diagnosis decision [15].

### **2.1.3 Decision Making Heuristics**

A current trend in the study of human judgment under uncertainty suggests that it is largely based on heuristic processes. A heuristic is a mental shortcut involving a variety of psychological processes that help humans assess information without using probabilities in their reasoning[16]. Heuristics are useful because we often don't know the true probability of each possible outcome, and so we use tactics like similarity, past experience, and examples to guide our decisions and save time in dynamic situations. This leads to the view that actual human information processing is not as accurate as normative models of classical rational decision theory which involve actual event probabilities. Heuristics explain our deviations from these optimal decision frameworks. They trade accuracy for speed and attempt to save time and cognitive resources by using shortcuts or samples of available information. However, some research has shown that well developed heuristics can be both fast and accurate. Under normative theory, the decision maker must select the best course of action based on the highest expected utility. Expected utility can be calculated by taking the "cost" of an outcome and multiplying it by the probability that particular outcome will occur. Since humans naturally avoid using probabilities in their decision making however, this expected utility is often difficult to determine, and that normative decision making

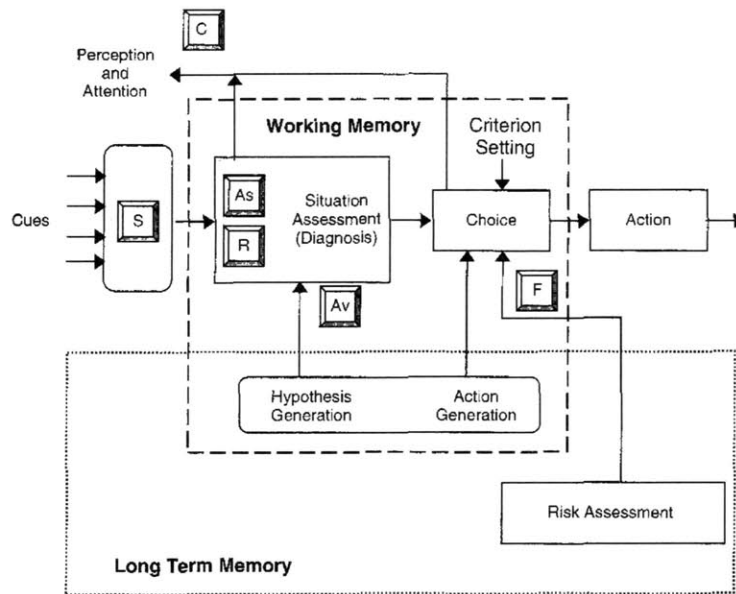


theory may not be the most accurate way to model real world human decision making [16]. Human problem solvers often select what they perceive to be the current best plan with no guarantee or even probabilistic assurance that it is the absolute best plan. From an information processing standpoint, decisions represent a many to one mapping of information to responses. The simplest decisions are go-no go scenarios, with more complexity entering the equation as the decision options become multiple choice. More complex decisions take more time and cognitive effort to go through the decision making process.

A basic set of information processing activities and the biases and heuristics that commonly affect them is presented below. The information processing model lays out the steps of the decision making process as[22]:

1. Cue sampling
2. Situation Assessment (Diagnosis)
  - a. Hypothesis Generation
3. Choice
  - a. Criterion Setting
  - b. Risk Assessment
  - c. Action Generation
4. Action

Within each of these steps are specific heuristic processes that affect the output of that particular stage. Figure 3 shows a diagram of this process, with square “keys” representing the effect of different decision making biases and heuristics on that stage.



**Figure 3: Representations of memory heuristics and where they act on the decision making process**

Different memory heuristics affect each step of the decision making process. The first step, cue sampling, is subject to the salience bias (S). The salience bias is a human bias where people focus their attention on things that are highly salient or prominent and overlook more subtly presented, modest information in the environment[23]. Situation assessment is affected by the “as-if” heuristic (As) and the representativeness heuristic (R). The “as-if” heuristic means that operators weight all incoming information equally when making decisions, with no one piece of information being considered more important or reliable than another. The representativeness heuristic biases operators to judge the probability of an outcome by how much it resembles available data. The inputs for situation assessment, like mental models, come from long term memory and are affected by the availability heuristic (Av). The availability

heuristic acts from the long-term memory because this is where recollections of previous events, mental models, and patterns are stored. It influences the operator to judge the likelihood of a choice as being optimal based on how quickly and easily they can recall situations to support that particular event[16, 24]. Other heuristics that affect the operator's decision making strategy originate from within the system. It has been suggested that automated systems introduce opportunities for new decision making heuristics and associated biases[25]. Automated systems introduce new highly salient cues that supplant the operator's traditional reliance on assessing patterns or combinations of familiar cues. This bias, known as an automation bias, reduces situation awareness and leads to overreliance on cues provided by an automated system.

### ***2.1.3.1 Operator Mental Models***

The prevalence of heuristics in decision making is largely connected to the mental models that the operator has established about the specific task they are attempting to perform, or others similar to it. Humans form mental models about the tasks they are doing and the environment. These models help humans integrate and comprehend the meaning of an aggregate of different sources of state information. A mental model is a set of well-defined, highly-organized yet dynamic knowledge structures developed over time from experience [26, 27]. These models are highly susceptible to heuristic biases, and are generated from an operator's personal

experience. They play a key role in operator decision making, and are actually a metric for assessing human performance. Accurate operator mental models are one of the foundations for achieving good situation awareness and subsequently making good decisions[28, 29].

Mental models are primarily developed through training and experience. This explains why novice operators have more difficulty making decisions and get overloaded with information more quickly[30]. In contrast, experienced decision makers assess and interpret the current situation (Level 1 and 2 SA) and select an appropriate action based on the different conceptual patterns stored in their long-term memory as mental models[31]. This stable source of information allows them to recognize specific situations quicker and more consistently. Specific environmental cues activate these stored mental models in experienced operators, which then guide their decision making process. Mental models, therefore, reflect the user's experience with and understanding of a system and its environment. Accurate mental models are one representation of operator experience and are helpful when other learned procedures fail as well as improving performance at novel tasks, because they draw on similarities between the current situation and different stored models. One of the goals of training is therefore to give operators sufficient experience to create accurate mental models and facilitate efficient system operation and decision making.

## **2.2 Human Decision Making**

Decision making is the act of choosing between alternatives under the conditions of uncertainty[16]. Decision making has been studied extensively, and decision making research generally falls into one of three classes:

- Rational or normative decision making approaches are focused on how people should make decisions according to some optimal framework. It includes factoring both the value of an outcome and its subjective likelihood into the choice. This type of research looks at how and why humans depart from these optimal strategies.
- Cognitive or information processing approaches are focused on the biases and processes used in decision making related to limitations in attention, working memory, strategy, or familiar decision routines such as heuristics. This type of research examines the causes of these biases
- Naturalistic decision making places great emphasis on how people make decisions in a realistic environment where they have expertise, and where the decisions are complex. This can be combined with other classes of research.

The decisions that are studied are typically classified according to several factors including their level of complexity and their quality. The environment has the largest influence on the actual decision. In the context of flight planning it has been shown that nearly half of the variability in pilots' problem solving behavior is due to environmental or domain features. The domain characterizes the type and complexity of the decision being made. Planning difficulty increases for tasks and environments where there are more choices available for action. Domains that generate good decision making, and those that are more likely to generate poor decision quality have certain identifiable characteristics[15]. Domains that lend themselves to good decision making are dynamic, repetitive, and have feedback available to the operator. Humans tend to perform better

in these environments when the decisions are about things, and are decomposable problems. Examples include chess, physicians, accountants, and weather forecasting. Characteristics of poor decision making domains are just the opposite. They are static, more uncertain, and incorporate less feedback. The decisions may be about people or behavior and are not easily decomposable, such as clinical psychology, personnel selectors, stock brokers, and court judges.

Aside from the domain, other key features of decision making include: the amount of uncertainty in the decision, expertise of the operator, and time allotted to make the decision. Uncertainty is associated with the consequences of a decision and is also termed risk. Information is defined as the reduction of uncertainty. Before the occurrence of an event we are less sure of the true state of the world than after that event occurs. This reduction in uncertainty comes from the information gained from the event. In decision making, this evidence is particularly important when decisions have more than two outcomes, which is often the case in real decision making tasks. Providing the operator with as much useful information as possible in order to reduce the uncertainty improves decision performance[8].

Expertise is a characteristic of the decision maker or operator that decreases the time they need to make a decision and helps alleviate time pressure. Time pressure can have a critical influence on the decision making process. The more time pressure the operator experiences, the more likely they are to use heuristic decision making methods

and the more likely they are to make a rash, incorrect decision. With time pressure, normal retrieval of relevant mental models is prevented; they do not have time to adequately search their long term memory. Operators are also influenced by stress and fatigue, which reduce both decision optimality and operator confidence.

Decision making can be studied from the standpoint of the actual decision or from the strategy employed to reach that decision. Judgment theory examines decision strategies using two different classifications of strategies. Judgments can be made using a compensatory strategy or a noncompensatory strategy. Compensatory strategies weigh both good and bad attributes of a decision option or choice, and allow these strengths and weaknesses to “compensate” for one another. These strategies generally make better use of available information, although struggle when the decision is being made with a high level of uncertainty. Noncompensatory judgment strategies do not consider a balance of the good and bad attributes of a choice and as a result may not consider all the information available to the decision maker. An example of a noncompensatory strategy is setting an absolute minimum for a certain characteristic. All alternatives that fall below that minimum would be rejected as valid choices. This is a common practice in aviation and simplifies a complex decision situation by reducing the cognitive effort required to make the decision[16].

### **2.2.1 Decision Making Performance**

### *2.2.1.1 Decision Quality*

The quality of a decision is typically based on the outcome. This quality metric is retrospective, and based on comparison to “expert” decisions. This is not necessarily an accurate measure of quality since it has been shown that experts don’t always make better decisions. It also assumes the expected value of the decision is common among all decision makers, which depends on assigning universally agreed upon values to the outcomes of a choice. Since this is not very realistic, it may be more important to look at the ways different environmental and information characteristics influence processing operations and outcomes. Decision quality is influenced by stress, which interrupts or exerts pressure on the decision maker and leads to a suboptimal decision making process, especially when the task involves a high level of spatial working memory [15]. Quality may be better determined by examining the process by which the decision was reached, and how consistently the operator applies this strategy, rather than the actual outcome.

### *2.2.1.2 Decision Performance*

Evaluating an operator’s decision making process or strategy is difficult because this information is challenging to elicit accurately. Humans are notoriously poor at accurately understanding and self-reporting cognitive activities. As a result, metrics available to examine decision making processes tend to be limited in their scope and effectiveness. While decision quality is an important consideration, other performance



metrics are commonly used which are more objective as to their definitions and easily measured. These common metrics are decision latency and accuracy. Latency is the time it takes for an operator to make a decision. Accuracy is simply whether or not the decision was correct as compared to the actual environment. The ideal decision would be both fast and accurate; however there is generally a tradeoff between these two metrics. Pilots who are good decision makers tend to take longer to understand a situation or decision problem and acquire evidence, even though they select and execute actions quicker[15]. Depending on the domain and the possible consequences, accuracy is generally considered the more important of the two metrics, although, in dynamic systems, excessive decision latency can also have an adverse affect on system performance which must be taken into consideration.

### **3 Lens Model**

The Lens model was selected as the human decision making model that both met the initial requirements, and provided the greatest amount of descriptive information about the human judgment and its corresponding environment. The Lens model is traditionally a static model. For this work, it was expanded to incorporate dynamic information as part of the human judgment strategy, and implemented to post process experimental data about failure detection and diagnosis decisions in the context of a lunar landing scenario.

#### **3.1 History and Formulation**

The Lens model is part of a class of normative decision making models that was developed and proposed by Ergon Brunswik originally for the study of perceptual psychology. Brunswik believed that psychology should seek to provide descriptions of behavior that actually occurred rather than discover laws of behavior[32]. This led him to develop the theoretical foundations for the Lens model in the early 1950's. Ken Hammond was the first to apply and popularize Brunswik's work through application to judgment and decision making theory, where the model is still primarily used today. The model's foundations are based on the concepts of representative design and probabilistic functionalism[5].

Representative design is concerned with the selection and inclusion of stimulus conditions in a scientific study. The goal is to ensure that any stimulus to be tested is

sampled from a representative population in the same way that subjects are sampled from a representative population. A representative population means that all facets of the population of interest that exist in the actual world are also represented in the study group. This is commonly practiced when selecting a subject pool, but experimenters typically only present subjects with a limited set of stimuli that is often unrepresentative of the true environment[32]. Brunswik believed that psychology generally focused too much on how subjects were sampled and not enough on ensuring that the environment was well represented in a variety of different contexts. The normal systematic design of experiments that is so common in the science communities to date strives to control variables through factorial design. This allows the results to be generalized to other subjects and populations of subjects but does not allow for generalization to other conditions or contexts within the same environment.

Probabilistic functionalism has two main premises:

1. Psychology should be concerned not just with the human organism but, more importantly, with the interrelations between the organism and its environment.
2. This relationship of organism to environment is based on uncertain or probabilistic relations among environmental variables. These variables are the proximal information or cues which are used to make inferences about the environment. These cues are not consistent, and not always available.

This second point leads to Brunswik's theory of vicarious functioning, which is essentially the redundancy of cue information in the environment. An organism may use any number of available means to achieve the end goal[33]. These theoretical

concepts are all encompassed in the basic formulation of the Lens model, shown graphically in Figure 4.

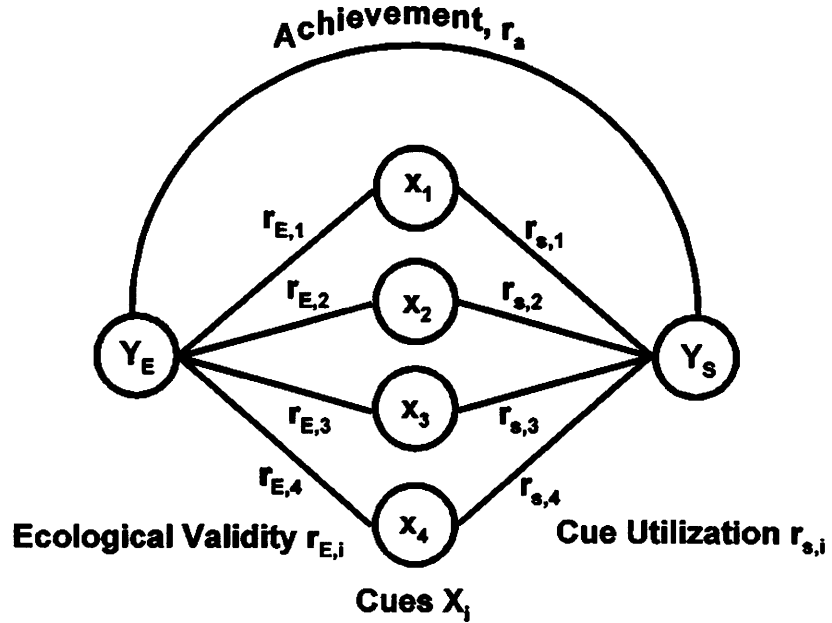


Figure 4: A visual representation of Brunswik's Lens Model

Several mathematical representations of the model have been developed, the formulation of the Lens model presented here is based on multiple regression and correlation techniques[34]. It begins on the left side of the model with the environmental variable,  $Y_e$ .  $Y_e$  is modeled as a linear function of a set of  $k$  cues,  $X_j$  where  $j = 1, \dots, k$ . Thus, the environment is modeled by the regression equation:

$$Y_e = \sum_{j=1}^k r_{e,j} X_j + \varepsilon_e$$

Where  $r_{e,j}$  is a vector of weights that represents the relative change of different cues with changes in the environmental state, and  $\varepsilon_e$  is the error term of the regression of  $Y_e$  on the cue values of  $X_j$ .

Similarly the human judgment about the environment, denoted  $Y_s$ , is modeled in the same way, yielding the formula:  $Y_s = \sum_{j=1}^k r_{s,j} X_j + \varepsilon_s$

Where  $r_{s,j}$  is a vector of weights that represents the relative change of different cues that the human judge ascribes to the environment in their decision strategy, and  $\varepsilon_s$  is the error term of the regression of  $Y_s$  on the cue values of  $X_j$ .

The cue weights for the judgment,  $r_{s,j}$ , and the environment,  $r_{e,j}$ , cannot be compared to one another initially because they are all in different units, depending on the cue the weight was generated from. In order to compare cue weightings to effectively understand the judgment strategy and environmental prediction from a certain set of cues, we must standardize the weights. These standardized coefficients are called beta weights and denoted  $\beta_{e,j}$  and  $\beta_{s,j}$ . These are relative weights that indicate the relative importance of one cue over another in the actual prediction of the environment and the corresponding judgment strategy.

Given the fact that the judgment variable and the environmental variable are based on the same set of observable cues present in the environment, a person's decisions should match the environment to the extent that the weights that the judge assigns to environmental cues match those used in the actual model of the environment. This is the same as identifying the correlation between  $\beta_{e,j}$  and  $\beta_{s,j}$  for all  $j = 1, \dots, k$ . This multiple correlation between the judgment and the criterion,  $\rho_{Y_e Y_s}$ , is also called achievement and denoted by  $r_a$ . This value is the primary performance metric

generated by the Lens model and is calculated based on the Lens model equation:

$$r_a = GR_eR_s + C\sqrt{(1 - R_e^2)(1 - R_s^2)} \quad [35, 36].$$

Where:

$r_a$	= Achievement:	correlation between judgment and criterion
$G$	= Knowledge:	correlation between the linear models of the judge and the environment
$R_e$	= Environmental Predictability:	multiple correlation between the environment and the cues
$R_s$	= Participant's Consistency:	multiple correlation between the cues and the judgment
$C$	= Unmodeled knowledge:	correlation between the residuals of both models

The primary values of interest in the model are the beta weights that represent the cue influence in both the environment and the judgment strategy and the achievement. The other Lens model parameters then provide supplementary information about the performance of the model, the environment, and the performance of the judge individually. Other alternative performance measures have been defined from the achievement equation. For situations where the unmodeled knowledge  $C = 0$ , the human component of the achievement equation can be expressed independently as the product,  $GR_s$ , named "linear cognitive ability". This neatly captures the ways in which judges match the environment and how consistent they are in their strategies. Similarly, the quantity  $GR_e$ , captures the validity of the judges strategy if it were to be

applied in a perfectly consistent manner and  $R_s = 1$ . This is what would happen if the human judge were replaced by their regression model[34].

While regression is the most common formulation of the model, other formulations have been suggested to overcome the inherent limitations of regression analysis. The use of the Lens model in judgment theory was popularized by Ken Hammond who is quoted as saying: “a [...] sin of commission on my part was to overemphasize the role of the multiple regression (MR) technique as a model for organizing information from multiple fallible indicators into a judgment. There is nothing within the framework of the lens model that demands that MR be the one and only model of that organizing process”[33]. Other methods have been used to get these judgment weights including fast and frugal heuristics[37], non-compensatory formulations [38], and Bayesian methods[39].

### **3.1.1 Limitations of Correlation and Skill Score**

Correlation is the basis for calculating all of the Lens model output parameters and so is an important piece of the formulation. The correct use and interpretation of correlation coefficients can significantly influence the interpretation of any Lens model results. Correlation is highly dependent on the assumptions made with respect to the data that is being correlated and on the basic principles forming this index of association. In situations where these assumptions are violated, the correlation

coefficient yields little valuable insight into the relationship between variables, in this case the human judgment and the corresponding environment[40].

It is unwise to assume that the correlation coefficient neatly captures the relationship between two variables in a single value. Many different types of relationships can have the same correlation coefficient and therefore the relationship should be examined more thoroughly. In order to accurately capture the relationships in a set of data, it is important that the data be sampled from a representative space in order to capture the relationships within a whole population and not just a segment [40]. Ensuring representative sampling is difficult and creates two primary biases in the calculation of correlation: scale errors and magnitude errors.

*Regression bias:* The regression bias is a scale error representing the degree to which the standard deviation of human judgments is not scaled to the standard deviation of the environment and imperfect achievement. A regression bias is common in uncertain environments. This bias occurs when an operator does not appropriately regress their judgments towards the mean but instead is biased by case-specific information. They do not adaptively balance both base-rate and case-specific information.

*Base rate bias:* A base rate bias is a magnitude error that exists when the mean of the distribution of the human judgment variable does not equal the mean of the distribution of the environment. This occurs because a judge must reason and make



predictions about a specific case and appropriately balance both case-specific information and base rate information about the whole population or environment. This means that the judgment and the actual environment distributions will be slightly offset, and the judges decisions are consistently under or over estimated[33].

Since correlation analysis is so critical to most Lens model formulations, the limits of correlation have led to the creation of another performance metric called the skill score. The skill score originated from weather forecasting[41], and was applied to human factors and situation awareness for judgment analysis[42, 43]. A diagram of the skill score as a measure of judgment quality is found in Figure 5.

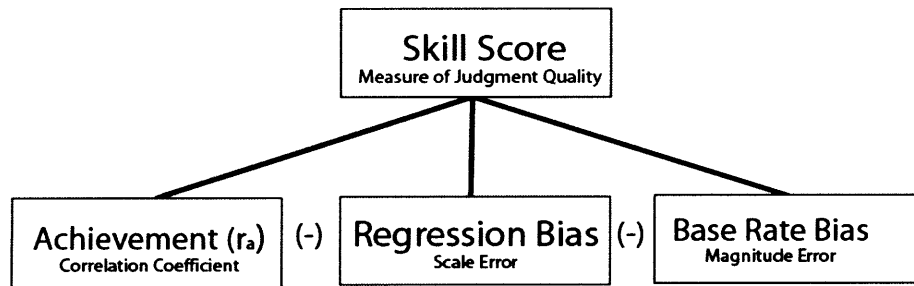


Figure 5: Diagram of Skill Score and its components

The skill score is a measure of “distance”, the squared Euclidian distance, between data sets rather than the shared shape of a distribution, which is how correlation is determined. It accounts for the regression bias and the base rate bias that are inherent in the calculation of correlation and can make its interpretation more ambiguous. The equation given for skill score is:  $SS = r_a^2 - \left(r_a - \frac{s_{judge}}{s_{env}}\right)^2 -$

$\left(\frac{\bar{Y}_s - \bar{Y}_e}{s_{env}}\right)^2$  where  $s_{env}$  and  $s_{judge}$  are the standard deviations of the judgment and the environment, and  $\bar{Y}_s$  and  $\bar{Y}_e$  are the means of the judgment and environment[33]. High quality judgments have a high correlation between the judgment and the environment but generally have low regression bias and base rate biases which leads to a high skill score.

### **3.2 Previous Applications**

The Lens model has been explored, enhanced, and developed for five decades. As a result there is a wealth of previous research to draw upon in a wide variety of domains. According to a meta-analysis of Lens model studies done in 2007, the Lens model has been applied to at least 259 different task environments in 78 different papers[34]. This work synthesized the Lens model statistics presented above and found overall high levels of achievement and noted that people can achieve similar decision performance in both noisy and predictable environments. One of its primary applications is for cognitive engineering and the Lens model has been applied in domains ranging from military command and control to music education, from laboratory experiments to field studies, and also studies containing an emphasis on learning. The Lens model is in a unique position to provide feedback for learning studies and determine what the most effective form of feedback is for a particular task[44]. It has business applications in marketing, accounting, and engineering, and

has continued to be used in social research areas like preventing violent behavior, moods, and consciousness studies.

This large body of previous work on the model and its extensions gives the model some credence for further use in more new domains. It has been used effectively on a wide variety of problems which has shown that despite its limitations it is a useful model for examining human judgment in an array of contexts. Few papers that contain Lens model data provided individual data and so assumptions cannot be made about variation within the different environments[34]. The model does appear however to be relatively robust to differences in both domain and task. The same meta-analysis provides generalized information from the aggregate of much of this past work.

### **3.3 Limitations**

While the Lens model provides us with an interesting way to examine judgments and environments there are four major limitations that it does not accurately model. First, it does not model a judge's adaptive use of cues not included in the model. This is true of most modeling efforts but still leaves the question of whether or not the most appropriate cues have actually been included. Secondly, the Lens model does not inherently capture the judge's ability to adaptively respond to non-linearity in the environment, and cue-criterion relationships. This means that any environmental variables that affect the criterion value in a nonlinear way will be represented in the quantity C, but the Lens model will not help determine how the operator might be

using those nonlinear cues and their relationships to any other cues in the model. These adaptations can significantly alter an operator's decision making strategy. In light of this, it is critical to recognize that environments with highly nonlinear components will have lower achievement values and the predictions of the Lens model will be less accurate. Next, the model does not explain any possible chance agreement between errors of both the human and environmental models. It is therefore important to understand that the Lens model does not capture or identify errors in the models. When selecting cues for the environmental model, if they can be selected or engineered (i.e. do not occur naturally), it is critical to select the best possible linear model of the environment with which to compare judges. This can be done by statistically investigating the effects of a variety of sets of "possible" cues to determine which ones are most important to include in the model[45]. One last important limitation of the Lens model is that it is primarily a static model without the capability to investigate dynamic cues. Though it has been used to try to evaluate decisions in time-dependent domains[13], this is generally done by ignoring the dynamically changing nature of information in the model directly and building models from snapshots of data at the instant of a judge's decision. Dynamically changing information is implied in the results, but not actually directly investigated in the model.

### **3.4 Implementation and Data Post Processing**

The Lens model was implemented in Mathematica's MATLAB 2009a as a generic m-file that can be used with any set of data in either text file, Comma Delimited (CSV), or Excel spreadsheet format. The implementation currently requires the following input data which was collected in CSV format from the experimental interface:

- Cues : These can be represented in any number of column vectors of data which correspond to the different observable cues or inferred dynamic information being used to predict the judgment and the environment.
- Actual Judgment: The actual judgment vector must also be a column vector that is the same length as the number of cue observations, one judgment response for every set of cues, this data is often binary or in some other scaled integer format.
- Actual Environment: The actual environment vector, represents the true condition of the environment for each set of cues, and has the same requirements as the actual judgment and should be in the same data format (binary, integer scale, etc.).

The model provides the following outputs, one value for each data file that is run through the model:

- Achievement: The measure of how well the judge did at predicting the environment, or how accurate they actually were. It is essentially a correlation coefficient but also accounts for nonlinear components of the environment and the different parameters listed below in its calculation.
- Skill Score: This metric of decision quality is described above in reference to the limits of correlation coefficients and their interpretation. It is a measure of quality which accounts for the correlation between two variables and also accounts for both the scale and magnitude errors inherent in its calculation.
- Judgment Strategy: This can be discerned by looking at the standardized weights assigned to each of the cue variables. These weights can be compared relative to each other to see where the human judge allocates more of their

confidence or importance in terms of the cues they believe best predict the environment.

- Environmental Predictability and Judgment Consistency: These measures are multiple correlations between the environment, the judgment, and their corresponding models generated in the multiple regression. This gives us some idea of how well the environment can be linearly predicted, and how consistently the judge applies their same judgment strategy each time they are faced with a decision
- Matching: This is a correlation between the judge and the environment. It describes how well the judge understood the linear component of the environment and adapted their judgment to match it.
- Unmodeled Knowledge: This accounts for any nonlinearities in the environment that cannot be explained by the linear combination of the cues presented. This is the correlation of the model residuals for both the judge and environment. It gives us an indication of how much outside information the judge was using in their decision making.

In this form, the model was used to post process data from the experiment to determine each subject's achievement and decision consistency, which were also performance metrics for their decision making. The experimental data files recorded the subject's decision as well as the values of each system parameter at each time of decision as well as calculating the inferred deviation of any of these system parameters from their corresponding normal states. First the detection and diagnosis data were separated for each subject so that the impact of the trend information could be evaluated for each decision type individually. Then these recorded parameter values became the actual judgment values and the cues or inferred states that were used to build the regression equation for the human judgment side of the model. The true

values of the environment were added to the CSV file manually by the experimenter after each session.

Once the environmental data was added, the data with trend information present and without trend information present were separated for each subject and then run separately through the model implementation. This generated two sets of Lens model output parameters for each subject, one that was achieved when trend information was present, one without it. These values were all aggregated into a single data file so that statistical analysis could be performed to explore the impact of the trend information on the Lens model output parameters and compare it to the results of the other human performance metrics like speed and accuracy for both failure detection and diagnosis.

## **4 Focus Area**

In order to utilize a model of human decision making and evaluate the impact of explicitly displayed trend information for system operators, it was necessary to narrow the field of decision making to a subset of decisions applicable to complex system design. Failure resolution is a critical task in human-system interaction and was selected as the focus area for this thesis, specifically two stages of the failure resolution process: detection and diagnosis. Some considerations for modeling human failure detection and diagnosis must include: the types of failure modes that can be modeled, the complexity of the model implementation, the decision performance, and the robustness in the presence of modeling errors[6]. In addition to these, several other studies suggest that participatory mode of the system operator should also be a consideration. Participatory mode is the amount of interaction the human operator has with the system. It is classified into two categories: operator and monitor, one in the control loop, one outside of it[12, 46].

### **4.1 Failure Types**

In order to understand failure detection and diagnosis, it makes sense to begin with the definition of a failure. A failure is any critical changes in the system parameters, or the inherent dynamics of the system[10]. In a complex system with hundreds of parameters that could change, this definition may be too narrow.



Therefore, a failure can also be defined as any unpermitted or uncommanded deviation of at least one characteristic property of a variable from acceptable behavior[7].

The types of failures that an operator may encounter are dependent on the specific domain and systems involved. All components in a system have some unique set of failure modes. However, there are some general failure types that have been previously investigated. Three primary types of failures are discussed below based on the time it takes the failure to develop:

*Step Failures:* One of the most common types of failures is a “hard over” failure or step failure. This is a failure which involves adding some kind of bias or significant error instantaneously to a parameter value. It is generally sudden and well above the detection threshold for humans and thus detection is often assumed. A step failure can be made more subtle by using an output-additive time function which transforms it into a ramp failure, by altering the mean of the observed process[9].

*Ramp Failures:* A more subtle type of failure is a ramp or degradation failure. This type of failure is closer to the threshold of detection and is more difficult to identify because it occurs slowly over the instrument lifetime and incipient. This failure is caused by the error measurement in a sensor or other instrument increasing very slowly, almost unnoticeably over time as the sensor generating the data ages or degrades and manifests itself in an indicator that is inconsistent with others[9].

*Intermittent Failures:* An intermittent failure is a repetitive on-off failure. This could be something like a communications link going down and then coming back online, cell phone reception, or other examples in signal transmission and receipt. It can be very difficult to detect because it is only apparent if the operator needs access to that particular piece of equipment or that information at a time when the instrument is not functional. For example, you only notice that you have poor cell phone reception when you need to make a call.

Another way to look at classifying failures is in terms of overall system impact. Up till now failure types have been discussed based on how they develop over time. Each of these types can have different effects depending on whether or not they can be resolved by the operator or the system. This “reparability” gives another dimension along which different failures can be investigated. Failures can either be permanent or transient. Permanent failures are typically the result of a hardware or software issue that cannot be fixed by the operator in real time during the current mission. The operator must operate with that failure or abort the mission. Transient failures are less serious in the sense that they are correctible during operation and are more often simply a disturbance. They require time and operator cognitive resources but do not usually necessitate a mission abort.

Failure distributions can be modeled in a number of different ways. Failure times can be distributed randomly among participants or components, or failure times

can be modeled as a stochastic process which is a function of the Mean Time between Failures (MTBF), for that specific component or system. Failures have commonly been investigated singly, as opposed to cascading failures or multiple single-point failures, assuming that system failures are independent which is unrealistic. In experimentation, failures are also modeled at higher failure rates than would normally occur because true failure rates can be low and infeasible for human subject testing. These unrealistic failure rates however, may induce expectancy or bias in the human detector to look for failures more than they normally would in a real world situation.

#### **4.2 Failure Detection**

Failure detection is of primary importance to the operator of a complex system. Key to balancing the performance, reliability, and cost of a system is the ability of that system or the human involved to perform accurate and timely failure detection and diagnosis. Engineering systems will eventually will encounter unexpected failures and environmental disturbances which must be detected to be resolved [11]. If the human is monitoring, their primary role is to be an accurate and efficient failure detection mechanism and problem solver who intervenes after a failure to ensure a safe outcome[12]. System failures can potentially result in loss of productivity, expensive equipment, and human lives[10].

Full failure resolution in a system follows a three step process: detection, diagnosis, and compensation[10]. The work in this thesis focuses on the detection and

diagnosis steps. Detection is defined as “determining if a malfunction has actually occurred in the system”[10]. Other authors incorporate the human operator, stating that failure detection is, “the process whereby a human operator decides that an event has occurred”[4]. This indicates the importance of decision making in the human failure detection process. Failure detection is believed to be largely domain and tasks dependent, but a general framework of four different process measurements that humans typically look for in order to detect failure has been identified[8]. These four include: the magnitude of changes between successive measurements, reversals in direction in successive measurements, simultaneous occurrence of large magnitudes and reversals, and localized magnitude changes. Based on these types of trends in system information, and their perceived accuracy of that information, operators must decide if they believe a failure has occurred[15].

There are circumstances where detection represents a source of uncertainty or potential bottleneck in performance because it is necessary to detect events that are subtle and near the threshold of human perception. Humans’ ability to do this effectively is based on their sensitivity as detectors in different situations. This is known as the sensitivity level. It is important to understand what these levels might be in a system so that it can be designed such that critical failures can be detected as quickly as possible.

There are a variety of things that affect detection sensitivity of an operator. One is the number of states of categorization. The process of detection may involve a binary go-no go decision or it may require the human to choose between three or four possible levels of uncertainty about an event, or detect more than one kind of event. This decreases detection performance by increasing the time it takes for the operator to evaluate multiple options. Sensitivity, like detection, is domain and task specific, influenced by operator training, display design, and a myriad of other factors such as physiological state, operator expectations, and experience. Participatory mode of the operator is another consideration. Though there are conflicting results, it is generally accepted that the amount of interaction the operator has with the system as part of the control loop, does have some effect on their detection performance. Sensitivity can also be impacted by the operator's level of preoccupation with secondary or side tasks or mental workload. Increased mental workload has been shown to decrease detection sensitivity[47]. However, despite all their possible limitations, humans have been shown to be relatively keen when contrasted with the detection resolution of machines[15].

Aside from sensitivity other measures of detection performance include latency, accuracy, and false alarms[6]. Detection latency is the time interval between the introduction of the failure and when the human can cognitively detect its presence. It has been argued that detection time is not the best measure of detection performance

and that identification and recovery times should be incorporated as well. Near the operators detection threshold, this is a valid and interesting parameter to evaluate[47]. Accuracy is a measure of how well a detector performs. This is often presented as the fraction of the failures that the human got correct. Another consideration that is closely tied to accuracy are false alarms, which are a special case of inaccuracy in detection which is of primary concern to designers in many different system domains such as medical monitoring. A false alarm occurs when a failure is detected when it is not actually present in the environment.

### **4.3 Failure Diagnosis**

Failure diagnosis is the next step in failure resolution and seeks to identify the cause of the failure. The more complex the system, the more difficult it is to diagnose failures and accurately evaluate system reliability, particularly where the human is involved. This is typically a categorical decision with multiple possible outcomes. The human must be able to discriminate between these different categories of failures, gather evidence about system performance, and receive feedback to accurately diagnose a failure. To do this, human operators use knowledge of cause and effect relations. The ways that faults propagate through a system and manifest themselves as observable symptoms generally follow physical cause and effect relationships. These relationships help operators to identify possible causes, assuming they are familiar with the system dynamics. Since the physical properties of the different system parameters are

connected quantitatively by the system dynamics and by time, these dynamics define the symptoms of each failure in a unique way. Diagnosis may require learning decision strategies and criteria for cross checking instrumentation as well as developing a set of rules to identify a failed component[48]. Diagnosis decisions then, are based on the observed symptoms of the systems current behavior up to the current time[7].

If the operator does not know the failure-symptom relationship prior to the failure occurring, they apply classification methods, such as neural networks, which classify symptoms into failure categories that make implicit sense from their knowledge about the system. If the causal relationship is known ahead of time, diagnostic reasoning strategies and logic are used to try and narrow down the possible failures. Rule based decision makers and Boolean algebra are models that fit this type of decision making process[7].

Like detection, diagnosis is affected by fatigue, stress, and complexity. Diagnosis is often a complex decision involving a large number of cues and a large number of possible outcome categories. Difficulty in making diagnosis decisions arises when the failure symptoms are not essentially unique, so that symptoms in the system dynamics could be attributed to multiple possible failures that look similar on multiple levels. Here operator experience, extensive knowledge of the system dynamics and functions, and accurate mental models are incredibly valuable for accurate failure diagnosis. Pattern recognition and associative memories are effective means for dealing with this

complexity; comparing the characteristics of failures with the signatures of previously known failures. Other tactics for failure identification include either hardware or analytical redundancy so that system parameters can be compared to a nominal value in order to determine what has failed[11].

Diagnosis performance is predicated on the initial step of failure detection, and is likewise measured by latency and accuracy. The diagnosis latency can be calculated as the time between the detection of a failure and the identification of its cause. For accuracy, inaccurate diagnoses take several forms. First, the operator both detected and diagnosed a failure that was not present. Secondly, the operator detected a failure correctly, but diagnosed it incorrectly. Next, the operator detects a failure but never made a diagnosis as to what occurred. And lastly, the operator failed to detect or diagnose a failure that was present. Some mistakes are mistakes only in detection, some are errors in diagnosis only, and some may be errors in both detection and diagnosis.



## 5 Methodology

This chapter discusses the failure detection and diagnosis experiment that was conducted to examine the effect of trend information on human failure detection and diagnosis performance and generate data to demonstrate the Lens model implementation.

### 5.1 Motivation

This experiment is designed to investigate the effect of explicitly displaying trend information about system parameters to operators on their performance in failure detection and diagnosis tasks. Understanding what constitutes normal for the system is critical to failure detection and diagnosis so that deviations from this normal state can be detected and identified. The system operator must have both good situation awareness and a mental model for what normal conditions should be to perform well at these tasks. Understanding the past trend of system parameter behavior is part of maintaining awareness of the current system state, as well as projecting those states into the future. Trend information also allows the operator to learn patterns and signatures for specific failure modes. These specific mental models and improved pattern recognition are important factors in accurate and timely failure diagnosis. Typically, these trends must be stored in the operator's working memory which is known to be limited in both duration and capacity. Displaying them explicitly removes this extra

cognitive workload from the operator, theoretically improving detection and diagnosis performance.

## **5.2 Research Questions and Hypotheses**

### **5.2.1 Research Questions**

The experiment was designed to explore the following research questions:

- Does explicitly represented parameter information in the form of past trend behavior improve the detection and diagnosis performance, in terms of reduced latency and increased accuracy, of the system operator?
- Does the addition of explicit trend information improve the achievement of system operators?
- Does the addition of trending information increase the consistency of operators' decision making?

### **5.2.2 Hypotheses**

*Adding trend information will decrease latency and increase accuracy for both the detection and diagnosis decisions.* Explicit trend information was expected to decrease detection latency and increase accuracy because deviations from the nominal behavior of a system should be more quickly recognizable if the past nominal behavior is immediately visible versus stored in working memory. Similarly for diagnosis, it makes sense that patterns would be more quickly and accurately identified if the past behavior is explicitly displayed. The corresponding reduction in the demands on memory was expected to help subjects be both faster and more accurate decision makers.

*Trend information will improve a subject's ability to consistently use the same judgment strategy in detecting and identifying failures.* The increase in consistency should occur for several reasons; first it has been shown that additional cues (in this case trend information) increase an operator's ability to make decisions consistently to a certain extent before they become overwhelmed by information overload[49]. This information provides a critical crosscheck that will help confirm or deny a subject's hypothesis about which failure has occurred. This extra evidence reduces uncertainty and therefore should increase both accuracy and consistency.

*Explicit trend information will increase the achievement of subjects when compared to the recall of this information from working memory.* This makes sense when we look at the hypotheses above and the equation for achievement in the Lens model:  $r_a = GR_eR_s + C\sqrt{(1 - R_e^2)(1 - R_s^2)}$ . Achievement is another measure of subjects' accuracy which takes into account other model parameters. It has been shown that  $R_s$  and  $G$  are positively correlated, and for a dynamic task it is assumed that the trend information is an important component of the actual environment,  $G$ . Given this assumption, and the belief that explicit historical information will increase the consistency of judges and their accuracy, an increase in achievement is expected.

### **5.3 Experiment Overview**

#### **5.3.1 Background**

The focal point for all experimental failures was the reaction control system (RCS) in the context of a lunar landing scenario. This subsystem is responsible for the attitude control of the spacecraft in the roll and yaw directions during descent. It was selected because failures of this system would all be manifested in some way on the instruments that might be available in the cockpit.

#### *5.3.1.1 Lunar Landing Spacecraft*

Due to the nature of the detection task, it was important to present subjects with multiple trajectories, in order to prevent them from becoming overly familiar with one trajectory signature and thereby make the detection task trivial. Three different trajectories were used based on the angle of their approach to the surface: 15 degrees, 30 degrees, and 45 degrees. The 30 degree approach angle was defined as nominal because it most closely approximated the approach angle used for lunar landing during the Apollo space program. The 15 and 45 degree trajectories were then defined as shallow and steep, respectively. All three trajectories had the same acceleration profile of 1.2 lunar gravities and began at a slant range of 1000 km from the desired landing point. Below in Figure 6 is a notional diagram of the three trajectories used. The figure also depicts the portion of the trajectory that was selected for the experiment from the end of the pitch over maneuver to the start of the terminal descent to the surface.

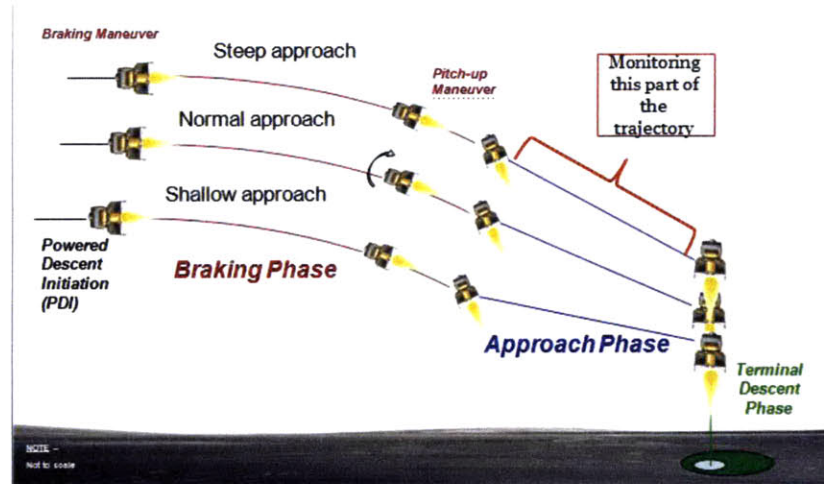


Figure 6: Diagram of three different lunar trajectories selected for experiment and highlighted which portion of the trajectory is incorporated in the experimental task

The trajectory is important to the operator because different trajectories exhibit different burn patterns during this phase. These burn patterns are important to an operators understanding of “normal” trajectory conditions, which means that fuel is a primary indicator of RCS system failures. Steep trajectories burn less fuel overall but come in much higher and faster than shallow trajectories. Shallow trajectories begin closer to the lunar surface and have larger burns required in order to complete the spacecrafts large rotation and pitch over maneuvers early on in the approach phase. Normal and steep trajectories have larger burns towards the end of the approach phase as they attempt to maintain proper attitude during deceleration for the terminal descent phase.

### 5.3.1.2 *Spacecraft Subsystems: RCS*

The RCS is primarily responsible for spacecraft attitude control. Guidance receives position information from the spacecraft navigation filter and commands the RCS thrusters accordingly in order to maintain a stable attitude. Spacecraft attitude is a three dimensional vector quantity comprised of three angles: pitch, roll, and yaw. Pitch is a forward and backward rotation around the Y-axis. In the spacecraft, positive pitch is a pitch backward or a pull up of the “nose” of the spacecraft. Roll is rotation about the X-axis where positive roll is roll to the right, negative roll to the left. Yaw is a rotation around the Z-axis. Positive yaw is defined to be motion to the right, and negative yaw to the left. In the lunar lander pitch is controlled by the spacecrafts gimbaled descent engine, called thrust-vector control (TVC) and so is largely unaffected by firings from the RCS thrusters.

The configuration for this test case was drawn from conceptual designs for the next lunar lander[50]. In this configuration, there are 16 RCS thruster located in four clusters of four around the spacecraft. Two thrusters on each cluster are oriented in the Z-axis and can affect pitch behavior when not overshadowed by the descent engine. The other two thrusters on each cluster stick out at an angle to the X and Y axes and can affect both the roll and yaw angles of the spacecraft which are coupled due to the dynamics and this physical orientation. In order to maintain the proper stable attitude during descent to the lunar surface, the thrusters fire in short bursts in pairs on opposite

sides of the spacecraft. The spacecraft axis convention and thruster configuration are shown below in Figure 7. The thrusters are fixed and cannot be individually pointed. Each thruster can produce approximately 100 lbs of force.

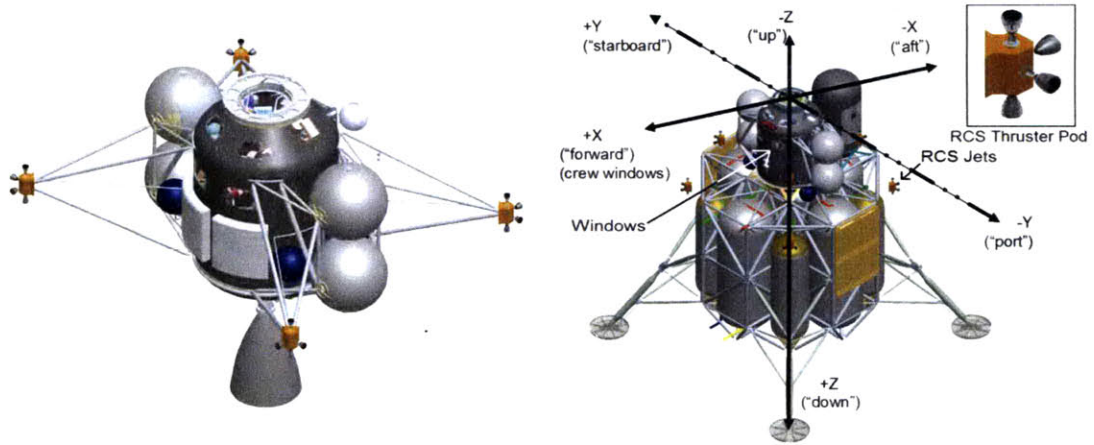


Figure 7: Square configuration of RCS thruster pods around lunar lander (left), full lunar lander with axis definition and close up of RCS Thruster pod (right)

*Fuel System:* The RCS uses hypergolic fuel technology for propulsion, which consists of a liquid fuel and oxidizer that combust on contact with each other, requiring no ignition mechanism and allowing an indefinite storage period as long as they are kept separate. The fuel for the RCS is liquid monomethylhydrazine, designated MMH. The oxidizer is nitrogen tetroxide designated NTO. The thruster system is fed by fuel pumps in order to keep the fuel/oxidizer ratio at the correct level. A conceptual schematic of the RCS that was modeled for the experiment can be shown below in Figure 8.

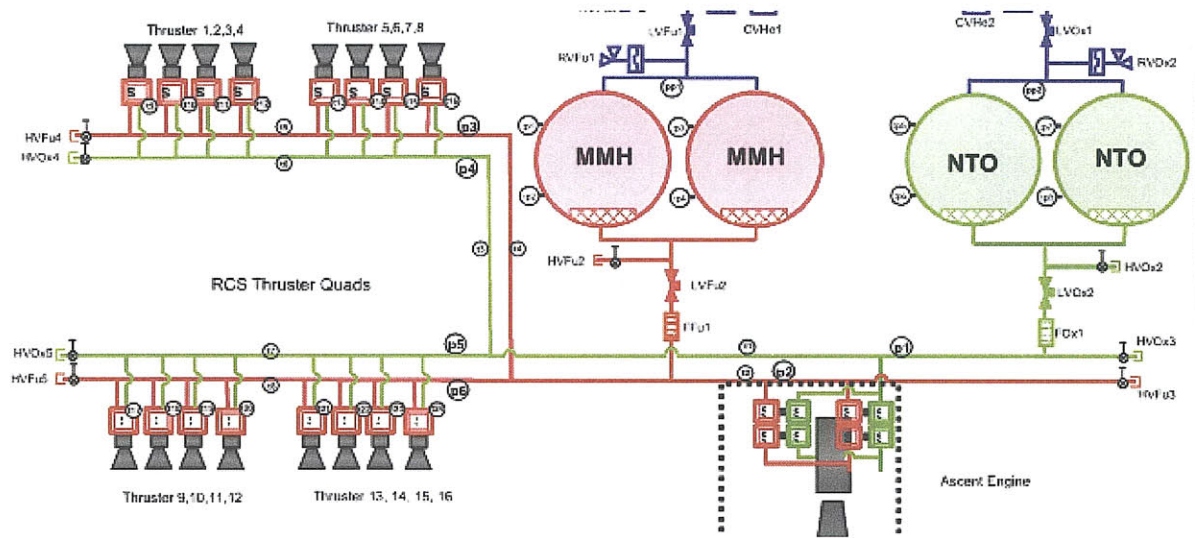


Figure 8: Conceptual schematic of the RCS including fuel tanks, thrusters and fuel pumps

The system is designed to consume approximately 460 kg of total RCS fuel for both the descent to the surface and the return ascent to lunar orbit[50]. Half of this should be remaining at touchdown in order to make a safe return to lunar orbit. Four different failure conditions plus normal operating conditions were presented to subjects during the experiment and are discussed in detail below.

*Fuel pump failure:* This failure would result in the loss of an entire cluster of four thrusters. Each cluster of thrusters is equipped with one fuel pump installed in the fuel line, which branches to feed each of the four thrusters individually. If a fuel pump were to seize or stop working, this would result in fuel starvation for an entire cluster of thrusters, rendering them inoperable. The spacecraft would exhibit deviations in roll and yaw in the direction of the failed cluster of thrusters as other thrusters on opposite sides of the spacecraft continue to fire as planned and are not compensated for by those



thrusters which are now non-operational. The fuel behavior would remain normal with the exception of a few additional burns necessary to correct the deviations in attitude that result from the loss of the four thrusters. With system weight a prime consideration, it is unlikely that there are auxiliary pumps on board, and thus a failed fuel pump would require increased compensation from the other thrusters in order to maintain proper landing attitude.

*Fuel leak:* This failure could result in a mission abort, as fuel is a critical resource for the return to lunar orbit. A leak will appear only in the cockpit fuel gauge. It should not cause any attitude deviation, but will affect the spacecraft's center of mass, which can affect its dynamics. If a leak occurred, the fuel gauges would drop linearly without any commanded burns. Typical fuel system behavior is a step function, with each thruster that fires creating a small step down in the fuel level. Depending on the trajectory, there may be long periods of time where there is no fuel decrease because there are no burns occurring.

*Valve Failure:* This failure mode could result in severe vehicle instability and possibly a loss of the mission. A valve failure is a very rare failure mode that could be caused either by the build up pressure pockets in the fuel lines around a valve or a short circuit in the valve. Either of these conditions, would cause intermittent thruster firings. These intermittent thruster firings would induce a mild oscillation in the roll or yaw directions which would slowly increase in magnitude as guidance attempts to correct

the oscillation. Since the intermittent firings are a random behavior, guidance would always be one step behind in its corrections increasing the magnitude of the oscillation slowly over time. The additional thruster firings and the compensation to damp out the oscillation and return to stable flight would cause fuel to decrease in a step fashion that depended on the number of random thruster firings.

*Thruster Failure:* This failure, modeled as a continually firing thruster, would result in a persistent deviation in roll and yaw that would be compensated for by continual firings from opposing thrusters. This would cause not only attitude deviations but a decrease in fuel similar to a fuel leak since at a minimum the stuck thruster would be continuously burning fuel. This failure was intentionally selected to be similar in its symptoms to a fuel pump failure. Deviations in roll and yaw for these two failures could look almost the same to the human operator. The distinguishing factor between the two failure modes is the fuel behavior. The gradual linear fuel decrease that occurs when a thruster is stuck on does not occur in a fuel pump failure, which burns fuel normally.

All of the data for each of these failures was taken from a high fidelity MATLAB simulation of the descent trajectory of a lunar landing spacecraft. The dynamics were altered in a simplified linear fashion to reflect the given failure mode. All of the failures could be detected and identified using three parameters: roll, yaw, and fuel quantity. RCS failures can have a significant impact on the success of the mission. Depending on

the failure, the consequences of such system failures can be relatively minor such as additional compensation from other thrusters, or could be serious and require immediate landing or abort to preserve the crew and the spacecraft.

### 5.3.2 Task

The experimental task was a supervisory control task, where the subject was monitoring the progress of an autopilot conducting the final approach to landing of a lunar lander spacecraft to a specific landing aimpoint. Each scenario started with the subject being presented a screen as shown below in Figure 9.

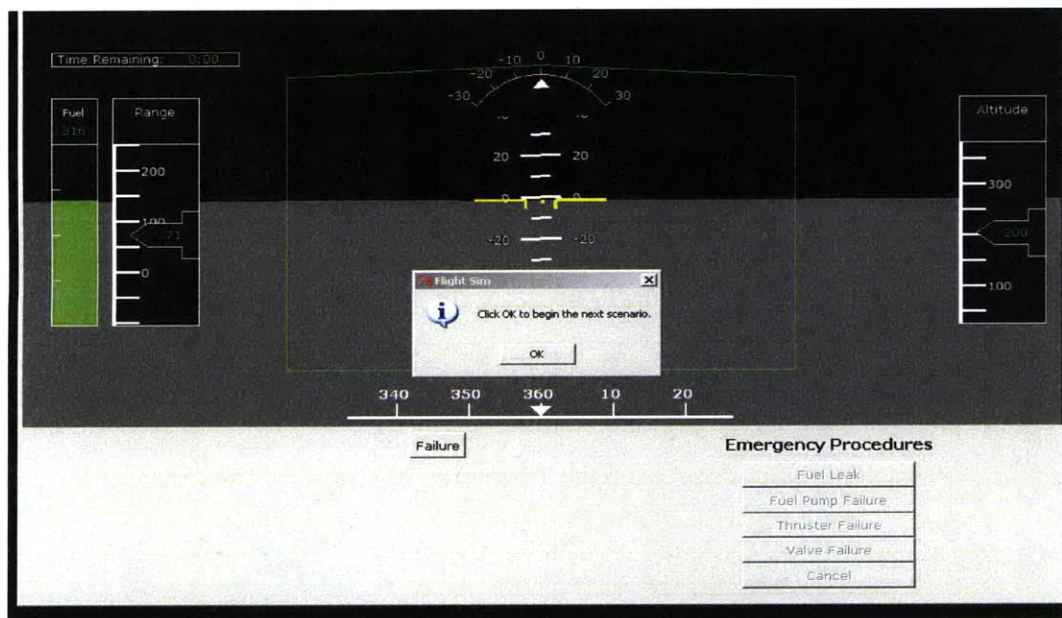
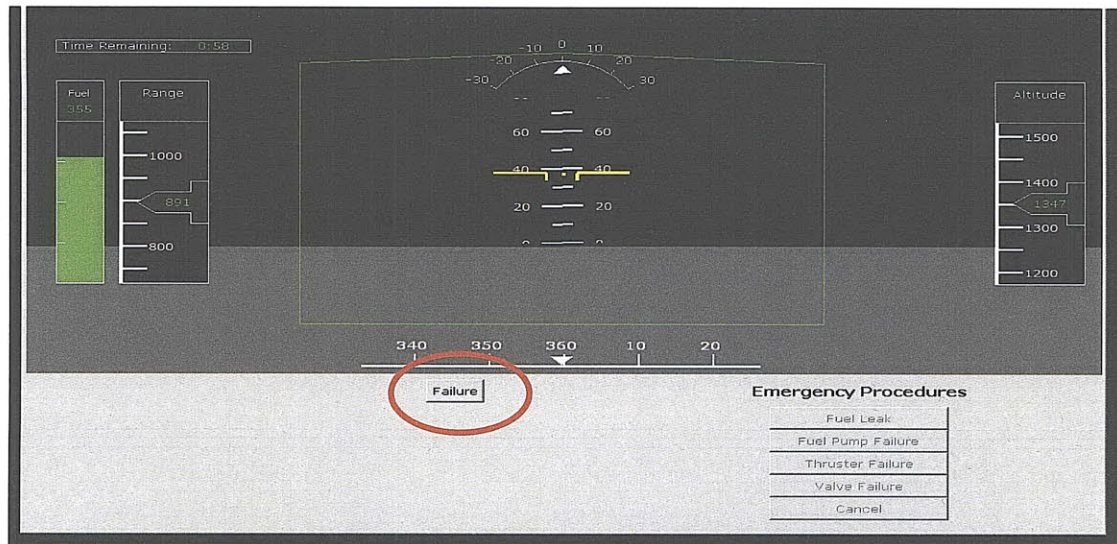


Figure 9: Display presented to subject at the start of each scenario

This screen allowed the subject to start each scenario when they were ready by clicking the “OK” button. The scenario would then begin to dynamically update. The subject’s responsibility was to detect and identify failures that occur in the RCS as described

above. It was also possible for no failure to occur and the system to perform normally. If a failure was detected the subjects were instructed to press the “Failure” button indicated by the red circle in Figure 10.



**Figure 10: Failure button to be selected when subjects detected a failure**

Once the failure was detected, subjects were then asked to diagnose the failure as one of the four predetermined types or to cancel their detection decision if they felt they had made an error. This was done by a series of five buttons on the emergency procedures panel which activated after a the failure button was pressed. The emergency procedures panel is shown in the red oval below in Figure 11.

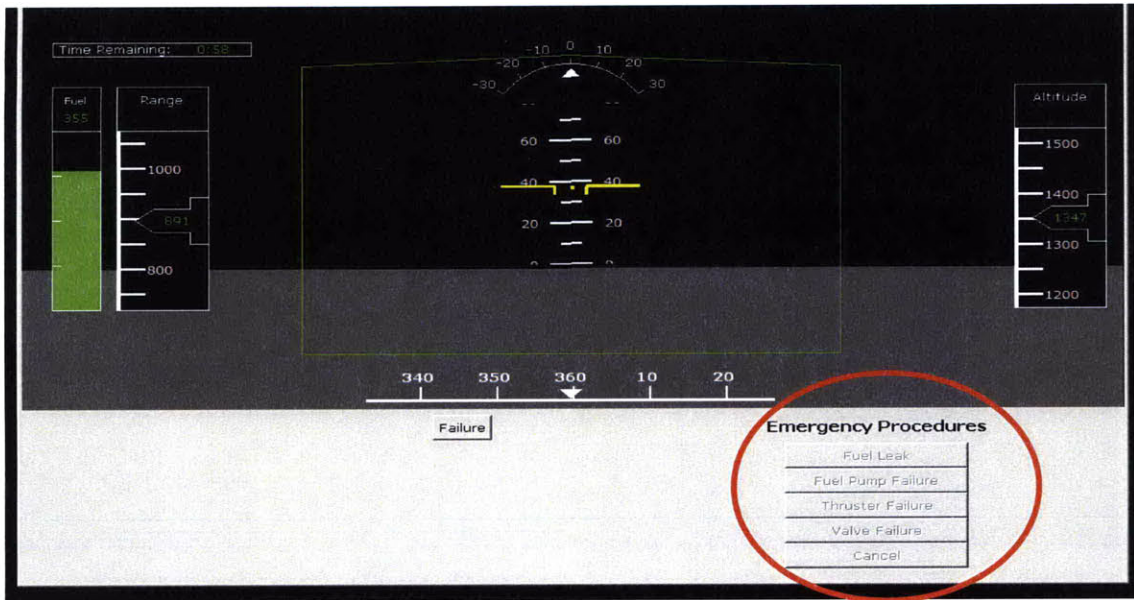
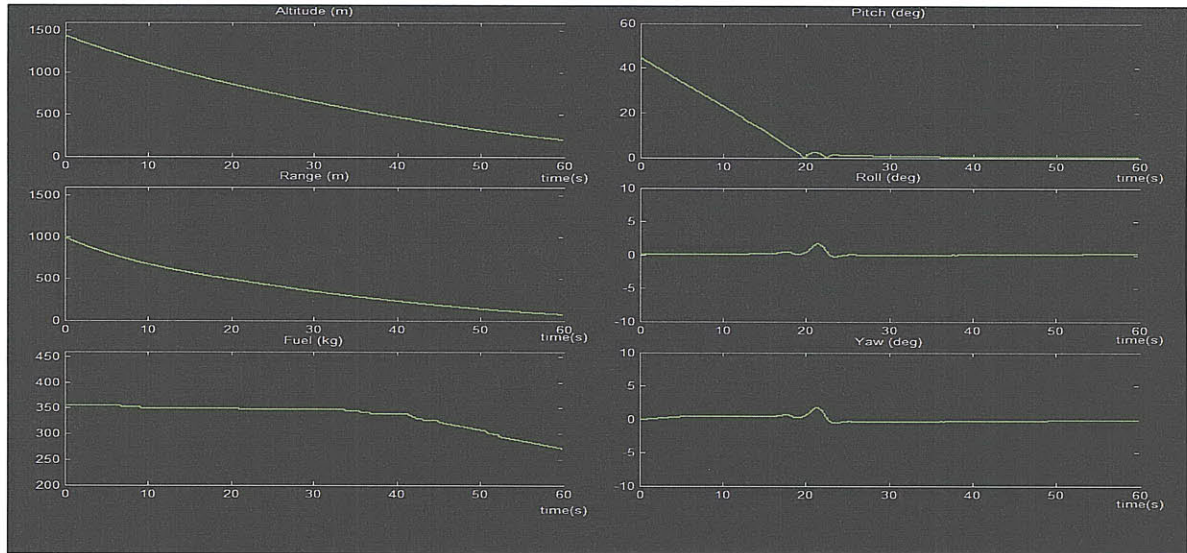


Figure 11: Failure identification panel where a subject selected a button to diagnose a failure after detection

Once the failure was diagnosed, the scenario ended automatically and the next scenario for that particular subject was automatically loaded by the simulation. The subjects were returned to another screen where they could select "OK" when they were ready to begin the next scenario as shown initially in Figure 9 above.

A trend information display shown below in Figure 12, was displayed in half of the scenarios to compare the impact of this additional information on subjects' performance.





**Figure 12: Trend information display shown to each subject during half the scenarios**

Three different trajectories were used to prevent the detection task from becoming trivial through subjects rapidly learning the failure signatures and normal trajectory pattern for a single trajectory over many scenarios. For each trajectory 20 scenarios were developed, 4 of each failure condition plus normal, for a total of 60 scenarios overall. The monitoring task was 60 seconds in duration for each scenario. Scenarios were blocked in groups of 30 based on the two information conditions described below, and the subjects were counterbalanced and randomized as to whether they received one or the other condition first. Each subject saw all 60 scenarios. Within the blocks of 30 scenarios, scenarios were evenly divided between trajectory and failure type such that there were two of each combination of these factors, and the scenarios were randomized within these blocks. Failures occurred randomly between 10 seconds and 50 seconds so that subjects still had enough time to make a decision if the failure occurred late.

Below are short descriptions of the experimental factors:

### Trend Information Conditions:

- Not Present: In this condition the operators was presented with only the primary flight display as a means of detecting and diagnosing system failures. The trend screen was present on the monitor but is blank grey. The primary flight display is shown in Figure 13 below.

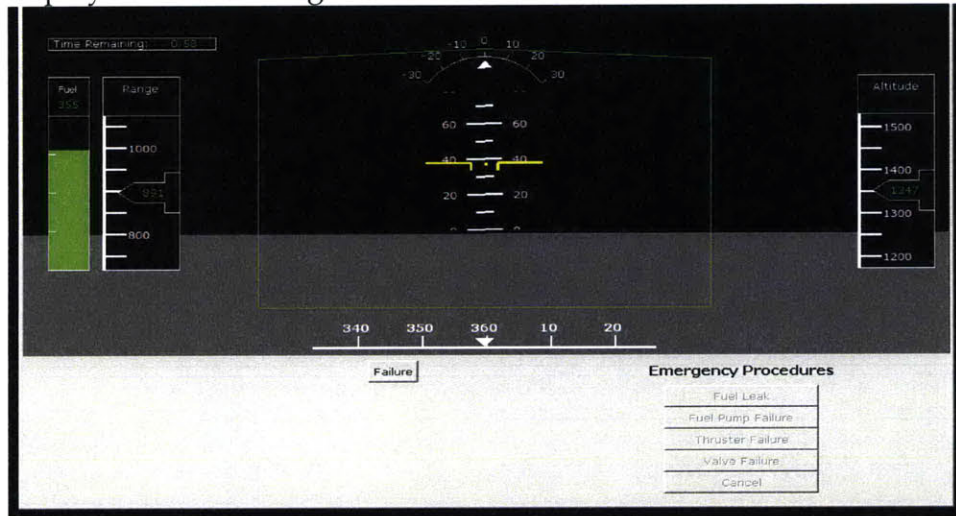


Figure 13: Primary flight display shown to subjects for every scenario

- Present: The trend information was added to the primary flight display on a secondary monitor which the subject may cross reference or utilize at their discretion in order to detect and diagnose failures. The trend information was updated at the same rate as information on the primary flight display. The trend display is shown in Figure 14 below.

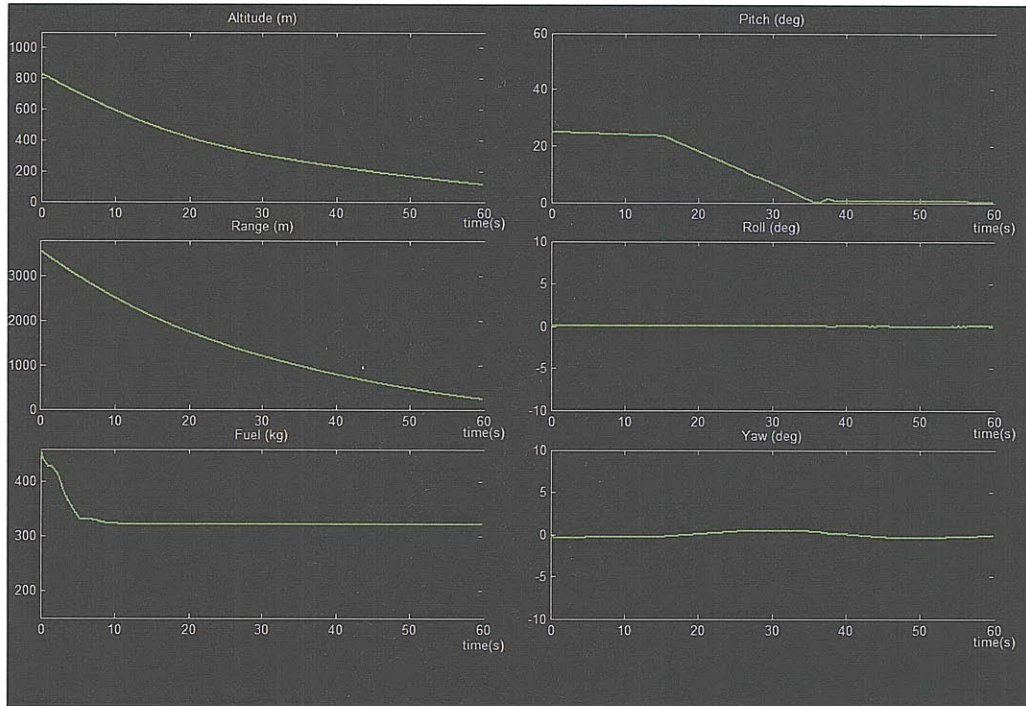


Figure 14: Trend display at scenario completion

### System States:

- Failure: The system encounters one of the four failure conditions described above, with corresponding symptoms displayed on either the primary flight display or both displays depending on the information condition.
- Nominal: The scenario proceeds from start to finish with no failure. This is a control condition to ensure that people are not able to cheat the experiment by instantly declaring a failure at the start. It also attempts to remove subjects' expectancy that a failure will occur in every case which could bias the results.

### Trajectory Types:

- Shallow: The shallow trajectory is a long range, low altitude trajectory which uses more fuel in a large burn at the beginning of each scenario.
- Steep: The steep trajectory is a short range, high altitude trajectory that is much more fuel conservative, but requires an additional large burn at the end in order to maintain stability as it decelerates.



- Nominal: The nominal trajectory has a blend of these characteristics, with a small burn exhibited at the scenario outset, and a larger burn towards the end similar to the steeper trajectory in order to stabilize during deceleration.

#### Independent Variables

- Trend Information Condition
  - o Not Present
  - o Present
- System States
  - o Failure
    - Fuel Leak
    - Fuel Pump Failure
    - Thruster Failure
    - Valve Failure
  - o Normal
- Trajectory
  - o Shallow
  - o Steep
  - o Nominal

#### Dependent Variables

- Detection/Diagnosis Latency
- Detection/Diagnosis Accuracy
- Judgment Consistency
- Lens Model Achievement

Below in Table 1 is the test matrix for all trajectory types. There are ten cases for each trajectory type and one replication of each of these conditions to produce 60 total scenarios.

**Table 1: Experimental test matrix for all three trajectory types**

Not Present	Present
Leak	Leak
Fuel Pump	Fuel Pump

Thruster	Thruster
Valve	Valve
Normal	Normal

### 5.3.1 Subjects

After conducting an a priori sample size calculation for a power of 0.8, a minimum of 24 subjects was required. The total number of subjects tested was 28, which should provide robust statistical results. The subject pool included 19 males and 9 females, ages 21-57, recruited from the Charles Stark Draper community. In order to ensure that no experience bias affected the results, because the display was based on a modern glass cockpit, subjects with little to no flight experience were preferred. Two subjects with flight experience were tested and their data was cross checked and no significant difference was found. Subjects were also asked to report their use of flight simulators and any experience that they had with spacecraft systems.

### 5.4 Procedure

Subjects first completed an informed consent form and a background questionnaire that gathered their demographic information. Next, they completed a 20 minute, computer-based PowerPoint tutorial that outlined the experimental tasks, explained the software interfaces, described the RCS, and covered all possible system states (See Appendix C: Experimental Consent Forms and Training Materials). Subjects were given a cheat sheet that listed the failures and their symptoms in bullet format,

and told that they could take notes on this sheet while they were viewing the tutorial if they so chose.

The subjects then completed a series of five interactive practice sessions (one with each failure type plus normal) in the experimental task environment. The practice sessions were active trainings in which subjects detected and diagnosed a system failure with the experimenter present for feedback and so they could ask any questions they had regarding the interfaces and the information presented to ensure they were comfortable before proceeding. During practice, important functionalities of the interfaces were explained and the subject was given the opportunity to view any of the trials over again if they desired.

Each subject then completed 60 full task scenarios in blocks of 30 with a 5 minute break in between blocks. After the last set of scenarios, subjects took part in a post-experiment interview to gather feedback about the experiment, their judgment strategies, the information they felt was most helpful, and how much learning they felt took place. These interviews were audio recorded for future data processing. The experiment was conducted over a period of 12 days, each session lasting approximately two hours.

#### **5.4.1 Apparatus and Graphical User Interface**

The experiment was conducted at Draper Laboratory. Participants monitored the scenarios on two 17" monitors situated directly next to one another and pressed a

button on the lower portion of the primary flight display when they detected and identified a failure. A standard keyboard and mouse were available on the desk. The room contained one other desk for the experimenter to observe and was otherwise cleaned of all other information. The data was refreshed to the screen at a rate of 25 Hz, displayed to participants using the graphical user interface described in detail in Appendix B. On the screens, the primary flight display was located flush with the upper right corner of the left monitor, and the trend information display was located flush with the upper left corner of the right monitor. This was done so that they would be close enough to scan but also not collocated on the same screen.

The graphical user interfaces (GUI) were implemented in the Python programming language. Each of the GUIs was implemented so that all of the available data and time was recorded in a data file for later analysis. The GUIs included button press functions that allowed the user to start each scenario at their own discretion when a pop up box indicated that they could move to the next scenario. Buttons were also included at the bottom of the display to indicate that a failure had been detected, and then diagnose the failure as one of the four predetermined failure types or cancel their detection decision.

#### **5.4.2 Data Collection**

The values of each parameter presented to the subject along with the time stamp, and decision at each decision point, indicated by a button press, were recorded in a text

file for each subject and each scenario. These parameter values provide the cue data that will be used in the Lens Model for data post processing to determine judgment consistency, achievement, and subject's judgment strategy[5]. Latency is the difference from the time of failure injection, which was fixed for each individual scenario, to the time the subject identified a problem via a button press on the interface. Detection accuracy is based on whether a failure was actually present or not and the subject's response. Diagnosis latency is the time difference between the detection of a failure, and when it is categorized as one of the four possible failure types. This is calculated by the difference in time between the two button presses on the interface. Diagnosis accuracy is determined by which button was pressed, compared to the true failure condition.

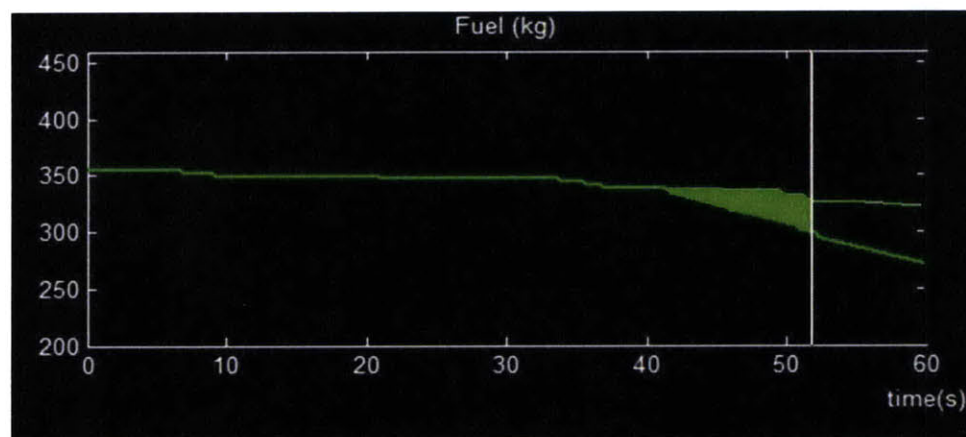
## 6 Data Analysis and Results

### 6.1 Purpose for Analysis

The purpose for the analysis was to investigate the hypotheses regarding the effect of trend information on operator failure detection and diagnosis performance. Performance is defined by operator latency and accuracy. Latency is the time difference between either the introduction of a failure or its detection, and the corresponding detection or diagnosis. Accuracy is determined by whether the subject's decision matched the actual state of the environment. Lens model parameters like decision consistency and judgment achievement, will also be evaluated as performance metrics. Improved performance results in decreased latency and increased accuracy, achievement, and decision consistency.

Another purpose for the analysis was to calibrate the Lens model by using it to post process dynamic data. This involved coding the dynamic trend information into the model. This has not been done previously even in dynamic cases where the Lens model has been applied. In order to represent the dynamic nature of trend information in the model, it was assumed that subjects infer the size of the deviation from normal that occur during a failure from the trend display. It was assumed, and supported by experimental observations of subject behavior, that larger, more dramatic deviations from an operator's normal mental model would typically generate a quicker and more urgent response because they are dramatic and hence appear more serious to the

operator. In order to evaluate this intermediate cognitive processing of the trend information and its impact on subject decisions, the dynamic information for use in the model was reduced to a ratio of the total visual display space which is taken up by any deviations from normal. A figurative representation of this is shown in Figure 11. This type of display was not actually shown to subjects in the experiment.



**Figure 15: The area of a deviation is the integral between the normal trajectory signature (top curve), and the failure case (bottom curve) from the scenario start to the time of decision indicated by the white line.**

Each approach type has a normal signature for each parameter that it follows when no failure is introduced. These normal scenarios are the baseline normal for this calculation. The percentage of the display taken up by the deviation is then determined by taking the integral between the current trajectory and the normal scenario. This area is divided by the area of the entire trend graph axes for that particular parameter to give the percentage of visual area of the deviation. This representation of dynamic trend information strives to capture the impact the deviation had on the operator's response.

## **6.2 Analysis and Results**

The analysis was designed to investigate the effects of trend information on subjects' detection and diagnosis latency and accuracy, as well as Lens model achievement and consistency data. Analysis was conducted using a combination of SPSS 15.0 Statistical Software Package and MATLAB. Other factors that also influenced the results included the approach type, failure type, the order the subject saw the different scenario pairs, and the whether they saw the trend information in the first block or the second block of experimental trials. These different components of the experimental design were evaluated to determine if they had a statistical effect on the results. Each of the steps of this analysis is discussed in detail below for each type of data collected. Additional analysis results in support of the discussion and conclusions can be found in Appendix G.

### **6.2.1 Latency Data**

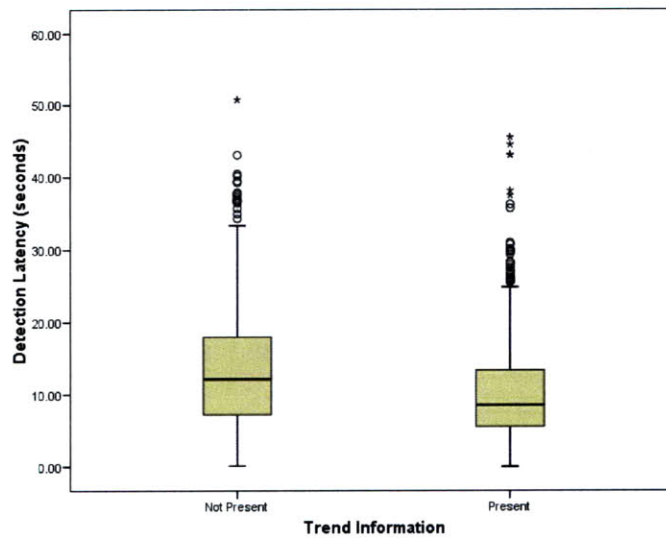
The first dependent variable was latency. Initially, the raw data was examined to predict the results of the statistical analysis. After drawing these initial conclusions it is important to support them with the appropriate statistical tests. Since latency is a continuous variable, parametric statistical techniques are most appropriate. The primary technique was either a Univariate or repeated measures ANOVA (Analysis of Variance), which are special cases of regression. First the data were cleaned so that, outliers and other invalid cases didn't skew the results. Then the data was examined



for normality, homogeneity of variance, and independence. If all these criteria were met, a Univariate ANOVA model with multiple factors was used to determine the significant factors and their interaction effects. If the assumptions were not met, data transforms were applied to help with normality and homogeneity. If the data could not be improved sufficiently, then nonparametric testing was used. If independence was violated, the simpler Univariate model was replaced with a repeated measures ANOVA model. This ANOVA is used when each subject is exposed to multiple levels of an experimental factor, meaning the observations for each experimental factor are not independent since more than one factor was observed by the same subject.

#### *6.2.1.1 Detection Latency*

For detection, all misses, false alarms, and correct rejections were cleaned from the data. Misses are scenarios where a failure was present but it was not detected. False alarms are defined as detecting a failure where none was present. Correct rejections are normal scenarios where no failure was present and the subject never detected a failure. Figure 12 below shows a pair of box plots of latency data broken up by whether it was achieved with or without trend information present. From this plot, there appears to be a significant effect for trend information which appears to decrease the overall mean latency. Significant effects for the type of failure and the approach type were also expected, as these affect the subjects' perception of "normal", which is a critical for developing a mental model to detect deviations from normal system behavior.



**Figure 16: Data suggesting that trend information has an effect on detection latency.**

This raw latency data failed to meet two primary assumptions: normality, homoscedasticity. The cube root transformation was used to give the data both a normal distribution and homoscedasticity. The results from running a Univariate ANOVA on this transformed data are presented in more detail in Appendix G. As expected, history, failure type, and approach type all had highly significant effects ( $p < 0.0005$ ). The experimental block, which indicated whether subjects saw trend information first or second, did not have a significant effect ( $p = .924$ ) and neither did the order in which the subjects saw each pair of scenarios ( $p = .624$ ), indicating that there was no significant learning effects for detection.

### 6.2.1.2 Diagnosis Latency

Any cases where a diagnosis decision was not made were eliminated from the diagnosis data. These were cases where subjects detected a failure but did not make a diagnosis before the scenario time ran out. In examining the raw diagnosis latency data, Figure 11 shows the latency as it was broken up by whether trend information was present or not.

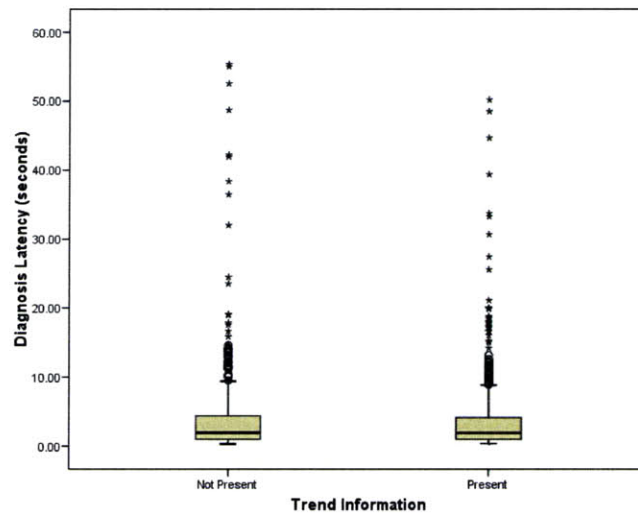


Figure 17: Data suggesting that trend information has no significant effect on diagnosis latency

This box plot of the data does not suggest a significant effect for trend information. Other expected effects include the failure type and approach, which are important since some failures and trajectories were intentionally selected to be easier to diagnose than others, and hence should have a significantly lower latency. The data is homogeneous, but not normal or independent. Therefore it is inappropriate to use a repeated measures ANOVA was used. ANOVA is robust to departures from normality

but not homogeneity of error variance, so no data transformation was necessary. Detailed results are shown in Appendix G. As expected there was no significant effect for the presence of trend information ( $p=0.722$ ). As expected, there was a significant effect for both failure type ( $p<0.0005$ ) and approach ( $p=0.024$ ).

### **6.2.2 Accuracy Data**

Accuracy is determined by whether the subjects' decisions correctly matched the actual state of the environment. Since this decision only has two possible categories, correct or incorrect, accuracy data ends up being binary, and does not meet ANOVA assumptions and so nonparametric testing was used. For simple single factor analysis with more than two levels, like failure type and approach type, the most appropriate test is the Kruskal-Wallis H test, which is essentially a nonparametric ANOVA. However, this test only evaluates one independent variable at a time. It was used primarily to get a sense of what to expect from further more inclusive analysis. In order to create a nonparametric model with multiple factors and interaction effects, binary logistic regression must be used. The regression model was built using a stepwise forward likelihood ratio procedure which checks at each step for variables that add significant effects to the model and adds those variables into the regression model until it converges on a solution that is the optimal predictive model for the given data.

### 6.2.2.1 Detection Accuracy

The inherent trends in this data are best observed by a bar graph of the percent correct versus the presence of trend information as seen in Figure 14 below.

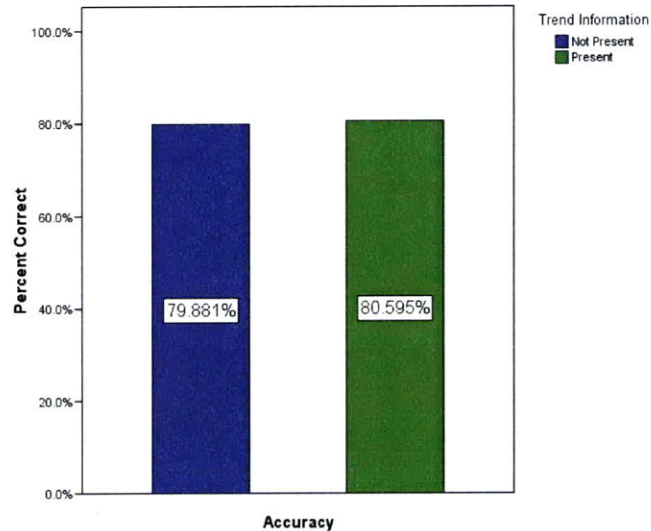


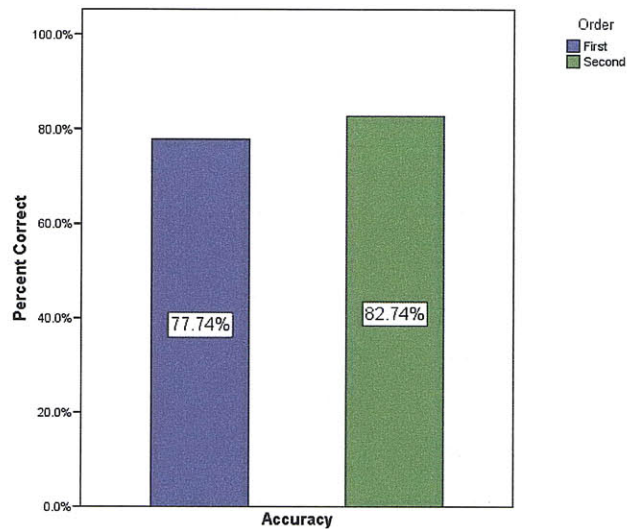
Figure 18: Detection accuracy results showing no significant effect for trend information

Since these percentages are fairly close together whether or not trend information was present, no significant effect on detection accuracy was expected. Significant effects were expected for both failure type and approach type because these factors are important to how subjects differentiate between normal and failure conditions. Some failure types and approach types have specific symptoms or characteristics that make them easier or harder to detect which would give the failure and approach type significant influence. A table of the initial Kruskal-Wallis tests performed is shown below in Table 2.

**Table 2: Factor Effects for Detection Accuracy Data (Significance:  $p < 0.05$ )**

Factor	Trend	Failure Type	Approach	Block	Order
Kruskal-Wallis Significance	0.622	<b>0.000</b>	<b>0.018</b>	0.806	<b>0.010</b>

As expected, trend information did not show a significant effect, while failure type and approach type were both significant. While the block was not significant, the order that participants saw the pairs of scenarios did have a significant effect which suggests that participants learned to detect a certain failure/trajectory combination better with time, i.e. they were more accurate at detection the second time they saw the same conditions, shown in Figure 15.



**Figure 19: Results indicating a significant effect on decision accuracy for order**

The regression model confirmed these Kruskal-Wallis results and also showed which specific trajectories and failure types were significant over the others.

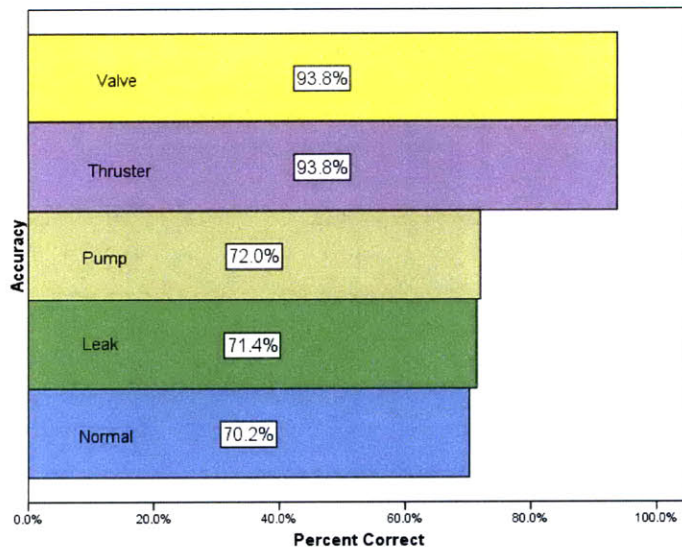


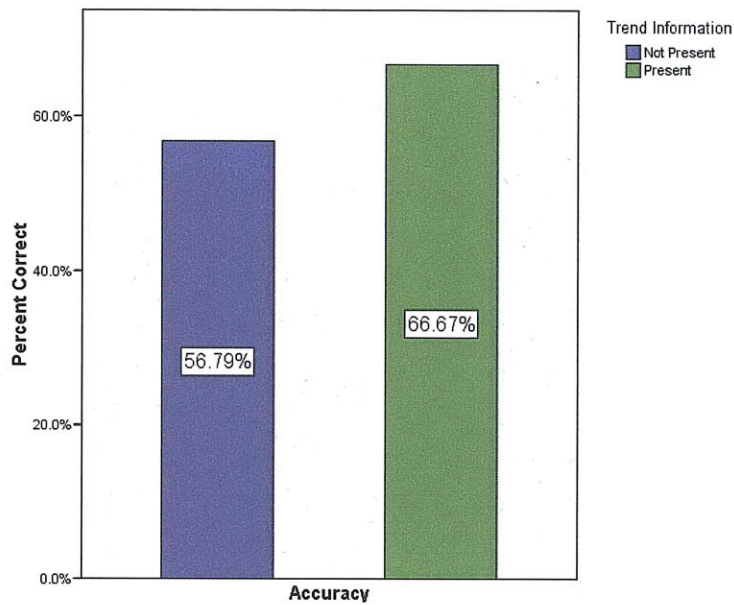
Figure 20: Results showing the effect of failure type on detection accuracy

As for individual failures, thruster and valve failures were significantly easier to detect accurately than the other types of failures. The shallow trajectory appears to have a negative effect on detection accuracy, indicated by a negative coefficient estimate ( $B = -0.489$ ) for this trajectory type. The regression analysis agreed with the Kruskal-Wallis tests that the trend information had no significant effect. The detailed table for binary logistic regression, using a forward stepwise model is presented in Appendix G.

### 6.2.2.2 Diagnosis Accuracy

The diagnosis accuracy data shows a different trend as shown in the bar chart of the percentage correct versus the presence of trend information in Figure 21 below.





**Figure 21: Diagnosis accuracy data showing a significant effect for trend information**

The trend here implies a large difference between trend information conditions. When trends are present, there are almost 10% more correct diagnoses, indicating that trend information has a significant positive effect on diagnosis accuracy. Significant effects for failure type and approach are expected, since some failures are intentionally easier to detect and diagnose than others. A table showing the results of the Kruskal-Wallis tests is shown below in Table 3.

**Table 3: Factor Effects for Diagnosis Accuracy Data (Significance:  $p < 0.05$ )**

Factor	Trend	Failure Type	Approach	Block	Order
Kruskal-Wallis Significance	0.000	0.000	0.092	0.079	0.079

As expected there were significant effects for both trend information and failure type. Both the block and the order that the scenarios were presented were marginally



significant, indicating that for the diagnosis task specifically, there may be some important learning effects over time.

A table of the regression results in Appendix G shows that two failure types, thruster ( $B=-1.804$ ) and pump failures ( $B=-1.101$ ), had negative effects on accuracy, while valve failures were easier to diagnose accurately. This is supported by the bar chart below in Figure 22.

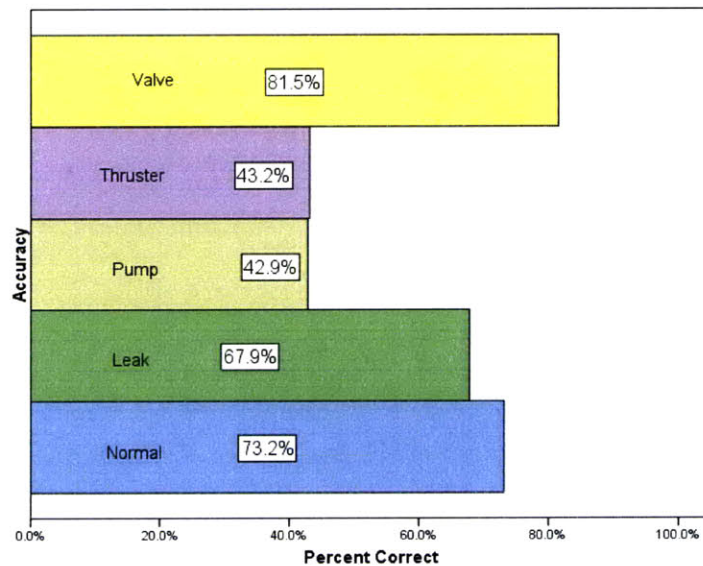


Figure 22: Results showing the effect of failure type on diagnosis accuracy

Pump and thruster failures are very similar in their symptoms and so this drop in accuracy seems plausible. There are also significant two way interactions in the model that suggest that history was significant in helping diagnose some failure types but not others, specifically pump and valve failures.

### 6.3 Lens Model Achievement and Consistency Data

The achievement and consistency values were collected by running each subject's data, with and without trend information separately, through the implemented Lens model code. These two parameters are part of the implementation output along with other Lens model parameters discussed earlier. This data was then evaluated using a repeated measures ANOVA.

### 6.3.1 Detection Achievement and Consistency

The Lens model achievement data is shown below in Figure 23. There is no clear trend which is supported by the lack of statistical significance for detection accuracy.

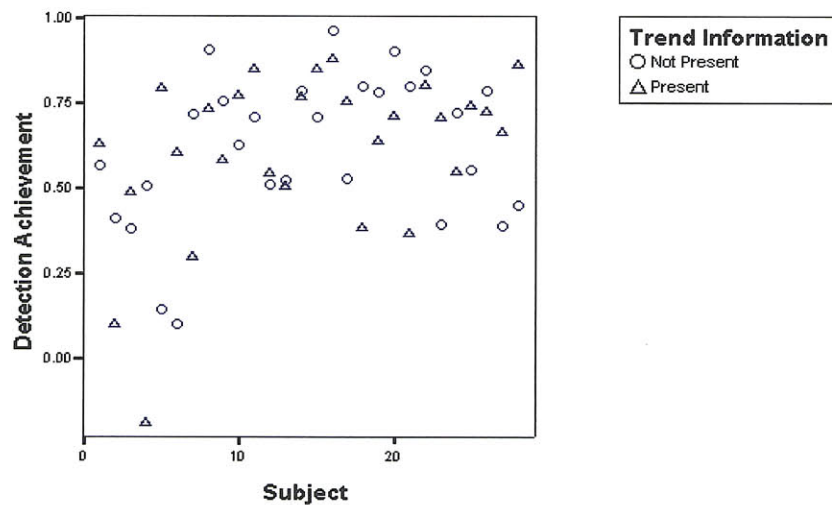


Figure 23: Detection achievement data showing no effect for trend information

Below in Table 4 is the significance of trend information from a repeated measures ANOVA on Lens model achievement, consistency and other parameters.

Table 4: Table showing the effects of trend information on Lens model output parameters for detection (Significance:  $p < 0.05$ )

Parameter	Achievement	Consistency	Predictability	Linear Knowledge	Unmodeled Knowledge	Skill Score

RM ANOVA Significance	0.995	<b>0.001</b>	<b>0.000</b>	0.447	0.531	0.578
--------------------------	-------	--------------	--------------	-------	-------	-------

There was no significant effect on achievement, which is not surprising given that achievement is similar to overall accuracy, and trend information did not make a difference for detection accuracy. Consistency and environmental predictability on the other hand did show significant improvement with the addition of trend information.

### 6.3.2 Diagnosis Achievement and Consistency

Given that trend information had a significant effect on diagnosis accuracy significant effects for both achievement and skill score are expected. Since it also seemed to improve operator judgment consistency for detection, it seems reasonable to expect a significant effect for diagnosis as well which is predicated on detection. Figure 24 shows the achievement data.

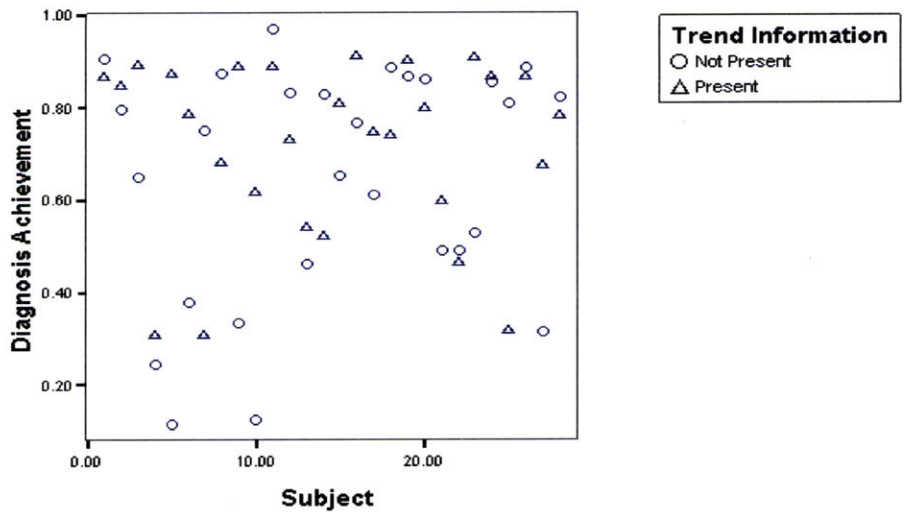


Figure 24: Diagnosis achievement data suggesting a possible effect for trend information

Table 5 shows the ANOVA results for the Lens model parameters for diagnosis.

**Table 5: Table showing the effects of trend information on Lens model output parameters for diagnosis (Significance:  $p < 0.05$ )**

Parameter	Achievement	Consistency	Predictability	Linear Knowledge	Unmodeled Knowledge	Skill Score
RM ANOVA Significance	0.170	<b>0.032</b>	0.054	0.153	0.474	0.177

It was surprising to find no effect on achievement and skill score. For consistency there is still a significant effect, and so it seems that trend information improves a subject's ability to apply the same diagnosis strategy consistently versus no trend information. Environmental predictability was marginally significant, indicating that trend information improves the overall predictability of this decision environment.

#### **6.4 Detection False Alarm Data**

The last question raised from the data concerned the number of false alarms. It is important to evaluate whether trend information lowers the false alarm rate. In order to answer this question, the number of false alarms for each participant with trend information and without was calculated. Graphing these values as shown in Figure 25 below, suggests that there may be a significant effect on the number of false alarms. Trend information actually produced an increase in the number of false alarms for 75% of subjects.

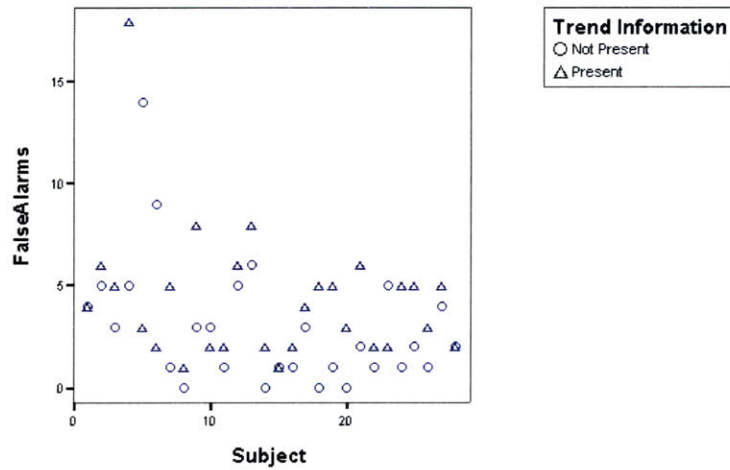


Figure 25: Results suggesting a trend towards increasing numbers of false alarms with trend information

The full results of the ANOVA can be found in Appendix G. The effect of trend information on the absolute number of false alarms was marginally significant ( $p=0.082$ ). Given the overwhelming majority of subjects who experienced a negative trend, this is surprising. With more observations it is possible that this effect could become significant, showing an overall negative impact on the false alarm rate.

## 6.5 Discussion

### 6.5.1 Detection

For detection trend information explicitly displayed to the subjects was helpful in reducing detection latency. It did not make any significant difference on the accuracy of detection or the occurrence of false alarms.

It is worth examining each of the results a little closer to understand what the results might suggest on a larger scale. For detection latency, the original hypothesis

was supported by a significant effect. Plots of the mean latency for each failure and approach type shown below in Figure 26 indicate that trend information decreased detection latency for all failure types for steep trajectories, and all failure types except fuel leaks for both nominal and shallow trajectories. This was surprising for nominal approaches since their fuel signatures are similar to a steep approach.

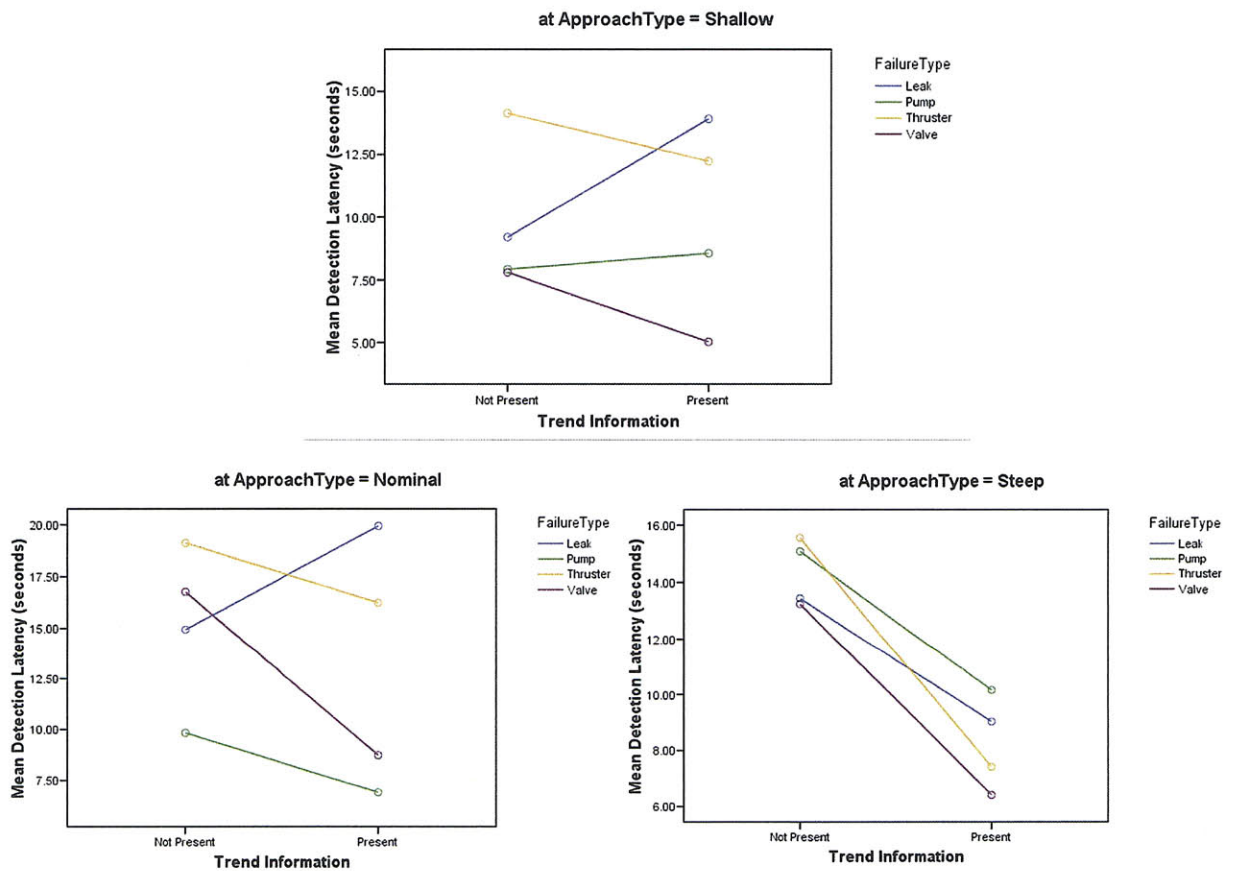


Figure 26: Results showing the combined effects of trend information, approach, and failure type on detection latency

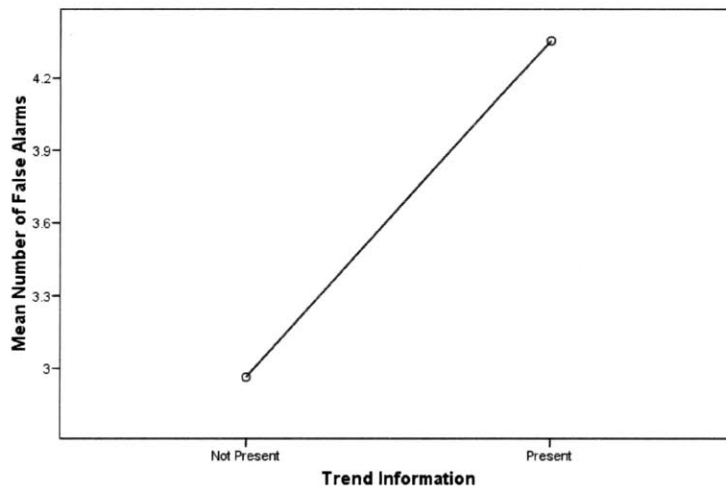
Step and nominal trajectories burn less fuel and have longer periods of time where no burns occur which would make detecting a leak easier. Shallow trajectories,

on the other hand, exhibit a large RCS burn at the very beginning of the section of the trajectory that is being monitored and more burns in general. Although the subjects were trained that this would occur, and reminded of it by the experimenter in practice, this was often falsely detected as a leak early on in the experimental trials as they were learning what constituted normal system behavior. Leaks also appear as more gradual deviations from normal which would have encouraged subjects to watch the historical information longer to confirm their suspicions before detecting the failure.

In terms of detection accuracy, the trends did not have an effect. This indicates that participants can gain an accurate mental model of what “normal” looks like for this task, and apply it, with or without the additional information. It does not improve their ability to see and understand deviations from normal, just the time it takes for that to occur. The deviations in system parameters would have been visible at the same time to the subject whether they were monitoring the trend display or the primary flight display, as long as they recognized the deviation in both cases, the accuracy would not be affected.

For false alarms, though the effect was marginally significant, it is important to look at the direction of that effect. A plot of the mean number of false alarms, shown in Figure 27 suggests that the effect of trend information is negative. Subjects typically had more false alarms with the trend information present than without it.





**Figure 27: Results suggest an increase in the number of false alarms with trend information present**

This is not surprising given that in the trend information condition, subjects are presented with the historical trends of each parameter on the flight display on a second display of equal size. This is twice as much information as they have in the other condition with only the primary flight display. Therefore, it is possible that there are more “trigger” states which could cause them to see something they think isn’t right and declare that a failure has occurred even if that is not true. The additional information could cause them to over analyze the data being presented and second guess their default condition of normal. Also, because the failure rate in this experiment must be so much higher than in a real world system in order to prevent boredom and gather valuable decision data, the subjects are primed to look for a failure condition, even though they know that normal scenarios are present. The importance of this trend must be evaluated in light of the consequences of false alarms versus



missed detections. There is a tradeoff between these two conditions in the sensitivity of any detector. Higher numbers of false alarms can indicate simply that the detector is more sensitive and should result in a corresponding decrease in missed detections which can be equally problematic depending on the system and task at hand.

### **6.5.2 Diagnosis**

It seems that historical trend information explicitly displayed to the subjects was helpful in improving diagnosis accuracy. However, it did not make any significant difference on the diagnosis latency.

The significant effect on diagnosis accuracy is expected because diagnosis is a complex decision amongst several possible outcomes. Additional information that explicitly displays the symptoms of a failure over time should be valuable to an operator in correctly assessing whether or not the failure was of a particular type. Accurate diagnosis requires the human to keep track of failure symptoms, and store some sort of mental record of how the system parameters have evolved over time. The trend display eliminates this extra cognitive effort, decreasing the possibility for error and improving accuracy.

When trying to understand lack of effect on latency, it is helpful to begin by looking at a plot that shows the mean latency with and without trend information for each failure type and approach type shown below in Figure 28.

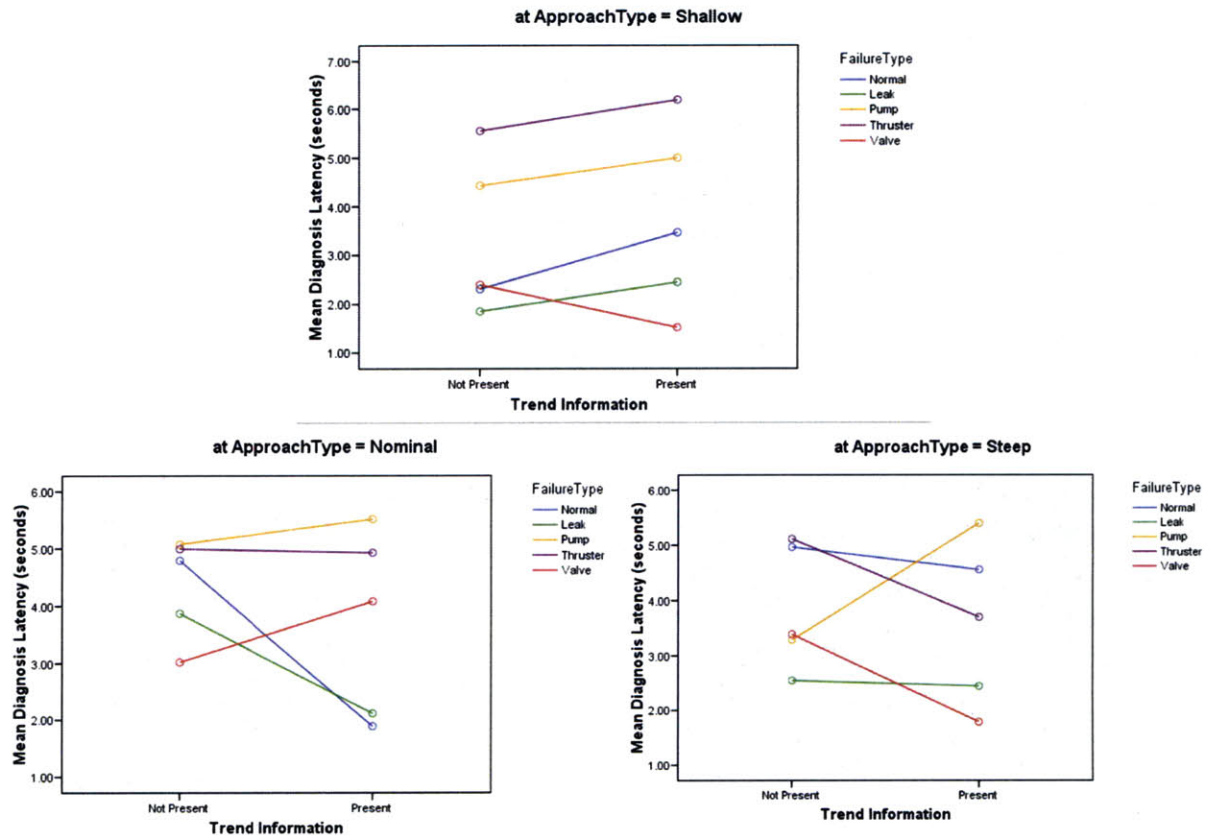


Figure 28: Results of the combined effects of trend information, approach, and failure type on diagnosis latency

Latency appears to get slower when historical information is provided for shallow trajectories and some failure types on other trajectories. The majority of subjects commented in post experiment interviews that they preferred the trend display over the primary flight display because they felt like they better understood the evidence available. It seems logical then that in order to make a more accurate diagnosis, they would use this extra information and observe the system behavior for a longer period of time to be certain before making a decision. However, if they observed it for too long, the scenario time ended without them having made a decision and there

was the risk of the operator getting “tunnel vision” and forgetting about other tasks and constraints like time.

### **6.5.3 Lens Model Parameters**

Lens model parameter data was derived by post processing the cue information and inferred deviation information through the implementation of the Lens model both with and without trend information present. This was done separately for both detection and diagnosis. The two key parameters were achievement and decision consistency, however other parameters that showed significant or marginally significant effects also deserve consideration.

For detection, it is not surprising that there was no effect for achievement which is effectively another measure of decision accuracy, since there was no effect for trend information on detection accuracy. However the significant effects on consistency and environmental predictability make a positive case for trend information. Increasing judgment consistency is important for operator training, and indicates that if operators could be trained to use the most accurate judgment strategy, then they could apply it more consistently with trend information present. This improves system safety and increases the likelihood of mission success.

The data from the Lens model for diagnosis showed significant effects for consistency only. Predictability, linear knowledge, achievement, and skill score, were all marginally significant. The improvements in consistency and predictability are

valuable, if operators can be trained to utilize a more accurate judgment model, they can predict the environment more accurately and more consistently. Since achievement measures how well the human judge does at predicting the environment, this data should have reflected similar results to the diagnosis accuracy data. It was surprising that it was not significant. One possibility for this is that the regression model used in the Lens model didn't have a sufficient amount of data to capture the full effect of the trend information. For detection, because of the nature of the decision, there are a full 30 cases for each subject with and without trend information. This meets the rule of thumb for 5 cases per cue used[33]. For diagnosis, because some cases were detected and never diagnosed, there were often fewer cases and hence did not meet this regression rule of thumb for many of the subjects, potentially yielding a less accurate model.

## 7 Conclusions and Future Work

### 7.1 Conclusions

A representation of the Lens model was selected to investigate the performance impacts of displaying trend information to system operators in failure detection and diagnosis. The experimental results revealed that the addition of trend information has some human performance benefits for failure detection and diagnosis tasks.

While both speed and accuracy can be viewed as important performance metrics, improving one side of this tradeoff while the other remains constant still yields overall performance benefits. For detection, having trend information available yielded faster decisions, but did not make operators more accurate. For diagnosis, the presence of trend information led to slightly longer decision times, but made operators more accurate at diagnosing the problem. These two significant results are important for several reasons. First, because diagnosis is predicated on detection; the faster that a failure can be detected, the sooner diagnosis, and the entire failure resolution process can begin. Detection is the first step in the process, which suggests that faster detection could lead to overall faster failure diagnosis and resolution. In general, the faster a failure can be resolved, the smaller its impact on the overall mission and system performance. Next, accurate diagnosis is critical to successful human failure resolution. If a failure is incorrectly diagnosed, then incorrect procedures for resolving that failure may be executed. This can lead to other emergent system problems and fail to resolve

the failure condition. This could cause the abort of the mission or the loss of crew or equipment. Therefore, it can be argued that the performance improvements in detection latency and diagnosis accuracy which are achieved by explicitly providing operator's trend information about spacecraft system states are valuable results. The addition of trend information displays could improve the performance and safety of spacecraft systems with humans as monitors.

One consideration in adding trend information is that it seems to produce a trend toward higher numbers of false alarms. Although the increase was not statistically significant, it resulted in a higher false alarm rate for 75% of experimental subjects. This trend towards higher false alarm rates may outweigh the human performance benefits in domains such as health care, where false alarm rates are a very important consideration. For many mechanical systems however, such as spacecraft systems, the improvements in detection latency and diagnosis accuracy outweigh the possibility of higher false alarm rates.

Lastly, the Lens model was found to be a useful and generalizable decision making model for human failure detection and diagnosis. The Lens model fulfills all the basic modeling requirements for the project and provides more information about the decision environment and decision strategy than any of the other decision making models from literature that were considered. It was calibrated through human subject

experimentation and expanded its ability to incorporate an operator's perception of dynamic trend information in their decision making process.

## **7.2 Future Work**

The experimental results regarding the impact of trend information on operator failure detection and diagnosis performance could also generate some interesting avenues for future work. The trend towards increased false alarms was surprisingly not significant and these ambiguous results could be further examined to understand why this occurs. One possible explanation is that a higher false alarm rate is the result of an increase in cognitive workload. An increase in the amount of information available, in this case trend information, may increase the operator's workload because they now have more information to monitor and process. This increase in cognitive workload may be the cause of the noticeably higher false alarm rate when trend information is present. Conversely, a higher false alarm rate may also perpetuate a cycle by then increasing the operator's workload as they attempt to diagnose and resolve failures that are not actually present.

The Lens model parameter results were also ambiguous. Parameter results, such as achievement, did not show significance where the human performance counterpart did show statistical significance for trend information. There are several possible explanations for this work, but one of them could become an interesting space for future work. It is likely that this was partially the result of the most valuable cues not

being selected either for presentation to the operator or for inclusion in the model. Another set of cues may have generated a better regression model and may have made the performance impact, specifically on achievement, more clear. It would be valuable to have a framework for selecting representative cues that could provide the best descriptive model of the environment, which would then theoretically provide the best possible set of information to the operator for decision making. This process is currently mostly trial and error and often requires significant resources to conduct multiple experiments and determine the best feasible combination of cues. However, a representative set of cues is critical in order to get an accurate picture of the human decision making performance provided by Lens model parameters.

The Lens model has been used for decades in judgment theory and decision making research and is a feasible and robust model that has recently been extended to dynamic decision making tasks. Future work could utilize this model, when integrated with other models of human performance and system dynamics, to examine system failure, human failure, and combinations of the two. These complex system models could provide insight into the downstream effect on overall system reliability and performance with operator input as components in the system. These complex system models could then be used as early-stage design tools that can aid in the design of intelligent complex systems, by allowing engineers to create hypothetical test cases to



evaluate system tradeoffs that affect both the human operator and the electro-mechanical system configuration.

## **Appendix A: Lens Model Implementation Validation**

The Lens model was implemented in Mathematica's MATLAB 2009a as a generic m-file. This implementation platform will allow it to be plugged more easily into the Simulink modeling component of MATLAB and integrated with existing system design and analysis tools for use in later closed-loop human system modeling.

In order to validate the implementation of the model, a test case was generated based on a paper published by Bisantz A. M., et al. (2000) that involved a Navy aircraft identification task in a dynamic task environment. A secondary validation was to have the actual implementation, mathematical assumptions, and validation results examined by a mathematician and MATLAB expert.

The first validation involves running a test case of known work using the model code and another statistical software tool (SPSS) for comparison of results. Bisantz, et al. (2000) was one of the closest works in literature, using the Lens model, which claims a dynamic task environment. It is valuable to generate similar case data and run the mathematics of the model in both programs to see if the MATLAB model, a widely used statistical software package called SPSS, and the paper results all agree on the model coefficients and correlation outputs.

### **Naval Aircraft Identification Example**

After mathematical validation, a test case from a known work was generated to validate the model code. The goal was to create a synthetic data set that was coded

according to the same logic and task scenario as the data presented in the paper and then compare results using both computer aided tools, and the results presented in the paper[13]. The authors coded the cues in their Lens model based on standardized data values of five different variables that described an aircraft approaching an aircraft carrier: range, radar, IFF, speed, and altitude. The data was coded in the following way[13]:

- Radar: coded as -1,0, or 1 depending on whether the aircraft had a hostile, ambiguous, or friendly radar signature. If the radar was not turned on or not requested by the participant, it was coded as 0.
- Altitude: coded as 0 if the aircraft was flying at 27,000 ft or higher (indicative of commercial or non hostile aircraft) and 1 if the aircraft was flying lower than 27,000 ft (indicative of a military or hostile aircraft)
- IFF: coded as -1,0, or 1 depending on whether the IFF signature was hostile, ambiguous, or friendly. The IFF was coded as 0 if it was not requested by the participant or if the aircraft was identified further than 150 nm out from the aircraft when IFF is unavailable. If the IFF was not turned on and the aircraft was within 150 nm of the ship, it was coded as -1 as all aircraft exhibiting this behavior were hostile.
- Speed: coded as 0 if the speed was more than 600 kn and 1 if the speed was less than 600 kn. Aircraft flying slower than 600 kn were generally commercial airlines and friendly, while aircraft exceeding 600 kn were military with the exception of helicopters.
- Range: coded as 0 if the aircraft was within 150 miles of the ship and 1 if it was further out than 150 miles. This bit of data was only included because if an aircraft was within 150 miles and did not have the radar on then it was considered hostile.

Since there are five cues present, following a general rule of thumb for multiple regression, 50 synthetic data cases were generated, 10 per cue, in order to ensure an accurate representation of the judgment and environment using regression analysis[33]. Once the data creation was complete, a multiple regression analysis was conducted in

SPSS in order to see what the regression coefficients would be. This was then compared to the results presented in the Bisantz et al (2000) paper and the output of the MATLAB model using the same data.

Table 6 shows the beta-weights for the fully correct case using each of the different methods and compared to the values presented in the paper for comparison:

**Table 6: Beta-weights calculated for synthetic validation case and compared to Bisantz, et al. (2000)**

Cue	SPSS	MATLAB model	Bisantz
Altitude	0.450	0.450	0.45
Speed	-0.201	-0.202	-0.2
Range	0.047	0.048	0.05
IFF	0.207	0.207	0.21
Radar	0.353	0.353	0.35

The slight differences in outputs are attributed to rounding errors and also the fact that the data set that was created will not have the same cue values as the actual data from the documented.

## **Appendix B: Graphical User Interface Details**

### **B.1 User Interface Development**

The user interfaces used in experimentation were designed to follow the current display “convention” or standard industry practice for aviation, which is in a sense the field “closest” to spacecraft design. This jump was considered to be acceptable since the astronauts in command of a lunar lander would also be aircraft pilots and these types of displays would be familiar to them. Modern aircraft displays have also already been investigated in numerous human factors experiments in literature. Spacecraft displays are often modification of aircraft displays to begin with.

The displays were designed to be as simple a representation of the critical system information as possible. Two user interfaces, seen below in Figure 29: Primary Flight Display shown to all subjects for each trial and Figure 30, were implemented: a primary flight display and a display of system parameter trend information.

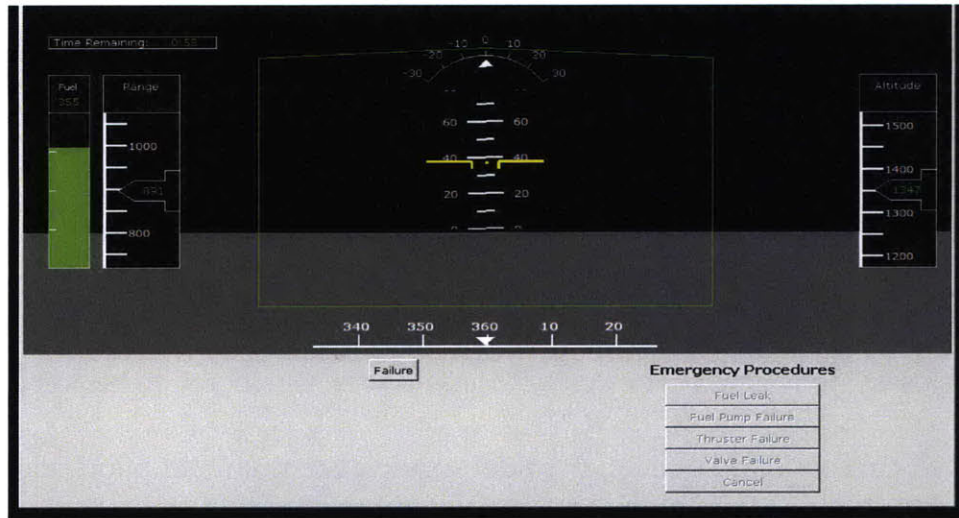


Figure 29: Primary Flight Display shown to all subjects for each trial

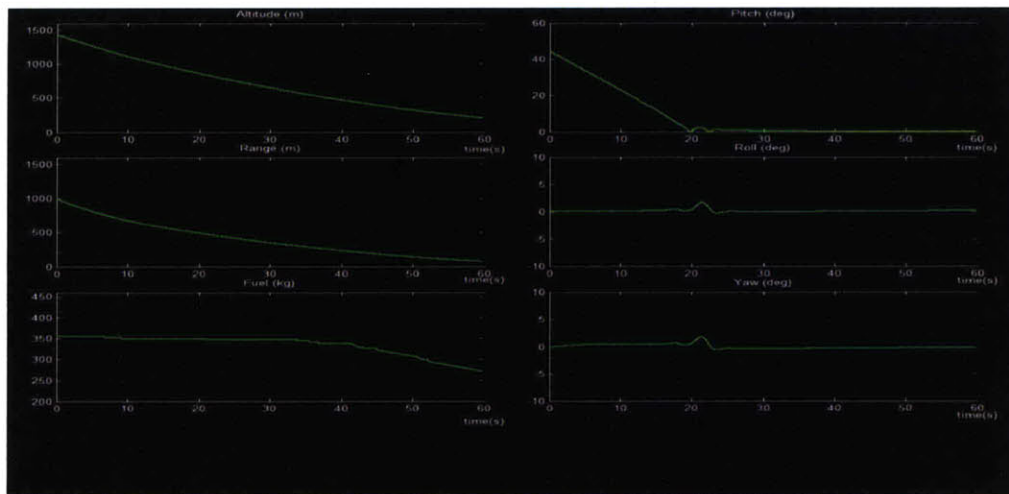


Figure 30: Trend Information display with same system parameters represented, shown to subjects for half of the trials

In order to truly test the effect of the additional benefit of trend information, all information was represented equally on both displays. If there was a question of which display had better resolution, the primary flight display was always favored to have higher resolution since this was the current conventional display and the effect of

display type or display resolution was not the research question. Green was selected as the primary color on the display in order to provide a good visible contrast on the black and grey background and also to avoid using colors such as red which are traditionally used to indicate warnings and problems. The font used in all the displays is Verdana 9 pt. This was selected based on its recommendation in the Federal Aviation Administration's Human Factors Standard as the best "readable" font for digital displays.

## **B.2 Information Displayed**

The information presented on the displays is dynamic and changes throughout the scenarios according to data generated from a high fidelity lunar landing simulation which was modified in a simplified linear fashion to reflect specific failures. There are six system state parameters represented on both the primary flight display and the trend information display which are: Altitude, Range, Pitch, Roll, Yaw, and Fuel Quantity.

The purpose of the experiment is to investigate the use of dynamic trend information in human decision making strategies. It is important here to make a distinction between implicitly and explicitly displayed trend information. Implicitly displayed information is not directly given to the operator as a numeric value but instead is inferred as a result of watching the instantaneous values of a dynamic system parameter vary over time. This information is said to be "achieved" since it is mentally

derived from the value of a separate parameter which is continually changing. This achieved information must be saved in the operators working memory in order to create the trend of what has happened in the system in the past. Explicitly displayed information is information which is given to the operator visually. The trend behavior is visually available to the pilot rather than inferred by watching the altimeter update its value continuously. Let's use the altitude as an example. The altimeter provides explicit information about the current value of the spacecraft's altitude. It also provides implicit rate information because as the spacecraft descends, the pilot can watch the values on the altimeter change and infer whether he is descending slowly, rapidly, or holding a steady altitude based on this observation. Exactly how fast he is descending would require a more exacting mathematical calculation, however implicit rate information may still be useful in his understanding of the current state of the world. Each of these pieces of information must be stored in the pilot's short term memory in the order that they made the observations for the pilot to be able to recall that they were approximately 200 meters higher approximately 4 seconds ago. If the operator is provided with explicit trend information, this would be akin to having a profile view of the current descent profile of the spacecraft. In the context of this experiment, one display contains explicit trend information and one does not. Implicit trend information is always available to the operator.

### **B.3 Variable Definitions**



Altitude: Altitude above the lunar surface is altitude relative to the surface in the positive Z-direction (surface=0 meters), determined by a radar altimeter. The unit of this measurement is meters. Altitude ranges from maximum values of approximately 1100 m for shallow trajectories to approximately 1700 m for steep trajectories.

Range: Range is the horizontal or ground distance from the spacecrafts current location to the designated landing point in the positive X-direction. The designated landing point is the reference point at 0 meters. The unit of this measurement is meters. Shallower trajectories have longer ranges of 3800 m, while steeper trajectories have much shorter ranges of 1600 m.

Pitch: Pitch is the measurement of backward-forward rotation around the Y-axis of the spacecraft. It is measured in degrees with positive (increasing) pitch being indicated to the pilot as backwards motion in the negative X-direction. During the spacecraft's pitch over maneuver the spacecraft is essentially going from 90 degrees of pitch, on its back to 0 degrees of pitch in preparation for the terminal descent to the surface.

Roll: Roll is the measurement of rocking motion about the X-axis of the spacecraft. It is measured in degrees with positive (increasing) roll being indicated to the pilot as rocking to the right, in the positive Y-direction. In general roll should remain approximately 0 and less than 2 degrees with the exception of one large roll

towards the end of the trajectory in order to give the crew a visual view of the designated landing point.

Yaw: This is essentially a “heading” measurement for the spacecraft. Yaw is the measurement of left and right motion around the Z-axis. It is measured in degrees with positive (increasing) yaw being indicated to the pilot as a swing to the right in the positive Y-direction. In general roll should remain approximately 0 and less than 2 degrees with the exception of a large yaw motion coupled with the large roll in order to view the designated landing site.

Fuel: This is indicated in the cockpit as fuel remaining and is a measure of the weight amount of fuel remaining in the RCS fuel tanks in kilograms. The lander begins its descent with a total of 460 kg of RCS fuel and requires at least half of this for the return journey. Fuel should decrease in small and possibly irregular steps depending on when the spacecraft guidance computer requires burns from different sets of thrusters. Each thruster uses approximately a quarter kg of fuel for a normal firing to correct minor attitude errors.

## **Appendix C: Experimental Consent Forms and Training Materials**

### **CONSENT TO PARTICIPATE IN NON-BIOMEDICAL RESEARCH**

#### **The Effect of System Parameter Time History on Human Judgment Performance**

You are asked to participate in a research study conducted by Professor John Hansman, Ph.D, and Rachel Owen from the Department of Aeronautics and Astronautics at the Massachusetts Institute of Technology (M.I.T.) The results of this study will be included in Rachel Owen's Master's Thesis. You were selected as a possible participant in this study because of your access to Draper Laboratory and your level of experience in the experimental domains. You should read the information below, and ask questions about anything you do not understand, before deciding whether or not to participate.

- **PARTICIPATION AND WITHDRAWAL**

Your participation in this study is completely voluntary and you are free to choose whether to be in it or not. If you choose to be in this study, you may subsequently withdraw from it at any time without penalty or consequences of any kind. The investigator may withdraw you from this research if circumstances arise which warrant doing so.

- **PURPOSE OF THE STUDY**

This study is designed to investigate how humans use explicitly displayed time history information about system parameters in order to help them detect and identify a failure in a system. It is critical that human operators be able to quickly and accurately detect system failures in complex systems for safety and reliability. This research intends to explore human judgment strategies and the effect of specific information on the accuracy, timeliness and consistency of those judgment strategies.

- **PROCEDURES**

If you volunteer to participate in this study, we would ask you to do the following things:

You will start by completing an informed consent form and a background questionnaire that gathers basic demographic information. Next, you will complete a computer-based

PowerPoint tutorial that outlines the experimental tasks, and explains the interfaces you will be using. This tutorial should take approximately 15-20 minutes.

Next you will complete a short interactive practice session in the using the computer software to ensure you are familiar with the protocol and the displays. The practice session will be an active training in which you will be asked to detect a system failure with the experimenter present so you can ask any questions they have regarding the interfaces and relationships among the information presented.

For the experimental trial, you will be asked to monitor an automated landing approach in a lunar lander simulator. Half of the trials will contain time history information and half will not. You should use all the information available to you to make a judgment about whether or not the system is operating correctly. You will be asked to press a specific button on the interface to indicate when you detect any kind of system failure. You will then do your best to identify the failure out of four possible failures you learned during training. These two tasks must be completed as quickly as possible.

You will then complete 60 full experimental task scenarios. An experimental scenario lasts approximately 1 minute. There will be a 5-10 minute break after the first 30 scenarios. You may take additional breaks as you feel necessary. After the last set of scenarios, you will take part in a post-experiment interview in order to gather feedback about the experiment, your judgment strategies, and the task scenarios which will last approximately 15 minutes.

This study will last approximately 2 hours depending on how quickly you choose to go through the scenarios.

- POTENTIAL RISKS AND DISCOMFORTS

There are no anticipated physical or psychological risks in this study. There is the potential discomfort of dry eyes due to your being asked to monitor a computer screen for an extended period of time. You may take breaks whenever you feel it necessary.

- POTENTIAL BENEFITS

While there is no immediate foreseeable benefit to you as a participant in the study, your efforts will help provide insight into human judgment strategies, and how information can be used by operators and monitors of complex systems in detecting system failures.

- PAYMENT FOR PARTICIPATION

No financial payment will be given for your participation. You will not receive a charge number but you will receive two free AMC movie tickets.

- CONFIDENTIALITY

Any information that is obtained in connection with this study and that can be identified with you will remain confidential and will be disclosed only with your permission or as required by law.

You will be assigned a subject number which will be used on all related documentation to include databases, summaries, and results. Only one separate master list of subject names and numbers will exist, password protected on the internal server at Draper Laboratory accessible only to members of the research team.

There is a post-experiment interview which will require a separate consent form. Please read and sign the attached document entitled "Consent to Participate in an Interview".

- IDENTIFICATION OF INVESTIGATORS

If you have any questions or concerns about the research, please feel free to contact the student investigator, Rachel Owen, by telephone at (617) 258-4494 or via email at [rlowen@draper.com](mailto:rlowen@draper.com). Her MIT faculty sponsor is Professor John Hansman who may be contacted at (617) 253-2271, email, [rjhans@mit.edu](mailto:rjhans@mit.edu), and his address is 77 Massachusetts Avenue, Room 33-303, Cambridge, MA 02139

- EMERGENCY CARE AND COMPENSATION FOR INJURY

If you feel you have suffered an injury, which may include emotional trauma, as a result of participating in this study, please contact the person in charge of the study as soon as possible.

In the event you suffer such an injury, M.I.T. may provide itself, or arrange for the provision of, emergency transport or medical treatment, including emergency treatment and follow-up care, as needed, or reimbursement for such medical services. M.I.T. does not provide any other form of compensation for injury. In any case, neither the offer to provide medical assistance, nor the actual provision of medical services shall be considered an admission of fault or acceptance of liability. Questions regarding this policy may be directed to MIT's Insurance Office, (617) 253-2823. Your insurance carrier may be billed for the cost of emergency transport or medical treatment, if such services are determined not to be directly related to your participation in this study.

- RIGHTS OF RESEARCH SUBJECTS

You are not waiving any legal claims, rights or remedies because of your participation in this research study. If you feel you have been treated unfairly, or you have questions regarding your rights as a research subject, you may contact the Chairman of the Committee on the Use of Humans as Experimental Subjects, M.I.T., Room E25-143B, 77 Massachusetts Ave, Cambridge, MA 02139, phone 1-617-253 6787.

SIGNATURE OF RESEARCH SUBJECT OR LEGAL REPRESENTATIVE
---

I understand the procedures described above. My questions have been answered to my satisfaction, and I agree to participate in this study. I have been given a copy of this form.

\_\_\_\_\_  
Name of Subject

\_\_\_\_\_  
Name of Legal Representative (if applicable)

\_\_\_\_\_  
Signature of Subject or Legal Representative

\_\_\_\_\_  
Date

SIGNATURE OF INVESTIGATOR
---------------------------

In my judgment the subject is voluntarily and knowingly giving informed consent and possesses the legal capacity to give informed consent to participate in this research study.

\_\_\_\_\_  
Signature of Investigator

\_\_\_\_\_  
Date

## CONSENT TO PARTICIPATE IN INTERVIEW

### **The Effect of Rate Information Display on Human Judgment Performance**

You have been asked to participate in a research study conducted by Rachel Owen from the Department of Aeronautics and Astronautics at the Massachusetts Institute of Technology (M.I.T.). The purpose of the study is to understand the effect of rate of change information on the humans ability to make accurate and timely decisions. The results of this study will be included in Rachel Owen's Master's thesis. You were selected as a possible participant in this study because of your access to Draper Laboratory and your level of experience in the experimental domains. You should read the information below, and ask questions about anything you do not understand, before deciding whether or not to participate.

- This interview is voluntary. You have the right not to answer any question, and to stop the interview at any time or for any reason. We expect that the interview will take about 15 minutes.
- You will not be compensated financially for this interview.
- Unless you give us permission to use your name, title, and / or quote you in any publications that may result from this research, the information you tell us will be confidential.
- We would like to record this interview on audio cassette so that we can use it for reference while proceeding with this study. We will not record this interview without your permission. If you do grant permission for this conversation to be recorded on cassette, you have the right to revoke recording permission and/or end the interview at any time.

This project will be completed by 30 June 2010. All interview recordings will be stored in a secure work space until 1 year after that date. The tapes will then be destroyed.

I understand the procedures described above. My questions have been answered to my satisfaction, and I agree to participate in this study. I have been given a copy of this form.

*(Please check all that apply)*

[ ] I give permission for this interview to be recorded on audio cassette.

I give permission for the following information to be included in publications resulting from this study:

my name    my title    direct quotes from this interview

Name of Subject \_\_\_\_\_

Signature of Subject \_\_\_\_\_ Date \_\_\_\_\_

Signature of Investigator \_\_\_\_\_ Date \_\_\_\_\_

Please contact Rachel Owen at [rlowen@draper.com](mailto:rlowen@draper.com) or (617) 258-4944 with any questions or concerns.

If you feel you have been treated unfairly, or you have questions regarding your rights as a research subject, you may contact the Chairman of the Committee on the Use of Humans as Experimental Subjects, M.I.T., Room E25-143b, 77 Massachusetts Ave, Cambridge, MA 02139, phone 1-617-253-6787



## Interview Materials:

The purpose of this interview is to thoroughly understand the decision making process and information used to make that decision. We will randomly choose one scenario from each of the experimental conditions. All questions posed by the experimenter will be in quotations and all other actions will be italicized. After the subject responds, follow on questions may be needed to clarify a subject's response; these questions cannot be predicted, but their content will only be about the specific decision making process or display content. All experimenter and subject responses will be recorded.

*Start the audio recording. Experimenter will state subject ID for coding responses.*

*"For this part of the experiment, we're going to select a subset of scenario from the experiment and discuss your response and how you made your decision. We're also going to discuss your response to the experiment and any external factors you thought influenced your behavior. Before we begin it would be appropriate to let you know that these responses are confidential and it would help us greatly if you would be as honest as possible in order to help us understand and improve our judgment models." Show subjects the scenario card and bring up the interface and simulation for that scenario.*

*Walk me through a scenario and what you were looking at, and thinking about as you made your decisions. (24, 19,38)*

- Can you tell me why you made the decisions you made (remind them what their decision was, failure, no failure and where the failure occurred)?
- On a scale from 1-10, where 1=no clue and 10=absolutely certain, how confident were you in your detection of a failure? Why?
- On a scale from 1-10, where 1=no clue and 10=absolutely certain, how confident were you in your identification of that failure? Why?
- Which of the displayed information did you find most useful in making your detection decision? Diagnosis?
- It seems like your general decision strategy was: \_\_\_\_\_(in non technical terms the experimenter will summarize the decision strategy that the subject has expressed). Would you agree?
- What difficulties did you face in the decision making process?
- Was there any other information you would have found particularly helpful in aiding your decision?

*The following questions will be asked of each participant after the scenarios are discussed.*

- How did you feel in terms of your overall accuracy, were you confident in your decisions? Did you feel you had enough information to make the decision?
- Did you find the task to get easier over time? (learning)
- Did you find that you preferred one type of information over another for your decision making?
- How did you cope with failure cases that looked very similar in both display cases?
- At what point during the experiment did you feel comfortable with what was considered “normal”?
- Did you feel like there were any limitations that prevented you from performing your best?

Reminder: Please no discussion of any classified material in your responses.

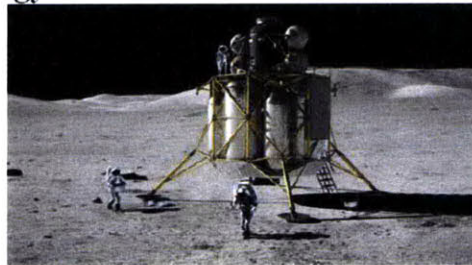
Training Materials:

# Failure Detection and Identification in the Lunar Lander

Rachel Owen  
C.S. Draper Laboratory  
3/1/10

## Experimental Domain

- Domain: Lunar Lander
- The experiment is investigating failure detection and identification while monitoring a lunar landing under automatic guidance and control technology





## Your Task

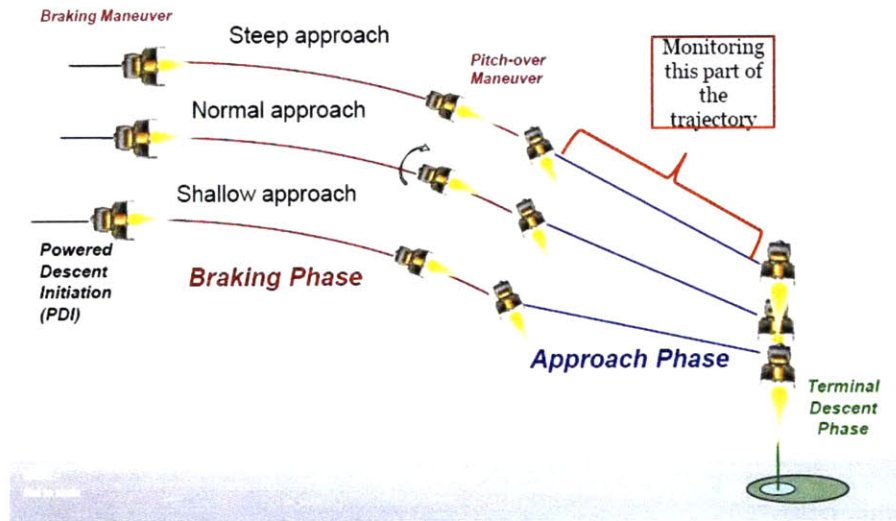
You are an astronaut aboard a NASA mission to the lunar surface. The spacecraft lands autonomously using the auto pilot and guidance computer. Your task is to monitor a portion of the descent trajectory, alert the rest of the crew, and initiate emergency procedures in the event that there is a system failure at this point in the mission.



## Task Details

- You will be monitoring a one-minute segment of the spacecraft's approach phase as it nears its final descent to the lunar surface.
- The lunar lander can approach the surface in any one of 3 ways
  - Steep Trajectory
  - Nominal Trajectory
  - Shallow Trajectory
- You will see all three trajectory types during the experiment

## Spacecraft Approaches



## Trajectory Details

- The pitch-over maneuver indicated on the diagram is a steady pitch transition from a pitch of approximately 60 degrees to an upright orientation which is indicated by a pitch of approximately 0 degrees.
- This is the orientation the spacecraft needs to descend to the surface.
- There should be very little yaw and roll motion throughout the trajectory other than normal disturbances due to the environment.
  - = Disturbances greater than 2 degrees may indicate a system condition



## Possible System Failures

- Failures may occur in the spacecraft's fuel and reaction control system (RCS) within the last critical minute of the approach phase. It is this system and its affects on the spacecraft that you will be interested in observing.
- The reaction control system is a set of thrusters positioned around the spacecraft that produce small impulses and control its orientation or attitude in space. It is fueled using a liquid methane and nitrogen tetroxide mixture.
- The RCS system has 460 kg of total fuel. Only 230 kg of this is allotted for the descent portion of the mission.



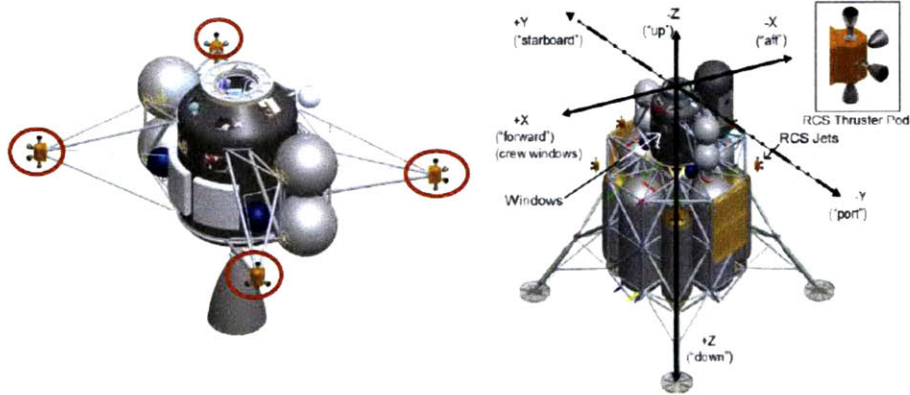
## Possible RCS Failure Cases

- Two failures will be associated with the RCS fuel system, and two with the actual thrusters themselves
- **Fuel System Failures:**
  - Fuel Leak
  - Fuel Pump Failure
- **Thruster System Failures:**
  - Thruster Failure-Stuck On
  - Valve Failure
- Now to look at the RCS system configuration....



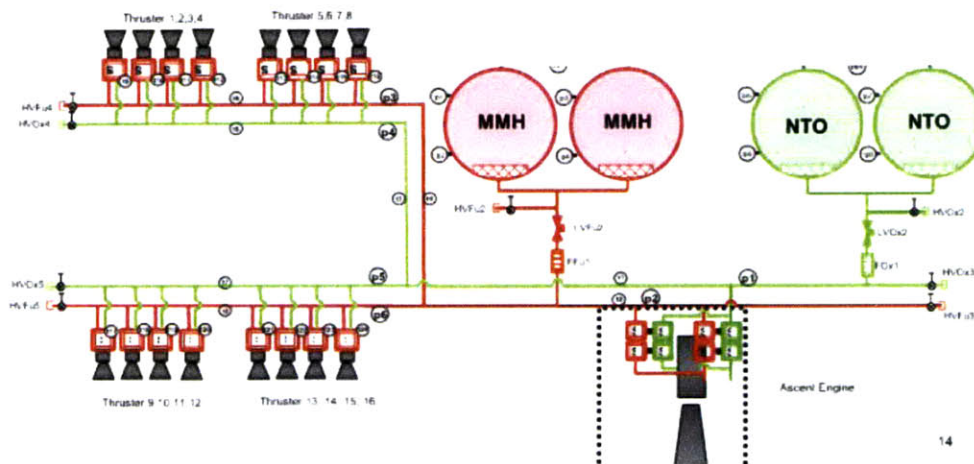
## Physical Thruster Configuration

- The spacecraft has 4 clusters of 4 RCS thrusters positioned in a square configuration
- Thrusters are fixed and cannot tilt individually
- Failures can affect a single thruster or an entire cluster

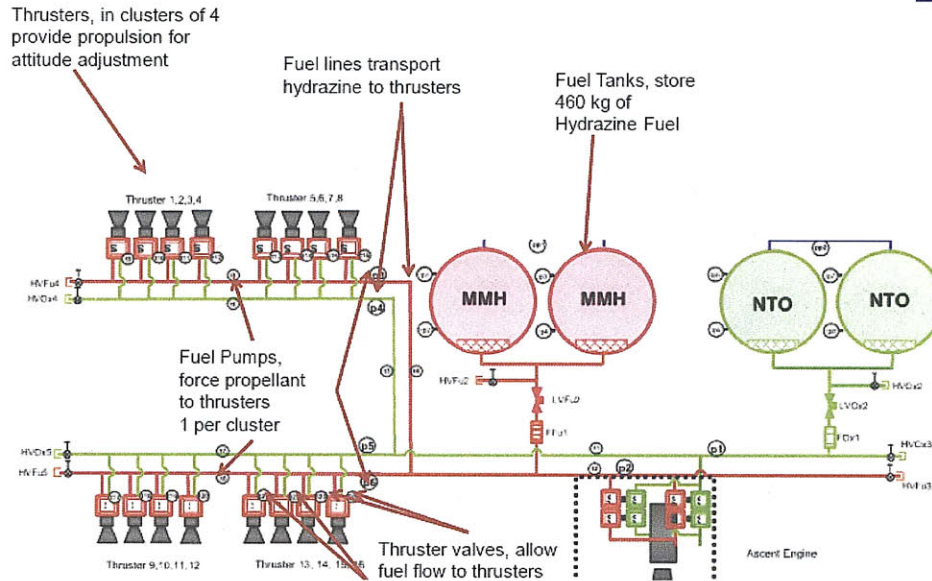


## RCS Propulsion Schematic

This system is mission critical for both the spacecraft's descent to the surface and the return to lunar orbit



## Key System Components



## Typical System Behavior

- Thrusters are aligned around the spacecraft to ensure that proper angles and attitude can be maintained. They typically fire on and off infrequently or at long intervals to correct spacecraft drift.
- The RCS fuel system typically sees fuel decrease in very small short bursts as thrusters are fired for fractions of a second to make small attitude adjustments. Fuel levels usually decrease in small step increments with wide intervals where no burns occur. Shallow trajectories come in closer to the lunar surface and at a longer range and hence require significantly more fuel burn initially than do steep trajectories. This does NOT necessarily constitute a failure.
- Attitude for this portion of the trajectory should be largely centered around zero for roll and yaw.
- Altitude and Range should be decreasing as the spacecraft descends and moves towards its designated landing point.
- Shallow trajectories burn a large amount of fuel at the beginning and then stabilize, steeper trajectories have a larger fuel burn towards the end of the scenario. These are not fuel leaks.



---

## Possible RCS Failure Cases

- Two failures will be associated with the RCS fuel system, and two with the actual thrusters themselves
- Fuel System Failures:
  - Fuel Leak
  - Fuel Pump Failure
- Thruster System Failures:
  - Thruster Failure-Stuck On
  - Valve Failure

---

## Fuel Leak

- Problem: A rupture has occurred somewhere in a fuel line or valve that is allowing fuel to leak out of the system
- Symptoms:
  - = Thrusters and fuel pumps are functioning normally; stable RCS output
  - = As a result, spacecraft attitude remains stable and on course
  - = Fuel will decrease at a consistent rate which is faster than normal. Even if thrusters are not firing, fuel level will still be decreasing

## Fuel Pump Failure

- **Problem:** A fuel pump on one cluster has seized and stopped working
- **Symptoms:**
  - An entire cluster of thrusters will no longer be firing because no fuel is being provided to them
  - Attitude will be noticeably affected because one full cluster of thrusters is non-functional. There will be changes in roll, and yaw as the other thrusters overcompensate and correct for the failed cluster
  - Fuel will actually decrease in a mostly normal pattern, with a few additional burns commanded by the guidance computer in order to correct the attitude deviations

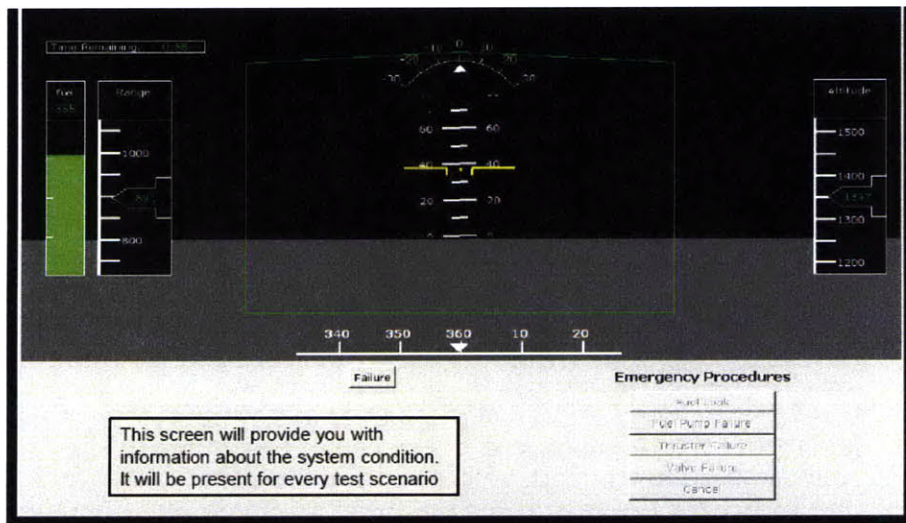
## Single Thruster Failure

- **Problem:** One thruster in a cluster of 4 is stuck on, it is continuously firing
- **Symptoms:**
  - Fuel will continue to decrease linearly at a faster than normal rate both because the thruster is continually firing and it must be compensated for by other thrusters
  - Spacecraft attitude will show a drift in the direction of the thrust in either roll, or yaw, or both, until control counteracts the failure

## Valve Failure

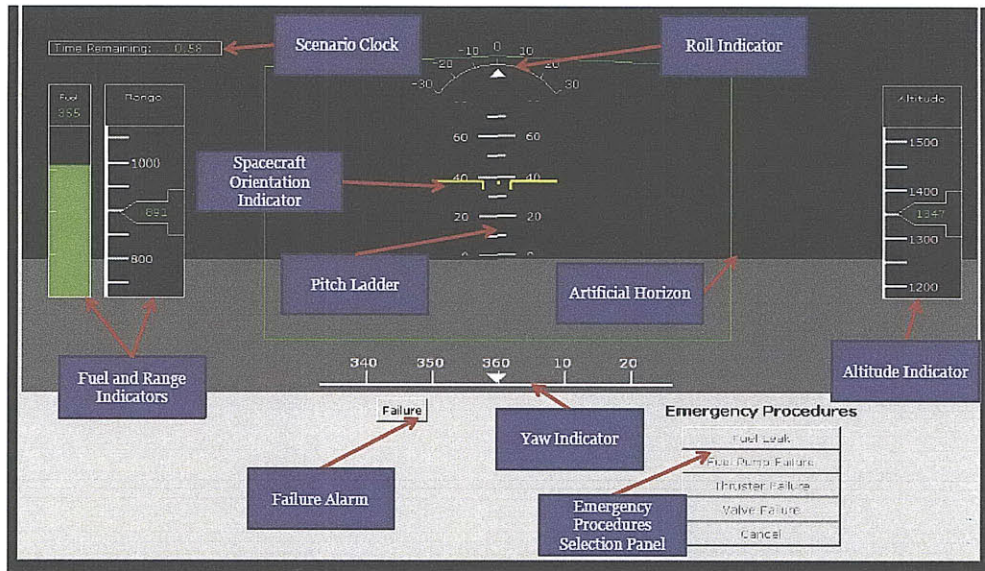
- **Problem:** A valve has created a series of random pressure pockets that causes a thruster to provide intermittent, (on/off), instead of consistent thrust.
- **Symptoms:**
  - Attitude will slowly begin to show a subtle rocking motion, or oscillation, in whatever axes the failed thruster fires on and off.
    - Because this behavior is random and unpredictable, the control algorithm will always be a step behind the failed thruster causing the magnitude of the oscillation to increase
  - Fuel will decrease faster than normal as the attitude oscillation becomes bigger and more thrusters operating for longer periods are necessary to correct the attitude

## Display: Primary Flight Screen





## Key Display Components

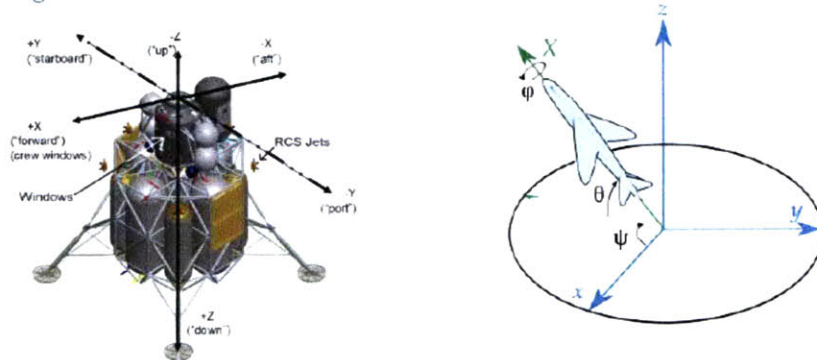


## Display Components: Avionics

- **Time (s):** Time remaining in the scenario is displayed by the scenario clock in the upper left corner of the display. Scenarios are 60 seconds long.
- **Altitude (m):** The current altitude, or vertical height above the lunar surface, is represented by the green values on the tape display as the tape moves accordingly behind the digital display.
- **Range (m):** The range display is set up to function exactly like the altimeter, but it provides information about the horizontal distance to the target landing site.
- **Fuel (kg):** The fuel gauge is a bar graph of fuel. The value of fuel remaining is displayed at the top of the gauge in the header. It is an indication of the total amount of RCS fuel remaining for the rest of the descent trajectory and also the ascent to lunar orbit at mission completion, 230 kg is necessary for the return journey.

## Display Components: Attitude Indicators

- Attitude is represented in 3 dimensions: Pitch, Roll, and Yaw, all of which are measured in degrees.
  - = Pitch: Motion about the Y axis, tilting forwards or backwards, positive pitch is a backwards or “nose up” pitch
  - = Roll: Motion about the X axis, tilting side to side, positive rotation to the right
  - = Yaw: Motion about the Z axis, swinging back and forth, positive rotation to the right

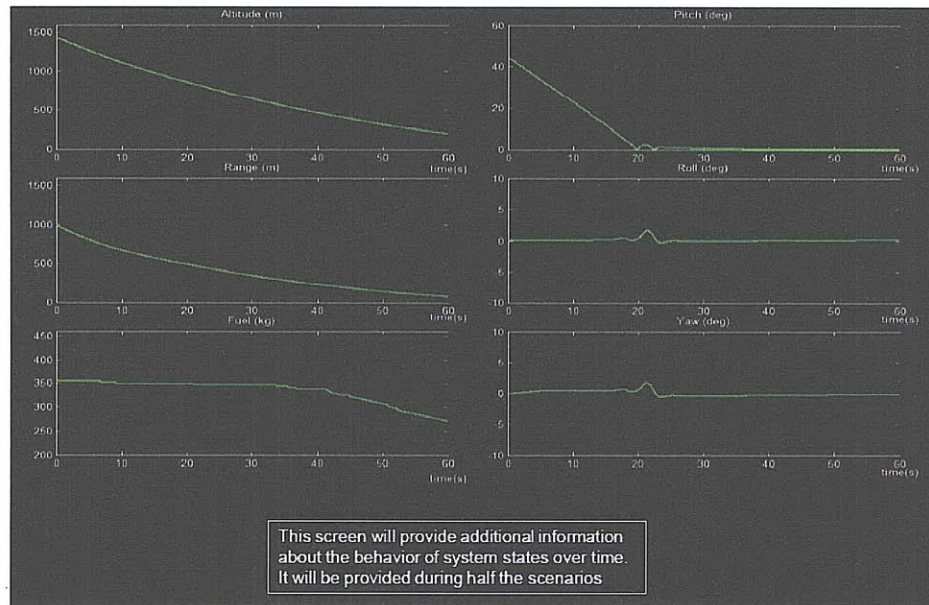


## Display Components: Failure Indicators

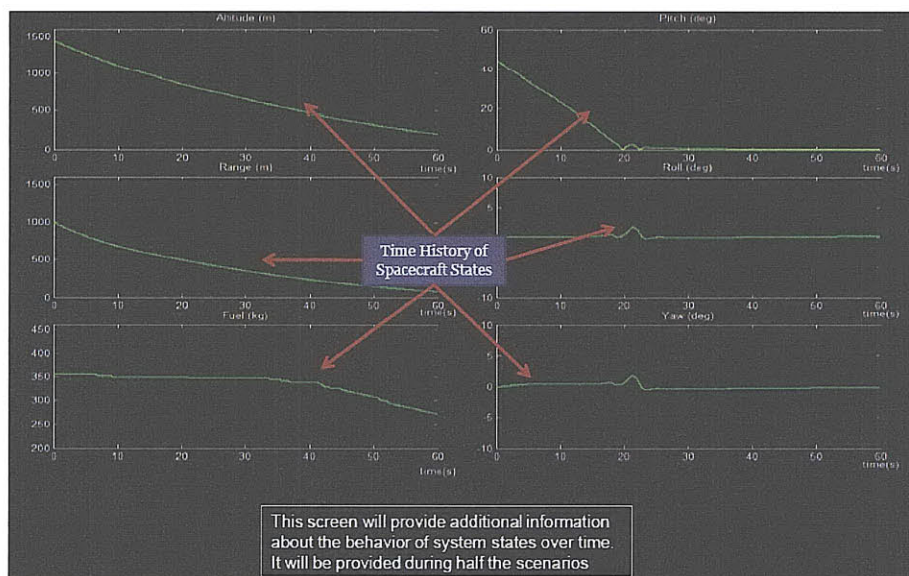
- **Failure Alarm:** The failure alarm is the grey button labeled “Failure”, and exists to alert the crew in the event of some kind of system failure. You are responsible for pushing this button whenever you detect a system failure.
- **Emergency Procedure Selection Panel:** Represented as a set of grey buttons in the lower right corner labeled with all possible failures. Detecting a failure is not sufficient for mission success. Failures must be identified so they can be solved. The crew is trained on emergency procedures for each of the given failure conditions. You are responsible for identifying the failure condition so that the correct emergency procedure can be implemented.



## Display: Time History of Parameters



## Key History Components



## Display Components: Time History

- **Time (s):** The time shown on the horizontal axis is the full scenario time, 60 seconds. The current time is indicated by the last point drawn on the chart. Time remaining in the scenario is shown by the scenario clock.
- **Parameter Information:** The units for these parameters will be the same as they are on the original avionics display. The curve will be generated from the values on the avionics display as the scenario time moves forward such that parameter behavior will be shown over time.

## How Each Scenario Will Take Place

Move to the Next Slide





## Scenario Start

Time Remaining: 0:00

Fuel: 316

Range: 200, 100, 0

Altitude: 300, 200, 100

Heading: -30, -20, -10, 0, 10, 20, 30

Failure

Emergency Procedures

- Fuel Leak
- Fuel Pump Failure
- Thruster Failure
- Valve Failure
- Cancel

Before beginning each scenario, the box above will appear to allow you to control the pace at which you complete the scenarios. Push ok to start the clock and begin

## Failure Detection

Time Remaining: 0:58

Fuel: 355

Range: 1000, 800

Altitude: 1500, 1400, 1347, 1300, 1200

Heading: -30, -20, -10, 0, 10, 20, 30

Failure

Emergency Procedures

- Fuel Leak
- Fuel Pump Failure
- Thruster Failure
- Valve Failure
- Cancel

If you detect a failure based on the behavior of any of the system states, use the mouse to press the failure alarm button as quickly as possible



## Failure Identification

The screenshot shows a flight simulator interface. At the top left, a box displays "Time Remaining: 0:30". On the left side, there are two vertical gauges: "Fuel" with a value of 385 and "Range" with a scale from 800 to 1000. On the right side, an "Altitude" gauge shows a scale from 1200 to 1500, with a current reading of 1247. The central display is a cockpit instrument panel with a heading scale from 340 to 20 and a pitch scale from 20 to 30. Below the heading scale, a "Failure" label is visible. At the bottom right, a panel titled "Emergency Procedures" contains five buttons: "Fuel Leak", "Fuel Pump Failure", "Thruster Failure", "Valve Failure", and "Cancel". A text box on the left side of the bottom panel contains the following text:

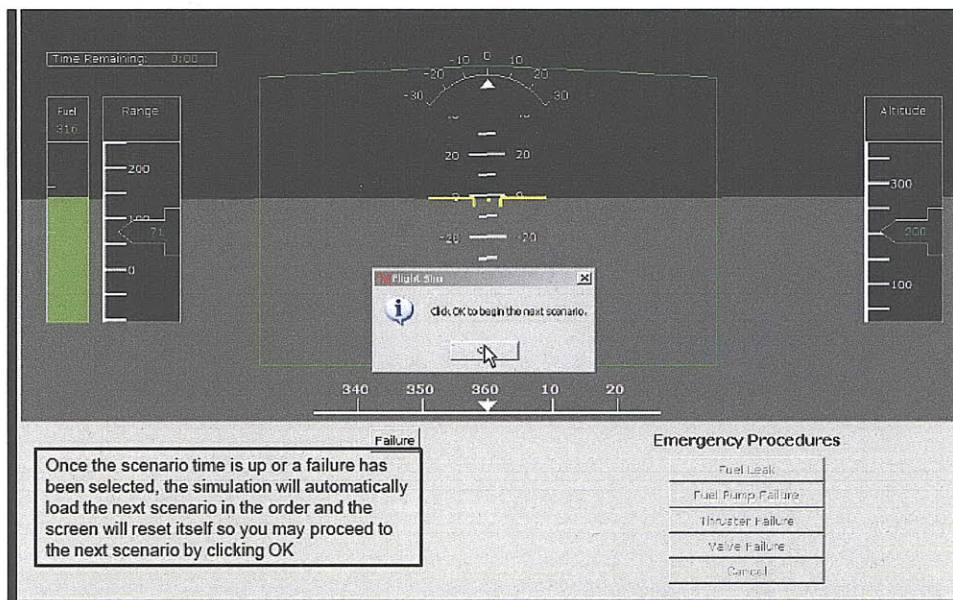
Once you have alerted the rest of the crew to the failure, you will see that the emergency procedure buttons have been activated. As quickly as possible select the appropriate emergency procedure for the failure you believe occurred. If you no longer agree with your detection decision, press "Cancel".

## End of Scenario

This screenshot is identical to the one above, showing the same flight simulator interface. The "Time Remaining" is still 0:30. The gauges and scales are the same. The "Failure" label is still present. The "Emergency Procedures" panel is still visible. The text box on the left side of the bottom panel now contains the following text:

Remember: Once you select an emergency procedure, the scenario will terminate and move on to the next scenario so make your selection carefully. If you press the "Cancel" button, the scenario will continue to run.

## End of Scenario



## Important Things to Remember

- Time history information will only be given to you on half of the experimental trials, these trials will all be grouped together and may occur either first or second
- Failures can occur at any time during the scenario so remain attentive, ensure you have enough evidence for your decision
- Just as failures do not occur on every mission in real life, failures WILL NOT occur in every scenario, you must discern this each time from the system performance
- Only one failure will occur in each scenario. You only need to click the failure alarm and proper emergency procedure ONCE. Emergency procedures simply identify the problem to the crew, once you select an appropriate emergency procedure, the simulation will terminate and you will move on to the next scenario.

QUESTIONS?

## Appendix D: Demographic Survey and Results

### Demographic Questionnaire

Subject number: \_\_\_\_\_

Age: \_\_\_\_\_

Gender:            *M*   *F*

Are you color blind?   *No*    *Yes*

Occupation: \_\_\_\_\_

if student, (circle one):    *Undergrad*    *Masters*    *PhD*

department: \_\_\_\_\_

expected year of graduation: \_\_\_\_\_

Military experience (circle one):    *No*    *Yes*

If yes, which branch: \_\_\_\_\_

Years of service: \_\_\_\_\_

Do you have any flight experience?   *No*    *Yes*

If yes, how many hours? \_\_\_\_\_

Do you have any experience in flight simulators?   *No*    *Yes*

If yes, please describe briefly:

Do you have experience in engineering or operating spacecraft systems ?   *No*    *Yes*

If yes, please describe briefly:

Level of Education, please list your degrees and subjects:

Rate your comfort level with using computer programs.

*Not comfortable*

*Somewhat comfortable*

*Comfortable*

*Very Comfortable*

Subject Demographics:

Total Number of Subjects: 28

Male: 19

Female: 9

No Color Blindness

Age Range: 23-57

Mean Age: 38

Military: 1

Students: 4

Pilots: 2

Flight Simulators: 5

Spacecraft Experience: 10

Education:

Bachelors: 28

Masters: 17

PhD: 5

Comfort with Computers: No participants circled less than "Comfortable"

The effects on performance (latency and accuracy) were investigated for the following demographics: Gender, Students, Pilots/Flight Simulators, Spacecraft Experience. There were no significant effects of these different demographic trends on performance.

## Appendix E: Detailed Analysis Results Tables and Plots

The statistical tables and results were computed in SPSS 15.0 statistical software package. For each data type and each decision there are detailed descriptive statistics, assumption checks and full ANOVA results tables.

### E.1 Detection Latency

Descriptive Statistics for both the original latency and the cube root transform:

**Descriptive Statistics**

	N	Minimum	Maximum	Mean	Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Latency	1109	.00	50.73	11.9204	7.94810	63.172	1.283	.073	2.071	.147
CubeLatency	1109	.00	3.70	2.1692	.51578	.266	-.118	.073	.357	.147
Valid N (listwise)	1109									

Detection latency lacked both homogeneity of variance and normality when the assumptions were verified, as a result, a cube root transform was applied in order to meet these ANOVA assumptions.

Results of a Univariate ANOVA with five fixed factors: Trend information, Failure type, Approach, the order the subjects saw scenarios (Order), and whether they saw the trend information first or second (Block)

**Tests of Between-Subjects Effects**

Dependent Variable: CubeLatency

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	101.918(a)	95	1.073	5.636	.000
Intercept	4726.618	1	4726.618	24829.628	.000
Trend	8.797	1	8.797	46.211	.000

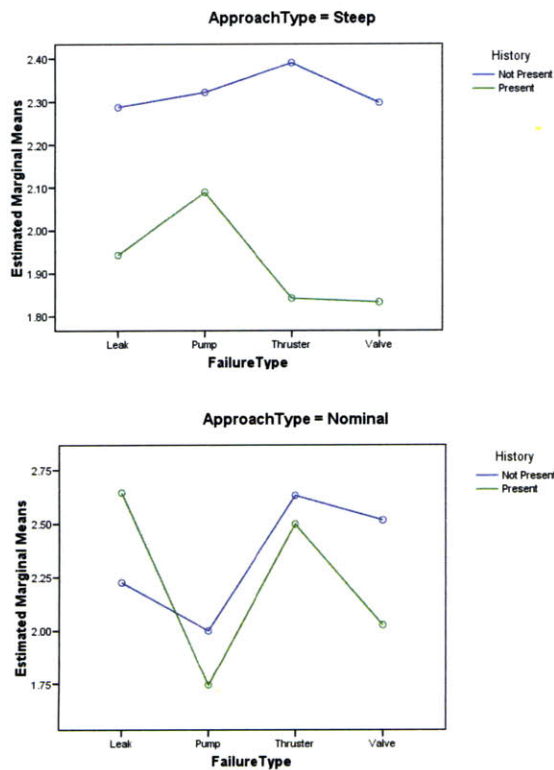


FailureType	18.774	3	6.258	32.874	.000
ApproachType	10.058	2	5.029	26.419	.000
Order	.046	1	.046	.241	.624
Block	.002	1	.002	.009	.924
Trend * FailureType	8.300	3	2.767	14.534	.000
Trend * ApproachType	6.056	2	3.028	15.905	.000
FailureType * ApproachType	19.082	6	3.180	16.707	.000
Trend * FailureType * ApproachType	5.262	6	.877	4.607	.000
Trend* Order	.002	1	.002	.010	.919
FailureType * Order	.376	3	.125	.659	.577
Trend * FailureType * Order	.951	3	.317	1.666	.173
ApproachType * Order	.445	2	.223	1.169	.311
Trend* ApproachType * Order	2.463	2	1.232	6.469	.002
FailureType * ApproachType * Order	1.063	6	.177	.931	.472
Trend * FailureType * ApproachType * Order	2.547	6	.424	2.230	.038
Trend * Block	1.638	1	1.638	8.604	.003
FailureType * Block	2.057	3	.686	3.602	.013
Trend * FailureType * Block	.390	3	.130	.683	.563
ApproachType * Block	.276	2	.138	.724	.485
Trend * ApproachType * Block	.237	2	.119	.623	.537
FailureType * ApproachType * Block	.920	6	.153	.806	.566
Trend* FailureType * ApproachType * Block	.451	6	.075	.395	.883
Order * Block	.163	1	.163	.857	.355
Trend * Order * Block	.042	1	.042	.223	.637
FailureType * Order * Block	.544	3	.181	.953	.414
Trend* FailureType * Order * Block	.247	3	.082	.432	.730
ApproachType * Order * Block	.073	2	.036	.192	.826
Trend * ApproachType * Order * Block	.159	2	.080	.419	.658
FailureType * ApproachType * Order * Block	.188	6	.031	.165	.986
Trend* FailureType * ApproachType * Order * Block	.795	6	.132	.696	.653
Error	192.837	1013	.190		
Total	5513.059	1109			

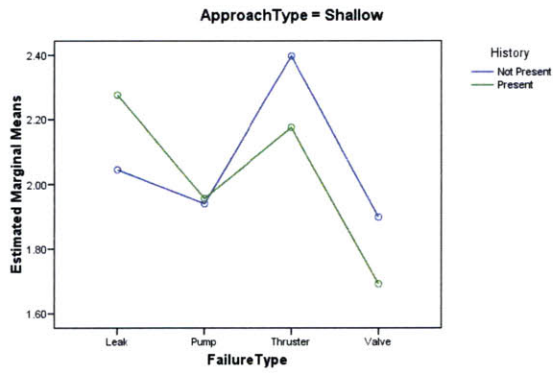
Corrected Total	294.755	1108		
-----------------	---------	------	--	--

a R Squared = .346 (Adjusted R Squared = .284)

Marginal means plots that show the mean cube root latency for each combination of failure type, approach type, and trend information (history), are useful in examining significant effects. For example, there is a distinct difference for all failure types for trend information on steep trajectories. This effect is not as clear for shallow and nominal trajectories.







## E.2 Detection Accuracy

Descriptive Statistics:

**Accuracy with Trend**

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Incorrect	160	9.5	19.3	19.3
	Correct	670	39.7	80.7	100.0
	Total	830	49.2	100.0	
Missing	System	857	50.8		
Total		1687	100.0		

**Accuracy without Trend**

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Incorrect	172	10.2	20.2	20.2
	Correct	678	40.2	79.8	100.0
	Total	850	50.4	100.0	
Missing	System	837	49.6		
Total		1687	100.0		

Forward Likelihood Ratio Binary Regression:

**Variables in the Equation**

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	Valve	1.508	.234	41.343	1	.000	4.516
	Constant	1.200	.065	344.457	1	.000	3.322
Step 2(b)	Thruster	1.502	.213	49.918	1	.000	4.489
	Valve	1.772	.236	56.392	1	.000	5.884

	Constant	.936	.070	178.640	1	.000	2.549
Step 3(c)	Thruster	1.570	.214	53.655	1	.000	4.804
	Valve	1.799	.237	57.736	1	.000	6.043
	Shallow	-.486	.133	13.414	1	.000	.615
	Constant	1.095	.084	169.027	1	.000	2.989
Step 4(d)	Thruster	1.576	.215	53.915	1	.000	4.835
	Valve	1.806	.237	58.012	1	.000	6.083
	Shallow	-.489	.133	13.481	1	.000	.613
	Order	.343	.128	7.135	1	.008	1.409
	Constant	.931	.103	81.821	1	.000	2.536

a Variable(s) entered on step 1: Valve.

b Variable(s) entered on step 2: Thruster.

c Variable(s) entered on step 3: Shallow.

d Variable(s) entered on step 4: Order.

### E.3 Diagnosis Latency

Descriptive Statistics from a repeated measures ANOVA:

**Descriptive Statistics**

	N	Minimum	Maximum	Mean	Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Age	28	23.00	57.00	37.8214	11.19211	125.263	.285	.441	-1.466	.858
Gender	28	.00	1.00	.6786	.47559	.226	-.809	.441	-1.456	.858
Trend	28	.96	8.23	3.0148	1.82849	3.343	1.284	.441	1.066	.858
NoTrend	28	.64	8.16	2.7908	1.94215	3.772	1.565	.441	1.877	.858
Valid N (listwise)	28									

A repeated measures ANOVA was conducted on the diagnosis latency data because the independence of the observations could not be supported.

**Multivariate Tests(b)**

Effect		Value	F	Hypothesis df	Error df	Sig.
Approach	Pillai's Trace	.248	4.297(a)	2.000	26.000	.024
	Wilks' Lambda	.752	4.297(a)	2.000	26.000	.024
	Hotelling's Trace	.331	4.297(a)	2.000	26.000	.024
	Roy's Largest Root	.331	4.297(a)	2.000	26.000	.024
Failure	Pillai's Trace	.556	7.512(a)	4.000	24.000	.000

	Wilks' Lambda	.444	7.512(a)	4.000	24.000	.000
	Hotelling's Trace	1.252	7.512(a)	4.000	24.000	.000
	Roy's Largest Root	1.252	7.512(a)	4.000	24.000	.000
Trend	Pillai's Trace	.005	.132(a)	1.000	27.000	.719
	Wilks' Lambda	.995	.132(a)	1.000	27.000	.719
	Hotelling's Trace	.005	.132(a)	1.000	27.000	.719
	Roy's Largest Root	.005	.132(a)	1.000	27.000	.719
Approach * Failure	Pillai's Trace	.498	2.484(a)	8.000	20.000	.047
	Wilks' Lambda	.502	2.484(a)	8.000	20.000	.047
	Hotelling's Trace	.994	2.484(a)	8.000	20.000	.047
	Roy's Largest Root	.994	2.484(a)	8.000	20.000	.047
Approach * Trend	Pillai's Trace	.004	.046(a)	2.000	26.000	.955
	Wilks' Lambda	.996	.046(a)	2.000	26.000	.955
	Hotelling's Trace	.004	.046(a)	2.000	26.000	.955
	Roy's Largest Root	.004	.046(a)	2.000	26.000	.955
Failure * Trend	Pillai's Trace	.318	2.796(a)	4.000	24.000	.049
	Wilks' Lambda	.682	2.796(a)	4.000	24.000	.049
	Hotelling's Trace	.466	2.796(a)	4.000	24.000	.049
	Roy's Largest Root	.466	2.796(a)	4.000	24.000	.049
Approach * Failure * Trend	Pillai's Trace	.400	1.666(a)	8.000	20.000	.169
	Wilks' Lambda	.600	1.666(a)	8.000	20.000	.169
	Hotelling's Trace	.666	1.666(a)	8.000	20.000	.169
	Roy's Largest Root	.666	1.666(a)	8.000	20.000	.169

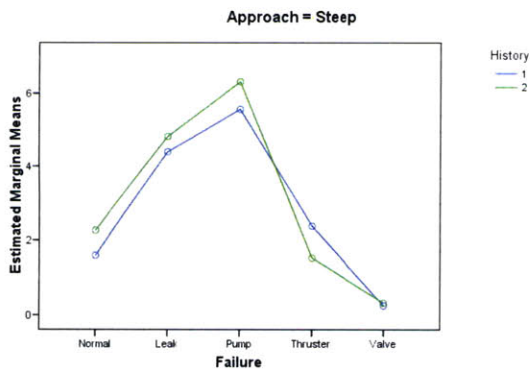
a Exact statistic

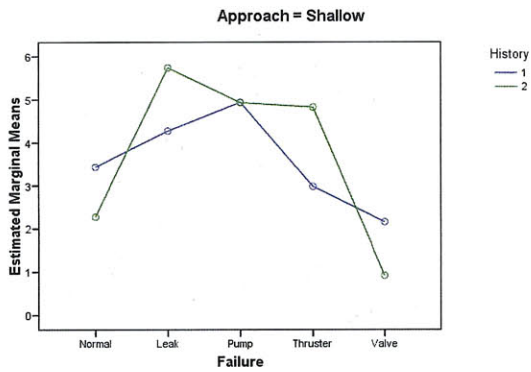
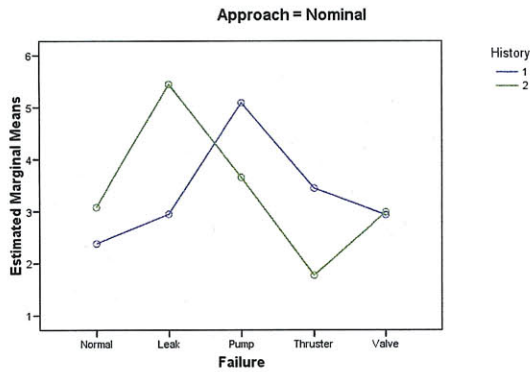
b Design: Intercept

Within Subjects Design:

Approach+Failure+History+Approach\*Failure+Approach\*History+Failure\*History+Approach\*Failure\*History

Marginal means plots that indicate the effects on diagnosis latency of combinations of trend information (history), failure, and approach types. There is no clearly different effects for trend information for any one trajectory, though specific failure types show differences, such as a leak, where history present seemed to make subjects consistently slower.





## E.4 Diagnosis Accuracy

Descriptive Statistics:

**Accuracy with Trend**

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Incorrect	277	16.4	33.4	33.4
	Correct	553	32.8	66.6	100.0
	Total	830	49.2	100.0	
Missing	System	857	50.8		
Total		1687	100.0		

**Accuracy without Trend**

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Incorrect	366	21.7	43.1	43.1
	Correct	484	28.7	56.9	100.0
	Total	850	50.4	100.0	
Missing	System	837	49.6		
Total		1687	100.0		

Forward Likelihood Ratio Binary Logistic Regression:

**Variables in the Equation**

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	Thruster	-1.049	.125	70.364	1	.000	.350
	Constant	.700	.058	145.956	1	.000	2.013
Step 2(b)	Pump	-1.208	.130	86.669	1	.000	.299
	Thruster	-1.273	.130	95.883	1	.000	.280
	Constant	1.009	.070	210.435	1	.000	2.743
Step 3(c)	Pump	-1.962	.191	105.259	1	.000	.141
	Thruster	-1.230	.132	87.185	1	.000	.292
	HistPump	1.430	.241	35.339	1	.000	4.181
	Constant	.997	.070	205.055	1	.000	2.711
Step 4(d)	Pump	-1.840	.192	91.759	1	.000	.159
	Thruster	-1.101	.134	67.932	1	.000	.333
	HistPump	1.442	.240	35.936	1	.000	4.227
	HistValve	1.068	.245	19.017	1	.000	2.909
	Constant	.864	.074	137.406	1	.000	2.374

- a Variable(s) entered on step 1: Thruster.
- b Variable(s) entered on step 2: Pump.
- c Variable(s) entered on step 3: HistPump.
- d Variable(s) entered on step 4: HistValve.

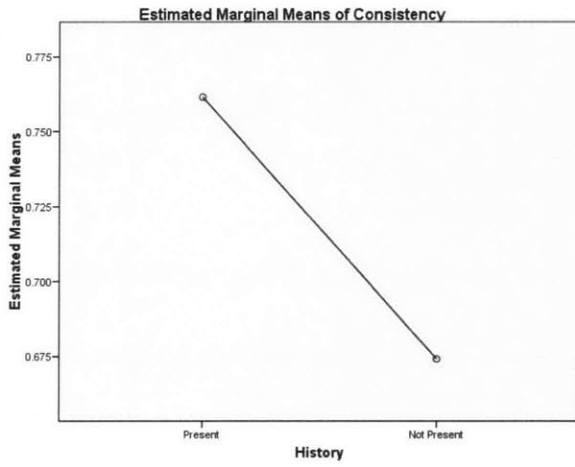
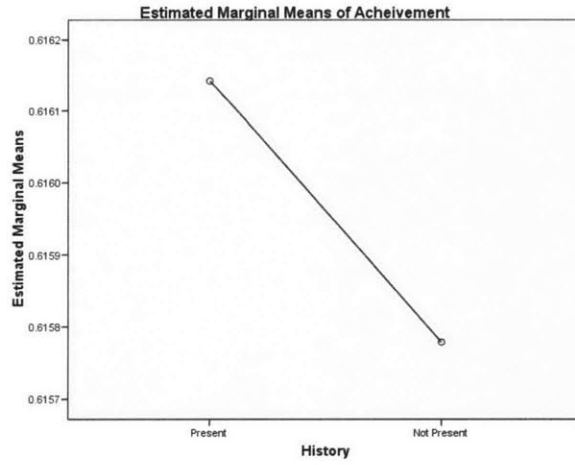
**E.5 Detection Achievement and Judgment Consistency and False Alarms**

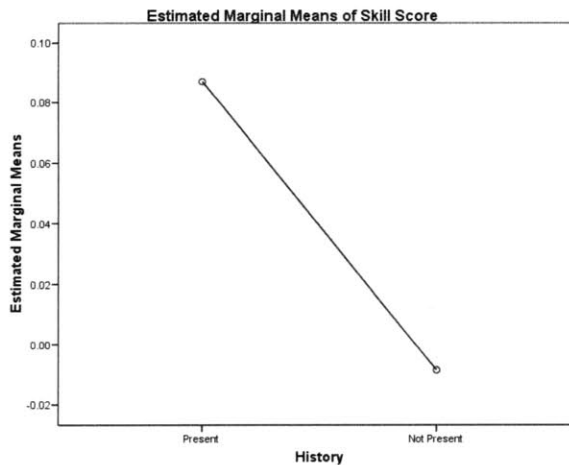
Descriptive Statistics from a repeated measures ANOVA:

**Descriptive Statistics**

	N	Minimum	Maximum	Mean	Std. Deviation
Achievement	56	-.19	.96	.6160	.23043
Consistency	56	.39	.91	.7178	.11250
Valid N (listwise)	56				

Marginal means plots of Lens model parameters achievement, decision consistency, and skill score that indicate higher values with trend information present, but not significantly.





False Alarms:

False alarms were examined using a one-sample T-test because they are continuous data and met all the statistical assumptions. The results showed marginal significance. With the mean value of false alarms being only 1.4 out of 60 trials.

**One-Sample Statistics**

	N	Mean	Std. Deviation	Std. Error Mean
Difference	28	1.3929	4.08556	.77210

**One-Sample Test**

	Test Value = 0					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Difference	1.804	27	.082	1.39286	-.1914	2.9771

## E.6 Diagnosis Achievement and Judgment Consistency

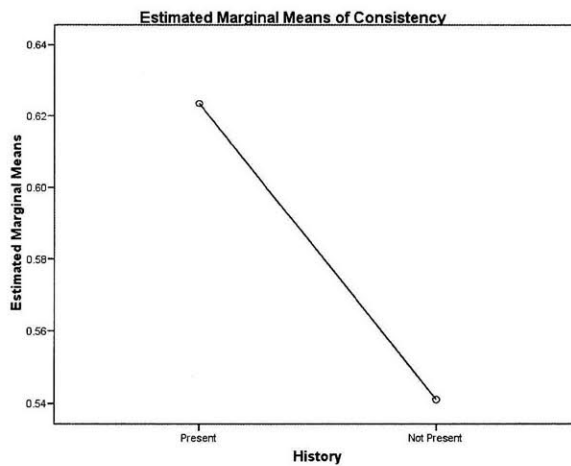
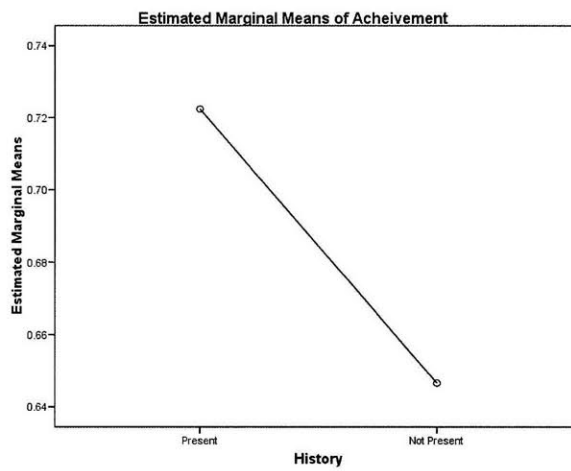
Descriptive Statistics from a repeated measures ANOVA:

**Descriptive Statistics**

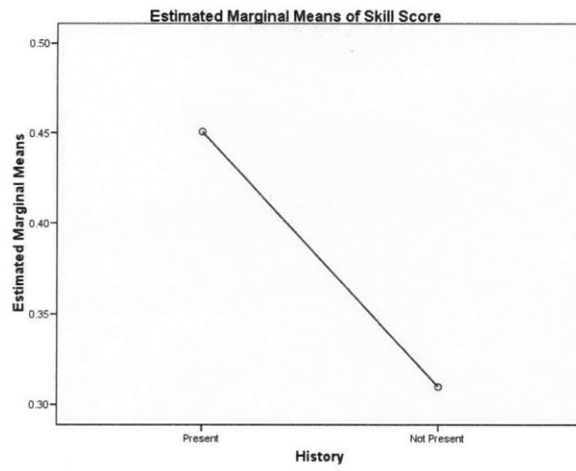
	N	Minimum	Maximum	Mean	Std. Deviation

Achievement	56	.12	.97	.6845	.22612
Consistency	56	.26	.79	.5822	.13371
Valid N (listwise)	56				

Marginal means plots showing the direction of trends for achievement, decision consistency and skill score. All of these seemed to increase, though not significantly when trend information was present.







## Appendix F: Linear Multiple Regression Overview

The primary method for determining the judgment strategy and the environmental state in the Lens model formulation is by multiple linear regression. Multiple linear regression is a statistical technique that attempts to model the relationships between variables in the complex context of real world data. This mathematical model simplifies and organizes observed relationships so that investigators can better understand the actual processes at work in the data. The model relates a dependent variable, generally denoted  $Y$ , and multiple independent variables denoted  $X_i$  where  $i = 1, \dots, n$  and,  $n$  represents the number of variables in the model[40]. A general two dimensional example which includes a regression trendline is found in Figure 31 below.

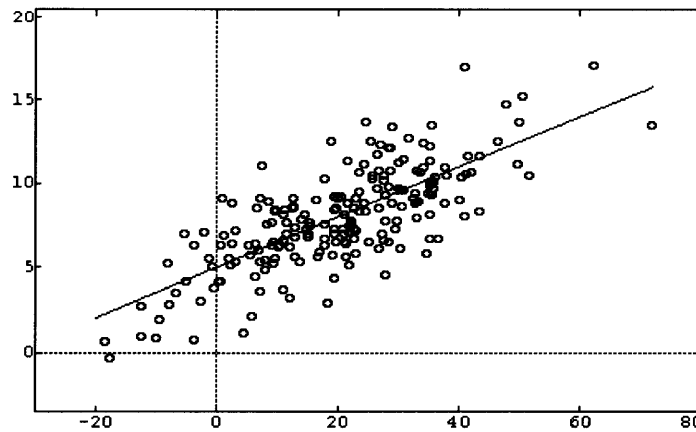


Figure 31: Example of multiple regression plot with regression model trend line

We find the formula for multiple linear regression twice in the Lens model formulation above in the form,  $Y_e = \sum_{j=1}^k r_{e,j} X_j + \varepsilon_e$  . Linear regression models can be useful for a number of applications but are largely used in prediction and forecasting for data sets and circumstances where the linear additive model is a good fit. It is also useful in determining relationships and sensitivity of a dependent variable to a set of independent variables that can be manipulated.

In general to get a model that is considered accurate, the rule of thumb is that five data points per independent variable, are necessary, and ten is better[33]. This means that the operator must make that same decision in the same task environment, given those same cues at least 5 times for each cue present to get a valid regression model of their decision strategy and of the environment. Other rules of thumb have been suggested. Norusis (2006) recommends 10 to 20 cases per independent variable. A third rule of thumb is:  $N \geq 50 + 8m$ , where  $m$  is the number of independent variables[51]. The bottom line appears to be, multiple regression requires a large number of cases. This is true because multiple linear regression attempts to fit a trend line, as shown in the figure above, to a set of data in multiple dimensions based on one of any number of estimation procedures. The more data the estimator is given for each independent variable, generally the more accurate the estimate of the regression coefficient will be. If the sample size is not large enough there is a significant risk that the model will be overfitted, which means it will not be generalizable to any another

data sample but only applicable to the sample set that originally created it[40]. Though there are many estimation methods, only one of these estimation procedures are discussed below as that is the procedure used by the mathematical tools that were used in the generation and post processing of the data for this thesis.

*Ordinary Least Squares:* (OLS) is the simplest estimation procedure used in linear regression and thus a very common technique. It is conceptually simple and computationally straightforward. Ordinary least squares seeks to minimize the sum of the squared difference between the observed and predicted values of the dependent variable, or the residuals, of a data set and leads to a closed form analytical solution for the estimated value of the regression coefficients. This method of coefficient estimation produces coefficients that are considered unbiased and consistent as long as the assumptions of linear regression are satisfied. This is the method used to compute the regression coefficients in the MATLAB model of the Lens model that will be discussed in future chapters. This is also the estimation method used by SPSS statistical software which was used to all of the experimental data post processing.

This estimation procedure generates estimates for the regression coefficients. There are two types of coefficients that are usually generated in regression analysis: unstandardized and standardized. Unstandardized coefficients are the model coefficients referred to in the Lens model formulation above as  $r_{e,j}$  and  $r_{s,j}$ . The magnitude of these coefficients does not necessarily tell us how good they are at

predicting the dependent variable because the size of the coefficient is dependent on the units of measurement of that particular independent variable. This means that the regression coefficients cannot be compared because they all have different units of measurement. The standardized coefficients are referred to generally as “beta” weights and are represented by  $\beta_{e,j}$  for the environment and  $\beta_{s,j}$  for the judgment strategy. The benefit to standardized coefficients is that they can then be compared to one another outright. Standardized coefficients are the result of standardizing both the dependent and independent variables to have a mean of 0 and a standard deviation of 1. The value of the coefficients still depends largely on the other variables included in the model, but they can now be compared to one another in the particular model in question. These coefficients are restricted between -1 and 1 unless there are two or more correlated independent variables in the model equation[40].

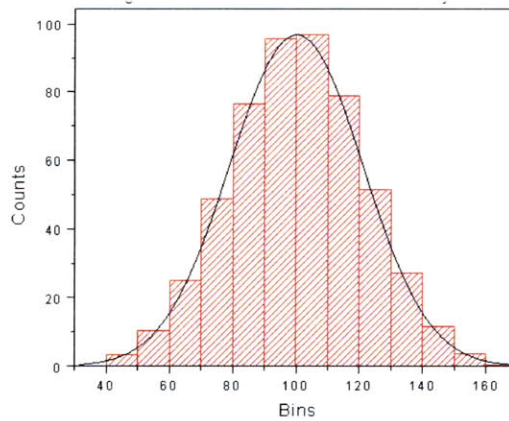
Linear regression employs a number of assumptions that must be satisfied in order for this method to be used on a data set. These assumptions are:

- Independence of error terms: The independent variables and hence their residuals are not correlated, each observation is independent of another.
- Linearity: There is a linear relationship between the dependent variable and the independent variables.
- Normality: The error terms for each independent variable follow a normal distribution
- Homoscedasticity: The error terms for each independent variable have the same variance.

One other assumption that can be helpful to check for is multicollinearity. If the goal of the regression analysis is either inference or predictive modeling, the performance of ordinary least squares estimates can be poor if this trait is present, unless the sample size is large. Multicollinearity is a statistical phenomenon where one or more independent variables are highly correlated. This can result in the condition of heteroscedasticity, and cause the regression assumptions to be violated[40].

In order to verify that each of these assumptions holds, there are a variety of tests that must be performed on the data before any analysis with regression can be done. The first step is generally to generate the residuals of the data which is simply the difference between the observed and the predicted value of the dependent variable. These are also dependent on the units of measurement of a given variable and so a truer reflection of the error variance can be determined from standardizing or studentizing the residuals.

*Normality:* The best way to assess normality of the residuals is to plot them in a histogram. A histogram of the data will show the number of different pieces of data that fall into different buckets. If this diagram is symmetric and generally mirrors the shape of the normal distribution curve with the majority of the data values falling in the center, the data can be assumed to be normal. See Figure 32 below for an example of a normal histogram.



**Figure 32: Example normal histogram with distribution curve**

*Linearity:* In order to determine if the variables have a linear relationship to the dependent variable it is best to simply create a scatter plot of the residuals for the dependent variable against the predicted values or observed values of the independent variables or predictor variables in the regression equation. If the errors seem to be distributed evenly above and below zero along the entire sample, the relationship is probably linear. It is possible to visualize this by graphing the dependent variable and each of the independent variables as well, but this will not give any indication other than a best judgment about exactly how linear the relationship actually is. An example of a data set which exhibits a linear relationship is shown in Figure 33 below.

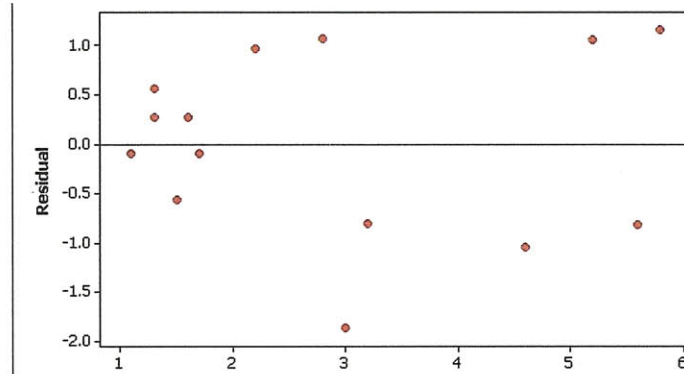


Figure 33: Plot of residuals vs. the independent variable for a data set that shows a likely linear relationship

*Homoscedasticity:* The easiest way to check for equality of variance is to examine the spread of the residuals over the range of the independent variable. If the spread of the residuals seems to increase as the predicted values increase then the variance is most likely unequal. A diagram of what a spread plot would look like is shown below in Figure 34. A statistical test known as the Levene test may also be used to verify that the variances the error terms of an independent variable are equal across all its factor levels. If this statistical test is significant, the variances are unequal and the data violate the assumptions for linear regression.



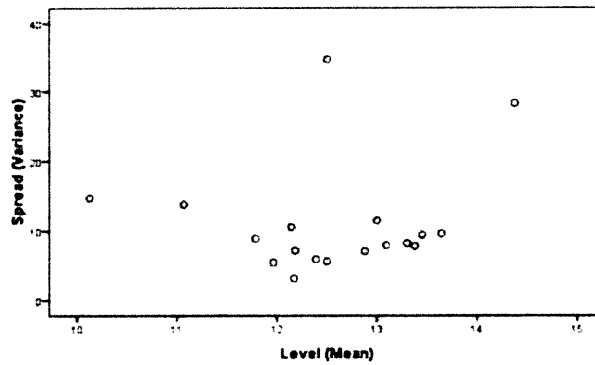


Figure 34: Example spread and level plot for determining equal variance

*Independence:* Checking for independence requires that you plot residuals in the order they were obtained if the observations were collected in a sequence. If there is a pattern, there may be a violation of the independence assumption. The most appropriate way to check for independence of observations however is using the Chi-squared test. If the Chi-squared test results are significant, this means that the null hypothesis that the variables are randomly correlated is rejected and the variables are not independent[40].

## References

1. Laboratory, C.S.D., *The Use of Markov and PARADyM Modeling to Evaluate and Optimize Designs for NASA Crewed Spacecraft*. 2008, NASA Engineering and Safety Center: Cambridge, MA.
2. Fuld, R.B., Liu, Y., Wickens, C.D. *The Impact of Automation on Error Detection: Some Results from a Visual Discrimination Task*. in *Human Factors Society*. 1987.
3. Parasuraman, R., T.B. Sheridan, and C. Wickens, *A Model for Types and Levels of Human Interaction with Automation*. *IEEE Transactions on Systems, Man, and Cybernetics*, 2000. **30**(3).
4. Rouse, W.B., *Models of Human Problem Solving: Detection, Diagnosis, and Compensation for System Failures*. *Automatica*, 1983. **19**(6): p. 613-625.
5. Brunswik, E., *Representative Design and Probabilistic Theory in a Functional Psychology*. *Psychological Review*, 1955. **62**(3): p. 193-217.
6. Willsky, A.S., *A Survey of Design Methods for Failure Detection in Dynamic Systems*. *Automatica*, 1976. **12**(1): p. 601-611.
7. Isermann, R., *Supervision, Fault-Detection and Fault-Diagnosis Methods- An Introduction*. *Control Engineering Practice*, 1997. **5**(5): p. 639-652.
8. Greenstein, J.S., Rouse, W. B., *A Model of Human Decisionmaking in Multiple Process Monitoring Situations*. *IEEE Transactions on Systems, Man, and Cybernetics*, 1982. **SMC-12**(2): p. 182-193.
9. Gai, E.G., Curry, R. E., *A Model of the Human Observer in Failure Detection Tasks*. *IEEE Transactions on Systems, Man, and Cybernetics*, 1976. **SMC-6**(2): p. 85-94.
10. Polycarpou, M.M., Vemuri, A. T., *Learning Methodology for Failure Detection and Accommodation*. *IEEE Control Systems*, 1995: p. 16-24.
11. Mehra, R., Rago, C., Seereeram, S., *Autonomous Failure Detection, Identification and Fault-tolerant Estimation with Aerospace Applications*. *IEEE*, 1998: p. 133-138.
12. Wickens, C.D., Kessel, C., *The Effects of Participatory Mode and Task Workload on the Detection of Dynamic System Failures*. *IEEE Transactions on Systems, Man, and Cybernetics*, 1979. **9**(1).

13. Bisantz, A.M., et al., *Modeling and Analysis of a Dynamic Judgment Task Using a Lens Model Approach*. IEEE Transactions on Systems, Man, and Cybernetics, 2000. **30**(6): p. 605-616.
14. Wickens, C.D. and J.G. Hollands, *Engineering Psychology and Human Performance*. 2000, Prentice-Hall: Upper Saddle River, NJ.
15. Wickens, C.D. and J.G. Hollands, *Engineering Psychology and Human Performance*. 3rd ed. 2000, Upper Saddle River, N.J.: Prentice Hall.
16. O'Hare, D., *Aeronautical Decision Making: Metaphors, Models, and Methods*, in *Principles and Practice of Aviation Psychology*, P.S. Tsang, Vidulich, M. A., Editor. 2003. p. 201-237.
17. Donmez, B., Pina, P.E., Cummings, M.L. *Evaluation Criteria for Human-Automation Performance Metrics*. in *Intelligent Systems Workshop*. 2008. Gaithersburg, MD.
18. Vidulich, M.A., *Mental Workload and Situation Awareness: Essential Concepts for Aviation Psychology Practice*, in *Principles and Practice of Aviation Psychology*, P.S. Tsang, Vidulich, M. A., Editor. 2003. p. 115-146.
19. Endsley, M.R., *Toward a Theory of Situation Awareness in Dynamic Systems*. Human Factors, 1995. **37**(1): p. 32-64.
20. Orasanu, J. and T. Connolly, *The re-invention of decision making*, in *Decision Making in Action: Models and Methods*, G. Klein, et al., Editors. 1993, Ablex Publishing: Norwood, N.J.
21. Endsley, M., *Measurement of situation awareness in dynamic systems*. Human Factors, 1995. **37**(1): p. 65-84.
22. Wickens, C.D., Flach, J. M., *Information Processing*, in *Handbook of Human Factors and Ergonomics*, G. Salvendy, Editor. 1988, Wiley.
23. Takano, K., Reason, J., *Psychological Biases Affecting Human Cognitive Performance in Dynamic Operational Environments*. Journal of Nuclear Science and Technology, 1999. **36**(11): p. 1041-1051.
24. Kahneman, D., Tversky, A., *On the Psychology of Prediction*. Psychological Review, 1973. **80**(4): p. 237-251.
25. Mosier, K.L., et al., *Automation bias: decision making and performance in high-tech cockpits*. The International Journal of Aviation Psychology, 1998. **8**(1): p. 47-63.

26. Glaser, R., *Expertise and Learning: How Do We Think About Instructional Processes Now That We Have Discovered Knowledge Structures?*, in *Complex Information Processing: The Impact of Herbert A. Simon*, H.A. Simon, Klahr, D., Kotovsky, K., Editor. 1989, Psychology Press.
27. Kozlowski, S.W.J., *Training and developing adaptive teams: Theory, principles, and research.*, in *Decision Making Under Stress: Implications for Training and Simulation*, J.A. Cannon-Bowers, Salas, E., Editor. 1998, APA Books: Washington.
28. Endsley, M.R., Jones, W.M., *Situation Awareness Information Dominance and Information Warfare*. 1997, Logicon Technical Services Inc: Dayton, OH.
29. Sarter, N.B., Woods, D.D., *Situation Awareness: A Critical but Ill-Defined Phenomenon*. *The International Journal of Aviation Psychology*, 1991. 1(1): p. 45-57.
30. Endsley, M. *Situation Awareness, Automation, and Free Flight*. in *FAA/Eurocontrol Air Traffic Management R&D Seminar*. 1997. Saclay, France.
31. Sefarty, D., MacMillan, J., Entin, E.E., Entin, E.B., *The Decision Making Expertise of Battlefield Commanders*, in *Naturalistic Decision Making*, C.E. Zsombok, Klein, G., Editor. 1997, Psychology Press.
32. Brannick, M.T., *The Lens Model*, University of Southern Florida, College of Arts and Sciences.
33. Miller, S., *Judgment Analysis: The Lens Model*. 2008, University of Illinois at Urbana-Champaign: Urbana, Illinois.
34. Karelaia, N., Hogarth, R. M., *Determinants of Linear Judgment: A Meta-Analysis of Lens Model Studies*. *Psychological Bulletin*, 2008. 134(5): p. 404-426.
35. Hursch, C.J., K.R. Hammond, and J.L. Hursch, *Some methodological considerations in multiple-cue probability studies*. *Psychological Review*, 1964. 71: p. 42-60.
36. Tucker, L.R., *A suggested alternative formulation in the developments by Hursch, Hammond, & Hursch and by Hammond, Hursch, & Todd*. *Psychological Review*, 1964. 71: p. 528-530.

37. Dhami, M.K., Harries, C. , *Fast and frugal versus regression models of human judgment*. *Thinking and Reasoning*, 2001. 7: p. 5-27.
38. Rothrock, L., Kirlik A., *Inferring rule-based strategies in dynamic judgment tasks*. *IEEE Transactions on Systems, Man, and Cybernetics*, 2003. 33: p. 58-72.
39. Keeley, S.M., Doherty, M. E., *Probabalistic and multiple regression modeling of the biologist's decision processes adn its implications for medical diagnosis*. *Bulletin of Mathematical Biology*, 1971. 33(3): p. 439-449.
40. Norusis, M.J., *SPSS 15.0 Statistical Procedures Companion*. 2006, Upper Saddle River, NJ: Prentice Hall.
41. Murphy, A.H., *Skill scores based on the mean square error and their relationships to the correlation coefficient*. *Monthly Weather Review*, 1988. 2417(24).
42. Kirlik, A., Strauss, R., *Situation awareness as judgment I: Statistical modeling and quantitative measurement*. *International Journal of Industrial Ergonomics*, 2006. 36(5): p. 463-474.
43. Strauss, R., Kirlik, A., *Situation awareness as judgment II: Experimental demonstration*. *International Journal of Industrial Ergonomics*, 2006. 36(5): p. 475-484.
44. Juslin, P.N., Karlsson, J., et al., *Play it Again with Feeling: Computer Feedback in Musical Communication of Emotions*. *Journal of Experimental Psychology: Applied*, 2006. 12(2): p. 79-95.
45. Miller, S., Kirlik, A., et al. *Supporting Joint Human-Computer Judgment under Uncertainty*. in *Proceedings of the Human Factors and Ergonomics Society*. 2008.
46. Masalonis, A.J., Byrne, E. A., et al. *Instrument Failure Detection and Workload in Simulated General Aviation Flight During Manual and Automated Lateral Tracking*. in *International Symposium on Aviation Psychology*. 1997. Washington, DC.
47. Ephrath, A.R., Curry, R. E., *Detection by Pilots of System Failures During Instrument Landings*. *IEEE Transactions on Systems, Man, and Cybernetics*, 1977. SMC-7(12): p. 841-848.

48. Bortolami, S.B., Duda, K. R., Borer, N. K., *Markov Analysis of Human-in-the-Loop System Performance and Reliability*. 2008, The Charles Stark Draper Laboratory: Cambridge, MA.
49. Cooksey, R.W., *Judgment Analysis: Theory, Methods, and Applications*. 1996: Academic Press.
50. Mueller, E., Bilimoria, K. D., Frost, C., *Effects of Control Power and Inceptor Sensitivity on Lunar Lander Handling Qualities*, in *AIAA Space*. 2009: Pasadena, CA.
51. Tabachnick, B.G. and L.S. Fidell, *Using Multivariate Statistics*. 4th ed. 2001, New York: HarperCollins. 880.