INFORMATION TO USERS

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

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

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

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

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI
A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700  800/521-0600
EXPERIENCE AND AUTOMATION
EFFECTS ON FAILURE DETECTION

DISSERTATION
Presented in Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy in the
Graduate School of the Ohio State University

By
Sheryl Lee Chappell, B.S., M.A.
The Ohio State University
1996

Dissertation Committee:
Professor Richard J. Jagacinski, Advisor
Professor Harvey G. Shulman
Professor Thomas E. Nygren

Approved by
Richard J. Jagacinski
Advisor,
Department of Psychology
ABSTRACT

The information available to the pilots in traditional aircraft is also present in modern aircraft, however, the information may not be utilized in the same way by the pilots. As pilots gain experience with the information provided in modern cockpits, they selectively adjust their attentional focus. Supplemental information that is redundant to primary information, but easier to use, results in the primary information being ignored, or sampled very infrequently, for example relying on the flight director to the exclusion of the glideslope and localizer.

A task created for these studies required subjects to perform a tracking task and report anomalies in a set of redundant cues. The subjects moved a cursor up a screen passing through the horizontal midpoint of eight pairs of squares. An additional pair of more closely spaced dots marked this midpoint accurately for the first 200 trials and on 90% of the screens on the next 200 trials. On 10% of the trials the dots were offset (left or right) from the center of the squares. These studies examined the reliance on the easier-to-use dots. It was shown that subjects perform poorly in the detection of an offset of the dots, when this information has proved to be reliable. The redundancy that is designed into the system and is required to be used is actually ignored.

The long-standing finding in the vigilance literature is that operators typically change their decision criteria over time. This change results in a reduced probability of signal
detection. With this task, after experience without a failure subjects had the same decision criterion, yet were less likely to detect the failure due to omitting the check for the presence of an offset.

The reliance on the more easily used information was examined in automated and manual control conditions. The findings show that with adequate feedback and low workload, automated control does not degrade nor contribute to failure detection.

These findings have critical implications for flight deck design, pilot training, and testing. The implications are discussed in detail and suggestions are offered to improve failure detection.
Dedicated to the memories of

Dr. William D. Reynard and Dr. Anil V. Phatak,

for all they taught me about aviation and more importantly about life.
ACKNOWLEDGMENTS

I owe an infinite amount of gratitude to my parents, Bonnie J. Chappell-Hoover and Gerald M. Chappell, and my sisters, Laura L. Chappell and Janice M. Clough, for their loving support and continued encouragement.

I am indebted to Dr. Richard J. Jagacinski, my advisor, who provided megabytes of wisdom and guidance across many miles, time zones, and years. I also wish to express my gratitude to Dr. Harvey G. Shulman and Dr. Thomas E. Nygren for serving as committee members and providing valuable input along the way.

Finally, I would like to express my appreciation to the Flight Management and Human Factors Division of the National Aeronautics and Space Administration for supporting this research.
VITA

1978          M.A. Experimental Psychology, Ohio State University
1974          B.S. Psychology, Ohio State University

1996-present  TransQuest Inc., Usability Engineering Division, Senior Human Factors Engineer
1980-1996     National Aeronautics & Space Administration, Ames Research Center, Flight Guidance and Human Factors Division, Principal Scientist
1981-1987     Embry-Riddle Aeronautical University, Moffett Field California Campus, Instructor of Statistics & Mathematics
1985-1987     Chapman College, Moffett Field California Campus, Instructor of Statistics
1982-1984     Manudyne Systems Incorporated, Los Altos, California, Consultant
1978-1980     Battelle Memorial Institute, Transportation Systems Section, Researcher
1974-1978     Ohio State University, Department of Aviation, Graduate Research Associate

PUBLICATIONS

RESEARCH PUBLICATION


FIELD OF STUDY

Major Field: Cognitive/Experimental Psychology
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Dedication</td>
<td>iv</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>v</td>
</tr>
<tr>
<td>Vita</td>
<td>vi</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xv</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xviii</td>
</tr>
<tr>
<td>Chapter 1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>The Experienced Pilot in the Automated Cockpit</td>
<td>1</td>
</tr>
<tr>
<td>Literature Review</td>
<td>7</td>
</tr>
<tr>
<td>Behavior Taxonomies</td>
<td>7</td>
</tr>
<tr>
<td>Rasmussen's taxonomy adapted for specific applications</td>
<td>10</td>
</tr>
<tr>
<td>Comparison of taxonomies on the perception-action dimension</td>
<td>12</td>
</tr>
<tr>
<td>The Stages in Learning a Skill</td>
<td>13</td>
</tr>
<tr>
<td>The Behavior of Experts</td>
<td>15</td>
</tr>
<tr>
<td>Examples of expert behavior</td>
<td>16</td>
</tr>
<tr>
<td>Skilled complex behaviors</td>
<td>17</td>
</tr>
<tr>
<td>Skilled components of the perception-action sequence</td>
<td>17</td>
</tr>
<tr>
<td>Skilled behavior by pilots</td>
<td>18</td>
</tr>
<tr>
<td>Failure Detection</td>
<td>19</td>
</tr>
<tr>
<td>Vigilance</td>
<td>20</td>
</tr>
</tbody>
</table>
Experience and Failure Detection .............................................................. 21
Automation and Failure Detection ............................................................. 23
The Use of Redundant Cues ..................................................................... 24
Implications for Cockpit Design ............................................................... 25
Implications for Training .......................................................................... 28
Summary of Literature Review ............................................................... 30

Chapter 2. Research Objectives .............................................................. 32

Determine the Form of the Learning Curve for this Task...................... 32
Determine Differences in the Failure Detection by Experienced and
Inexperienced Operators ........................................................................ 33
Observe Performance Changes for Experienced Subjects after the
First Failure .............................................................................................. 36
Observe any shift in detection by experienced subjects after
the first failure ....................................................................................... 36
Measure changes in MTs after the first failure for
experienced subjects ........................................................................... 37

Determine the Effect of Automation on Failure Detection and the
Interaction with Experience Level ......................................................... 38
Measure MTs and Errors for Targets with and without Failures .......... 40
MTs to detected failures ......................................................................... 40
Errors for detected failures .................................................................... 41
Measure Tracking Error and MT Resulting from Capture Errors ....... 41
Measure Differences in MT for Targets at 40 and 50 Degrees .............. 42

Chapter 3. Experiment 1 .......................................................................... 43

Experimental Method ............................................................................. 43
Subjects .................................................................................................. 43
Apparatus ............................................................................................... 43
Experimental Design ............................................................................. 44
The Task ................................................................................................. 46
Experimental Procedure ........................................................................ 52

Experiment 1 Results ............................................................................. 54
The Effect of Practice on Movement Time ............................................... 54
Effects of Experience Level ...................................................................... 56
  Movement times for experienced and inexperienced subjects .... 56
  Failure detection by experienced and inexperienced subjects .... 56
  Improvement in failure detection by experience level .............. 57
  Rate of detection versus movement times for experienced
  subjects ......................................................................................... 58
  Sensitivity and decision criteria by experience level ............... 60
  False alarms for experienced and inexperienced subjects ......... 62
Performance by Experienced Subjects after the First Failure ............. 65
  Detections by experienced subjects after the first failure......... 65
  Movement times before and after the first failure for
  experienced subjects ................................................................. 66
Movement Times and Errors for Offset and Normal Targets .......... 67
  Movement times for targets with and without offsets ............ 67
  Tracking errors for failure and normal targets ...................... 67
  Cursor positions for failure and normal targets ..................... 68
Capture Errors and Movement Times ................................................. 69
  Capture errors .......................................................................... 69
  Movement times for capture errors ........................................ 71
Movement Times for Targets at 40 and 50 Degrees ...................... 72
Effects of Automation and the Interaction between Control Mode
and Experience Level ......................................................................... 73
  Failure Detection for Subjects Using Manual and Automatic
  Control ....................................................................................... 73
  Sensitivity and decision criteria by control type .................... 74
  False alarms by subjects with manual and automatic control.... 74
Questionnaire Information .................................................................. 75
Experiment 1 Discussion .................................................................... 77
  The Effect of Practice on Movement Time ................................. 77
  Effects of Experience Level ....................................................... 77
  Movement times for experienced and inexperienced subjects .... 77
  Failure detection by experienced and inexperienced subjects .... 78
Experiment 2 Results ................................................................. 92

Detections and false alarms .......................................................... 92
Sensitivities and decision criteria for failure detection ...................... 93
Optimal and measured decision criteria for Subject 3 ...................... 95
The noise and signal distributions .................................................. 96
Prior probability and payoff effects on sensitivity ......................... 97
Prior probability and payoff effects on decision criteria .................. 98

Experiment 2 Discussion ............................................................. 99

Chapter 5. Experiment 3 .............................................................. 101

Experiment 3 Method ................................................................. 103

Subjects...................................................................................... 103
Apparatus................................................................................... 103
Experimental Design ................................................................. 103
The Task..................................................................................... 104
Experimental Procedure ........................................................... 106
Expected Results ...................................................................... 107

Experiment 3 Results ................................................................. 109

Detections, False Alarms, and Positive Responses in Block 1 .......... 109
Detections in Block 2 ................................................................. 112
  Detection performance by group .............................................. 112
  Measures of sensitivity and decision criteria ............................ 114
  Detections for the four groups ............................................... 114
  Detection rates over time ..................................................... 115
  Block 1 positive response rates and Block 2 detection rates ....... 116
  Detection of the first failure ................................................. 117
  Detection in Experiment 1 versus 3 ....................................... 118
  Detection in Experiment 2 versus 3 ....................................... 119
False Alarms in Block 2 .............................................................. 121
  False alarm rates by group ................................................... 121
  False alarm rates over time for the four groups ....................... 121

xii
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>The control loops for manual and automated aircraft</td>
<td>3</td>
</tr>
<tr>
<td>2.</td>
<td>Experimental conditions for Experiment 1</td>
<td>45</td>
</tr>
<tr>
<td>3.</td>
<td>Stimuli with cursor trajectory through the line midpoints</td>
<td>47</td>
</tr>
<tr>
<td>4.</td>
<td>Primary cues (squares) with secondary cues (dots) depicting tolerance area</td>
<td>47</td>
</tr>
<tr>
<td>5.</td>
<td>Trial stimuli with a failure, a set of dots</td>
<td>49</td>
</tr>
<tr>
<td>6.</td>
<td>The dimensions of a set of squares</td>
<td>50</td>
</tr>
<tr>
<td>7.</td>
<td>Response screen with rating scales and target numbers</td>
<td>51</td>
</tr>
<tr>
<td>8.</td>
<td>Mean movement times (seconds) for the 16 experienced subjects, trials 1-200</td>
<td>55</td>
</tr>
<tr>
<td>9.</td>
<td>Mean movement times (seconds) for experienced subjects (trial 201-400)</td>
<td>57</td>
</tr>
<tr>
<td>10.</td>
<td>Mean detection rate for experienced and inexperienced subjects over time by pairs of targets</td>
<td>59</td>
</tr>
<tr>
<td>11.</td>
<td>Proportion of subjects detecting failure, time to first and all gates</td>
<td>61</td>
</tr>
<tr>
<td>12.</td>
<td>The Receiver Operating Characteristic plot for each</td>
<td>63</td>
</tr>
<tr>
<td>13.</td>
<td>The Receiver Operating Characteristic plot for each inexperienced subject (n = 16) for Trials 1-200</td>
<td>64</td>
</tr>
</tbody>
</table>
14. The mean false alarm rate for experienced and inexperienced subjects over time .......................................................................................................................... 65

15. Movement times (seconds) for detected offset (failure) and normal targets. ....... 68

16. Two types of cursor placement error, capture error and noncapture error .......... 70

17. Mean movement times (seconds) with standard deviations for correct cursor placement and capture errors to detected offset stimuli ........................................... 72

18. Experimental conditions for Experiment 2 ................................................... 90

19. Payoff matrix to encourage a response that an offset had been present ............ 91

20. Payoff matrix to encourage a response that an offset had not been present ....... 91

21. The probability of detection versus the probability of a false alarm for each payoff and offset probability for Subject 1 ................................................................. 94

22. The probability of detection versus the probability of a false alarm for each payoff and offset probability for Subject 2 ................................................................. 95

23. The probability of detection versus the probability of a false alarm for each payoff and offset probability for Subject 3 ................................................................. 96

24. Optimal β and ß measured for Subject 3 under the four conditions of signal probability and payoff ................................................................. 97

25. Plot of the z scores of the probabilities of false alarms and detections for the three subjects in the four conditions ................................................................. 98

26. Experimental design for Experiment 3 ........................................................... 105

27. The predicted mean movement times (seconds) in trial Blocks 2 and 3 for the three subject groups. Using loading effects on movement time as a convergent measure for determining Group 1's change in detection performance ................................................................. 108
28. The positive response ("yes") rate for Group 1' versus the combined Groups 2' and 3' for trials in Blocks 1 and 2..........................113

29. The probability of detection versus the probability of a false alarm for each subject in Group 1.................................116

30. The probability of detection versus the probability of a false alarm for each subject in Group 2........................................117

31. The probability of detection versus the probability of a false alarm for each subject in Group 3........................................118

32. Detection rates by the four groups across the five pairs of failures in Block 2........120

33. Block 2 false alarm rates for the four groups in twenty-trial sets.......................122

34. The false alarm rate for Group 1' and Groups 2' and 3' combined for Block 1 and Block 2..........................................................124

35. The mean movement times (seconds) for each gate in Blocks 2 and 3 for the three subject groups. Both Block and Group effects were significant........126

36. The mean movement time (seconds) for each gate for the four groups for Block 2 and 3.........................................................128

37. The mean error rate for the four groups for Block 2 and 3.................................130
LIST OF TABLES

Table Page
1. The proportions of failures detected by experience level and control type ............73
2. The questionnaire ratings of subjects in the experienced, inexperienced, manual, and automatic conditions from very easy (1) to very difficult (10) ............76
3. Detection rates, false alarm rates, d' and β for the four conditions of signal probability and payoff.................................................................93
4. Detection rates, false alarm rates, and positive response rates, for each subject in the 100 trials of Block 1.................................................................110
5. Detection rates, false alarm rates, and positive response rates, for each subject in the 100 trials of Block 2. (The subjects denoted with an * performed at a chance detection rate and were placed in Group 4.) ........................................115
6. Detection rates, false alarm rates, and positive response rates, for each subject in the four groups for the 100 trials of Block 2................................................119
CHAPTER 1

INTRODUCTION

The Experienced Pilot in the Automated Cockpit

Technology is progressively changing the role of the human from an active controller to a high-level decision maker and supervisor of automation (Danaher, 1980; Sheridan, 1987; Wiener & Curry, 1980). The implications of this shift in roles have been considered by many (Bainbridge, 1987; McCormick, 1976; Woods, 1988). Perhaps, as Kelley suggested, "Appropriate automation of inner loops can result in better control over the higher-level goals of the system." (1967, p. 253). However, as the use of automation increases, an operator's ability to detect a failure has been shown to diminish (Wickens, 1984; Wickens & Kessel, 1979, 1981). Parasuraman, Molloy, and Singh (1993) caution against the assumption that there is redundancy in the human-machine system as a result of the human monitoring the automation. They point out that research shows that, "the presence of automation affects the efficiency of human monitoring." (p. 21). This decrement may be the result of inadequate or inappropriate feedback provided by the automated system, especially in anomalous situations (Norman, 1990). In aviation, as in other domains, there has been a strong plea to keep the pilot 'in the loop' (Wiener, 1988), i.e., provide an operating environment in which the operator has sufficient feedback to
maintain awareness of the situation. Many questions arise as this goal is pursued. Does the information available to a pilot executing manual control differ from the information available when the pilot controls the aircraft using the autopilot? Does the pilot utilize different information as the automation becomes more trusted? Is the information optimized for automatic control?

The manual control loops are very well understood. There are pilot models for the inner control loops for pitch, roll, yaw and power, and for outer control loops, e.g., heading, vertical flight path and speed. What is not well understood is how pilots control aircraft when these control loops are nested deep within the automation and the pilot's task is now one of supervisor, rather than direct controller (Figure 1). The information contained in the feedback loops of automated cockpits is not well delineated. The information utilized by pilots is even less understood. The time constants are unmeasured. Manual control provides immediate feedback after a control input is made. The feedback from automatic control input to computerized flight management systems may not occur until the end of a very long flight. The level of control varies within subtasks. The skills required may be quite different from those needed to fly traditional aircraft. While the task has changed, the pilot still must monitor and react to complex dynamic stimuli. Failures must be immediately detected and properly responded to. The research described here addressed failure detection under automated and manual control.

As pilots gain experience with the information provided in modern cockpits, they may adjust which information they attend to. Supplemental information that is redundant to primary information, but easier to use, may result in the primary information being ignored or at best sampled very infrequently (Moray, 1986). For example, the basic information or 'raw data' for an instrument approach to landing is contained in the localizer for lateral
Figure 1. The control loops for manual and automated aircraft.

control and the glideslope for vertical flight path angle. These indicators depict the angular difference between the aircraft's current position and the desired position. However, most transport aircraft also have flight directors that give pilots information regarding the desired pitch and roll of the aircraft to null lateral and vertical position errors. Pitch and roll are parameters that pilots have direct control over. Not surprisingly, the flight director has come to be frequently used as the primary source of flight guidance information during an approach. The problem lies in the fact that an aircraft can be well off the desired vertical
profile and lateral course and yet the flight director will show no discrepancy between the
desired and current status, if the commanded vertical flight path and crab angles are being
flown. The only source of the amount of deviation is the localizer and glideslope; the
deviation may go undetected if the pilot does not cross-check this information.

Another decrement to attention to primary information sources is attentional capture
(Kahneman & Henik, 1981). This phenomenon occurs when information is distributed
across different perceptual groups. For example, pilots landing a simulated aircraft did not
see the airplane positioned on the runway when using a head-up display for the
presentation of the aircraft’s instrumentation, e.g., glide slope and localizer offsets, even
though this information was presented on the same display as the simulated outside scene
found that with higher levels of workload, subjects were more likely to demonstrate
attentional capture for both objects in the outside scene and in the head-up display.
McCann and Foyle (1996) determined that the relative motion of the objects was a critical
grouping factor. They determined that scene linking superimposed symbology eliminated
the attentional capture of the information presented on the head-up display. Presenting
flight guidance as road signs and pathways in the sky, allows the outside information to be
incorporated in the pilots attentional capture.

In another example alerting information has become the primary source of information,
rather than the backup, as intended by designers (Wiener & Curry, 1980). The primary
source of information for altitude control is the altimeter. Upon receiving an altitude
clearance, the pilot must remember the altitude and watch the indication on the altimeter
until reaching the target altitude. (A climb from takeoff to cruise altitude is approximately
25 minutes for most transport aircraft.) The altitude alert is designed as a backup to alert
pilots first that they are approaching the target altitude and, if necessary, that they have exceeded it. This alert is a device the pilot sets to the target altitude. This alert provides a visual digital display of the target altitude, and a visual display (light), and in some aircraft a tone, when within approximately 800 feet of reaching the target altitude. There is an additional light and tone if the altitude is exceeded by approximately 300 feet. The cues provided by the altitude alert are much more salient than those found on the altimeter. Also, the permanent digital display is less volatile than the pilot’s memory of the clearance altitude. For these reasons, the altitude alert has become the primary source of information when climbing or descending to a clearance altitude. The designed redundancy and backup do not exist in practice.

The above examples demonstrate possible negative side effects of a person’s familiarity with a task and developed confidence in the reliability of the secondary information. This tendency to reduce the information requirement is a pragmatic and a logical step in an operator’s attempt to reduce mental demands and maximize the efficiency of the task. This step has extreme consequences for human-machine interface design. The challenge falls on the designers to provide primary cues that are salient and easily usable. The addition of effective secondary cues is also often necessary to provide backup redundancy.

These studies addressed this phenomenon. As in the above examples, the operators redefined the task and ignored the primary cues as they observed their redundancy with the secondary cues. The experiments measured the detection of a disagreement between the primary cues and the easier-to-use secondary cues.

These studies measured the likelihood of failure detection by experienced and inexperienced operators, for the task investigated. When the secondary cues became
unreliable, the task required the use of the primary cues. The omission of the step of comparing the secondary and primary cues was more prevalent in highly experienced operators than in novices. Performance differences in both speed and accuracy were measured as the operators shifted from skilled responses (to the secondary cues) to unskilled responses (to the primary cues). Effects such as interference were identified.

These studies also lend insight into a human's ability to detect failures when a machine is performing the task. Previous studies have shown that changes in a system could be better detected under manual control conditions and that monitoring of automatic control is subject to distractions from other tasks (Wickens & Kessel, 1981; Parasuraman, Molloy, & Singh, 1993). The role of the human, active performer or monitor of automation, was an independent measure in the first study, to determine the effects of the operator's role on failure detection. This variable was crossed with two levels of task familiarity, experienced and inexperienced. Interactions in these two factors were investigated.

In summary, the main thrust of these studies was to determine: 1) The failure detection performance by experienced versus inexperienced operators. 2) The number of failures detected by operators manually performing the task versus those monitoring a machine's performance of the task. 3) The existence of any interaction between the four conditions of experienced versus inexperienced behavior with manual versus monitoring behavior.
Literature Review

The following section reviews the literature on the behavior taxonomies and the effects of experience on behavior, including erroneous behavior. Research on failure detection is summarized. The implications of expertise and automation are explored. The challenge of designing and training for error-free expert performance is described.

Behavior Taxonomies

Many behavioral taxonomies have been developed to provide insight and structure to the understanding of human performance. These taxonomies are pertinent to this research in the treatment of levels of experience. The taxonomies are generally aimed at describing either human error or the learning process. Taxonomies that address the steps involved in proceeding from a perception to an action, fall under both categories. Error taxonomies typically have a dimension that describes the perception-action process and one describing the types of errors that occur at the different stages of that process. Although labeled in different ways, many scholars of human performance describe the differences in behaviors as a result of experience. It is theorized that all types of behaviors can traverse the dimension from controlled, effortful, conscious behavior, at the novice phase, to the automated, effortless, unconscious behavior of the expert. The location of the behavior on that dimension is based on the operator's experience and not on the type of task.

Reason (1987) developed a classification framework that combines basic error tendencies and information-processing stages to create primary error groupings. These
could then be combined with situational factors to form predictable error classifications. Basic error tendencies range from ecological constraints, change enhancements, resource limitations, schema properties, to strategies and/or heuristics. The information processing domains are: (1) sensory registration, (2) input selection, (3) volatile memory, (4) long-term memory, (5) recognition processes, (6) judgmental processes, (7) inferential processes, and (8) action control. The matrix of basic error tendencies and information processing domains forms categories of primary error groupings. These are: false sensations, attentional failures, memory lapses, inaccurate recall, misperceptions, errors of judgement, reasoning errors, and unintended words or actions.

Rouse and Rouse (1983) also have an error taxonomy that accounts for stages in the perception-action cycle. The behavioral categories are: (1) observation of system state, (2) choice of hypothesis, (4) testing of hypothesis, (5) choice of goal, (6) choice of procedure, and (7) execution of procedure. The error categories are specific to each of the categories in the perception-action cycle. Generally, errors at each stage of behavior are described as the incomplete, incorrect, unnecessary, or lack of performance of the stage, e.g., choice of procedure.

Norman (1981) did not classify the early stages of the perception-action process. His taxonomy addresses the sequence from (1) the formation of the intention to (2) the activation of a schema or action sequence; this schema then gets triggered and the result is the action. Errors that are characterized by unintended actions, or slips are specifically addressed. A collection of slips is divided into categories by the stage of occurrence in the intention-action sequence. The slips are further classified by the nature of their presumed source. Under category 1 are the errors made in the formation of the intention, e.g., erroneous classification of the situation and ambiguous or incomplete specification of the
intention. Category 2 encompasses errors resulting in the faulty activation of schemas. These include unintentional activation of an inappropriate schema and loss of activation of appropriate schema. Category 3 errors are those resulting from faulty triggering of active schemas. Included in category 3 are: triggering at an inappropriate time, e.g., prematurely or in combination with another schema, and failure to trigger, e.g., the action is preempted by competing schemas. Cognitive and perceptual feedback are required to recognize and correct slips. However, this is complicated by the fact that intentions are high level and actions are lower level, e.g., driving home is a series of movements of the steering wheel making corrections to the course.

Rasmussen (1983) categorized behavior by the stages in the perception-action process along three different paths. The distinction is made between skill-based, rule-based, and knowledge-based behavior. This progression is from well defined, well-practiced behaviors to less well defined, new behaviors. Skill-based behaviors are automated, i.e., they require no short term memory or attention. These behaviors are highly practiced and have consistent and unambiguous stimulus-response relationships. Skill-based behaviors are very accurately and quickly performed. These behaviors involve only two steps: (1) feature formation and (2) automated sensory-motor patterns, e.g., the act of eating a sandwich. Rule-based behaviors are performed according to previously stored rules or procedures. They involve both of the steps of skill-based behavior with some intermediate steps. The full process is: (1) feature formation, (2) recognition, (3) association state or task, (4) stored rules for tasks, and (5) automated sensory-motor patterns, e.g., the act of carving a turkey. Knowledge-based behaviors are goal controlled. They include developing a plan and evaluating the plan by testing it against others either by trial and error or conceptually. Knowledge-based behaviors involve both of the steps of skill-based behavior, one step from rule-based behavior, and intermediate steps. The full process is:
(1) feature formation, (2) identification, (3) decision of task, (4) planning, (5) stored rules for tasks, and (6) automated sensory-motor patterns, e.g., preparing the Thanksgiving meal.

Rasmussen (1987) presents a three-step process for error analysis. Both intentions and judgements of the situations are assessed. The first step is to describe the events involving human malfunction. These begin with the causes of human malfunction, and lead to mechanisms of human malfunction, and internal human malfunction, and result in the external mode of human malfunction. The second step is to evaluate the element of the internal cognitive decision process that is responsible for the malfunction. This element comes from the steps delineated above as those involved in skill-based, rule-based, and knowledge-based behaviors. The final step in the analysis of errors is to characterize an event by the psychological mechanism of cognitive control: anatomical properties, physiological functions, psychological mechanisms, cognitive and affective, mental information processing, and subjective value formation.

Rasmussen's taxonomy adapted for specific applications. Pew, Miller and Feehrer (1981) adapted Rasmussen's (1983) taxonomy to analyze operator decisions in the nuclear control room. The changes they made are slight and insignificant to this discussion. Skill-based behaviors proceed directly from activation to the execution of a response, without conscious thought. The act is accomplished without the operator being aware of each of the steps being performed. An example in the control room is turning off an annunciator alarm. A rule-based behavior involves conscious recognition of a pattern of displayed variables resulting in a mental representation of the system state. The selection and execution of the appropriate procedure are based on that recognition. Knowledge-based behaviors involve the full range of human decision making: interpretation of the
system state, hypothesis generation, and evaluation of alternative strategies. Behaviors of this type occur when there is ambiguity about the system state or the appropriate procedures to be employed. These decisions require extensive knowledge of the system.

Skill based behaviors were used in the normal operations of the nuclear control room. Emergency states required rule-based behaviors, if there were rules that covered the emergency state, and knowledge-based behaviors if no rule could be applied. Suggestions were made to improve decision making. Skill-based behaviors could be improved by better displays, controls, and work space layout. These behaviors could also be improved by the introduction of computerized safety monitors and disturbance detection-classification systems. Rule and knowledge-based behaviors could be improved by better operator training on the details of plant operations, better displays, and by decision aids and rule support systems.

After reviewing many taxonomies, Jensen and Chappell (1983) also adapted Rasmussen's (1983) taxonomy for the analysis of pilot errors. In their taxonomy, skill-based behaviors included: vehicle control, distance and speed judgments, and geographic orientation. Rule-based behaviors included the management of communications, navigation, fuel, powerplant, vehicle configuration, displays, and the autopilot. Knowledge-based activities were self evaluation of skill, knowledge, physical and psychological condition, and navigation planning, hazard assessment, assessment of attention requirements, assessment of aircraft and ground system capabilities, mission priority adjustment, illegal operations, and communication. A category of behavior was added, monitoring, which included the monitoring of: traffic or obstacles, instruments, environment, navigation, and hazards.
Comparison of taxonomies on the perception-action dimension. It is the consensus of the authors cited that behavior classification by stage in the perception-action process is useful for understanding performance. In addition, all those analyzing errors, except Reason (1987) and Rouse and Rouse (1983), realize the importance of classifying errors according to the level of skill of the behavior, to use Rasmussen's (1983) terms, by knowledge-based, rule-based, and skill-based behaviors. In general, error analyses should always locate the operator within the skill-based, rule-based, and knowledge-based experience level before considering the erroneous steps in the perception-action process, since they differ. It would be meaningless to consider an error in planning when a person is experienced and operating at a skill-based level. Reason's behavioral categories have components of both Rasmussen's rule-based and knowledge-based behaviors, e.g., recognition and judgmental processes, respectively. Rouse and Rouse describe behavior similarly to Rasmussen's knowledge-based category including the testing of the hypothesis and the choice of a goal. Norman (1981) addresses only skill-based behaviors. His taxonomy starts with an intention and ignores errors in the determination of goals in decision-making or in problem-solving.

The two instances of the specific application of Rasmussen's (1983) taxonomy demonstrate the efficacy of distinguishing the type of behavior in evaluating human error. These applications illustrate that behaviors vary as to what steps are taken in the perception-action process as a result of the operator's experience with the situation and the possible outcomes. Pew, Miller and Feehrer (1981) observed skill-based behaviors such as turning off an annunciator alarm and rule or knowledge-based behaviors in emergency situations. Jensen and Chappell (1983) found pilots performed well-practiced behaviors in a skill-based manner, such as holding an altitude. They found tasks with more ambiguous stimulus-response relationships, e.g., powerplant management, to be rule-based.
Behaviors such as navigational planning required a knowledge-based level of processing, since each flight has unique conditions that have to be assessed.

The Stages in Learning a Skill

Unlike the taxonomies described above, the following authors did not develop categories of human error. These categories describe the stages in learning a skill. Fitts and Posner (1962) described learning a skill as a three-stage process: (1) the early, cognitive phase, (2) the intermediate, associative phase, and (3) the final, autonomous phase. Fitts and Jones (1945) had developed an error taxonomy of pilot errors. These error categories were closely tied to the observable event. The stages of learning or the stages of the perception-action process were not considered in the earlier work.

Schneider and Shiffrin (1977) described two behavioral categories, (1) controlled and (2) automatic. The distinction is made on a dimension from well defined, well-practiced behaviors to less well defined, new behaviors. Behaviors that are characterized as a controlled process are under short term memory control and require active attention. Therefore they are capacity limited. These behaviors can be performed in new situations and can be altered during execution. Behaviors that are characterized as an automatic process have the following properties: the action sequence always becomes active in response to a particular input and the sequence is activated automatically without the necessity for active control or attention.

Klein (1989) models the decision making behavior of experts. The decisions studied were made in situations with time pressure, shifting scenario features, and where the
decision maker had personal responsibility for decision outcomes. Experienced decision makers, even in emergency situations, e.g., fire fighting, did not use an analytical process such as generating a variety of optional actions and contrasting the strengths and weaknesses of each action. Experts (1) compare the situation to those they are familiar with, if it is recognized as typical they (2) assess the situation according to (a) plausible goals, (b) critical cues and causal factors, (c) expectancies, and (d) typical actions. The test of expectancies challenges the initial hypothesis. Note that the tendency to ignore data that serve to reject the initial hypothesis, found by other investigators, was not replicated by Klein's study of experts. The next step is (3) serial evaluation. This is the serial assessment of options until a satisfactory one is found. This is different from, and less time consuming than the comparison of a set of options. The final stage is (4) progressive deepening, i.e., the process of imagining how an option will be carried out within a specific situational context, a mental simulation or "instant preplay".

Fitts and Posner's (1962) early cognitive phase, intermediate associative phase, and final autonomous phase, are condensed to two phases by Schneider and Shiffrin's (1977), controlled and automatic behaviors. In the context of Rasmussen's (1983) taxonomy, behaviors progress from knowledge-based to rule-based to skill-based with experience. These authors disagree only in the labeling of the stages in this progression.

Klein (1989, 1993) addresses only experts performing well-practiced behaviors. The analogy can loosely be drawn from Klein's "situation comparison and assessment" to Rasmussen's (1983) "feature formation". Klein's serial evaluation and progressive deepening are not represented as separate processes in Rasmussen's skill, rule, or knowledge-based level of behavior. Perhaps Klein has discovered that there are more steps involved in skill-based behavior than has been previously observed.
The Behavior of Experts

Behaviors become more automated and less controlled with practice. However, this can only occur when an unambiguous relationship between the stimulus (the perception of the situation) and the response (the appropriate action) can be developed (Damos, 1988). The advantage of this progression with experience is the reduction in cognitive load, the increased speed of performance, and the increased accuracy (Wickens, 1984). The disadvantage is the lack of conscious monitoring that results in Norman's (1981) slips.

Operators in many domains are highly skilled and perform in very familiar, information-rich, dynamic environments. The behavior of experts in unambiguous, familiar task environments has been described as automatic (Bahrick & Shelly, 1958; James, 1890; Schneider & Shiffrin, 1977; Solomons, 1899), autonomous (Fitts & Posner, 1967), pattern recognition (Allport, 1980; Boy, 1991; Chase & Simon, 1973; Chu, Steeb & Freedy, 1980; Moray, 1986; Pau, 1981; Rouse, 1983; Wickens, 1984), intuition (Edwards & von Winterfeldt, 1986; Gardenier, 1981; Larkin, McDermott, Simon, & Simon, 1980), and recognition-based decision making (Klein, 1993). "Many authors have argued that expertise is synonymous with highly developed pattern recognition abilities." (Rouse, 1983, p. 614). Rasmussen (1983) categorized this type of behavior as "skill-based". He theorized that experience causes patterns in the environment to be linked to action sequences. These patterns are then perceived as time-space "signals" that trigger smooth, automated, highly integrated patterns of behavior. "Skilled performance rolls along without conscious attention or control" (p. 259). The response is automatically triggered by the situation (Allport, 1980). Fitts (1964) draws the analogy between highly practiced skills and reflexes; both run without verbalization or conscious content. This skilled
behavior has been characterized as fast, effortless, autonomous, obligatory, consistent, stereotypic, and unavailable to conscious awareness (Logan, 1988a).

The characteristics of skilled behavior permit many tasks to be performed accurately and quickly without an unmanageable workload. As the skill level increases, the requirement for cognitive involvement diminishes (Moray, 1986). Kelley (1967) suggests the failure of research in the past to identify relationships between levels of effort and performance may be largely attributed to the fact that well-practiced behaviors are performed with little cognitive demand.

Examples of expert behavior. Those studying the behavior of experts provide specific examples of the robustness of this phenomenon across domains and tasks. Larkin, McDermott, Simon and Simon (1980) report that grand master chess players can recognize approximately 50,000 patterns and that actions or strategies are likely to be evoked when these patterns are recognized. Moray (1986), in reference to supervisory control tasks, observes that as operators gain experience, their internal models evolve, and ultimately the behavior becomes automatic. When this occurs, the human is an efficient and optimal decision maker. He points out that operators are not aware of the strategies of highly practiced behaviors. Gardenier (1981) stated that some of the behavior in ship navigation and pilotage “seems to be intuitive” (p. 51). Klein (1989) observed the decision making of fire ground commanders and interviewed them as soon as possible after the fire-fighting. The stimuli attended to and the decisional boundaries of the environmental conditions were documented. These individuals were able to observe the complex stimuli of a fire and, if a pattern was recognized, their experience provided a direct mapping to the appropriate action. They characterized their behavior as “acting and reacting on the basis of prior experience” (p. 49).
Skilled complex behaviors. These brief examples are indicative of the variety of behaviors that can become skill-based. These examples also demonstrate that the complexity of the task does not appear to be predictive of whether the task can be accomplished with skill-based behavior (Fisk, Ackerman, & Schneider, 1987; Fisk & Schneider, 1981). Simple behaviors, e.g., turning a knob to open a door, readily become skill-based or automatized. The learning of complex behaviors is slower to progress from knowledge-based to rule-based to skill-based. Shiffrin and Dumais (1981) posit that skilled behaviors become automatized and that any behavior for which the stimuli and response are consistently mapped can become automated, that it is not the nature of the task, but rather the amount of experience, that determines the type of behavior used to accomplish the task.

Skilled components of the perception-action sequence. Wickens (1984) has described some automatized behaviors as being open-loop, provided they are: highly practiced, low in attention demand, unambiguous in the mapping of stimulus to response, and have a consistent outcome. Czerwinski, Lightfoot, & Shiffrin (1992) describe the parsing of the stimuli that evolves with practice. As behaviors become more learned, the relationship between the stimulus and response becomes less ambiguous (i.e., the relationship between the feature formation and the automated sensory-motor patterns). Skilled behavior is an interaction between enhanced evidence accumulation and strategic response criteria (Strayer & Kramer, 1994).

For many tasks, the entire process of pattern recognition, decision making, and action can become skill-based with practice. However, the nature of certain tasks precludes the response from being part of the skill-based behavior, either because no immediate response is required, or because the same situation may require different responses under different
conditions. For these tasks, the recognition of the situation can be skill-based but the performance of the response may be rule or knowledge-based. For example, Rasmussen (1981) found that for diagnosing process plant failures "pattern recognition ... can efficiently identify familiar system states and disturbances directly, but it is also used frequently during, e.g., topographic search to guide the tactical decisions" (p.248). Pew, Miller and Feehrer (1981) claim that behaviors in response to emergencies are rule or knowledge-based. This is perhaps due to a lack of experience with the emergency situations. Klein (1989) uses different labels but describes expert behavior in emergency situations as a more direct process between perception and action, or in Rasmussen's terms, skill-based: feature formation leading to automated sensory-motor patterns. Klein describes a process called "mental simulation" in which an expert recognizes a situation and then evaluates an action by imagining the performance and consequences of the action. Boy (1991) analyzed experts diagnosing faults on the orbital refueling system of the space shuttle and found "situational patterns" arose from the compilation of the "analytical knowledge" of the novice. He describes the situational knowledge as the "essence of expertise," which Boy says corresponds to Rasmussen's skill-based behavior. Boy has characterized the following of procedures (rule-based behaviors) as a useful method of operation until experience permits skill-based behavior. This is another example of a very complex skill-based behavior and is different from Klein's findings only in that Klein found continual reevaluation of the "pattern" and the action. Many symptom-diagnosis relationships are represented in the patterns the experts recognize. Missing data can be dealt with quickly by a "best fit" matching.

**Skilled behavior by pilots.** Skill-based behaviors are prevalent in flight deck operations. Pilots are highly skilled and the nature of the task requires a situation to be consistently mapped to an action as a matter of standard procedure. These procedures are
performed many times in exactly the same way. Since many of a pilot's tasks are highly practiced, in a consistent task environment, the argument has been made that these flight tasks are being performed automatically (Lauber, 1989). As behaviors become more practiced and able to be performed through pattern recognition, cognitive load is reduced and the speed and the accuracy of the performance are increased (Wickens, 1984). This level of skill enables pilots to perform many tasks simultaneously, without excessive workload. Training methods and pilot-aircraft interface design can positively impact the probability of tasks being performed in a skill-based manner.

**Failure Detection**

In the aviation domain, the detection of and response to an anomalous condition are important operator functions. The literature on both failure detection (e.g., Rasmussen & Rouse, 1981) and on alerting systems (e.g., Berson, Po-Chedley, Boucek, Hanson, Leffler, & Wasson, 1981) is directed toward the human's ability to identify the existence of a failure and respond appropriately. Grossberg (1982) argues that detection is due to an alerting mechanism being triggered when a situation fails to match the expected template. Allport (1987) posits that the alerting mechanism subsequently brings about a disengagement in the previous mode of behavior and an engagement in the new appropriate behavioral mode. Detection rates vary as a result of the type of event. Detection of simple events was found by Thackray and Touchstone (1989) to show no decrement over time in an air traffic control task. Detection of complex events did show a decrement with time on the task. Detection is more difficult in the case of gradual degradation (Moray, 1986). Sometimes detection is dependent on the operator observing that the appropriate stimuli or
the pattern is absent. It is often difficult for the operator to delineate the information used for failure detection.

Computerized monitors can improve monitoring performance (Parasuraman & Wisdom, 1985). However, the accuracy of an automated detection system, for alerting the human of a failure, can have serious implications on the performance of the human plus the alerting system. Sorkin and Woods (1985) analyzed the performance of the human with the alerting system, a combination they referred to as the "alerted-monitor system". They pointed out that if the operator has knowledge of the a priori probability of a signal, knowledge of the probability of a signal given that the alert has sounded, and knowledge of the costs of misses and false alarms, the optimal decision criterion can be determined. They showed that the operator-alerter system performance is severely impacted by a high level of false alarms from the automated alerting system.

Vigilance

Closely related to the literature on failure detection is the literature on vigilance. See Parasuraman, 1986 for a review. This area of research concentrates on the performance of operators in conditions where signals (failures) are rare. Detection performance has been shown to be low when an operator has experienced a long period of no or relatively few signals. This phenomenon has been labeled the vigilance decrement. This declining hit rate observed in perceptual vigilance experiments is often the result of an increasingly strict decision criterion rather than of decreasing sensitivity. The two major variables affecting the decision criterion in vigilance tasks are monitoring time (time on task) and a priori signal probability. The signal probability is usually low. The subjects criterion increases
(toward greater strictness) with time on task. Criterion shifts in vigilance and monitoring tasks can be interpreted in terms of changes in expectancy and subjective probability.

The criterion increment is associated with a decrease in both hits and false alarm rates; operators respond positively less often. Increasing the signal probability leads to a lowering of the decision criterion. The criterion increment is smaller if observers are given sufficient training with signal probabilities that are the same as those used in the main task. However a residual criterion increment may still occur.

**Experience and Failure Detection**

Operators in a familiar situation, who have experienced a high degree of reliability in the past, may be less likely to detect a failure. This detection decrement may exist because they have stopped looking for the failure or because they have changed their decision criterion, such that the situation must be more extreme to be classified as a failure. Pilots reporting to the National Aeronautics and Space Administration's (NASA) Aviation Safety Reporting System (Reynard, Billings, Cheaney, & Hardy, 1986) describe failures to recognize key information in situations they had a high level of experience with. Their lack of detection led to many different safety consequences. One captain, who had flown into an airport for over 20 years, landed on the wrong runway (Report number 209611). Other pilots also blame familiarity as a cause of their failure to prevent safety incidents, stating: "Familiarity breeds error." (Report number 244894) and "Familiarity breeds laziness." (Report number 251272).

Parallels can be drawn between skilled behaviors in the cockpit and the laboratory research on automaticity (e.g., Dulany & Gordon, 1992; Fisk, Ackerman, & Schneider,
1987; Grose & Damos, 1988; LaBerge, 1981; Logan, 1976, 1988b; Logan & Stadler, 1991; Schneider & Fisk, 1982, 1983; Schneider & Shiffrin, 1977, 1985; Solomons & Stein, 1896). Much is known about the development and the stimulus-response conditions under which automaticity can occur in discrete laboratory tasks. Traditional automaticity paradigms change the stimulus-response mapping from consistent to varied, and observe the performance effects. Detection of a change in the task is not required. What has not been investigated is the magnitude and direction of skill effects on the detection of a subtle change in a task (Moray, 1986). The change must be detected; the operator must realize that the skilled behavior is no longer appropriate and begin to perform the new behavior.

Once detection has occurred, response may be more difficult when an inappropriate response has become very well practiced and must be suppressed. The operator must perform the appropriate actions for the failure situation. The information required for familiar tasks may be different from that needed for failures. As an example, transport pilots are trained to interrupt skilled, well-practiced action sequences when a procedure other than the normal skilled behavior is required (Norman, 1992). They do this by constructing barriers that prevent the well-practiced actions. When a generator has failed on a Boeing 727 aircraft, a different checklist is required for the approach to landing. This is a very busy phase of flight and strong habit patterns are relied upon to accomplish all the necessary tasks. It is the procedure of some airlines to put a paper cup over the flap handle so that it feels and looks different and alerts the crew not to lower the flaps as they normally would, without performing the additional checklist. This need to interrupt well-practiced behaviors is not generally a part of interface design requirements (Chappell, 1991). Because of this design deficiency, creative pilots have developed cognitive aids to enhance their performance of flight deck duties. As Norman reports “Most of the aids in the cockpit are casual and informal, or invented by crews in response to their own experience with
error.” (1992, p. 165). A better understanding of these errors can assist the development of design guidelines aimed at preventing inappropriate, well-practiced behaviors. The studies described herein examined errors in adaptive performance once failure detection had occurred.

**Automation and Failure Detection**

As systems are becoming increasingly automated, the operator's involvement with the task is changing. Satchell (1993) labels this detachment between the pilot and the automated aircraft "peripheralisation". The operator's main task in an automated system is fault detection (Moray, 1986). The operator's mode of control, from manual to automated, has been shown to affect performance in detecting a failure. Automation has the side effect of inducing complacency (Parasuraman, Molloy, & Singh, 1993; Satchell, 1993). The work of Wickens and Kessel (1979, 1981) suggests that the manual mode of control will result in better failure detection than when monitoring machine control. Other work suggests the opposite result. Ephrath and Curry (1977), Ephrath and Young (1981) and Fuld, Liu, and Wickens (1987) demonstrated better failure detection when subjects monitored an automated control than when manually controlling. This improved performance was attributed to the decreased workload in the automation condition. Ephrath and Young (1981) suggest that workload is the significant determiner of whether a task should be manual or automatic, to facilitate failure detection. If workload is low, detection will be best with manual control. If workload is high, detection will be best if the task is automated. Previous research leaves many questions to be resolved in determining the optimum level of automation to ensure failure detection. The present studies utilized a self-
paced manual task that removes the issue of workload by allowing subjects to perform the
detection task and the manual control task at their desired pace and even serially.

Research on failure detection has not explored the interactions between control mode and
experience level. In view of the differences in cognitive operation among operators at
different levels of skill, this factor may significantly influence the optimum level of
automation. Experienced operators may find the workload to be low and therefore show
better failure detection under manual control, while inexperienced subjects may detect better
while in the automated control mode. This research addressed the effect of experience level
on failure detection under automated and manual control.

**The Use of Redundant Cues**

Skill or experience level may affect the operators' use of primary and secondary sources
of information, when the two sources are redundant. Bahrick and Shelly (1958) state that
redundant elements in tasks are taken advantage of to change the method of responding to
become automatic. Slovic and Lichenstein (1971) report that when two cues were
completely correlated, subjects typically used both and weighted them equally. When the
cues disagreed, however, the subjects either focused on only one or used other cues to
resolve the conflict. Wyer (1970) defined redundancy as the extent that the probability that
the cues are found together exceeds the probability that the cues are found separate. In his
study, he found an interaction between favorableness of the cues and redundancy. When
two cues were less redundant and highly unfavorable, the subjects' ratings were
significantly more negative. When the cues were less redundant and were favorable the
ratings were more positive, although not significantly. That is, the evaluations were more
polarized when the cues were less redundant. Busmeyer, Myung and McDaniel (1993) manipulated the validity of one cue and observed the effect on another cue. They quantified a cue competition effect. They showed that by increasing the validity of one cue, the effectiveness of another cue was decreased, even when the cues were independent. J. A. Modrick (personal communication, September 13, 1991), in an unpublished study, found that when a cue correlated 60% with another source of information, it was over-used for prediction. When it correlated only 30%, it was under-used.

When redundancy is combined with a difference in the saliency of the two sources of information, the more salient information will engage the operator's attention and disproportionately affect judgements. This is termed the salience bias in the decision making literature (Taylor, 1982). When the information is no longer redundant, the discrepancy between primary and secondary sources of information must be detected. These studies investigated differences between experienced and inexperienced operators in the detection of this discrepancy.

**Implications for Cockpit Design**

A better understanding of how experts perform familiar tasks will enable the design of efficacious information and control systems. To account for the different types of behaviors being performed by individual operators, a human-machine system design should consider the level of experience of each operator and present the information optimally for their type of behavior. A design that is optimal for a novel task may be inappropriate for a pattern recognition task. The nature of the task should determine the type of information display (Casner, 1991). Conversely, the way information is presented
can influence the mode of problem solving (Rouse, 1983; Hammond, Hamm, Grassia, & Pearson, 1987). When the operator must recognize a familiar situation, the term 'pattern recognition' reflects the importance of a spatial presentation of information for certain tasks. Experts are able to recognize a spatial pattern and associate that pattern with the appropriate action. Woods (1988) cites an illustration of this relationship from a nuclear power plant, where annunciator alarm tiles representing a component's status were contained in panels that were organized according to functional units. When the tiles and panels were replaced with a computerized system that chronologically listed the status changes, the spatial organization of the information was lost. Operators were no longer able to determine the overall problem based on the pattern of lights. The annunciator tiles had to be replaced.

Pew, Miller and Feehrer's (1981) suggestions for the nuclear powerplant could be applied to many systems. They suggest that skill-based behaviors correspond to most normal operations and that these behaviors can be improved by better displays, controls, and work space layout, and the introduction of computerized safety monitors and disturbance detection and classification systems. These improvements serve to enhance "pattern recognition". They suggest that rule and knowledge-based behaviors are improved by better operator training on the details of plant operations, better displays, and by decision aids and rule support systems. Vicente and Rasmussen (1988) state that interfaces for skill-based behaviors should “Allow the operator to act directly on the display so that interaction can take place via time-space signals.” p. 255. Information displays should utilize the expert’s ability to recognize spatial patterns and quickly associate those patterns with the appropriate response. Note that information presented through other sensory modalities can have other patterns, e.g., temporal patterns in audition.
It is important for the human interface to provide the pilot with reliable triggering events that are associated with the skill-based behaviors. The absence of unique stimuli that trigger even a well-learned behavior, can lead to the omission of that behavior. Lauber (1989) describes the scripts of behaviors that are part of normal flight procedures and the sequential triggering events that many behaviors rely upon. Without a specific event to trigger the action, an airline crew failed to perform the taxi checklist and attempted to takeoff without extending the flaps and slats, resulting in a fatal accident. Triggering events can be structured into the pilot-aircraft interface, and the operating procedures (Degani & Wiener, 1990) to promote the performance of skill-based behaviors.

Another implication for system design is that warning systems should not activate unnecessarily, so that the response to cancel the warning, e.g., silencing the landing gear unsafe horn, does not become skill-based. If this behavior is skill-based the horn would be quickly and accurately silenced without recognition of the unsafe condition. The high number of gear-up landings may be better understood by measuring the behavior of silencing the warning horn than by asking the pilots why they did not hear the warning.

In some situations, a different response is required than that which has become associated with a specific situation. Unfortunately, it may be difficult to prevent the activation of a preprogrammed action sequence (Norman, 1981). The operator must detect the need for an alternative action. Langer (1989) offers techniques for promoting this state of “mindfulness”. There are several implications for the human-machine interface when it is desirable to prevent skill-based performance. The prevention may not occur unless a salient stimulus blocks the highly practiced behavior. In the cockpit, pilots have employed alerting mechanisms that interrupt and thus prevent action sequences. As in the example presented above, when receiving an aircraft with a failed generator, airline crews are taught
to put a coffee cup over the flap handle. This is done so that later, during the approach, the flap handle will look and feel different and alert them not to lower the flaps according to the normal landing checklist. This use of a salient stimulus to block the normal behavior was adapted after many instances of pilots failing to recall the need for the non-normal checklist. The designer of the human-machine interface should provide for appropriate barriers to well-practiced behaviors, when those behaviors are inappropriate.

Implications for Training

The goal of training programs is to promote safe efficient operation of the human-machine system. However, the emphasis of most of the research addressing training has been placed on the behavior of novices. This section focuses on experts performing highly practiced tasks. The literature on highly practiced tasks suggests methods for promoting skilled behavior through training. Training programs should concentrate on the practice of emergency procedures so that these behaviors become skill-based, to reduce cognitive load, reduce reaction time, and increase response accuracy (Chappell, 1991). The stimulus in the emergencies, e.g., warnings, should be unambiguous since this is a requirement for automatic behavior. Procedures should be developed to provide triggering events to stimulate well-practiced behaviors (Chappell, 1991).

For many tasks, skill-based performance is desirable. Shiffrin and Dumais (1981) posit that the development of an automatically triggered association between a stimulus and the appropriate response is a major component of skill acquisition in both cognitive and motor tasks. Damos (1988) suggests that “training subjects to automated processing levels is the
only effective method currently available to reduce an operator's information processing load and minimize the decrement in system performance" (p.3).

For some flight tasks, training is most efficient if practice occurs initially on separate parts of the entire task. This permits expertise to develop for the subtask, without being confounded by attention to the entire task or the interaction between its subtasks. Part-task training allows for more consistent training, which promotes the development of skill (Fisk & Gallini, 1989). Once the component tasks are well learned they can be incorporated into complex integrated behaviors that are executed with a single intention (Rasmussen, 1983). However, it is important to later provide practice on the combined tasks in the full-mission context. The addition of concurrent tasks has been found to result in a context effect, causing a slight decrement in the performance of the primary task (Fisk & Rogers, 1988).

Operators of highly automated systems assume the role of monitors to a greater extent than operators of less sophisticated systems. As a result, they may become distracted or complacent and therefore less likely to detect a system anomaly, especially when the anomaly is not announced by an alert. Fisk and Schneider (1981) demonstrated that the magnitude of a vigilance decrement could be greatly reduced with the following training method. The critical stimuli for anomaly detection should consistently appear only as target stimuli, the elemental features of target stimuli should be kept as consistent as possible, pilots should be trained with high target probabilities and few non-target searches, operators should train under moderate speed stress, operators should be provided feedback to maintain highly motivated performance, and operators should be given extended over training (e.g., 10,000 failure detections). They also suggest that short refresher detection tests might be helpful in maintaining a high probability of failure detection.
Summary of Literature Review

In summary, this literature review began by presenting examples of behavior taxonomies. As discussed, the taxonomy developed by Rasmussen (1983) has been used successfully in several applications. Other taxonomies, developed to describe the stages of learning a skill, have been mapped into Rasmussen's categories. The theory has been proposed that the type of behavior is due to the experience of the performer and not the nature of the task itself. Operator experience leads to a strong association between the stimulus and the appropriate response, which is necessary for a behavior to become skill-based or automatic. Studies of experts performing very complex behaviors automatically were cited as evidence for this theory. The literature on failure detection was reviewed, along with the effects of automation on failure detection.

Implications of the operator's skill level for system design and training were discussed. Performance can be improved when training and system design promote skill-based behavior. This type of performance enjoys reduced cognitive load and increased speed and accuracy. Along with the procedures that are taught to operators, it is important to train under conditions that promote the development of skill-based behavioral sequences that are triggered by the appropriate stimuli. Training techniques that have been shown to be effective in producing expert performance include: part-task training, training for monitoring, and techniques for the transfer of knowledge.

In addition, it is important to design an interface that promotes recognition of the triggering stimuli and prevents the skill-based behavior when it is not appropriate. Methods for promoting pattern recognition through human-machine interface design include the use of spatial presentations of information and providing triggering events. In
instances where the familiar, well-practiced behavior is not appropriate and it is desirable to prevent the response, a human-machine interface may require salient cues to alert the operator of the need for suppressing the highly practiced response.
CHAPTER 2

RESEARCH OBJECTIVES

A task was created for these studies that required the subject to place a cursor through the midpoint of an imaginary line between end points that were the primary cues for the task. The midpoint and the tolerance area around it were depicted by secondary cues, which were occasionally misleading, i.e., offset from the center between the primary cues. The subject's task was to detect the presence of an offset in the secondary cues, in addition to performing the tracking task. When the secondary cues were misleading, they were to be ignored and the primary cues used for the placement of the cursor. The task is described in detail in Chapter 3.

Determine the Form of the Learning Curve for this Task

The first objective of the study was to determine the form of the learning curve for this task. It was expected that skill would develop with practice and that this would be reflected in a reduction in movement times with successive trials. The learning curve for the movement times (MTs) was expected to follow the power law of practice (Bahrick, Fitts, & Briggs, 1957; Fitts & Posner, 1967; Logan, 1988b; Newell & Rosenbloom, 1981) which
has been shown to hold for many tasks. The underlying function describing the change in MT with practice was determined. The number of trials required for asymptotic performance was also determined.

Determine Differences in the Failure Detection by Experienced and Inexperienced Operators

In the operational environment, operators become experienced. The effect of experience on failure detection is of great interest. If the development of skill produces a decrement in detection performance, training programs and interface designs must attempt to minimize this effect.

The literature makes conflicting predictions about the effect of experience level on failure detection. After experience has been developed in the absence of a failure, some studies suggest that experienced subjects may be less likely to detect the failure for three reasons:

1) The experienced subjects may have consciously or unconsciously redefined the task. They may omit the task of comparing the secondary cues to the primary cues, since the secondary cues have proved to be redundant. Tracking and detection behaviors compete for attention resources (Sorkin, Kantowitz, & Kantowitz, 1988). This task redefinition allows subjects to spend more resources on the tracking task (Moray, 1986); therefore they can reduce movement time, but failure detection is compromised. This task redefinition is a form of speed-accuracy tradeoff (Bryan & Harter, 1899; Fitts, 1966; Wickens & Kessel, 1981). The result would be worse detection performance by experienced subjects than by inexperienced subjects. If the task of checking for the failure is omitted or performed with less frequency, the experienced subjects are expected to show less sensitivity to the failure,
due to their lack of attention to it. This change in sensitivity was termed "operator sampling strategy" by Sorkin and Woods (1985). The signal and noise distributions would overlap almost entirely; the expected value of the sensitivity measure $d'$ would be near zero. The operational significance of this performance difference is that experienced operators may let a failure progress because they are not monitoring key information, even though that monitoring is a primary component of their task. If experienced subjects add the task of checking for a failure after the first failure, the MTs in the manual condition may increase. See the sections below for a discussion of this effect.

2) The task of moving the cursor through the dots may have become automatic for the experienced subjects (Bahrick & Shelly, 1958; Logan, 1976; Moray, 1986; Schneider & Shiffrin, 1977). They may have a strong stimulus-response mapping that allows them to acquire the target with less cognitive involvement than inexperienced subjects. If the task of checking the alignment between the primary and secondary cues is not included in the automatized behavior, experienced subjects may be less likely to detect the failure than inexperienced subjects (Iba, 1991; Norman, 1981b). This effect should also be manifested in a lower sensitivity to the failure, since the experienced subjects are not performing the task of checking for the failure. The presence of automaticity may be evidenced after the detection of the first failure. Subjects may shift from automatized to controlled behavior, resulting in an increase in movement times. They may occasionally exhibit automatized responses, even though they correctly identify the presence of a failure. See the sections below on MTs to targets with failures, and on capture errors.

3) Another reason to predict that experienced subjects will exhibit poorer detection performance is that they have previously experienced 200 trials in which no failure occurred. Their estimate of the probability of a failure will be very low, and therefore they
may have more strict decision criteria, according to signal detection theory (Green & Swets, 1973; Parasuraman, 1986; Swets, 1973; Wickens & Kessel, 1981). This adjustment in their decision criteria will result in a decrement in their detection of the failure.

The opposing view to these theories is that highly experienced subjects will detect the failure more readily than novices (e.g., Fisk, Lee, & Rogers, 1991). Some posit that experts have more resources available to detect the failure (e.g., Schmidt, 1975), since they require less attention to perform the tracking task. Experts may develop an expectation of the pattern of the information and detect when the stimuli do not match that pattern. They may not be aware of this expected pattern. For example, in a report to NASA's Aviation Safety Reporting System (Report number 140270), an airline pilot on an approach descended below the proper altitude. Observing the outside scene, he/she stated, "I didn't feel right about this, although I wasn't sure why." Roth and Woods (1988) found that experts were able to adjust their attention to guard against cognitive tunneling, i.e., they were able to divide their attention across the required components of the task. If this adjustment occurs, the experts can attend to the task of checking the alignment of the primary and secondary cues and detect the failure more frequently than inexperienced subjects.

There may be no difference between experienced and inexperienced operators in the number of failures detected. Parasuraman (1986), in a review of the vigilance literature, reports trained professionals and practiced volunteers are just as likely to show a vigilance decrement as are naive, unpracticed, or untrained subjects. Shiffrin and Dumais (1981) posit that uncontrolled processes are monitored by controlled processes and therefore the
level of awareness will not differ in the two conditions, and consequently, neither will the failure detection performance.

There is always a concern when doing laboratory research on failure detection that the artificially high number of failures in the laboratory may not be reflective of an operator's behavior in an actual operation, where the failure rates are very low. An encouraging note comes from the research of van de Graaff and Wewerinke (1983) who used extremely different prior probabilities of failures, .2 and .8. Their subjects performed simulated instrument approaches using the autopilot. They found that the prior probability did not affect the time to detection or the magnitude of the failure at the time of detection. This finding encourages extrapolation from laboratory research to real world systems.

**Observe Performance Changes for Experienced Subjects after the First Failure**

*Observe any shift in detection by experienced subjects after the first failure.* Subjects, particularly those that are experienced, may be more likely to detect a failure after the first failure occurs. This effect may be evidenced by an increase in the detection rate for subsequent failures. This improvement can be the result of subjects adding the task of checking for the failure, if they had previously omitted it, as discussed above. An increase in detection performance may also be the result of subjects slowing down to improve their accuracy and may be accompanied by an increase in MT (described below). A change in detection can also be the result of subjects adjusting their estimate of the probability of a failure, and therefore their decision criteria.
If experience in detecting the failure improves detection performance, the training implication is obvious. The necessary frequency of exposure to failures in order to retain a high level of detection performance can be determined for a task. However, if this increment in detection performance is accompanied by slower action, the consequences can be significant.

**Measure changes in MTs after the first failure for experienced subjects.**

If experienced subjects have previously omitted the task of checking the accuracy of the secondary cues, they may start to perform that task after a failure occurs. This may result in a change in MTs for experienced subjects using manual control. As the phases of detection and adaptation progress (Phatak & Bekey, 1969), an increase in the movement time may occur. Skilled behavior may be replaced by unskilled behavior. The presence of this change is of interest. Even if subjects have been performing the task of checking the reliability of the secondary cues, they may slow down to be more careful after the first failure. As suggested above, experienced subjects may have to adjust their method of responding from the automatized responses that developed as a result of extensive practice with no failures, to controlled responses. This may result in an increase in MTs for the failure rows, as discussed below. There may also be a general increase in MTs on all targets after the first failure is detected. These arguments predict an accompanying improvement in the detection performance after the first occurrence of a failure. See above for a discussion of this possible finding.

If there is an increase in the MTs of experienced subjects after the first failure is detected, it may be manifested in the time to reach each target, or just the first one in a trial. The trial begins with all primary and secondary cues presented. A subject can choose to check each set of squares and dots for a failure before beginning the first movement of the trial. An
increase in the times to the first gate only, suggests that subjects have added, or are
devoting more time to, the task of checking for a failure. An increase in all MTs, with no
disproportionate increase in the first one on a trial, could be due to a change in the nature of
the response from automatized to controlled, or the subjects checking for the failure as they
pass through each gate, rather than at the beginning of the trial.

**Determine the Effect of Automation on Failure Detection and the Interaction with Experience Level**

The major goal of this research was to determine whether the level of automation affects
failure detection and whether this effect was different for experienced and inexperienced
operators. Two control modes were tested, manual and fully automated. In the automated
condition, the operator monitored the cursor movement as it moved up the screen. With
automated control, operators may be less aware of the characteristics of the task, i.e., less
'in the loop', and therefore they may be less likely to detect a failure. They may have less
information available to them when simply observing the task being performed, or the
information may be less easily interpreted. The operators may rely on the automation to
perform the task and not be as sensitive to the presence of a failure, especially when there is
a great competition for their attention. Parasuraman, Molloy, and Singh (1993) found that
'automation-induced complacency' is more easily detectable in multitask environments
when operators are responsible for many functions. Alternatively, operators may be more
able to attend to the presence of a failure due to the reduction in task requirements in the
automated condition or the lack of distraction from the manual task. The literature is not
definitive with regard to the detection of failures while controlling manually versus
monitoring automated control. Some studies, e.g., Wickens and Kessel (1979, 1981),
reveal an advantage of manual control for failure detection. Ephrath and Curry (1977), Ephrath and Young (1981) and Fuld, Liu, and Wickens (1987) found the opposite effect of control mode. More specifically, Fuld, Liu, and Wickens (1987) found that the operators under automatic control tended to show greater sensitivity to the failure. Also of note, they found that when controlling manually, the operators adopted a conservative, non-optimal response criterion.

Note that in the present study the automated control was non-adaptive. In this study the cursor position relative to the secondary cues did not provide information regarding the presence of the failure, nor did the cursor path relative to the primary cues. In the previous research cited here, the automated control was also non-adaptive; the automation did not compensate for the failure, i.e., the resultant flight path was off nominal. In those experiments however, the non-adaptive control provided additional information on the presence of the failure because the error between the desired and achieved flight path was observable. In a complementary finding, Young (1969) found superior failure detection performance with manual control versus a monitored automated adaptive control. The adaptive automation resulted in an accurately flown approach. Even though the failure had occurred, there was no observable consequence in the automated condition.

Experience level may confound the effect of control mode on failure detection. If experienced subjects tend to be less in the loop, they may be even less involved under automatic control. Inexperienced operators may get more benefit from the reduction in resource requirements under automatic control.

As systems become more automated, there is a great deal of focus on the need to keep the operator informed when using automation, which is most of the time in many
applications. The operator must be able to readily detect a failure in both manual and automated operations (Norman, 1990). This study provides insight into the effect of automation on failure detection. What is more important, the study examines the interaction of experience level and automation effects. This is important, since many operators are highly skilled and have built up a great deal of experience with reliable automation.

**Measure MTs and Errors for Targets with and without Failures**

**MTs to detected failures.** Even though the movement distance is identical, movement times can be greater for targets with failures than those without, for several reasons. To determine the target area for a failure target, the subject must visualize the midpoint and then allow 7.5 mm on either side. It is very likely that this process requires more time than observing the secondary cues, on the normal trials. In addition, on the targets with failures the secondary cues can be misleading, and thus the skilled behavior in response to these cues is inappropriate. If the subjects' performance of the task has become automatized, they have to begin to perform it as a controlled task, i.e., ignore the secondary cues and attend to the primary information. It may be more difficult to ignore the misleading cues after a great deal of practice. Experienced subjects may encounter interference from the previously learned behavior when responding to the new stimuli (Fisk, Lee, & Rogers, 1991), much like the Stroop effect (Jonides, 1981; Kahneman & Treisman, 1984; Stroop, 1935). See the discussion below on capture errors, another effect of switching to controlled behavior. An opposing view is that skill is transferable (Pew, 1974) and the subject can transfer the skill of using the secondary cues to define the target, to using the primary cues and ignoring the secondary ones.
If the difference in difficulty is the predominant cause of an increase in MTs for failure targets, there will be no interaction of MTs for failure and normal targets with experience level. If there is an interaction, i.e., if experienced subjects show relatively more of a MT increase for failure rows, this suggests that experienced operators require more time to perform a task that cannot be performed in the way that is well practiced. The argument for a shift from automatized to controlled behavior will be supported.

**Errors for detected failures.** The more ambiguous target area for targets with misleading secondary cues may result in greater errors for these trials. Alternatively, subjects may be more cautious in responding to failure targets, resulting in fewer errors. The amount of error may differ for experienced and inexperienced subjects.

**Measure Tracking Error and MT Resulting from Capture Errors**

When subjects are highly practiced at a task, the learned response to stimuli may interfere with the intended response, such that they perform the practiced task, even though it is no longer appropriate. This is more likely to occur when the stimuli in the new task differ very little from the practiced task. Fisk, Lee, and Rogers (1991) suggest that experts have difficulty both ignoring information that they had been trained to automatically attend to (the secondary cues), and attending to previously ignored information (the primary cues). Logan (1988b) describes the 'obligatory retrieval' of highly practiced stimulus-response sequences that interfere with another task. Norman (1981a) has labeled these 'capture errors' stating that the more practiced sequences capture control and result in the faulty activation of behaviors. If this interference occurs, errors in cursor movement should result. The subject will respond to the secondary cues instead of the primary ones,
and will not be able to ignore the secondary cues. The cursor path that results in this type of error is shown in Figure 16. This capture error is expected to occur more often for experienced than inexperienced subjects. If this response is made, a faster MT will also likely occur, as this response is well practiced.

This phenomenon has serious operational ramifications. The operator may make an error despite 'situational awareness' of its inappropriateness. Since these capture errors are extremely well practiced and performed with little conscious involvement, both training and procedural restraints are inept at reducing these capture errors. Logan and Stadler (1991) showed that these capture errors are made to generalized stimuli, not just to those that are practiced. Interface design may show some promise for combating this unwanted behavior. The example discussed earlier of placing a barrier over a control, to prevent its being used improperly, is a positive measure to prevent capture errors.

**Measure Differences in MT for Targets at 40 and 50 Degrees**

The line from one target to the next forms either a 40 or 50 degree angle from vertical. Half of the targets on a screen are at 40 degrees and the other half are at 50 degrees. Previous research with target acquisition tasks has shown that performance is faster when the target is at a position closer to vertical or horizontal (Jagacinski & Monk, 1985; Radwin, Vanderheiden & Lin, 1990). This study examines the MTs for 40 and 50 degree targets for this task.
CHAPTER 3

EXPERIMENT 1

Experimental Method

Subjects

The subjects in Experiment 1 were 16 adult males and 16 females, with no specific qualifications. There were four experimental groups, each with four males and four females. The mean age of the 32 subjects was 28.3 years and their ages ranged from 18 to 45. Twenty-eight of the subjects were right handed and four were left handed. They were screened for normal (20:20) corrected near vision. The subjects were paid for the two or four hours of participation.

Apparatus

The subject was seated at a table facing a cathode ray tube display at a distance of approximately 35 cm. An electronic tablet was on the side of the hand that the subject
chose to use to control the pen. The pen was connected to the tablet by a wire. The subject controlled the cursor on the screen by moving the pen across the tablet. The pen did not need to be touching, but had to be within 1 cm of the tablet to register on the screen. The position of the pen on the tablet was mapped to the position of the cursor on the screen (zero-order control). There was a time delay between pen movement and cursor movement of 0.25 seconds. (In actual implementation, the time delay varied slightly on some updates and had a standard deviation of 0.017 seconds.) This delay was introduced to make the task more difficult. The operating area of the tablet was 27.8 cm wide by 26.7 cm deep. The screen was 34.9 cm wide by 27.0 cm high with a resolution of 1280 by 1024 pixels. The gain of the tablet to the screen was 1.2 to 1 horizontally and 1 to 1 vertically. The position of the cursor was shown as the end of a red line of 0.7 mm width on the black screen. This line depicted all previous positions of the cursor on the screen for a trial. A Silicon Graphics Iris UNIX workstation was used to drive the display, process the subjects' input, and collect the data.

**Experimental Design**

This study employed a two by two, between-subjects design. The four conditions in this experiment were two types of control: manual and automated, and two levels of experience: experienced and inexperienced (see Figure 2). There were eight subjects, four males and four females, participating in each condition. The subjects were assigned to the experimental conditions in a counter-balanced order. The subjects in the experienced conditions reported on Day 1 for two hours and received 200 manual trials with no failure present. However, they were instructed that they were to detect a failure on those trials. The experienced subjects returned the next day for the second half of the trials (two hours)
with no indication that there was a change in procedure. The inexperienced subjects only reported for two hours. Both groups received 200 trials in the final portion of the experiment. These trials were performed with automation by half of the experienced and inexperienced subjects and manually by the other half. There was a failure present on 10% of the final 200 trials.

<table>
<thead>
<tr>
<th></th>
<th>Manual</th>
<th>Automatic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experienced</td>
<td>8 Subjects x 200 manual trials (1600 movements) + 200 manual trials 20 failures</td>
<td>8 Subjects x 200 manual trials (1600 movements) + 200 automatic trials 20 failures</td>
</tr>
<tr>
<td>Inexperienced</td>
<td>8 Subjects x 200 manual trials 20 failures</td>
<td>8 Subjects x 200 automatic trials 20 failures</td>
</tr>
</tbody>
</table>

Figure 2. Experimental conditions for Experiment 1.
This experiment employed a between-subjects design for the detection of a failure, however other factors, such as movement times and errors were examined for both between- and within-subjects effects using repeated-measures.

**The Task**

An experimental task was created for this study, to address questions about failure detection. This task had three necessary qualities: it was unfamiliar to the subjects, it could be learned rapidly, and the secondary cues could be degraded, permitting the measurement of failure detection. While this task did not accurately represent an actual flight task, the elements of a continuous tracking and a discrete failure detection task were present, as they are in the cockpit tasks.

The subjects were required to move a cursor from the bottom to the top of a screen passing through the midpoints of horizontal lines (not visible), formed by two squares (Figure 3). A unique configuration of eight pairs of squares appeared on each screen. Each set of squares was 10.6 cm apart horizontally. The squares were yellow filled and were 6 mm in width. The vertical distance from the start position to the top set of squares was 26 cm. There was an accuracy tolerance limit around the line midpoints. If the cursor passed within this area, this constituted perfect accuracy. This tolerance area was 1.5 cm (9/16 inch) wide. When the secondary cues were present, the tolerance area was bounded by two yellow filled dots that were 2.5 mm wide (Figure 4). The distance between the midpoint of consecutive sets of squares, i.e., from one target to the next, was 4.7 cm. Both the target width and the distance were held constant. The angle from one target to the next was ±40 degrees from vertical for four targets on each screen and ±50 degrees for the other four, ordered randomly. The direction of consecutive targets, right or left, was
randomly assigned, with the constraint that the targets fall within the area of the screen. The subjects used the dots to determine the midpoint of the line between the squares, provided the dots were aligned with the midpoint. The presence of the tolerance, defined by the dots, made the task relatively easy to perform accurately. Subjects were instructed to perform as quickly as possible with perfect accuracy. Appendix A contains the subjects' instructions.

![Figure 3. Stimuli with cursor trajectory through the line midpoints.](image)

![Figure 4. Primary cues (squares) with secondary cues (dots) depicting tolerance area.](image)

In the automated mode, the position of the cursor was controlled by the computer. Subjects monitored the automation to assure that the cursor was being placed correctly through the midpoints of the squares. They were instructed to press the pen to the tablet if they detected an error in the position of the cursor as it passed through the squares. The
position of the cursor was never in error. The path of the cursor in the automated condition
followed a previously recorded, manually performed path, so that the stimuli were similar
to the manual condition, a technique used by Fuld, Liu, and Wickens (1987). Each subject
in the automated condition observed the same 200 trials. The fifty-trial blocks were
presented in a counter-balanced order across subjects.

Failure detection was required in both the manual and automated control modes. The
failure was created by shifting one set of dots to the right or left of the midpoint (Figure 5).
The subject had to perform the task of checking that the dots had the same midpoint as the
squares. The failure occurred in 10% (20) of the trials. The failure trials appeared in a
random order in every 50 trials. For each screen on which a failure occurred, only one pair
of dots was offset; it was randomly selected from the eight. The direction of the failure
(right or left) was randomly assigned on each trial. The dots were shifted from the center
by 6.25 mm (25 pixels) (see Figure 6). This distance was determined in a pilot study to be
the amount of offset that was detected half of the time. The automated control positioned
the cursor through the center of the squares, while remaining between the misaligned dots.
This eliminated the cursor position with respect to the dots as a cue for failure detection.
The possibility of a failure was discussed with the subjects and demonstrated in the training
session described below.

The trial onset was self-paced and began when the subject positioned the cursor in a 6
mm box at the bottom center of the display and pressed the pen to the tablet. The manual
trials were entirely self-paced. The subjects were able to search for a failure before
beginning the tracking task, which many reported doing. This aspect of the task removes
the issue of workload in the manual condition. In the automated mode, the cursor moved
through the gates at the average rate of a manual tracker. This task differs from most signal
detection tasks in that it is self-paced in the manual mode and in the automatic mode there is ample time to detect the failure. The trial automatically ended and the stimuli and cursor trail were removed after the subject or the automation moved the cursor through the horizontal line defined by the top squares.

At the end of each trial the subjects were asked which target was offset, if any, and how sure they were of the correctness of their response. After the trial stimuli disappeared from the screen, the failure identification screen was presented (Figure 7). The screen consisted of an 8.1 cm horizontal bi-polar rating scale at each vertical location where the eight targets had been and beside the word 'none' at the bottom of the screen. The words 'not sure' were to the left of the line and the words 'very sure' to the right. The target number, from one to eight, appeared to the left of the rating scale. The subjects placed the cursor (now a
red dot) vertically where the offset target had been and horizontally along the rating scale and depressed the pen to the tablet to make their response. (For scoring purposes, a response that was either on or within one target of the correct target number was scored as a correct response.)

Once the response was made, the screen was replaced by a feedback screen that displayed the number of seconds to complete the trial, the trial number, and if a failure had been present, the target number that had the failure. No target number appeared if a failure was not present. Feedback on tracking accuracy was delivered verbally by the experimenter for each training trial only. No tracking accuracy feedback was given for the experimental trials. However, the presence of the red line trail of cursor movement, left on the screen after passing through each gate, allowed the subject to readily evaluate the accuracy of the movements.
The software that presented the task and recorded the data permitted specification of a number of parameters by the experimenter. The presence of dots and of a failure could be selected for a block of trials. The number of pixels that the failure is offset from the center was also experimenter-selectable. This study used an offset of 25 pixels. The number of trials in a block was the final parameter specified by the experimenter.

The data for each block of trials were written to a file. The first time the cursor passed through the vertical coordinate corresponding to each set of stimuli, the time and horizontal position were recorded. Also recorded were the coordinates of each square and dot and the midpoint of the squares. If a failure existed, the target number containing the misleading
stimuli and the amount of offset were recorded. The failure row selected by the subject and
the confidence rating were written to the data file as well. The data collection rate was 20
hertz.

**Experimental Procedure**

Training preceded the experimental trials. The subject began by reading the instructions
in Appendix A. The first five trials were completed with primary cues (squares) only
(Figure 3). This was done to emphasize the importance of the primary cues. Next, five
trials were performed with the dots depicting the target area. The experimenter emphasized
that the subject was to move the cursor as quickly as possible, while maintaining perfect
accuracy. It was explained that if the cursor went through any part of the target area, the
movement was completely correct. If the cursor touched the dots, the movement was
considered in error. These instructions were meant to constrain the subject's speed-
accuracy tradeoff. The next five trials of the training demonstrated the presence of the
failure; a set of dots was not in the center of the squares. The subjects were told that one of
the pairs of dots was not centered and were asked to identify it and to remember which row
it appeared in. The subjects were instructed on the use of the confidence rating scales and
encouraged to utilize the entire range of responses. The subjects were paid 10¢ for every
correctly identified normal and failure trial and were debited 10¢ for every false alarm and
missed failure.

Subjects in the automatic control condition were given an additional five trials of training
on the use of the automatic control. The subjects watched as the computer moved the
cursor through the squares, up the screen. They were instructed to press the pen to tablet
when the cursor position was out of tolerance. The training was to ensure that the task was understood properly and was not intended to contribute significantly to the level of practice.

The experienced subjects reported for two consecutive days, for two hours each day. They received the training and manually performed the first 200 trials on the first day. Each experienced subject received the same number of trials (1600 movements). This number was determined in a pilot study to be adequate to produce a high level of skill. The trials provided sufficient practice to produce skilled asymptotic performance of the task for all subjects. The trials were presented in two blocks of 100 per block. Each block was approximately 40 minutes in duration. The subjects were given approximately a ten minute break between the two blocks of trials. The first 200 trials performed by the experienced subjects had aligned dots (Figure 4); no failure detection was necessary, although the instructions to the subjects did not suggest this. They were required to make a confidence rating at the end of each trial as to how sure they were that a failure did or did not exist.

On Day 2 the experienced subjects performed a second set of 200 trials, either manually or with the automation, depending on their assignment to the control conditions. Half of the experienced subjects (four males and four females) were assigned to each condition. This session was again broken into two blocks of 100 trials, with a break between.

The sixteen inexperienced subjects (eight males and eight females) received the training and performed the 200 experimental trials in one two-hour session. They too were assigned, in a counter-balanced order, to either the manual or automatic control condition.

At the end of the experiment, the subjects were asked to complete a questionnaire (Appendix B). Subjects reported the hand used for the task, age, and gender. On a ten-
point scale, from very easy to very difficult, the subjects then were asked to rate the use of
the pen and tablet, the task of placing the cursor through the center of the squares, the task
of monitoring the automated cursor movement, if applicable, and the task of detecting a
failure. In the final question, they were asked whether they had any previous experience
with the task.

After completing the questionnaire, the subjects were debriefed. The experimenter
discussed the independent and dependent measures of the experiment and the goal of
providing guidance for cockpit design, flight training, and aircraft operating procedures.
Potential further research was also described.

Experiment 1 Results

The following section describes the results of this study. The results are organized
according to the research questions that were addressed, as described in Chapter 2.

The Effect of Practice on Movement Time

The effects of practice on motor performance were measured by the mean movement
time for each trial. Movement time was defined as the difference from the time the cursor
passed through the vertical coordinates of successive sets of stimuli. There were eight
movements per trial (screen). The mean MT over the number of trials was analyzed. A
regression analysis was performed on these data to determine the goodness of fit of an equation describing the underlying function. The MTs were expected to decrease as the number of trials increased, according to the power law of practice. The mean MTs (in seconds) are shown in Figure 8 for the 16 experienced subjects for the 1600 movements in trials 1-200. These data are described by the equation:

$$MT = 3.8102 \times \text{trial number}^{-0.14992}$$

This function is also shown in Figure 8. The goodness of fit of this equation is high; the coefficient of determination ($r^2$) is 0.931.

Figure 8. Mean movement times (seconds) for the 16 experienced subjects, trials 1-200.
Effects of Experience Level

Movement times for experienced and inexperienced subjects. The experienced subjects returned for a second set of 200 trials, some of which had failures. Their mean movement times and those of the inexperienced subjects (trials 1-200) are shown in Figure 9. The experienced subjects had reached asymptote before starting this block of trials; their MTs did not change appreciably over the last 200 trials. The inexperienced subjects reached the skill level of the experienced subjects, as measured by the MT. The mean MTs for the experienced and inexperienced subjects were 1.90 seconds and 2.14 seconds, respectively. The standard deviations were 0.86 and 0.58 seconds. The movement times of these groups across the 200 trials were not significantly different.

Failure detection by experienced and inexperienced subjects. The detection of a failure was the key dependent variable in this study. The number of subjects detecting the failure by experience level was used to test whether the detection performance of experienced subjects differed from the detection by inexperienced subjects. The expected probability of a failure being detected was .50, based on a pilot study.

For the first failure, the number of subjects detecting its presence differed significantly for the experienced and inexperienced groups ($X^2 (1, n = 32) = 4.12, p<.05$). Fewer experienced subjects detected the initial failure. Five of the sixteen experienced subjects (31%) detected the first failure, while twelve of the sixteen inexperienced subjects (75%) were successful in detection.

The proportion of all failures detected by experienced ($M = 0.730, sd = 0.188$) and inexperienced subjects ($M = 0.678, sd = 0.206$) did not differ. The arc sine transformed
proportions were also not significantly different for the two experience levels. The subject with the highest detection rate was successful 95% of the time and the subject with the lowest rate detected only 20% of the failures.

**Improvement in failure detection by experience level.** The proportion of detections for each failure for experienced and inexperienced subjects were measured over
time. These data appear in Figure 10. There was a significant improvement over time in the detection of the offset by the inexperienced subjects ($F(19, 285) = 1.61, p < 0.05$). The detection performance by experienced subjects also improved over time ($F(19, 285) = 2.67, p < 0.001$).

After the first two failures, the detection performance appears to remain fairly constant for the experienced subjects. An analysis of the proportion of detections by each subject for failures 3 and 4 through 19 and 20 showed no significant improvement in detection performance over those trials.

**Rate of detection versus movement times for experienced subjects.** The movement times for experienced subjects were explored as a measure of the amount of processing on the detection task. If, as the subjects' detection performance improved, they were spending more time looking for the failure, this would be reflected in their time to complete the tracking task. Some subjects reported doing the detection task first for all the targets before beginning the movements and some subjects reported doing the detection as they went through each set of dots and squares. One subject reported switching strategies from the former to the latter and this was observable in his movement times.

The time was measured between the onset of the stimuli (i.e., the time of the presentation of the new screen of squares and dots) and the time the cursor passed through the first set of squares. This time to the first gate was expected to reflect the time required to detect the failure, if the subject checked for failures at the beginning of each trial, which some subjects reported doing. These data were compared with the proportion of detections over the last 200 trials for the experienced subjects. Three analyses were performed involving different levels of data aggregation. The mean time to the first gate for each
subject for a block of 20 trials was compared with their proportion of detections on pairs of failures. There was no correlation between the movement times and the detection rate over the twenty-trial blocks. The mean time to the first gate across subjects for twenty-trial blocks was compared with the mean detection rate for pairs of failures across all subjects. These data appear in Figure 11. This correlation was also not significant. Each subject's
mean time to the first gate across all trials was compared with his/her overall detection rate. This correlation too was not significant.

The mean time to complete the tracking on the entire screen was calculated across subjects for twenty-trial blocks (shown in Figure 11). These times were compared with the proportion of detections on pairs of offsets. There was a negative correlation between movement times and detection performance over time \( r (8) = -0.743, p < .05 \). The experienced subjects improved their detections as they continued to improve their tracking speed.

The tracking performance of the experienced subjects was tested for a speed-accuracy tradeoff. The mean time to complete the eight gates on a screen for all experienced subjects by twenty-trial blocks was compared to the mean number of tracking errors for non-offset gates in those blocks (see Figure 11). There was no correlation between these parameters, suggesting no speed-accuracy tradeoff in the tracking task.

**Sensitivity and decision criteria by experience level.** In this study, the confidence ratings of the responses to the presence of a failure were collected to calculate the sensitivity to the failure. The subjects typically indicated either that they were very sure that a failure had been present or that they were very sure that none had been present. When a failure was indicated, 55% of the confidence ratings were 'very sure'. When none was indicated, 83% of the confidence ratings were 'very sure' that no failure was present. Despite their instructions, the subjects did not use the full range of the confidence scales. For this reason, the area under the Receiver Operating Characteristic curve \( (A_p) \) (Davison & Jagacinski, 1977; Donaldson, 1993; Green & Swets, 1973) was not an appropriate metric of the subjects' sensitivity to the failure. The subject's sensitivity to the failure was
Figure 11. Proportion of subjects detecting failure, time to first and all gates, and proportion of trials with errors over time for experienced subjects.
therefore measured by the distance between the means of the noise and signal distributions ($d'$). The $d'$ measure was compared for the experienced and inexperienced subjects who had at least one false alarm (11 and 8 respectively of 16 subjects in each group). The detection rate versus the false alarm rate, i.e., Receiver Operating Characteristic (ROC) plots, for the experienced and inexperienced subject groups are shown in Figure 12 and 13. No difference in sensitivity to the failure ($d'$) was observed between experienced ($M = 2.62, sd = 0.69$) and inexperienced subjects ($M = 2.63, sd = 0.78$).

One decision criterion measure used was beta ($\beta$). The decision criteria were compared for the experienced and inexperienced subjects who had at least one false alarm. The $\beta$ indicated that there were no differences in the decision criteria of the experienced ($M = 11.48, sd = 7.38$ for 11 subjects) and inexperienced subjects ($M = 11.21, sd = 7.46$ for 8 subjects).

False alarms for experienced and inexperienced subjects. A second measure of decision criterion was the false alarm rate (Swets, Tanner & Birdsall, 1961). The number of times a subject indicated that a failure was present when none of the targets was offset was compared for the experienced and inexperienced subjects. The proportion of false alarms was low for both groups. The experienced subjects had a false alarm rate of 0.010 ($sd = 0.013$). The false alarm rate of the inexperienced subjects was 0.010 ($sd = 0.012$). There was no effect of experience level on the number of false alarms.

This measure was evaluated across the trials for experienced and inexperienced subjects. For the experienced subjects in the 200 trials with no failures the false alarm rate was the positive response rate. The average false alarm rates across 50-trial blocks for the sixteen experienced and the sixteen inexperienced subjects are shown in Figure 14. The false
alarm rates decreased over time for the experienced subjects ($F(7, 105) = 2.17, p < 0.05$). (Note that the block effect significance level was the same when the false alarms were tallied over blocks of 100 trials ($F(3, 45) = 2.96, p < 0.05$).) As can be seen in Figure 14, the false alarm rate for the experienced subjects became nearly constant from trial 101 through 400. There was no significant increase or decrease in the false alarm rate for trials 101 to 400.
The false alarm rates for the experienced and inexperienced subjects were examined over trials 201 to 400 in 50-trial blocks. There was no effect of time and no experience level effect for the false alarm rates in these trials. There was also no interaction of the block number and experience level for the false alarm rates during these trials.
While the false alarm rates were low for the last 200 trials, they were not at zero. The mean false alarm rates for Trials 201–400 were determined to be significantly different than zero for experienced subjects ($t(df = 3) = 4.159, p < .05$) and inexperienced subjects ($t(3) = 4.086, p < .05$, one-tailed t-test for two sample means assuming unequal variance).

**Performance by Experienced Subjects after the First Failure**

**Detections by experienced subjects after the first failure.** After the first failure, the experienced subjects' detection performance improved to the level of the inexperienced subjects. The number of experienced subjects detecting the second failure...
was double that of the first (10 of 16, 62%). This improvement was not statistically significant, however. All the subjects who detected the first failure also detected the second one. The mean detection rate for the experienced subjects across all 20 failures was 0.709 (sd = 0.181) and the mean overall detection rate for the inexperienced subjects was 0.678 (sd = 0.206).

**Movement times before and after the first failure for experienced subjects.** To determine if the first failure resulted in experienced subjects slowing down to improve their detection performance, the MTs before and after the first failure were analyzed. Only MTs to normal (non-failure) targets were included, to remove any confounding effects of target type. The experienced subjects did not show a significant increase in movement time as a result of their improved detection performance after the first failure. The mean MT for the eight targets before the first failure for the experienced subjects was 2.19 seconds (sd = 1.42). The mean MT for the eight targets immediately following the first failure was 2.35 seconds (sd = 1.69).

The time was measured between the onset of the stimuli and the time the cursor passed through the first set of squares, as a measure of the time required to detect the failure. On the two trials before the first failure, the mean time to the first gate was 3.45 seconds (sd = 2.60) and the mean time to the first gates on the two trials after the failure was 3.87 seconds (sd = 3.11). The increase in times to the first target after the first failure by experienced subjects approached significance ($F(1, 7) = 4.75, p=.066$). The analysis of the times to the first gate for the five trials before and the five after the first failure yielded less significance.
Movement Times and Errors for Offset and Normal Targets

Movement times for targets with and without offsets. The movement times for targets with failures versus aligned targets were compared (Figure 15). The movement times to targets with detected erroneous secondary cues (offset dots) were greater than movement times to normal targets ($F(1, 14) = 21.86, p < 0.001$). The mean movement time to detected offset targets was 3.05 seconds ($sd = 1.93$). Fifteen normal targets were selected at random for each subject and the MTs were compared to the offset targets; the mean for the sample of normal targets was 1.90 seconds ($sd = 1.21$). An interaction between MTs for failure and normal targets with experience level was tested; there was no interaction between target type and experience level for movement times.

Tracking errors for failure and normal targets. Movement accuracy was defined as the distance from the tolerance limit around the center of the two squares to the point where the cursor crossed between the squares. Accuracy was expected to be very high, since perfect accuracy was required. Subjects manually controlling the cursor did not perform without error. The mean number of times per gate that the cursor was not placed within the tolerance area around the midpoint for the manual control subjects was 0.019 of the gates and the standard deviation was 0.015.

The proportion of errors was greater for detected failures than for normal targets ($F(1, 14) = 9.66, p < 0.01$, for the arc sine transformed proportions). Subjects were less likely to position the cursor through the center of the squares, when the dots were offset and the offset was detected. The mean proportion of errors for targets with an offset was 0.192 ($sd = 0.238$). The mean proportion of errors for normal targets was 0.018 ($sd = 0.014$).
Figure 15. Movement times (seconds) for detected offset (failure) and normal targets.

There was no effect of experience level on error rate, nor was there an interaction between experience level and target type (normal/offset) for error rate.

**Cursor positions for failure and normal targets.** The placement of the cursor through the squares was not as exact for the targets where the dots could not be used to determine the midpoint of the squares ($F(1, 15) = 25.65, p < 0.001$). The mean distance from the center point to the cursor position, as it passed through the horizontal line defined by the squares was 18.20 pixels ($sd = 13.34$), when the dots were not centered and the offset was detected. The mean distance between the center point and the placement of the
cursor when the dots correctly defined the midpoint of the squares was 8.72 pixels \((sd = 9.16)\).

**Capture Errors and Movement Times**

*Capture errors.* Even though the failure was detected, as evidenced by the subject's response on the confidence rating scales, there may have still been an error, due to the subject's unintentional response to the well-practiced dot stimuli, a capture error. Responses to detected failures were measured as the number of times the cursor was not placed in the center of squares but was inside the dots, versus placement out of the center of the squares on the other side of the dots (as shown in Figure 16). Note that a capture error was not measurable as an error unless the cursor was placed through the area between the dots that was not also in the center of the squares. The amount of overlap of these two areas was half of the target width, 7.5 mm. To test for the presence of capture errors for detected failures, an analysis was performed on the proportion of errors between the dots and not between the dots (noncapture errors). The sixteen manual control subjects had an average of 1.12 errors each \((sd = 1.84)\) on an average of 12.92 detected failure targets. No difference in the type of error for detected failures was found. The mean number of cursor placement errors per subject for detected failure stimuli was 1.56 for capture errors \((sd = 2.00)\) and 0.69 \((sd = 1.62)\) for errors on the outside of the dots.

The presence of an experience level effect was examined. The eight experienced subjects manually controlling the cursor averaged 2.25 capture errors \((sd = 2.55)\) (noncapture \(M = 1.25, sd = 2.19\)). The inexperienced subjects had a mean number of capture errors of 0.87 \((sd = 0.99)\) (noncapture \(M = 0.12, sd = 0.35\)). The experience level
effect and the interaction of experience level with error type were not statistically significant.

The proportion of capture errors relative to the number of detected failures by each subject was analyzed for effects of experience level. Experienced subjects had 18 capture errors for 111 offsets detected versus 7 capture errors for 108 detected offsets by inexperienced subjects. This difference was not significant with an analysis of variance or a nonparametric test. The experienced subjects' proportion of detected failures that had noncapture errors also did not differ from the proportion for the inexperienced subjects. Experienced subjects had 10 noncapture errors for 111 offsets detected versus 1 capture error for 108 detected offsets by inexperienced subjects.
To account for the difference in error rate across subjects, the proportion of tracking errors on offset dots that were capture errors was compared between experienced subjects and inexperienced (the number of capture errors divided by the sum of the number of capture and noncapture errors on offset gates, for each subject). A Wilcoxon-Mann-Whitney analysis yielded no difference between experienced and inexperienced subjects.

Movement times for capture errors. The mean MTs for each subject were analyzed for movements with capture errors versus the offsets correctly responded to. There was a difference in the movement times between the erroneous and correct movements to detected offset stimuli \((F(1, 14) = 10.71, p<.01)\). When movements were erroneously made to pass the cursor through the dots, i.e., the capture errors, the movement times were faster \((M = 2.09 \text{ seconds}, sd = 0.76)\) than for correctly made movements \((M = 3.06 \text{ seconds}, sd = 1.14)\). See Figure 17 for the means and standard deviations of these movement times. The amount of variability in the cursor positions on the correctly made responses \((sd = 8.49 \text{ pixels})\) was greater than for the capture error responses \((sd = 6.02 \text{ pixels})\) \((F(1, 165) = 1.99, p<.05)\), therefore, the faster times cannot be attributed to less accuracy.

Movement times for capture errors were tested for the presence of an experience level effect. There was no effect of experience level on the MTs to detected offsets with correct movements and capture errors combined. There was no interaction between the MTs with correct cursor placement versus capture errors and the experience level.
Figure 17. Mean movement times (seconds) with standard deviations for correct cursor placement and capture errors to detected offset stimuli.

**Movement Times for Targets at 40 and 50 Degrees**

The distance from one target to the next (A) and the target width (W) were held constant. The angle between successive targets was either 40 or 50 degrees to the right or left of vertical. Previous studies have shown that MTs are affected by the angle of the movement (Jagacinski & Monk, 1985; Radwin, Vanderheiden & Lin, 1990). Movements to targets at angles that were greater from vertical or horizontal were slower. The targets in this experiment were either 40 degrees from vertical or 40 degrees from horizontal so no difference in movement times due to angle was expected. The last ten trials (80 movements) of the first block of 200 trials for the experienced subjects were examined for this effect, where no failures were present and learning had reached asymptote. A test of
the MTs for targets at ±40 degrees and ±50 degrees showed no difference in MTs for these targets ($M = 1.77$, $sd = 1.18$ seconds for ±40 degrees and $M = 1.71$, $sd = 1.03$ for ±50 degrees).

Effects of Automation and the Interaction between Control Mode and Experience Level

**Failure Detection for Subjects Using Manual and Automatic Control.** The control mode effects on the probability of failure detection were tested. The type of control, manual or automatic, did not affect the subjects' ability to detect a failure. The proportion of failures detected by the sixteen subjects controlling the cursor manually was 0.646 ($sd = 0.199$). Those using the automatic control had a probability of detecting the failures of 0.741 ($sd = 0.176$).

The detection rates were examined for interactions between the experience level and the control mode. There was no interaction in these two factors. See Table 1 for the proportion of failures detected by the four subject groups.

<table>
<thead>
<tr>
<th>Experience</th>
<th>Control</th>
<th>Count (Ss)</th>
<th>Mean</th>
<th>S. D.</th>
<th>Max.</th>
<th>Min.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experienced</td>
<td>Manual</td>
<td>8</td>
<td>0.699</td>
<td>0.151</td>
<td>0.900</td>
<td>0.400</td>
</tr>
<tr>
<td></td>
<td>Automatic</td>
<td>8</td>
<td>0.719</td>
<td>0.217</td>
<td>0.950</td>
<td>0.400</td>
</tr>
<tr>
<td>Inexperienced</td>
<td>Manual</td>
<td>8</td>
<td>0.594</td>
<td>0.237</td>
<td>0.900</td>
<td>0.200</td>
</tr>
<tr>
<td></td>
<td>Automatic</td>
<td>8</td>
<td>0.762</td>
<td>0.136</td>
<td>0.900</td>
<td>0.500</td>
</tr>
</tbody>
</table>

Table 1. The proportions of failures detected by experience level and control type.
To determine whether the lack of effects of control type on failure detection was a result of the number of subjects tested, an additional twelve inexperienced subjects were run. Six of the new subjects used manual control and six used automatic control. These subjects were combined with the original 32 and the control type effects were examined. There were no differences in the proportion of failures detected by the 22 subjects using manual control (M = 0.653, sd = 0.042) and the 22 using automatic control (M = 0.684, sd = 0.045). In addition, there were no significant differences in sensitivity or decision criteria when these subjects were added.

**Sensitivity and decision criteria by control type.** Subjects' sensitivity to the failure, as measured by the distance between the means of the signal and noise distributions, was compared for the subjects manually controlling and those with automated control. The sensitivity to the failure was measured by $d'$ for those subjects with at least one false alarm. The sensitivity for the manual condition ($M = 2.67, sd = .80$) did not differ significantly from the sensitivity of the subjects in the automatic control condition ($M = 2.57, sd = 0.64$). An analysis of the interaction between control type and experience level yielded no significance in sensitivity to the failure. The decision criterion ($\beta$) was measured for those subjects with at least one false alarm. This also did not differ for subjects with manual control ($M = 11.97, sd = 7.24, n = 10$ of $16$) and those with automatic control ($M = 10.70, sd = 7.54, n = 9$ of $16$).

**False alarms by subjects with manual and automatic control.** The number of targets identified as failures, that were in fact not offset, was compared for subjects in the manual and automated groups. The false alarm rate was $0.048$ ($sd = 0.185$) and $0.019$ ($sd = 0.037$), respectively. There was no effect of the type of control on the proportion of
false alarms. There was also no interaction between the type of control and the experience level on the false alarm rate.

**Questionnaire Information**

On the post-experiment questionnaire (see Appendix B) subjects rated the different aspects of the experiment on a ten-point scale from very easy (1) to very difficult (10). They rated the pen and tablet as somewhat easy to use ($M = 3.03$, $sd = 1.77$, range 1-7). The task of placing the cursor through the center of the squares was rated near the middle of the same ten-point scale ($M = 4.28$, $sd = 1.87$, range 1-8). The task of detecting the failure was also rated half way between very easy and very difficult ($M = 4.94$, $sd = 2.17$, range 1-8). The subjects in the automated condition rated the task of monitoring the cursor movement by the automatic control as somewhat easy ($M = 2.94$, $sd = 2.02$, range 1-7).

These data were used to detect any differences in the subjects' ratings due to their different experimental conditions. See Table 2 for these ratings by experience level and control type. The subjects in the manual condition rated the task of detecting that the dots were not centered as significantly more difficult than those in the automatic condition ($F(1, 28) = 4.34$, $p < 0.05$, for the arc sine transformed ratings). The mean rating of the subjects in the manual control group was 5.69 and the subjects with the automatic control rated the detection task as 4.19 on the scale from very easy to very difficult (1-10). There were no interactions of the control condition and the experience level.
Table 2. The questionnaire ratings of subjects in the experienced, inexperienced, manual, and automatic conditions from very easy (1) to very difficult (10).

(* Denotes statistically significant differences.)

The subjects were assigned to the different experimental conditions in a counter-balanced order. The subjects' ages were analyzed, to ensure there were no differences in the groups. Both the experience level and the control type showed no significant differences in the age of the subject; the interaction of experience level and control type was also not significant. The mean age of the subjects was 28.34 years (sd = 9.64). The eight subjects in the experienced manual control group had a mean age of 32.62 (sd = 11.70). Those in the experienced automatic control group averaged 26.25 years (sd = 9.11). The mean ages of the subjects in the inexperienced manual and automatic control conditions were 26.50 (sd = 11.06) and 28.00 years (sd = 6.28), respectively. None of the subjects in this study reported having any previous experience with this task.
Experiment 1 Discussion

The conclusions drawn from the results of Experiment 1 appear in this section. Again the organization follows the research questions delineated in Chapter 2.

The Effect of Practice on Movement Time

An important issue in this experiment was the creation of the experience level. The form of the learning curve was determined for this task. It was hypothesized that the time to move the cursor from one target to the next would decrease as the number of trials increased, according to the power law of practice. The experienced subjects made 1600 movements on 200 trials in their first session. As shown in Figure 8, these data follow a power function.

Effects of Experience Level

Movement times for experienced and inexperienced subjects. The MTs of the experienced subjects had reached asymptote prior to the first presentation of a failure. The inexperienced subjects were presented with failures in the first block of trials. The difference in MTs for the experienced and inexperienced subjects is not sustained throughout the 200 trials in which the failures occur (Figure 9). The experience effects on detection, described below, diminished prior to the confluence of the MTs of the experienced and inexperienced subjects.
Failure detection by experienced and inexperienced subjects. As described in Chapter 2, the literature predicts both that experienced subjects will detect more failures than inexperienced subjects, and that inexperienced subjects will have better failure detection performance. The crucial finding of this research is that the extensive number of trials with reliable secondary cues, before the first failure by the subjects in the experienced condition, made them less likely to detect that failure. Equally important is the fact that these subjects then quickly excelled in their detection performance, as evidenced by the high number of subjects detecting the second failure. Also, the proportion of detections by the experienced subjects, over the 20 failures in the 200 trials, did not differ from the inexperienced subjects. Both groups of subjects improved in their detection performance over time, however, the improvement by the experienced subjects was contained in the first two failures.

Rate of detection versus movement times for experienced subjects. The movement times did not reflect a change in the amount of processing required to improve the detection performance. If, as the detection improved the subjects were taking longer to look for the offset before the trial began, the time between the appearance of the stimuli and the cursor being passed through the first gate would be increasingly greater. The time to the first gate was not correlated with the increasing detection rate over the last 200 trials. The time to all gates was negatively correlated with the detection rate over these trials. The experienced subjects were able to improve their performance both on the detection and the tracking tasks over time. There was no indication of a speed-accuracy tradeoff in the tracking task for experienced subjects. The number of errors and the time to perform the tracking task were not correlated.
Sensitivity and decision criteria of experienced and inexperienced subjects. Since the experience level effects were not sustained after the first failure, it is not possible to use the traditional measures of signal detection theory to determine whether the initial inferior performance by the experienced subjects was the result of a failure to perform the detection task, and therefore a loss of sensitivity to the failure, or a different decision criterion.

Additionally, a measurement issue arose with the use of the nonparametric measure the area under the receiver operating characteristic curve ($A_g$) as a measure of sensitivity to the stimulus. Because subjects failed to differentiate their confidence in their ratings of the presence of a signal or a noise, the endpoints were the only data available for measurement of the curve.

False alarms for experienced and inexperienced subjects. The false alarm rate was used as a measure of the decision criterion. This was examined over time. The vigilance literature on monitoring with a low probability of a signal (e.g., Parasuraman, 1986) predicts that the experienced subjects would adjust their decision criteria to become more conservative. This would be evidenced by a reduction in the number of positive responses. The false alarm rate did decrease over time for the experienced subjects (shown in Figure 14). This is consistent with the literature and would predict that the experienced subjects would be less likely to detect a failure, which they were for the first failure. The finding that the false alarm rates for the experienced and inexperienced subjects were the same at the time the failures were introduced is contrary to the literature on the detection of low probability signals. The decline in false alarm rate for the experienced subjects occurred in the first 100 trials, after which the rate stabilized at the inexperienced subjects' false alarm rate. Since the false alarm rate is reflective of the decision criteria, the
probability of detections should have been the same for the experienced and inexperienced subjects. The decision criteria did not account for the finding that the experienced subjects were less likely to detect the first failure. Therefore it can be assumed that the experienced subjects' sensitivity to the failure was less at the time of the first failure. This finding supports the theory that the experienced subjects stopped performing the detection task as a result of their experience with no failures. Once a failure was experienced, either through detection or feedback, the experienced subjects apparently began to perform the detection or check more closely for the offset, as evidenced by their improved detection rates.

**Performance by Experienced Subjects after the First Failure**

*Detections by experienced subjects after the first failure.* Experienced subjects improved in detection after the first failure. The number of subjects detecting the second failure was twice that detecting the first failure. This improvement, though not statistically significant, suggests that experienced subjects were able to immediately adapt their behavior to incorporate the task of checking for the failure, if they had omitted it as a result of the extensive number of trials without a failure. Had they adjusted their decision criteria, there would have been an increase in the false alarms after the first failure, which did not occur. The experienced subjects were able to immediately compensate for the initial lack of detection.

*Movement times before and after the first failure for experienced subjects.* The experienced subjects manually controlling the cursor movement were able to improve their detection performance with no significant attenuation of their tracking performance. There were no increases in the movement times for the individual targets
after the first failure. The increases in MTs for the first target on a trial, after the first failure, were nearly significant, suggesting more time being spent on the detection task before beginning the tracking for that trial. The movement times of the capture errors, described below, suggest that the tracking task had become automatized for the experienced subjects. These subjects were still able to perform the tracking task at this high level of performance while incorporating the detection task.

These findings reveal an important aspect of experienced behavior, that given a reminder of the presence of failures, experienced operators will quickly adapt their behavior to the level necessary to perform the detection task, and that the improvement will be accomplished with very little penalty in the performance of the other tasks. Note that this study only examined the improved detection in the context of a tracking task, it is possible that some other types of tasks or higher workload environments may suffer a performance decrement resulting from the improvement in detection performance. However, movement time is a sensitive measure, and it is reasonable to expect that an increase in demands from the detection task would produce an increase in movement times, particularly since subjects reported performing the detection task before starting the tracking, when the screen was presented. This increase was very slight, not statistically significant, and was not sustained after the first couple of trials following the first failure.

**Movement Times and Errors for Failures and Normal Targets**

**Movement times for target types.** As expected, it took subjects longer to position the cursor through the middle of the squares when the dots were offset and not providing reliable information about the center point (as shown in Figure 5). The movement
distances and target widths were identical to the normal targets. The difference in MTs can be attributed to the difficulty of using the different stimuli. When the dots were centered the subject simply had to place the cursor through them without having it touch either dot. This understandably took less time than determining the midpoint of the squares, estimating the tolerance area around that midpoint and placing the cursor within that area. Subjects may have tried to place the cursor through the exact midpoint between the squares. This is unlikely, since the actual distance of the cursor from the center of the squares was greater for offset than for normal targets. (See the section below for a discussion of this measurement.) There were no interaction effects between experience level and MTs for failure and normal targets. This implies that the predominant cause of the increase in MTs for failure targets was the difference in difficulty in responding to the squares, rather than the dots.

**Tracking errors for failure and normal targets.** Subjects were less likely to place the cursor within the tolerance area, when they were not guided by the dots. In the instructions to the subjects, accuracy was stressed as a more important consideration than speed. More time was required to position the cursor between the squares, with the dots providing misleading information, but the subjects did let their accuracy degrade.

**Cursor positions for failure and normal targets.** To examine the subjects' performance in more detail, an analysis was performed of the actual position that the cursor went through the squares and dots. As just discussed, there were more errors on the targets for which the dots were offset. From the measurement of the cursor position it could be determined if the subjects were being more careful overall to place the cursor exactly in the center of the squares, when the dots were misleading. If this was the case, the increase in MT for correctly responded to failed targets would be explained, according
to the speed-accuracy tradeoff. The increase in MTs for failed targets cannot be accounted for by the accuracy of the subjects' cursor placement for the failed targets. The cursor was less in the center on these targets. The subjects had more errors and utilized more of the tolerance area when the dots were not centered. This finding supports the premise that the secondary cues (the dots) aided the subject in performing the task. This was a key component of the experimental task, since this is the condition in actual operational environments that was to be represented, e.g., the flight director which is easier to use than the localizer and glideslope.

**Tracking Errors and Movement Times Resulting from Capture Errors**

**Cursor position errors from capture errors.** Much can be learned about a task from findings such as the presence of capture errors. In this study there were capture errors on 10.3% of the detected offset targets. The dots were not ignored properly and the cursor was placed between them, rather than between the previously ignored squares. This suggests that there was some difficulty in inhibiting the practiced response. In unprompted comments in the debriefing, subjects reported this to be the case.

Capture errors can pose a significant safety problem. These well-practiced responses cannot be prevented simply by notifying the operator of their inappropriateness. Barriers need to be constructed to prevent the practiced responses to stimuli when those stimuli are misleading. A better solution is to take advantage of the practiced behavior by driving the normal, practiced stimuli with the accurate information. Fortunately, this solution is more often employed as systems become more sophisticated and redundant sources of information are available to drive the displays.
Movement times for capture errors. The MTs were measured for capture errors through the dots with detected failures. These MTs were predicted to be less than the MTs for detected failures accurately responded to. Subjects in the experienced group were expected to show more of this tendency than those in the inexperienced group. The capture errors were made more quickly than the correct responses, when the failure was detected. The presence of capture errors and the faster MTs when a capture error occurred suggest that the task had become automatized. The lack of experience level effect does not support or disconfirm this hypothesis.

Movement Times for Targets at 40 and 50 Degrees

Previous research on target acquisition has shown that targets close to vertical or horizontal are more quickly acquired. To understand the qualities of the experimental task more fully, the movement times were examined for the two target angles used in this study. Half of the targets were 40 degrees from vertical and the other half were 40 degrees from horizontal. If there had been an inherent difference in the MTs to these targets, it should be accounted for in examining the other movement time effects, e.g., MTs for targets that were normal and those with failures. Since no differences in MTs were found as a function of the target angle, the data were able to be combined and target angle did not have to be separated as an independent variable when looking at other effects.
Effects of Automation and the Interaction between Control Mode and Experience Level

Failure Detection for Subjects Using Manual and Automatic Control. The hypothesis was tested that control mode affects the probability of a failure being detected. Previous research has shown that two factors affect the detection performance of operators under manual and automatic control; these are the level of workload and the availability of relevant information for detecting the failure (Wickens & Kessel, 1979). If workload is high, detection is better under automatic control (Ephrath & Young 1981; Fuld, Liu, & Wickens, 1987), unless the task is very complex and compelling, in which case operators rely on the automation to perform the failure detection (Parasuraman, Molloy & Singh, 1993). In this study the nature of the manual task was such that workload was not a relevant factor in the manual control condition. Manual subjects were able to perform the detection and tracking tasks serially, and reported doing so.

The information available to detect the anomaly is the second factor that has been shown to affect an operator's ability to detect a failure under manual and automatic control. For example, if kinesthetic information is relevant to the failure detection, this information is only provided with manual control. In the work of Wickens and Kessel (1981), detection was better with manual control than under automatic control where the kinesthetic information was not available. The present research differs from the studies of Wickens and Kessel (1979, 1981) in that the failure detection and the adaptive control in this study relied on vision; proprioception was unlikely to have provided pertinent stimulation as it did in their work. In the present study, the incorrect response to the misleading cues may be identical to the correct response to the correct cues, or a slightly different movement
distance may occur (less than 2.5 mm). In the current study, the information relevant to failure detection was the same for subjects under manual and automatic control.

Note that the failure in this study did not rely on the comparison of present stimuli to remembered information. The failure was apparent from the static presentation. This differs from other research. For example, Wickens and Kessel (1979) changed the order of the control and therefore subjects had to compare the current control dynamics with those of the previous control.

**Sensitivity and decision criteria by control type.** There were no effects on sensitivity or decision criteria for control type or an interaction between control type and experience level. The absence of effects is consistent with the lack of control type effects for detections.

**False alarms by subjects with manual and automatic control.** The false alarm rate was the same for subjects manually controlling and those monitoring the automatic control. This was expected since the decision criteria were expected to be the same for the two groups. There was no interaction between control type and experience level on the number of false alarms.

**Questionnaire Information**

Even though the subjects manually controlling the cursor movement rated the detection task as more difficult than the ratings of those with automatic control, their detection performance did not differ. The subjects using the manual control were able to compensate
for the increased difficulty of performing the detection task along with manually positioning the cursor, without a performance loss in the number of failures detected.
CHAPTER 4

EXPERIMENT 2

The goal of Experiment 2 was to investigate the nature of the signal and noise distributions for this detection task. If the signal and noise are normally distributed, the two traditional measures in signal detection theory, $d'$ and $\beta$, could be used to further investigate the difference in detection found in the initial failure for experienced subjects in Experiment 1. In Experiment 2, the subjects were encouraged to change their decision criteria, while the signal strength was held constant.

Experimental Method

Subjects

Three subjects participated in Experiment 2, two males and one female. They each reported for two sessions of 2.75 hours. The subjects were paid for their time. Their ages were: 22, 25 and 34 years. All three subjects were right handed. One of the subjects had participated in Experiment 1.
**Apparatus**

The apparatus used in Experiment 2 was the same as that used in Experiment 1.

**Experimental Design**

There were four conditions in Experiment 2. Each subject was assigned to each of the four conditions in a counter-balanced order. There were two probabilities of a set of dots being offset on a screen and two payoff structures for responding that an offset had or had not been present (Figure 18). The signal strength, i.e., the amount of the offset of the dots, was held constant. The two probabilities of a trial having an offset were 0.1 and 0.2.

The payoff matrix to encourage the subjects to state that a set of dots had been offset (shown in Figure 19) gave them: $1.00 for correctly saying that an offset had been present, 1¢ for correctly stating that no dots were offset, -10¢ for saying that a pair of dots was offset when none was, and -50¢ for saying that no dots were offset when there had been an offset. The payoff matrix to encourage negative responses (shown in Figure 20) gave them: 10¢ for correctly saying that a set of dots was offset, 25¢ for correctly saying that the dots were not offset, -50¢ for saying there was no offset when there was, and -20¢ for saying there was an offset when there was not.

**The Task**

The task performed by the subjects was similar to that of Experiment 1, except no tracking was required. The subjects activated the stimulus screen by placing the cursor in
### Figure 18. Experimental conditions for Experiment 2.

<table>
<thead>
<tr>
<th>Payoff favoring &quot;Offset&quot;</th>
<th>Probability of offset = 0.1</th>
<th>Payoff favoring &quot;No offset&quot;</th>
<th>Probability of offset = 0.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>same 3 Subjects</td>
<td>same 3 Subjects</td>
<td>same 3 Subjects</td>
<td>same 3 Subjects</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>200 manual trials</td>
<td>200 manual trials</td>
<td>200 manual trials</td>
<td>200 manual trials</td>
</tr>
<tr>
<td>20 failures</td>
<td>40 failures</td>
<td>20 failures</td>
<td>40 failures</td>
</tr>
</tbody>
</table>

The subject's task was to determine if any of the dots was not in the center of the squares. Only one set of dots was offset from center, if any. The amount of offset was reduced from that used in Experiment 1, from 6.25 mm (25 pixels) to 3.75 mm (15 pixels). Once the stimuli had been viewed, the subjects moved the cursor through the top of the screen and the stimuli disappeared. The response screen used in Experiment 1 then appeared. Subjects again placed a red dot cursor on the line corresponding to the vertical location of the offset dots or the word 'None'. They rated how sure they were that the dots were offset, from not sure to very sure. When they responded that none of the dots was offset,
they also rated their confidence from not sure to very sure. After their response was made, a screen appeared to inform them of the row of dots that had an offset or that none of the rows had an offset.

**Experimental procedure**

The near-vision of the subjects in Experiment 2 was tested. All three had normal (20:20) vision. The subjects were then asked to read the task instructions in Appendix C, with the payoff structure for the first condition only. The subjects were given five practice trials, in which all screens contained one set of dots that was offset. They were aware that one of the pairs of dots was offset and responded by selecting the row they believed contained the offset. The feedback screen informed them of the correctness of their selection. They were
There were 200 trials in each of the four conditions. The new payoff structure and percentage of offset targets were presented at the beginning of each block of 200 trials. Two conditions were completed in each of the two sessions (400 trials total). Each set of 200 trials required approximately 75 minutes to complete. The screen blanked after each set of 50 trials and input was required by the experimenter to begin the next set of trials. The subjects took a 10 minute break at the end of the first 200 trials in the session.

Experiment 2 Results

Detections and false alarms

The detection rates, false alarm rates and $d'$ and $\beta$, for the four conditions of signal probability and payoff, appear in Table 3. The overall detection rate for the four conditions was 0.660 and the false alarm rate was 0.032. The false alarm rates were low for two of the three subjects. The mean false alarm rate for subjects 1 and 2 was 0.007. Subject 3 had a mean false alarm rate of 0.081.
Table 3. Detection rates, false alarm rates, \( d' \) and \( \beta \) for the four conditions of signal probability and payoff.

<table>
<thead>
<tr>
<th>( p(s) )</th>
<th>Payoff</th>
<th>( S )</th>
<th>Detect Rate</th>
<th>Mean Detect Rate</th>
<th>F/A Rate</th>
<th>Mean F/A Rate</th>
<th>( d' ) (optimal)</th>
<th>( \beta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>no signal</td>
<td>1</td>
<td>0.80</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.10</td>
<td>no signal</td>
<td>2</td>
<td>0.65</td>
<td>0.01</td>
<td>2.67</td>
<td>12.45</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.10</td>
<td>no signal</td>
<td>3</td>
<td>0.55</td>
<td>0.67</td>
<td>0.10</td>
<td>1.41</td>
<td>2.25</td>
<td>(6.75)</td>
</tr>
<tr>
<td>0.10</td>
<td>signal</td>
<td>1</td>
<td>0.65</td>
<td>0.01</td>
<td></td>
<td></td>
<td>2.92</td>
<td>22.68</td>
</tr>
<tr>
<td>0.10</td>
<td>signal</td>
<td>2</td>
<td>0.75</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.10</td>
<td>signal</td>
<td>3</td>
<td>0.50</td>
<td>0.63</td>
<td>0.04</td>
<td>1.70</td>
<td>4.24</td>
<td>(0.66)</td>
</tr>
<tr>
<td>0.20</td>
<td>no signal</td>
<td>1</td>
<td>0.90</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td>3.00</td>
</tr>
<tr>
<td>0.20</td>
<td>no signal</td>
<td>2</td>
<td>0.65</td>
<td>0.01</td>
<td>2.89</td>
<td>21.13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.20</td>
<td>no signal</td>
<td>3</td>
<td>0.57</td>
<td>0.71</td>
<td>0.11</td>
<td>1.45</td>
<td>2.20</td>
<td>8.09</td>
</tr>
<tr>
<td>0.20</td>
<td>signal</td>
<td>1</td>
<td>0.92</td>
<td>0.01</td>
<td></td>
<td></td>
<td>3.94</td>
<td></td>
</tr>
<tr>
<td>0.20</td>
<td>signal</td>
<td>2</td>
<td>0.72</td>
<td>0.03</td>
<td></td>
<td></td>
<td>2.46</td>
<td>4.71</td>
</tr>
<tr>
<td>0.20</td>
<td>signal</td>
<td>3</td>
<td>0.37</td>
<td>0.67</td>
<td>0.07</td>
<td>1.16</td>
<td>2.84</td>
<td>(0.29)</td>
</tr>
</tbody>
</table>

**Sensitivities and decision criteria for failure detection**

In Figures 21-23 each subject's proportion of detections and false alarms are shown for each payoff condition and probability of offset. The mean detection rate and false alarm rate for the three subjects when the probability of an offset was 0.1 and the payoff favored a negative response were: 0.667 and 0.037 respectively. The average \( d' \) and \( \beta \) for this condition were 2.04 (sd = 0.89) and 7.35 (sd = 7.21). The optimal value of \( \beta \) for this condition is 6.75. Optimal \( \beta = \) ((payoff for correct rejection - deduction for false alarm) x probability of no offset) + ((payoff for detection - deduction for miss) x probability of offset). For the 0.2 offset probability condition favoring a negative response, the mean
A detection rate was 0.669 and the false alarm rate averaged 0.036, with a $d'$ and $\beta$ of 2.17 ($sd = 1.02$) and 11.66 ($sd = 13.38$). The optimal value of $\beta$ for this condition is 3.00.

![Figure 21. The probability of detection versus the probability of a false alarm for each payoff and offset probability for Subject 1.](image)

When a positive response was favored financially and the probability was 0.1 that an offset would be present on a screen, the detection rate was 0.633 and the false alarm rate was 0.017. The mean $d'$ was 2.31 ($sd = 0.86$) and the mean $\beta$ was 13.46 ($sd = 13.04$). The optimal value of $\beta$ for this condition is 0.66. The mean detection rate was 0.675 for the three subjects for the final condition, which had a payoff matrix that also favored a positive response with an offset probability of 0.2. Their false alarm rate in this condition
Figure 22. The probability of detection versus the probability of a false alarm for each payoff and offset probability for Subject 2.

was 0.036. For the final condition, the $d'$ was 2.52 ($sd = 1.39$) and the $\beta$ was 5.21 ($sd = 2.66$). The optimal value of $\beta$ for the final condition is 0.29.

**Optimal and measured decision criteria for Subject 3**

The subjects did not conform to an optimal decision criterion for the four conditions of offset probability and payoff. The optimal and measured values of $\beta$ for subject 3 are shown in Figure 24. This was the only subject who had a false alarm in every condition, and the only subject with a substantial number of false alarms.
The noise and signal distributions

Signal detection theory states that the noise and signal are normally distributed for the individual observers. If the noise and signal distributions are normal, a straight line will be formed by the $z$ scores of the probability of a detection plotted against the $z$ scores of the probability of a false alarm, for each condition of prior probability and payoff (Swets, 1964). To test the assumption of normality, the $z$ scores of the probabilities of false alarms and detections were plotted for each subject for each condition (Figure 25). The lack of linearity in the values indicates that the noise and signal distributions are not normally distributed for this detection task.
Figure 24. Optimal $\beta$ and $\beta$ measured for Subject 3 under the four conditions of signal probability and payoff.

**Prior probability and payoff effects on sensitivity**

An analysis of variance was performed on the $d'$ and the $\beta$ statistics for the three subjects in each of the prior probability and payoff conditions (where available, $n = 6$). The offset stimulus was the same in all these conditions. As expected, there were no differences in $d'$ as a result of changes either to the prior probability or the payoff structure. The interaction of signal probability and payoff produced no differences in sensitivity.
Prior probability and payoff effects on decision criteria

The subjects also did not show a difference in their decision criteria, as measured by $\beta$, for the two conditions of signal probability. The subjects also did not vary their decision criteria significantly in response to the payoff structures favoring a positive or a negative response. The interaction of signal probability and payoff structure was not significantly different for $\beta$. 
Experiment 2 Discussion

Experiment 2 demonstrated the subjects' response to changes in the probability that a set of dots would be offset from the center of the squares and the payoff for saying that a set had been offset. The constancy of the subjects' sensitivity (as measured by $d'$) to the presence of an offset was as expected across the four conditions of signal probability and payoff, since the dots had the same offset in all the experimental conditions. The smaller magnitude of $d'$ was also expected as compared with Experiment 1, where the offset had been greater.

In general the subjects were not able to adjust their decision criteria to optimize their payoff under the four conditions. This finding is not consistent with Signal Detection Theory (Green & Swets, 1973; Swets, 1961; 1973; 1986; Swets, Tanner, & Birdsall, 1961). This characteristic of subjects' inability to properly adjust to the experimental conditions of both prior probability and payoff has been experienced in many other studies (see Slovic & Lichtenstein, 1971). This is probably best observed in the difference between the optimal and measured $\beta$ for Subject 3 shown in Figure 24.

An alternative to signal detection theory is found in the high-threshold theory (see, e.g., MacMillan & Creelman, 1991). This theory claims that there are a few discrete states of the human detecting a signal, rather than a continuum of noise and signal distributions. In the simple two-state model, a high threshold detect state occurs on some proportion of signal trials, and very rarely on noise trials. In the non-detect state, the observer simply guesses
that a signal is present on some proportion of the trials. This theory, unlike signal detection theory, is consistent with a non-zero detection rate being accompanied by a near-zero false alarm rate, which was found in these studies. However, high-threshold theory does not explain the phenomenon of an adjustment in detection rate without a change in the false alarm rate, as occurred in these experiments. In addition, this theory does not explain the very low number of false alarms, which remained even when payoffs and probabilities were changed.

To account for such a phenomenon, one would postulate that on some proportion of the trials, determined by the payoff matrix and/or the signal probability, the subjects simply do not observe the primary signals. If on these occasions, the subjects do not guess that the signal is present, the performance on such trials will be in the lower left corner of the ROC space (zero hits and false alarms). A mixture of such performance with high threshold performance could account for a shift in the hit rate while the false alarm rate remains near zero.

The non-normality of the signal and noise distributions was observed in Figure 25. This finding has methodological consequences, since the metrics $d'$ and $\beta$ are only valid under conditions of normally distributed noise and signals. These data also do not support the high-threshold theory of detection that would predict an upwardly concave curvilinear function of the $z$ scores.
CHAPTER 5

EXPERIMENT 3

The goal of Experiment 3 was to determine whether subjects who receive many trials in which the secondary cues have proved to be redundant and easier to use, start to ignore the primary cues and use only the secondary cues. Do they, in fact, change the task, such that the search for and therefore the detection of such disagreement between the primary and secondary cues goes undetected, or do the subjects simply shift their decision criterion for reporting such a disagreement?

In Experiment 1, the subjects who had experienced many trials without a failure were less likely to detect the first failure. Only five of the sixteen subjects detected the first failure. After the first failure, there was an improvement in detection. Twice as many subjects detected the second failure (this improvement was not statistically significant). There was also a nearly significant increase in time to the first gate on the screen for the first two trials after the first failure, which is the point in the trial that some of the subjects reported checking for the offset. These findings suggest that when the subjects devoted more time to the task of detecting the failure, they were successful in that detection.

Research on discrete movements, such as the Fitts' task (1954), has shown that requiring the subject to perform an additional task, such as choosing the proper target, will increase
the time to complete the task. Specifically, the reaction time is increased, i.e., the time before the movement is begun (Hart, Sellers, & Guthart, 1984; Wickens, 1984). This will result in an increase in the time the cursor passes through successive targets. Research has also shown the complement effect, that reducing the task demands results in shorter times to react and complete the response. Ona (1990) found that when attention could be focused on making the movement, the movement times decreased.

Experiment 3 was designed to test the finding that subjects who have experienced a large number of trials without a failure will tend to either stop performing the detection task, or devote less time to detection. If this is the case, then the time to do the movements should be similar between trials where the detection is required but not performed, and trials where the detection task is removed by the experimenter. The assumption was made that the detection task had not become automatized and was not being done in parallel with the tracking task; the movement times would reflect the task loading which occurred when the detection task was being performed.

The task was changed from Experiment 1 to remove confounding effects of subject strategies regarding when they searched for the failure. The subjects in Experiment 1 either checked all sets of dots on the screen to determine whether any were centered before beginning the tracking or they looked for the failure on each set of dots as the tracking was performed. The task in Experiment 3 precluded doing the detection of all targets before beginning the first movement.
Experiment 3 Method

Subjects

There were eight subjects in each of three groups, six males and two females. The average age of the subjects was 31.25 years \((sd = 9.0)\). The ages did not differ significantly between the groups. All but one of the subjects were right handed. As before, the subjects in Experiment 3 were paid for their time. None of the 24 subjects in Experiment 3 had participated in Experiments 1 or 2.

The apparatus used in Experiment 3 was the same as in Experiments 1 and 2. The subjects used the electronic pen and tablet as an input device to control the cursor on the screen. A keyboard was added for the subjects to press the space key after detecting an offset pair of dots. The key press was made after the screen had cleared, usually with the hand not used for the pen.

Experimental Design

In Experiment 3 the Experimenter changed the task. The detection task was omitted in the third block of trials for Groups 1 and 2. There were three subject groups with eight subjects in each group. Group 1 received a block of 100 trials without a failure, then a
block of 100 trials with a failure on 10% of the trials, followed by a block of 50 trials with no failure detection required. See Figure 26. The subjects in Group 2 received a block of 100 trials with a failure on 10% of the trials, then a block of 100 trials with again a 10% failure rate, followed by 50 trials in which no failure detection was required. The subjects in Group 3 received a block of 100 trials with a 10% probability of failure, followed by a block of 100 trials with the same failure rate and a final block of 50 trials, also with failures on 10% of the trials. The transition between Blocks 1 and 2 was not obvious to any of the subjects. The transition between all blocks was not obvious to the subjects in Group 3.

The 100 trials in Block 1 were sufficient for movement time performance to approach asymptote. (This number was determined from Experiment 1, where 90% of the movement time learning effect, in the first 200 trials for the experienced subjects, was accomplished by trial number 93.) This permitted the comparison of MTs for Blocks 2 and 3 without confounding learning effects.

**The Task**

As in Experiment 1, the task in this experiment required the subjects to place the cursor through the midpoint of the squares, using the dots as the target area. However, the set of squares and dots did not appear until just before the cursor passed through the preceding set of squares and dots. This change in the task was made to prevent the subjects from checking all the targets for the offset at the beginning of each trial before performing the tracking. This change did not preclude them from doing the failure detection before the movement. However, each pair of dots and squares had to be examined one at a time. This change was instituted to encourage detection while tracking, and therefore reduce the variability in strategies between subjects.
Figure 26. Experimental design for Experiment 3.

The screens that would contain the failures were assigned randomly at the start of each set of 50 trials, as in Experiments 1 and 2. The set of dots to be offset on the screen was also assigned randomly, except in Experiment 3 the offset never occurred in the first or second pair of dots and squares on the screen. This change was implemented so that the subjects would continue to search for a failure, at least through the first three gates.
subjects were not told that there could be only one set of dots offset on a screen or that the first two sets could not contain an offset.

The offset of the dots from the center in this experiment was the same as in Experiment 2 (3.75 mm), but less than in Experiment 1. If the dots were not centered between the squares, the subject had to detect this failure. The response required to indicate the detection of a failure was to press the space key on the keyboard if an offset had been present. This response was made after the tracking, while the time to complete the trial was displayed. No detection response screen was presented requiring a rating of how sure they were that an offset had or had not been present. No action was required if there was no offset. This change in the task from Experiments 1 and 2 served to de-emphasize the detection task, as in the operational tasks targeted by this research, i.e., the operator is not asked if a failure occurred. For the trials in Block 3 for Groups 1 and 2, the keyboard was removed, to reinforce the absence of the detection task. The final change in the task in Experiment 3 was that there was no feedback of an offset's presence.

**Experimental Procedure**

Experiment 3 had a two hour duration for each subject. The subjects were assigned to an experimental group in a counter-balanced order, allowing for gender. The training for all groups of subjects was the same as Experiment 1. The subjects' instructions appear in Appendix D. The payoff matrix used in Experiment 3 was structured to encourage positive responses. The subjects in all groups performed Blocks 1 and 2 consecutively, with no indication of a change in the task. They were offered a break after the first 100 trials. Most subjects took about a 10 minute break at that point.
The task in Block 3 for Groups 1 and 2 was different from Experiments 1 and 2. No detection was required and no failures were present. The keyboard was removed so that the subject was unable to indicate the presence of a failure. The trials in Block 3 for subjects in Group 3 were the same as those in Blocks 1 and 2, requiring failure detection.

**Expected Results**

On the basis of the findings of Experiment 1, the subjects in Group 1 were expected to have poorer detection performance in Block 2 than the other two groups, since they previously had 100 trials with no failure. If this performance decrement was due to a change in decision criterion, Group 1 would be expected to have a higher $\beta$ than Groups 2 and 3. The movement times for Blocks 2 and 3 would be expected to be the same as those of Group 2 (see Figure 27). If the poorer performance by Group 1 was due to an inattention to the detection task, or a reduction in time spent on the task, the subjects in Group 1 would be expected to have a lower $d'$ (approaching zero) than that of Groups 2 and 3. If the latter were true, the time to complete the movement from one target to the next in Block 2 would be expected to be less for Group 1, since they would be spending less time looking for the offset. This experiment uses the converging measurements of signal detection and movement times to understand performance differences between subjects who are expecting a failure to occur and those who are not expecting a failure, due to their previous experience in which no failures occurred.

If the subjects in Group 1 stop performing the detection, or at least devote less time to the detection task as a result of observing the reliable redundancy of the secondary cues in
Figure 27. The predicted mean movement times (seconds) in trial Blocks 2 and 3 for the three subject groups. Using loading effects on movement time as a convergent measure for determining Group 1's change in detection performance.

Block 1, then the expected change in movement times from Block 2 to Block 3 would be minimal, as shown in Figure 27. Group 1 was expected to have the fastest movement times in Block 2 and little change in MTs for Block 3, if they had stopped performing the detection task. If the subjects in Group 1 shifted their decision criterion to become more conservative, their times would get shorter from Block 2 to Block 3 (Figure 27). They would perform the detection task in Block 2. In Block 3, when detection was no longer required of them, their times would get shorter. Group 3 was expected to have little change between Blocks 2 and 3 with the longest movement times of all groups in Block 3, since
they were still required to perform the detection task. Group 2 was expected to have movement times the same as Group 3 for Block 2, because there was no difference in their experimental conditions up to that point. However, Group 2 was expected to have faster movement times in Block 3, where no detection was required. The movement times should provide converging evidence for either a decision criterion shift or a change in sensitivity or attention to the failure for Group 1 as compared to Groups 2 and 3.

**Experiment 3 Results**

**Detections, False Alarms, and Positive Responses in Block 1**

The subjects in Groups 2 and 3 had ten failures in the 100 trials of Block 1. Their detection rate was 0.431 (sd = 0.287) and their mean false alarm rate was 0.084 (sd = 0.158). The detection rate, the false alarm rate and the rate of positive ("yes") responses for each subject appear in Table 4. There was no difference in the false alarm rates for Group 1 and Groups 2 and 3 in Block 1.

The rate of positive responses, i.e., the number of trials in which the subject said there was an offset, was calculated for Block 1. The subjects in Group 1 had no failures, and Groups 2 and 3 had a 10% failure rate in the 100 trials in Block 1. This measure was used as an indication of the decision criterion, since \( \beta \) could not be used, due to the lack of normality of the signal and noise distributions as measured in Experiment 2. In addition, 5 of the 24 subjects had no false alarms, making the calculation of \( \beta \) for them impossible.
<table>
<thead>
<tr>
<th>Group</th>
<th>Subject</th>
<th>Detect Rate</th>
<th>Mean Detect Rate</th>
<th>F.A. Rate</th>
<th>Mean F.A. Rate</th>
<th>&quot;Yes&quot; Rate</th>
<th>Mean &quot;Yes&quot; Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.300</td>
<td>0.033</td>
<td>0.060</td>
<td>0.130</td>
<td>0.070</td>
<td>0.156</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0.900</td>
<td>0.044</td>
<td>0.040</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0.300</td>
<td>0.011</td>
<td>0.040</td>
<td>0.100</td>
<td>0.070</td>
<td>0.156</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>0.900</td>
<td>0.011</td>
<td>0.040</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>0.100</td>
<td>0.000</td>
<td>0.040</td>
<td>0.010</td>
<td>0.070</td>
<td>0.156</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>1.000</td>
<td>0.133</td>
<td>0.180</td>
<td>0.180</td>
<td>0.156</td>
<td>0.156</td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>0.400</td>
<td>0.156</td>
<td>0.180</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>0.700</td>
<td>0.525</td>
<td>0.178</td>
<td>0.230</td>
<td>0.116</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.200</td>
<td>0.011</td>
<td>0.030</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.100</td>
<td>0.011</td>
<td>0.020</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0.800</td>
<td>0.000</td>
<td>0.080</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>0.400</td>
<td>0.000</td>
<td>0.040</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>0.100</td>
<td>0.078</td>
<td>0.080</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>0.100</td>
<td>0.033</td>
<td>0.040</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>0.600</td>
<td>0.633</td>
<td>0.630</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>0.400</td>
<td>0.338</td>
<td>0.011</td>
<td>0.097</td>
<td>0.050</td>
<td>0.121</td>
</tr>
</tbody>
</table>

Table 4. Detection rates, false alarm rates, and positive response rates, for each subject in the 100 trials of Block 1.

The rate of positive responses for Group 1 was the false alarm rate, 0.070 (sd = 0.096, median = 0.020) on the 100 trials in Block 1. Groups 2 and 3 had a mean positive response rate of 0.119 (sd = 0.151). There was no significant difference between the positive response rates of Group 1 and Groups 2 and 3 in Block 1.
There was no correlation between the number of false alarms in Block 1 and the number of correctly identified failures in Block 2 for the subjects in Group 1. However, there was a positive correlation between the number of false alarms in Block 1, which represents the positive response rate, and the number of positive responses in Block 2 for the subjects in Group 1 ($r(8) = 0.885, p < .01$).

Although Group 1 received no failures in Block 1 of this experiment, six of the eight subjects may have believed that there were signals in Block 1 (suggested by their false alarm rates). Two subjects reported no failures and one reported 26, with the mean for the group 0.070 false alarms. This differs from Experiment 1 where the false alarm rate for the experienced subjects for trials 1-200 with no signals was 0.020 positive responses per trial and only nine of the sixteen subjects had any positive responses. Since there was no feedback as to the presence of an offset in Experiment 3, the misperception that an offset was present did not get corrected. Therefore, the majority of the subjects in Group 1 did not build an experience base without failures in Block 1.

One of the eight subjects in Group 1 who did not have any false alarms also did not see any of the failures in Block 2, nor report any false alarms in that block. That is, that subject never responded positively to the presence of a failure, throughout the phases of the experiment. This subject, as well as two others from Group 1, one subject from Group 2, and two from Group 3 all had detection performance no better than chance in Block 2. Since three of eight subjects in Group 1 fell into this category and only three of sixteen subjects from Groups 2 and 3 had performance equivalent to random responses, these proportions were tested for significance. The difference in the proportion of subjects from Group 1 compared to Groups 2 and 3 was not significant.
The six subjects whose detection performance was at or below the level of chance were collectively placed in a new Group 4. The remaining subjects in the other groups were labeled Groups 1', 2', and 3'. The Group 4 subjects had a mean positive response rate in Block 1 of 0.140. The positive response rate for Group 4, compared to those subjects who performed the detection task (Groups 1', 2', and 3' combined) did not differ significantly in Block 1. Group 1' was removed from the analysis since no failures were presented to them in Block 1 and Group 4 was then compared to Groups 2' and 3' combined. There was no difference in the rate of positive responses in Block 1 for Group 4 and Groups 2' and 3'.

The rate of positive responses for Group 1' versus Groups 2' and 3' in Block 1 and Block 2 are shown in Figure 28. The positive response rate for these groups did not differ in Block 1; the overall mean was 0.090 positive responses per trial. The positive response rates for Block 2 are discussed below.

**Detections in Block 2**

**Detection performance by group.** The detections were measured for the 100 trials in Block 2. The detection performance was expected to be the same for Groups 2 and 3, since they had experienced the same rates of failures in all preceding trials. The detection performance was expected to be worse for Group 1, since they had a preceding block of 100 trials without failures. The dependent measures to be evaluated were detection rate, false alarms, positive response rates and the proportion of subjects detecting the first failure. Note that $d'$ and $\beta$ were evaluated for subjects who had false alarms,
Figure 28. The positive response ("yes") rate for Group 1' versus the combined Groups 2' and 3' for trials in Blocks 1 and 2.

However the lack of evidence for normality in the noise and signal distributions found in Experiment 2 gives these measures questionable validity.

The Block 2 detection rates, false alarm rates, and the rates of responding positively to the presence of a failure, are shown for each subject in Table 5. In Figures 29-31 each subject's proportion of detections and false alarms are shown for Groups 1, 2, and 3. The mean detection rate for the eight subjects in Group 1 was 0.325 ($sd = 0.293$) and for Groups 2 and 3 was 0.275 ($sd = 0.246$). The detection performance of Group 1 was not
statistically different from the other groups. As can be seen from these relative operating characteristic (ROC) plots, some subjects' performance falls at or below chance. Five of these subjects had no detections in block 2. As described above, these subjects with performance equivalent or worse than if they were responding randomly, i.e., a $d' \leq 0$, were extracted from the three original groups for further analysis.

**Measures of sensitivity and decision criteria.** The measures of sensitivity to the failure and decision criteria are of questionable validity due to the lack of normality in the noise and signal distributions found in Experiment 2. The sensitivity, $d'$, showed no difference between Group 1 and Groups 2 and 3. The mean $d'$ for Group 1 was 1.45 ($sd = 1.11, n = 3$). The mean $d'$ for Groups 2 and 3 was 1.47 ($sd = 1.00, n = 14$). The measure of decision criterion, $\beta$, for Group 1 was not significantly different from that of Groups 2 and 3. $\beta$ averaged 5.78 for Group 1 ($sd = 6.34, n = 3$) and 8.08 for Groups 2 and 3 ($sd = 7.02, n = 14$). Note that no $d'$ or $\beta$ parameters were available for subjects that had either no detections or no false alarms, which included five of eight subjects in Group 1 and eight of 16 subjects in Groups 2 and 3.

**Detections for the four groups.** The detection rates were compared for the four groups formed when the subjects who were not performing the detection task were placed in Group 4. (These subjects are denoted with an * in Table 5 and the four groups appear in Table 6.) There was a significant difference in detection rates for these four groups ($F(3, 20) = 6.23, p < 0.01$). The Block 2 detection rates for Groups 1', 2', and 3' were 0.52, 0.39, and 0.25, respectively. The detection rate for Group 4 was 0.03. When Group 4 was removed from the analysis, Groups 1', 2' and 3' were not statistically different from each other in their detection rates. When Group 4 was compared to the other
<table>
<thead>
<tr>
<th>Group</th>
<th>Subject</th>
<th>Detect Rate</th>
<th>Mean Detect Rate</th>
<th>F.A. Rate</th>
<th>Mean F.A. Rate</th>
<th>&quot;Yes&quot; Rate</th>
<th>Mean &quot;Yes&quot; Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.500</td>
<td>0.067</td>
<td>0.110</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0.300</td>
<td>0.200</td>
<td>0.210</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0.500</td>
<td>0.000</td>
<td>0.050</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4*</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>0.700</td>
<td>0.000</td>
<td>0.070</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>6*</td>
<td>0.000</td>
<td>0.044</td>
<td>0.040</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>7*</td>
<td>0.000</td>
<td>0.011</td>
<td>0.010</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>0.600</td>
<td>0.325</td>
<td>0.011</td>
<td>0.042</td>
<td>0.070</td>
<td>0.070</td>
</tr>
<tr>
<td>2</td>
<td>1*</td>
<td>0.000</td>
<td>0.056</td>
<td>0.050</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.500</td>
<td>0.000</td>
<td>0.050</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0.100</td>
<td>0.000</td>
<td>0.010</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>0.900</td>
<td>0.000</td>
<td>0.090</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>0.100</td>
<td>0.000</td>
<td>0.010</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>0.300</td>
<td>0.033</td>
<td>0.060</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>0.200</td>
<td>0.033</td>
<td>0.050</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>0.600</td>
<td>0.337</td>
<td>0.144</td>
<td>0.033</td>
<td>0.190</td>
<td>0.064</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.400</td>
<td>0.000</td>
<td>0.040</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.200</td>
<td>0.022</td>
<td>0.040</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0.500</td>
<td>0.011</td>
<td>0.060</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>0.100</td>
<td>0.000</td>
<td>0.010</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>0.100</td>
<td>0.022</td>
<td>0.030</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>0.200</td>
<td>0.056</td>
<td>0.070</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>7*</td>
<td>0.200</td>
<td>0.289</td>
<td>0.280</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>8*</td>
<td>0.000</td>
<td>0.212</td>
<td>0.000</td>
<td>0.050</td>
<td>0.000</td>
<td>0.066</td>
</tr>
</tbody>
</table>

Table 5. Detection rates, false alarm rates, and positive response rates, for each subject in the 100 trials of Block 2. (The subjects denoted with an * performed at a chance detection rate and were placed in Group 4.)

three groups in combination the detection rates were significantly different ($F(1, 22) = 11.93, p < 0.01$).

**Detection rates over time.** The detection rates over time were compared for the four groups. These data appear in Figure 32. There was no main effect of time (failure
Figure 29. The probability of detection versus the probability of a false alarm for each subject in Group 1.

Pair). The groups did not differ in their detection rates by failure pairs across the 10 failures in Block 2, i.e., there was no group by failure pair interaction.

**Block 1 positive response rates and Block 2 detection rates.** The Block 1 positive response rate was used as an indication of the effective signal rate for each of the subjects. There was no correlation between the Block 1 positive response rates and the Block 2 detection rates for all subjects combined or for just Group 1 subjects.
Detection of the first failure. In Experiment 3 there was no difference in the number of subjects detecting the first failure between the subjects who had previous trials with a failure (Groups 2 and 3) and those who had no failures in previous trials (Group 1). The number of subjects detecting the first failure in Group 1 in this experiment was three of the eight (37%). The number of subjects in Groups 2 and 3 detecting the first failure was five, of the 16 subjects in these two groups (31%). This finding differs from Experiment 1.
Detection in Experiment 1 versus 3. The detection rates for Experiment 3 were worse than for Experiment 1 \((F(1, 54) = 40.59, p < 0.001)\). The mean detection rate for all subjects in Experiment 1 was 0.703 \((\text{sd} = 0.202)\) and the mean detection rate for all subjects in Experiment 3 was 0.314 \((\text{sd} = 0.249)\). This was expected since the offset was twice the distance in Experiment 1. There were no subjects in Experiment 1 with detection performance at or worse than chance. In Experiment 3, the detection performance of six subjects was at or worse than chance \((d' \leq 0)\).
### Table 6. Detection rates, false alarm rates, and positive response rates, for each subject in the four groups for the 100 trials of Block 2.

<table>
<thead>
<tr>
<th>Group</th>
<th>Subject</th>
<th>Detect Rate</th>
<th>Mean Detect Rate</th>
<th>F.A. Rate</th>
<th>Mean F.A. Rate</th>
<th>&quot;Yes&quot; Rate</th>
<th>Mean &quot;Yes&quot; Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1'</td>
<td>1</td>
<td>0.500</td>
<td></td>
<td>0.067</td>
<td></td>
<td>0.110</td>
<td></td>
</tr>
<tr>
<td>1'</td>
<td>2</td>
<td>0.300</td>
<td></td>
<td>0.200</td>
<td></td>
<td>0.210</td>
<td></td>
</tr>
<tr>
<td>1'</td>
<td>3</td>
<td>0.500</td>
<td></td>
<td>0.000</td>
<td></td>
<td>0.050</td>
<td></td>
</tr>
<tr>
<td>1'</td>
<td>4</td>
<td>0.700</td>
<td></td>
<td>0.000</td>
<td></td>
<td>0.070</td>
<td></td>
</tr>
<tr>
<td>1'</td>
<td>5</td>
<td>0.600</td>
<td>0.520</td>
<td>0.011</td>
<td>0.056</td>
<td>0.070</td>
<td>0.102</td>
</tr>
<tr>
<td>2'</td>
<td>1</td>
<td>0.500</td>
<td></td>
<td>0.000</td>
<td></td>
<td>0.050</td>
<td></td>
</tr>
<tr>
<td>2'</td>
<td>2</td>
<td>0.100</td>
<td></td>
<td>0.000</td>
<td></td>
<td>0.010</td>
<td></td>
</tr>
<tr>
<td>2'</td>
<td>3</td>
<td>0.900</td>
<td></td>
<td>0.000</td>
<td></td>
<td>0.090</td>
<td></td>
</tr>
<tr>
<td>2'</td>
<td>4</td>
<td>0.100</td>
<td></td>
<td>0.000</td>
<td></td>
<td>0.010</td>
<td></td>
</tr>
<tr>
<td>2'</td>
<td>5</td>
<td>0.500</td>
<td></td>
<td>0.033</td>
<td></td>
<td>0.060</td>
<td></td>
</tr>
<tr>
<td>2'</td>
<td>6</td>
<td>0.200</td>
<td></td>
<td>0.033</td>
<td></td>
<td>0.050</td>
<td></td>
</tr>
<tr>
<td>2'</td>
<td>7</td>
<td>0.600</td>
<td>0.386</td>
<td>0.144</td>
<td>0.030</td>
<td>0.190</td>
<td>0.066</td>
</tr>
<tr>
<td>3'</td>
<td>1</td>
<td>0.400</td>
<td></td>
<td>0.000</td>
<td></td>
<td>0.040</td>
<td></td>
</tr>
<tr>
<td>3'</td>
<td>2</td>
<td>0.200</td>
<td></td>
<td>0.022</td>
<td></td>
<td>0.040</td>
<td></td>
</tr>
<tr>
<td>3'</td>
<td>3</td>
<td>0.500</td>
<td></td>
<td>0.011</td>
<td></td>
<td>0.060</td>
<td></td>
</tr>
<tr>
<td>3'</td>
<td>4</td>
<td>0.100</td>
<td></td>
<td>0.000</td>
<td></td>
<td>0.010</td>
<td></td>
</tr>
<tr>
<td>3'</td>
<td>5</td>
<td>0.100</td>
<td></td>
<td>0.022</td>
<td></td>
<td>0.030</td>
<td></td>
</tr>
<tr>
<td>3'</td>
<td>6</td>
<td>0.200</td>
<td>0.250</td>
<td>0.056</td>
<td>0.019</td>
<td>0.070</td>
<td>0.042</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.000</td>
<td></td>
<td>0.000</td>
<td></td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>0.000</td>
<td></td>
<td>0.044</td>
<td></td>
<td>0.040</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>0.000</td>
<td></td>
<td>0.011</td>
<td></td>
<td>0.010</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>0.000</td>
<td></td>
<td>0.036</td>
<td></td>
<td>0.050</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>0.200</td>
<td></td>
<td>0.289</td>
<td></td>
<td>0.280</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>0.000</td>
<td>0.033</td>
<td>0.000</td>
<td>0.067</td>
<td>0.000</td>
<td>0.063</td>
</tr>
</tbody>
</table>

**Detection in Experiment 2 versus 3.** The detection rates for the subjects in Experiment 3 were worse than in the 10% failure conditions of Experiment 2 ($F(1, 17) = 5.62, p < 0.05$), even though the offset distance was the same. The mean detection rates of Blocks 1 and 2 for subjects in Groups 2 and 3 in Experiment 3 was 0.320 ($sd = 0.231$). For Experiment 2 subjects in the 10% offset probability condition, the mean detection rate was 0.650 ($sd = 0.109$). The $d'$ measure was not significantly different for the two
Figure 32. Detection rates by the four groups across the five pairs of failures in Block 2.

experiments. However, there was only a small number of subjects for which $d'$ was available.
False Alarms in Block 2

**False alarm rates by group.** There were failures reported when none was present in Block 2 for all groups. The false alarm rates for each subject are shown in Table 5. The false alarm rate was identical for Group 1 and Groups 2 and 3. The eight subjects in Group 1 had a mean false alarm rate of 0.042 with a standard deviation of 0.068 for the 100 trials in Block 2. The mean false alarm rate for the sixteen subjects in Groups 2 and 3 was also 0.042 and their standard deviation was 0.076.

**False alarm rates over time for the four groups.** Table 6 shows the false alarms for each subject in the groups formed when those not performing the detection task were placed in Group 4. These four groups were analyzed for changes in their false alarm rates over time for Block 2. It was hypothesized that the false alarm rates would decrease over time, due to subjects becoming more conservative in their decision criteria. It was further hypothesized that this decrease would vary across the experimental groups. The false alarm rates were calculated for each set of 20 trials in Block 2 (Trials 101-200). These rates are shown for each of the four groups in Figure 33. The false alarm rate decreased significantly by twenty-trial sets ($F(4, 80) = 2.84, p < 0.05$). There was no group effect in the false alarm rates and the groups did not differ in the effect of time across Block 2, i.e., there was no group by trial set interaction. Note that the limited number of subjects in each group and the low number of false alarms results in little power in the statistical comparison of these groups over the trial sets.

In Figure 33 it is apparent that the major decrease in false alarm rate occurred between Trials 101-120 and Trials 121-140. When the first 20 trials in Block 2 (Trials 101-120) were removed and an analysis of variance performed, the effect of time was no longer
Figure 33. Block 2 false alarm rates for the four groups in twenty-trial sets.

significant. In this analysis as well, there was no group effect and no group by trial set interaction. It is noteworthy that this decrease in false alarm rate occurred after the offsets began to appear at a 10% rate. This may be due to the lack of feedback in this experiment. It is also interesting to note that the decrease in false alarm rate occurred at approximately
the same number of trials as the decrease in Experiment 1. In that experiment there was feedback, but no failures were present.

**False alarm rate for the four groups by block.** The false alarm rate of Group 1' was compared by block to the rates of Groups 2' and 3' combined. There was no difference in the false alarm rates of Group 1' and Groups 2' and 3'. There was also no interaction of group by block. There was, however, a block effect. The false alarm rates decreased from Block 1 to Block 2 ($F(1, 16) = 7.36, p < .05$). These rates are shown in Figure 34.

**Positive Response Rate in Block 2**

The false alarm rate and detection rate were combined to determine if there was a difference in the groups' tendencies to say that a failure was present. The rate of responding positively in Block 2 was $0.070 (sd = 0.067)$ on the average for the subjects in Group 1. The mean positive response rate was $0.065 (sd = 0.073)$ for the Group 2 and 3 subjects (see Table 5). The positive response rates did not differ between Group 1 and Groups 2 and 3 combined.

The positive response rates for Block 2 for Groups 1' versus Groups 2' and 3' are shown in Figure 28. The difference in the positive response rates in Block 2 was not significant. For Blocks 1 and 2 for Groups 1' versus Groups 2' and 3' there was no group
Figure 34. The false alarm rate for Group 1' and Groups 2' and 3' combined for Block 1 and Block 2.

effect, no block effect, and no interaction of group by block. There was also no difference in the rate of positive responses in Block 2 for Group 4 and Groups 2' and 3' combined.
Movement Times

The movement times were examined for the normal targets, i.e., targets containing no offset of the dots. Experiment 1 showed that movement times to offset dots were greater. For this reason, these movement times were excluded for the analysis across blocks. The movement times for all eight targets on a screen were combined for analysis. A second analysis was conducted on the movement time of the fourth and fifth targets on a screen. These center targets tended to have less variability and it was thought the movement times to these targets may more accurately reveal the underlying movement times of the subjects, absent of starting and stopping effects.

Block effects. The mean movement times to move through all eight gates on a screen for Blocks 2 and 3 are shown in Figure 35, for each group. In Block 3, the subjects in Groups 1 and 2 were not required to perform the detection task. This was expected to reduce the movement times. For Groups 1 and 2 the movement times were shorter for the last 45 screens in Block 3 ($M = 1.29$ seconds, $sd = 0.20, n = 720$) than for the last 45 screens without an offset in Block 2 ($M = 1.81$ seconds, $sd = 0.54, n = 720$) ($F(1, 14) = 10.44, p < 0.01$). When only the data from Group 1 were included in the analysis, the finding was the same, the movement times decreased significantly when the detection task was removed ($F(1, 7) = 17.76, p < 0.01$). As expected, the MTs did not vary on Blocks 2 and 3 for Group 3, since there was no difference in their task in the two blocks of trials. When the MTs for all three Groups were included in an analysis of variance, the reduction in MTs for Block 3 was also significant ($F(1, 21) = 12.25, p < 0.01$).

The mean movement times for targets four and five were also significantly different between Block 2 and Block 3 ($F(1, 21) = 13.69, p < 0.01$). The mean movement time for
Block 2 was 1.57 seconds ($sd = 0.44$). The mean movement time for the fourth and fifth movements in Block 3, where no detection was required for Groups 1 and 2, was 1.37 seconds ($sd = 0.33$).

**Group effects.** The movement times for Blocks 2 and 3, differed across groups ($F(2, 21) = 3.79, p < 0.05$). Group 1 had the lowest mean movement time for the eight targets on a screen ($M = 10.52$ seconds, $sd = 1.66, n = 720$). Group 2 had the highest mean
movement time ($M = 13.47$ seconds, $sd = 3.96, n = 720$). Group 3 was in the middle ($M = 11.86, sd = 4.04, n = 720$).

The mean movement times for targets four and five followed the same pattern as the total movement times ($F(2, 21) = 4.29, p < 0.05$). Group 1 averaged 1.28 seconds ($sd = 0.23$). Group 2 had a mean movement time of 1.63 seconds ($sd = 0.46$). The Group 3 mean movement time for the fourth and fifth targets on a screen was 1.50 seconds ($sd = 0.40$).

When the subjects were considered as four groups, there were no differences between the groups in their movement times. The subjects in Groups 1' and 2' were combined and compared with the subjects in Groups 3' and 4, i.e., the subjects whose task did not change from Block 2 to 3. There was no group effect for these combined groups.

The mean movement time for all trials in Block 2 for each subject was subtracted from the mean movement time for that subject in Block 3. The resultant time was their movement time adjusted for the time to complete a screen when the task was the same (Block 2). Thus the movement times could be compared by eliminating group differences that were present when the tasks were the same. There was no group effect when Groups 1' and 2' were compared to Groups 3' and 4.

**Block and group interaction.** There was no interaction of block and group (Groups 1, 2, and 3) effects on total movement times, for Groups 1, 2, and 3. The interaction of block and group was nearly significant for the mean movement time of the fourth and fifth targets ($F(2, 21) = 3.21, p = 0.061$). The pattern was the same for the movement times to the fourth and fifth targets as for the mean movement time for all eight
targets (see figure 35). The slope for Group 1 was between that of Group 2 and Group 3. There was no interaction of the four groups with the block effect (see Figure 36). There was also no interaction of Groups 1' and 2' combined and Groups 3' and 4 and block.

![Figure 36. The mean movement time (seconds) for each gate for the four groups for Block 2 and 3.](image)

Figure 36. The mean movement time (seconds) for each gate for the four groups for Block 2 and 3.
Movement Errors

The number of movement errors was tabulated to examine the presence of a speed-accuracy tradeoff for the movements made through normal targets. The average error rate for the 1190 movements to normal targets in Blocks 2 and 3 for all subjects was 0.030, the standard deviation was 0.026 (n = 24 subjects), the subject with the minimum error rate had a rate of 0.003, and the subject with the maximum rate had 0.083.

There was a significant increase in the number of errors from Blocks 2 to Block 3 (F(1, 21) = 10.37, p < .01). The mean error rate in Block 2 for all subjects was 0.022 (sd = 0.021). The mean error rate in Block 3 was 0.038 (sd = 0.028).

The speed-accuracy tradeoff of movement time and error rate was tested by the correlation of these two parameters. There was a significant negative correlation between each subject's mean movement time and error rate for Blocks 2 and 3 (r(46) = -0.532, p < .01).

The number of errors was nearly significantly different for the three experimental groups (F(2, 21) = 3.34, p = 0.055). Group 1 subjects had a mean error rate of 0.044 (sd = 0.033). Group 2's mean error rate was 0.019 (sd = 0.018). Group 3 had a mean error rate of 0.026 (sd = 0.019). The task did not change between Block 2 and 3, for the subjects in Group 3; detection was required in Block 3 as well. There was no interaction between groups and blocks for the number of errors. Analyzing the movement errors for the four groups, there was no group effect for the error rate and the group by block effect was not significant. These data are shown in Figure 37.
Figure 37. The mean error rate for the four groups for Block 2 and 3.
Experiment 3 Discussion

Detection

The lack of differences in detection between Group 1 and Groups 2 and 3 was contrary to expectation. From the results of Experiment 1, it was predicted that Group 1 subjects would be less likely to detect a failure than Group 2 and 3 subjects, especially the first failure. Group 1 subjects were expected to omit the detection task or devote less time or attention to it. This was not the case. (Note that Experiment 3 analyses in general did not have the power of Experiment 1 due to the smaller number of subjects, especially when broken into four groups.)

One possible explanation for Experiment 3 not replicating this finding of Experiment 1 is that the signal strength was low enough that the Group 1 subjects perceived failures in the first 100 trials where none was present. The lack of feedback in Experiment 3 did not permit these misperceptions to be corrected. The rate of positive responses, i.e., the subject said that an offset was seen, for Group 1 was 0.07 in Block 1. In contrast, the positive response rate of the experienced subjects in Experiment 1 for the first 200 trials with no failures was 0.01. The positive correlation of the rate of positive responses in Block 1 (where no failures were present) and Block 2 (where failures were present) for Group 1 is further support for the argument that subjects in Group 1 in Experiment 3 did not, in actuality, experience an initial period without failures.

For these same reasons, the smaller offset being harder to detect and the lack of feedback, there were conversely subjects in Groups 2 and 3 who effectively had no failures
in Block 1. Since they did not perceive any failures, they stopped performing the detection task. This was evidenced by their detection performance falling below chance in Block 2. Those subjects were combined with the three subjects in Group 1 who stopped performing the detection and collectively formed Group 4. These subjects were significantly less likely to detect a failure in Block 2, yet their positive response rates for Block 1 were no different, suggesting that neither their decision criteria nor their perceived probabilities of a failure were different as they began to search for the failures in Block 2.

The Block 1 positive response rate for Group 4 also did not differ significantly from Groups 2' and 3' combined who had failures in Block 1. It is noteworthy that the Block 2 positive response rate for Group 4 also did not differ from Groups 2' and 3'. Although the Group 4 subjects were not performing the detection task, as indicated by their detection rate falling below the chance level, they continued to make positive responses at the same rate as the groups who had received failures all the way through Blocks 1 and 2. It must be pointed out also that the Block 1 positive response rate for each subject, which can be considered the perceived signal rate, did not correlate with the detection rate in Block 2.

Of the subjects who continued to perform the detection task, i.e., the subjects in Groups 1', 2', and 3', there was no indication of a more conservative decision criterion for the subjects in Group 1'. The Group 1' subjects had no failures in Block 1 and yet their decision criteria, as indicated by their positive response rates, were the same as the subjects in Groups 2' and 3'. This is consistent with the experienced and inexperienced subjects in Experiment 1 and contrary to the literature on vigilance in low signal probability conditions.

A further difference between Experiments 1 and 3 is the lack of improvement in detection performance over time. The detection rate of both the experienced and
inexperienced subjects in Experiment 1 increased with subsequent failures. The detection rates of the four groups in Experiment 3 did not improve with exposure to the failures.

The detection performance in Experiment 3 can be compared with the 10% failure conditions of Experiment 2. The offset was the same in these two experiments. The poorer detection performance in Experiment 3 can be attributed to not only the lack of feedback, but also to the addition of the tracking task. In the debriefing, some subjects reported being so involved in the tracking task, that they did not look for the offsets, even though they knew it was very important to find them. This finding is similar to that reported by Sorkin, Kantowitz, and Kantowitz (1988), who found that the addition of a tracking task reduced $d'$. 

**False Alarms Rates Over Time**

The subjects in all groups became more conservative in their decision criteria over time, as evidenced by the decrease in the false alarm rate. This is the case when comparing false alarm rates from Block 1 to Block 2 and across the twenty-trial sets during Block 2. However this effect was concluded in the first 20 trials of Block 2. This adjustment in false alarm rate was also seen in the experienced subjects in Experiment 1.

When the subjects in Group 1 experienced failures during Block 2, they did not exhibit different decision criteria than the subjects in Groups 2 and 3, as measured by the false alarm rate. It is interesting that there was no group effect and no group by time interaction, even when the subjects that were performing at a chance level of detection were segregated into Group 4. There was also no difference between Group 1' and Groups 2' and 3' from
Block 1 to Block 2. This is further support for the finding in Experiment 1 that the subjects who had no failures in the first set of trials were operating at the same level of decision criteria as those who experienced failures all through the experiment.

**Movement Times**

**Block effects.** The improvement from Block 2 to Block 3 for movement times for Group 2 was as expected, since the detection task required time to perform and was removed for Block 3. Since the subjects in Group 1 did not relax their search for failures in Block 1 and 2, the movement times improved when the detection task was removed in Block 3. They performed similarly to the subjects in Group 2, as shown in Figure 35.

As expected, Group 3 showed no change in movement times between Blocks 2 and 3. The task was the same in these blocks for Group 3 subjects. As predicted, sufficient practice had been accomplished in Block 1, such that significant learning effects were no longer present.

**Group effects.** The subjects in the three groups differed in the speed with which they passed the cursor through the center of the squares. While the magnitude of the speed varied between the groups, the relative change between Block 2 and Block 3 for Groups 1 and 2 did not vary. There was no interaction between blocks and groups for the movement times of Groups 1 and 2. The two groups had the same incremental improvement in movement times from Block 2, where the detection task was being performed, to Block 3, which required no detection task.
Block and Group interaction. Considering the four Groups, the interaction of group and block was not significant, however, an interesting observation can be made. As can be seen in Figure 36, the Group 1' and Group 2' subjects had a decrease in movement times when comparing Block 2 with Block 3. The slopes of these two groups are similar. This is expected, since these subjects were performing the detection in Block 2 and not in Block 3. In contrast, the subjects in Group 3' and Group 4 had no change in their task from Block 2 to Block 3, the former by experimental design and the latter due to their not performing the detection task in Block 2 when failures were present. The slopes of these two groups are similar and Group 4's times were less than Group 3'. These findings are consistent with the hypothesis that Group 4 subjects had stopped performing the detection task in Block 2.

Movement Errors

The increase in movement errors for Groups 1 and 2 from Block 2 to Block 3 was not expected. The subjects perhaps were trying to reduce the movement times and therefore shifted their speed-accuracy tradeoff. This finding confounds the interpretation that the reduction in MT was due to the elimination of the detection task.
CHAPTER 6

GENERAL DISCUSSION

Summary of General Findings

Effects of experience on failure detection. This series of studies addressed the detection of a failure in the context of a tracking task. The failure was a disagreement between the primary cues and the secondary cues. The task, although created for this research, has traits representative of tasks in the operational environment in which secondary cues prove to be redundant and easier to use. After experience with the primary and secondary cues reliably in agreement, operators have been observed to stop checking for the disagreement between these cues. This was shown to be true in this research, even though the subjects were aware that the failure could occur and had experience detecting it.

In Experiment 1 the subjects who had previous experience with reliable secondary cues were less likely to detect the first failure than the subjects with no previous experience. This deficiency in detection was immediately overcome by the experienced subjects. After the first failure, they started to perform the detection task. The overall detection performance did not differ for the experienced and inexperienced subjects. Note that there was feedback regarding the presence of the failure in Experiment 1.
In Experiment 3 the same pattern was expected but did not emerge. The subjects in Group 1 only had previous experience with reliable secondary cues, and those in Groups 2 and 3, had experienced offsets previously. The same number of subjects in these two groups detected the first failure. The explanation that is most convincing is that the signal strength was weak enough to produce false alarms, and when these were not corrected by feedback, the subjects believed a failure had been present; the secondary cues were not perceived as redundant. The offset of the dots from the center of the squares was greatly reduced in Experiment 3; $d'$ averaged 1.47 in Experiment 3, versus 2.62 in Experiment 1. Six of the twenty-four subjects had $d'$ less than or equal to zero; their detection performance was no better than chance. The result was that no detection decrement occurred for subjects who had not experienced the failure. That is, the perceived failures were sufficient to create an experience of failures.

An examination of the differences in detection of the first failure across Experiments 1 and 3 further confirms this explanation. It is not surprising that the number of subjects in Group 2 and 3 in Experiment 3 that detected the first failure was less, five of sixteen as opposed to twelve of sixteen inexperienced subjects in Experiment 1. The number of subjects who detected the first failure in Group 1 (38%) was effectively the same as the numbers in Groups 2 and 3, (38% and 24% respectively) rather than being lower as predicted from Experiment 1. Furthermore, this was the same as the number of experienced subjects that detected the failure in Experiment 1 (31%). Seventy-five percent of the inexperienced subjects detected the first failure in Experiment 1. This finding suggests that the experience of the subjects in Groups 2 and 3 in Experiment 3 differed from that of the inexperienced subjects in Experiment 1. The experimental differences were the amount of offset and the presence of feedback.
From Experiment 1 it was learned that operators can quickly adjust their detection performance, such that they detect the second failure. This finding is encouraging. After the first failure and the feedback of its presence, they were subsequently able to perform the detection effectively, without a sustained increase in their time to perform the task.

Effects of automation on failure detection. There was no effect of automation on failure detection in this research. This finding shows that the use of automation need not impact the probability of failure detection. To ensure that automation does not degrade detection, the operator must be provided with equivalent information under manual and automatic control. This was true of the experimental task in this research. To ensure that manual control does not degrade detection or that automation complacency does not occur, the workload must not be so great as to preclude the operator from devoting resources to failure detection, as in these studies.

Effects of offset distance and feedback on failure detection. By comparing the detection rates of subjects in Experiment 1 with the rates in Experiment 2, in the 10% probability of failure condition, the effect of offset distance was assessed. Note however, that the presence of the tracking task in Experiment 1 and the different payoff conditions in Experiment 2 confounded this comparison. The detection rate for inexperienced subjects in Experiment 1 was .68 and the detection rate for Experiment 2 in the 10% failure probability condition was .65. The similarity of these rates suggests that the decrease in offset distance from 6.25 mm in Experiment 1 to 3.75 mm in Experiment 2 and the presence of the tracking task had no effect on the detection rates.
The detection rates in Experiment 2 (with 10% failure probability) compared to the
detections rates of Groups 2 and 3 in Experiment 3, provide insight into the effect of
feedback on failure detection for this task. The distance of the offset of the dots was the
same in Experiments 2 and 3. (The note above regarding the lack of tracking and the
payoff structure in Experiment 2 applies to this comparison as well.) The detection rate for
Experiment 2 was .65. The detection rate for Groups 2 and 3 for Blocks 1 and 2 in
Experiment 3 was .35. This difference suggests that the presence of feedback affects the
detection rates. Note that there was also tracking required in Experiment 3.

**Effects on false alarm rate.** In all three experiments, false alarm rates were
exceptionally low for this task, as compared to other signal detection tasks. The false alarm
rates and the detection rates appeared to be relatively independent. In Experiment 1 the
experienced subjects' false alarm rates did not increase when they started to do the detection
task. Altering the payoffs and signal probabilities in Experiment 2 did not affect the false
alarm rates for this task. The low false alarm rates were not altered by changes in the signal
strength as evidenced by the comparison of the false alarm rates in Experiments 1 and 2.
The false alarm rates did not change with the presence of feedback (Experiment 1 versus
3). The similar false alarm rates for Experiment 2 versus 3 show that the presence of the
tracking task had no effect.

**Movement times.** The longer movement times and the greater number of errors for
the targets with the offset dots indicate that the task was more difficult when the center of
the squares was not aligned with the center of the dots. Not only were the squares further
apart than the dots, but the target width (tolerance around the center) was not depicted when
the dots were offset. The task was designed in this way to simulate the condition arising
when supplemental information is easier to use than the primary information source. An
additional factor contributing to this finding is the amount of experience the subjects had using the dots, thus making them difficult to ignore. The number of capture errors is further support for this factor.

**Capture errors.** Capture errors were made when the learned response to the cues was not appropriate. The targets were correctly detected as failures, but erroneously responded to by placing the cursor between the dots and beyond the tolerance area around the center of the squares. The movement times were faster for these responses than for offset dots correctly responded to. This finding suggests that there is a time cost associated with the necessity to ignore the dots as cues, since placing the cursor between them was the practiced behavior.

**Significance of Key Findings**

**Is there an Effect of Long Periods with No Failures?** In Experiment 1, the previous experience with reliable cues resulted in fewer detections of the first failure. After the first failure, the difference due to prior experience disappeared. The proportion of detections by experienced subjects for each failure improved over time. However, this improvement leveled quickly; there was no significant improvement in detection over time when the first pair of failures was omitted from the analysis. The proportion of detections by inexperienced subjects also improved over time. There was no improvement from the first pair of offsets to the second pair for them either.

In Experiment 3, three of eight subjects in Group 1 stopped doing the task, as measured by their detection performance being equal to or below chance in Block 2. Only three subjects out of the sixteen in Groups 2 and 3 combined stopped doing the detection task.
This difference was not statistically significant, however it is consistent with the findings of Experiment 1, where the subjects who had experience with reliable secondary cues did not detect the first offset. Note that with no feedback as to the presence of the offset in Experiment 3 there was no 'recovery' of detection performance after the first failure.

In summary, experience without failures led to poorer detection initially, but after one failure (with feedback) the detection decrement disappeared.

Is the effect of long periods without failures the result of subjects discontinuing the detection task, or shifting their decision criteria?

Experience without failures led to a decrease in false alarm rate suggesting a shift to a more conservative decision criterion. There was a significant decrease in experienced subjects' false alarm rates over the eight 50-trial blocks. This vigilance decrement is consistent with the expected effect of a shift to a more conservative decision criterion when no signals are observed. However, the false alarm rates stayed the same for experienced subjects after the first 100 trials, which did not have offsets, i.e., there was no effect of time (block effect) for trials 101 to 400 on false alarm rate. In addition, the false alarm rate for blocks 5-8 (trials 210-400) was the same for subjects with prior experience without offsets and those receiving offsets in their first block. (There was no interaction of block and experience.) Therefore, at the time the first offset was introduced the subjects whose prior experience caused them not to expect offsets had stable decision criteria and those criteria were at the same level as the inexperienced subjects, as reflected in the false alarm rate.

In Experiment 3, as many subjects in Group 1 stopped doing the detection task as in the other 2 groups combined. The positive response rate (correct detections and false alarms) for subjects in Group 1' was the same as Groups 2' and 3' in Block 1 and Block 2. The
false alarm rate for subjects in Group 1' was the same as Groups 2' and 3' in Block 1 and Block 2. There was no group by block interaction for either positive response rate or false alarm rate. In Experiment 3, Group 1 showed no hint of becoming more conservative in their decision criteria than the other groups, either when considered with all members of the group or when the subjects that had stopped doing the detection were removed.

In conclusion, the false alarm rates do not support the theory that the subjects experiencing a period of reliable secondary cues had more conservative decision criteria than the subjects who had failures throughout the session. Therefore, poorer performance in detecting the first failure can be attributed to subjects who have experienced reliable secondary cues no longer checking for the failure in that information.

Is time to complete the tracking and detection tasks a convergent measure of the amount of processing occurring in the detection task? The experienced subjects in Experiment 1 slowed down, although not significantly, on the time to pass the cursor through the first gate from the two trials before to the two trials after the first failure. A conflicting finding is that as experienced subjects' detection rates improved over time (across target pairs), they also got faster in the time to complete a trial. There was a negative correlation between the total time the subject took to make movements to all eight sets of squares on a screen for 20-trial blocks and their accuracy in detecting whether an offset was present on the screen. The time from the start of the trial to the cursor passing through the first set of squares on the screen, when grouped by 20-trial blocks, was not correlated with the improvement in detection rate over target pairs. The error rate was not correlated with the MTs over the 20-trial blocks. Therefore a speed-accuracy trade-off was not present to account for this improvement in movement times.
In Experiment 3, the MTs for the four Groups decreased significantly from Block 2 to Block 3. There was no group effect, nor any interaction effect of group and block. The slopes of the MTs from Block 2 to Block 3 are parallel for Groups 3' and 4, and this slope is not as great as the slopes for Groups 1' and 2'. However, these differences in slopes were not statistically significant. This finding is consistent with the fact that the subjects in Group 4 effectively had the same task in Blocks 2 and 3, since they did not perform the detection task in Block 2 and it was not required in Block 3. The subjects in Group 3' had the detection task in both blocks. The subjects in Groups 1' and 2' had the detection task in Block 2 and no detection task in Block 3, therefore their slopes were greater.

Groups 1' and 2' and Groups 3' and 4 were combined for analysis of their MTs over the last two blocks. There was no group effect or group by block interaction. The block effect was the same as the four-group analysis. The difference in MT from Block 2 to Block 3 was analyzed for these two combined groups to control for the differences in Block 2 times, since the task was the same for all groups in Block 2. There was no group effect for this analysis.

The number of errors became greater from Block 2 to Block 3. There was a negative correlation of each subject's MTs for Block 2 and Block 3 with their error rate. This suggests that there is a speed-accuracy trade off that accounts for at least part of the decrease in MTs in Block 3 in Experiment 3.

In summary, the times to the first gate before and after the first failure in Experiment 1 are suggestive, but not conclusive, of the experienced subjects either adding the detection task, or at least devoting more time to the performance of the detection task. In Experiment 3 the slopes of the MTs from Block 2 and 3 are also supportive, but not definitive proof.
that MTs are a converging measure of the amount of processing occurring in the detection task. The time to complete the tracking and detection task combined suggests that the subjects who have experienced reliable secondary cues have omitted the task of detecting a failure in those cues. When a failure is detected or feedback of a missed failure occurs, the detection task is added.

**Are movements through squares with offset dots performed differently by experienced and inexperienced subjects?** Although the experienced subjects had a greater number of capture errors than the inexperienced subjects in Experiment 1, the differences were not significant. The movement time through gates with offset dots when a capture error was committed was significantly faster for all groups than the time to make a correct response. The experienced subjects' movement times were less for capture errors than the inexperienced subjects, but this difference was not significant. The results are only suggestive of differences between the movements through squares with offset dots performed by experienced and inexperienced subjects.

**Measurement Issues**

The lack of sustained differences in detection by experienced and inexperienced subjects in Experiment 1 did not afford the measurement of differences in the signal detection measures of $d'$ and $\beta$. The goal of Experiment 3 was to resolve this issue. The feedback was removed and the offset was reduced in an attempt to prolong the difference in detection. This attempt was not successful. Since these differences were not found, it was impossible to use these measures to resolve whether the difference in detection on the first failure was due to a shift in the decision criteria or a change in the strategy of task.
performance to devote fewer resources to the detection task, since it had proved to be unnecessary.

There were several other problems with the application of signal detection measures in these experiments. The subjects tended to have very few false alarms, making the calculation impossible for both the sensitivity to the signal and the decision criterion, $d'$ and $\beta$. This was the case in conditions with and without feedback and with two different signal strengths. In Experiment 2, when the payoffs and signal probabilities were changed to favor a less conservative decision criterion, subjects did not increase their false alarm rate. Another problem with the application of signal detection measures was that the signal and noise distributions were not normally distributed. Both $d'$ and $\beta$ measures were invalid for this task.

A nonparametric alternative to these measures was tried in Experiment 1 and 2, the use of confidence ratings to generate the area under the ROC curve, $A_g$, as a measure of the subject's sensitivity to the stimulus. The subjects did not use the full range of the rating scales, despite instructions to do so. Their responses clustered at very sure of signal and very sure of no signal. This caused the measured ROC function to have nearly no horizontal component and therefore the area under the curve was not a reasonable measure. The use of the rating scales had the additional drawback of emphasizing the detection task, since the rating scales were presented at the end of each trial and had to be responded to. This is unlike the operational world where usually no reminder exists that there could be a disagreement between primary and secondary information.

The approach was then taken to evaluate the decision criteria by examining the positive response rates. By testing for differences in positive response rates, the decision criteria
could be assessed. Since there was no difference found in the decision criteria of the experienced and inexperienced subjects, the difference in these two groups in detection performance of the first failure was concluded to be the result of differences in their sensitivity to the failure due to their attention to the detection task.

**Further Research**

To further study the effect found for the first failure in Experiment 1, a study could be performed with the same amount of offset as Experiment 1, with the feedback regarding the presence of an offset given at the end of the block, rather than at the end of each trial. This may prolong the difference found in the experienced and inexperienced subjects for the detection of the first failure. Another possibility is to have the feedback in the first set of trials when the signal is absent and not in the subsequent set where the signal is present.

To address the measurement issues discussed above, eye movement data could be taken to determine whether the subject was checking that the dots were centered between the squares.

In order to control the speed-accuracy trade off, feedback on tracking accuracy, in addition to the feedback on time to complete the tracking, could be presented. In addition the subject could be penalized for tracking errors.

This research, unlike previous studies, addressed failure detection with manual and automated control across levels of experience. The interaction of the experience level of the operator with the control mode was examined. While no interaction effects were found,
further research is warranted on this topic. These studies were not comprised of heavy workload conditions. However, situations where workload causes poor detection performance under manual control may be confounded by experience level. That is, with experience, this effect may be diminished due to the reduced workload for the experienced operator. Alternatively, experience that results in an unwarranted trust in the reliability of the secondary cues may exacerbate failure detection in automatic control conditions in which there is typically a reduction in feedback. Further research on failure detection, under conditions where these two factors exist, may reveal an interaction between control mode and the operator's experience with failure detection. Unfortunately in the operational environment, these situations are confounded. Highly reliable information sources have reduced the need for, and therefore the probability of, operators checking the reliability of the information.

**Implications of this Research**

**Experience and failure detection.** The most important finding of these studies is that experienced operators do not detect failures in easier-to-use secondary information that has previously been redundant to primary information. After determining that the redundancy is reliable over a large number of trials, they will not detect a disagreement between the primary and secondary information, even when they are aware of the potential for this failure and familiar with its appearance.

This finding is predicted by studies of operators with long periods without failures (see Parasuraman, 1986 for a review). However in these studies, the vigilance decrement reported in the literature does not account for the poor performance of experienced subjects
in detecting the first failure. Although the experienced subjects adjusted their decision
criteria initially, the false alarm rates for the experienced and inexperienced subjects were
the same at the time of the first failure. This finding challenges the long-standing literature
on failure detection. The results of the present studies support the conclusion that these
subjects had a reduced sensitivity to the failure due to either ceasing to look for the offset,
or devoting less time and/or attention to the detection.

It is noteworthy that knowledge that a failure will occur was shown by Wickens and
Kessel (1981) to have little effect on operators' detection performance. They measured
failure detection time for small and large changes to control dynamics. The subjects'
knowledge that a failure would occur had no effect on the time to detect a failure. They did
find some differences in the character of the control behavior when the changes to the
control dynamics were small. In the research presented here, all subjects had prior
knowledge that a failure would occur and training in the detection of and response to the
failure. The experienced subjects in Experiment 1, however, after experiencing 200 trials
without a failure, were less able to detect the presence of a failure until the first one
occurred and the feedback was presented. The knowledge that a failure is possible is not
sufficient to maintain a high level of performance in detecting its presence. Research has
shown that artificial failures presented with some regularity in embedded training eliminate
the effect of long periods without a failure by providing operators experience with the
detection task (Parasuraman, 1986).

This deficiency in detection implies that one cannot assume that primary information is
being utilized when that information has proved to be redundant to secondary information
that is more easily used. Only if the operator has recently experienced a failure of the
secondary information, is this assumption warranted. Humans should not be expected to
monitor very reliable information (Parasuraman, Molloy, & Singh, 1993). Pilots reporting to NASA's Aviation Safety Reporting System recount failures to verify information when they have experienced a high degree of reliability, especially when workload is high.

"I allowed myself to get too busy during the descent to make essential cross-checks to confirm the FMS [flight management system] was working as advertised. The correction: always double check the FMS data against other available navigational data to insure that your programming is correct and that the aircraft is following accurate FMS guidance. Overconfidence in the FMS and increased workload ... is no excuse for sound pilotage and the maintenance of situational awareness." (Report number 272508)

There are two approaches to tackling this deficiency in monitoring performance: 1) The most effective way is through the design of the system. System design should compensate for the lack of monitoring performance by operators experiencing failures very infrequently. The system, rather than the operator, should monitor the sources of information and alert the operator of a lack of correspondence. Alerting systems that become the primary source of information will have to be designed as such. 2) When a system is incapable of producing salient information for the detection of differences between primary and secondary information, procedures should be adopted to stimulate operators to verify the primary and secondary information. For example, this can be accomplished with checklists.

Fortunately, these studies show that operators who have had recent experience with a failure are likely to detect it. This rapid recovery in detection performance by the experienced subjects is encouraging. This finding has consequences for training programs. Training programs must provide experience in detection where it is not frequently available.
in the operational environment. It is important to practice the detection of the failure, not just the response to it, especially for subtle failures that the operators are not likely to encounter in normal operations. As was seen in Experiment 3, merely the perception that a failure has occurred is sufficient to overcome the reliance on the accuracy of the secondary information. This exposure needs to be done with some frequency to retain a high probability of detection.

**Performance without reliable secondary cues.** Information that comes from a primary source cannot be expected to substitute equally for the secondary information that is used regularly, without an increase in response time. If the secondary information is that which is used, the reason may be that it is easier to use. If this information must be ignored and the primary information responded, a decrement in performance must be anticipated. This constitutes justification for the requirement to always supply the operator with the practiced stimuli, be it primary or secondary cues, driven by the correct information. At a minimum, the misleading cues should be removed to eliminate the requirement to ignore the practiced stimuli.

**Capture errors.** The capture errors observed in Experiment 1 have grave implications for safety. Subjects were not able to ignore the dots and place the cursor through the center of the squares, even when they correctly identified the dots as being offset. The subjects were aware of the required response to the failed stimuli, yet were unable to always overcome the practiced response. They were more likely to make a capture error than to place the cursor outside the center when there was no offset. In fact, this erroneous response was accomplished more quickly than the correct response. In environments where a conditioned response to an erroneous stimulus could be unsafe, the prevention of the response must be very effective. To merely require that the response not be made,
without creating a powerful stimulus to prevent the capture error, is insufficient, as seen in these studies.

Systems must be designed such that these inappropriate responses can be prevented, ideally by taking advantage of the practiced behavior to complete the proper action. This can be done by driving the familiar cues with reliable information. An operator should be able to easily change the source of the information to drive the familiar cues with valid inputs. For example, if the flight director indication from one of the flight management computers is faulty, a backup flight management computer should be selectable to drive the flight director. Fortunately, the greater flexibility of newer systems often affords this capability. If this is not possible, or desirable for some reason, the operator must be able to inhibit misleading information, especially when that information is normally utilized and the response may have become automatized, making the misleading cues difficult or impossible to ignore. This need is further supported by the finding that there is a time cost associated with ignoring the cues normally responded to.

If the information cannot be removed, the practiced response must be blocked. As mentioned in Chapter 1, procedures can be put in place to prevent capture errors, for example, the controls normally used for the task may be disabled and replaced by controls that look and feel sufficiently different so that the stimuli cannot be responded to in the practiced manner. This approach will, however, result in an increase in workload. This tactic should only be considered if the stimuli cannot be corrected or inhibited.

**Failure detection under automated and manual control.** In this research and previous work, failure detection rates were the same using manual and automatic control. To produce equal detection performance under manual and automatic control, two
conditions must be satisfied. The information available to the operator must be equivalent, and the workload must be low enough that the operator is able to devote resources to the detection task.

In Experiment 1 there were no differences between the detection by subjects using manual control and those using automatic control. Due to the nature of this task, workload was not a factor in the failure detection under the two control conditions. The workload was not too high under manual control. The subjects were not better able to perform the failure detection using automation, as was found by Ephrath and Curry (1977), Ephrath and Young (1981) and Fuld, Liu, and Wickens (1987). The task also did not permit the subjects to rely on the automation for failure detection as in the study by Parasuraman, Molloy, and Singh (1993). The workload and the nature of the task were such that failure detection could be performed serially, rather than time-shared. Subjects reported checking all the targets for failures when the screen appeared, before beginning the tracking. Therefore the manual task did not load the detection task and automation of the tracking did not unload detection. The similarity of detection rates in Experiment 1 with manual and automatic control and the similarity of detection in Experiments 1 and 2, with and without tracking respectively, indicate that the workload in the tracking task did not affect the detection performance.

There are important implications to these findings. First, the ability to perform the detection task independent of the workload associated with other tasks ensures that the detection performance will not degrade under higher workload conditions such as when using manual control. To prevent the operator from being overcome by other tasks, he or she must be specifically required to stop the other, perhaps more compelling, tasks to observe the failure. This principal must be addressed both in the design of the system and
in the operational procedures. The design should allow tasks to be allocated to automation to periodically alleviate the demands on the operator, ensuring better failure detection. In addition, the operational procedures should be structured to require the operator to actively perform the detection task, perhaps momentarily forsaking other tasks. This can be accomplished through checklists or other reminders. Detection should not be assumed to occur due to the reasons described above under experience level effects.

The second implication of this research, regarding manual and automatic control, is that information available to operators using automation needs to be equivalent to the information for those manually controlling. The nature of the task may preclude this under normal conditions, e.g., if kinesthetic information is necessary for failure detection and is not available under automatic control. There are two options for providing adequate information for failure detection under automatic control. This can be done by conducting studies to determine the cues that are utilized for failure detection with manual control, including kinesthetic information. If the same quality of information can be provided under automatic control, as in these studies, no decrement in detection performance under automatic control will be observed. Alternatively, procedures may be employed that require the operator to manually control the system periodically, to ensure the critical information is obtained. A testimony to the efficacy of this approach is that pilots will typically disconnect the autopilot when they see or feel something out of the ordinary, to gain the information that is only available through the manipulation of the controls.

Additionally, experience with manual control has been shown to transfer and improve detection under automatic control. To this end, Parasuraman, Molloy, and Singh (1993) suggest adaptive function allocation of automated tasks to the human for manual control, during non critical work periods. They have shown this to improve failure detection, even under subsequent automatic control.

153
In conclusion, since most operators in aviation and other domains are highly experienced and there is an increasing amount of automation, this research has important design implications for the optimum human interface. The findings from these studies can guide the designers of flight decks and other human-machine systems, to create an interface that promotes failure detection and prevents capture errors. In addition there is much that can be done to implement operating procedures that will improve the detection of a failure and impede capture errors, thereby improving safety.
LIST OF REFERENCES


158


APPENDIX A:
EXPERIMENT 1 INSTRUCTIONS TO SUBJECTS
Instructions to Subjects

Thank you for participating in this experiment. You will be asked to move a pen on a tablet to control a cursor on the screen. The object of the task is to move the cursor from the bottom to the top of the screen passing through the center points of horizontal lines defined by two squares. The trial begins when a box comes up at the bottom of the screen. Place the cursor in the box and depress the pen. You can then let up on the pen. A screen with eight pairs of boxes will appear. Move the cursor through the center of an imaginary line between the boxes, until you pass through the top set of boxes. In this experiment, being 100% accurate is more important than going quickly. As long as you pass the cursor within 1/2 inch of the actual center of the squares, it will be counted correct. Being more precise than that doesn't count. It doesn't matter what path the cursor takes from one set of boxes to the next, as long as you pass between the boxes in the center. The position gets recorded the first time through, you can't go back to fix an error.

As you do the task, you'll find that you are able to complete the screens more quickly. Go as quickly as you can, but be sure to keep 100% accuracy throughout the experiment. The clock starts as soon as the screen comes up and ends when you have passed through the last set of squares. The time to complete the screen will be displayed at the end of each trial. Now let's take a look at the screens and try a few for practice.

--- Experimenter performs one screen, subject performs four. ---

Now the screen will include dots that show you where the 1/2 inch target area is in the center of the squares. As long as the cursor passes between the dots, without touching them, the goal is met. Your score is the same as long as you are anywhere within the dots.

--- Subject performs five screens. ---

In very rare instances the dots will not be in the center of the squares. If you see this, it is important that you still move the cursor through the midpoint of the squares, not the dots. At the end of each screen you will be asked to indicate which set of dots was not centered, if any. You will also be asked how sure you are that the dots were or were not centered. Press the pen to the tablet at the place on the line corresponding to how sure you are of the correctness of your response, from not sure to very sure. Use the whole range of the scale, as appropriate. Take as long as you like to make this response, this is not timed.

You will begin with $5.00. For every correct response you make regarding dots that were or were not offset, you will receive 10¢. For every incorrect response, you will lose 10¢. See the table below.

<table>
<thead>
<tr>
<th>Said Offset</th>
<th>No Offset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Offset</td>
<td>+10¢</td>
</tr>
<tr>
<td>No Offset</td>
<td>-10¢</td>
</tr>
</tbody>
</table>

--- Subject performs five screens. ---

Now, if you feel these instructions make sense and you're ready to begin, we'll start the data collection.
Automated Trials

There will now be 4 sets of 50 screens in which the automatic controller moves the cursor. Your job is to make sure that the cursor goes through the center of the squares. If you detect an error in the location of the cursor as it passes through the squares, press the pen to the tablet. As long as you respond before the screen disappears, it will be counted as a correct response.

As before some of the dots may be offset from the center of the squares. You will again be asked to rate how sure you are that all of the dots were centered or were not centered.

--- Subject performs five screens. ---
APPENDIX B:
EXPERIMENT 1 SUBJECT QUESTIONNAIRE
Subject Questionnaire

Subject number __________

Right handed ___  Left handed ___

Age _____

Female ___  Male ___

Rate the use of the stylus and tablet:

Very Easy 1 2 3 4 5 6 7 8 9 10 Very Difficult

Rate the task of passing the cursor through the center of the squares:

Very Easy 1 2 3 4 5 6 7 8 9 10 Very Difficult

Rate the task of monitoring the cursor movement by the automatic control:

Very Easy 1 2 3 4 5 6 7 8 9 10 Very Difficult

Rate the task of detecting the dots that were not centered:

Very Easy 1 2 3 4 5 6 7 8 9 10 Very Difficult

Do you have any prior experience with this task?

yes _________  no _________
APPENDIX C:
EXPERIMENT 2 SUBJECT INSTRUCTIONS
Instructions to Subjects

Thank you for participating in this experiment. You will be asked to observe 8 patterns of dots and squares on a screen. The dots are in the center of the squares most of the time. The task requires you to detect if a set of dots is not centered in the squares. On most of the screens, they will all be centered, but if there is a set of dots not in the center, there will only be one on a screen. Remember which one it is. Look through the dots and squares as quickly as possible while making sure you have checked each one carefully. This part of you task is timed, but it is more important to be accurate than quick. When you are done looking at the screen, pass the cursor through the top of the screen and a response screen will appear.

You will be asked to indicate which set of dots was not centered, if any. You will also be asked how sure you are that the dots were or were not centered. Press the pen to the tablet at the place on the line corresponding to how sure you are of the correctness of your response, from not sure to very sure. Use the whole range of the scale, as appropriate. As long as you are within one row of the offset row, your response will be counted correct. Take as long as you like to make this response, this is not timed. After you make the response, the time to observe the screen will appear. If a set of dots was not in the center, its row number will also appear.
10% Offset, More Pay for Saying There Was Offset

There will be 200 screens. 20 screens will have an offset and 180 will not. You will begin with $20.00. Every time you correctly say that a set of dots was offset you will receive $1.00. If you correctly say that the dots were not offset, you will earn 1¢. Every time you say there was no offset and you're wrong, you will lose 50¢ and every time you say there was an offset and there was not, you lose 10¢. See the table below.

<table>
<thead>
<tr>
<th></th>
<th>Offset</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td>Offset</td>
<td>+$1.00</td>
<td>-50¢</td>
</tr>
<tr>
<td>None</td>
<td>-10¢</td>
<td>+1¢</td>
</tr>
</tbody>
</table>

If all correct: $21.80
If say offset every time: $2.00
If say none every time: -$8.20

Now, if you feel these instructions make sense and you're ready to begin, we'll start.
10% Offset, More Pay for Saying There Was None

There will be 200 screens. 20 screens will have an offset and 180 will not. You will begin with $20.00. Every time you correctly say that a set of dots was offset you will receive 10¢. If you correctly say that the dots were not offset, you will earn 25¢. Every time you say there was no offset and you're wrong, you will lose 50¢ and every time you say there was an offset and there was not, you lose 20¢. See the table below.

<table>
<thead>
<tr>
<th>Said</th>
<th>Offset</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td>Offset</td>
<td>+10¢</td>
<td>-50¢</td>
</tr>
<tr>
<td>None</td>
<td>-20¢</td>
<td>+25¢</td>
</tr>
</tbody>
</table>

If all correct: $47.00
If say offset every time: -$34.00
If say none every time: $35.00

Now, if you feel these instructions make sense and you're ready to begin, we'll start.
20% Offset, More Pay for Saying There Was Offset

There will be 200 screens. 40 screens will have an offset and 160 will not. You will begin with $20.00. Every time you correctly say that a set of dots was offset you will receive $1.00. If you correctly say that the dots were not offset, you will earn 1¢. Every time you say there was no offset and you're wrong, you will lose 50¢ and every time you say there was an offset and there was not, you lose 10¢. See the table below.

<table>
<thead>
<tr>
<th>Was</th>
<th>Offset</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td>Said</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Offset</td>
<td>+$1.00</td>
<td>-50¢</td>
</tr>
<tr>
<td>None</td>
<td>-10¢</td>
<td>+1¢</td>
</tr>
</tbody>
</table>

If all correct: $41.60
If say offset every time: $24.00
If say none every time: -$18.40

Now, if you feel these instructions make sense and you're ready to begin, we'll start.
20% Offset, More Pay for Saying There Was None

There will be 200 screens. 40 screens will have an offset and 160 will not. You will begin with $20.00. Every time you correctly say that a set of dots was offset you will receive $0.10. If you correctly say that the dots were not offset, you will earn $0.25. Every time you say there was no offset and you're wrong, you will lose $0.50 and every time you say there was an offset and there was not, you lose $0.20. See the table below.

<table>
<thead>
<tr>
<th>Said</th>
<th>Offset</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td>Offset</td>
<td>+10¢</td>
<td>-50¢</td>
</tr>
<tr>
<td>None</td>
<td>-20¢</td>
<td>+25¢</td>
</tr>
</tbody>
</table>

If all correct: $44.00
If say offset every time: -$28.00
If say none every time: $20.00

Now, if you feel these instructions make sense and you're ready to begin, we'll start.
APPENDIX D:
EXPERIMENT 3 SUBJECT INSTRUCTIONS
Instructions to Subjects

Thank you for participating in this experiment. You will be asked to move the pen on the tablet to control a cursor on the screen. The object of the task is to move the cursor from the bottom to the top of the screen passing through the center points of horizontal lines defined by two squares. The trial begins when a box comes up at the bottom of the screen. Place the cursor in the box and depress the pen. You can then let up on the pen but don't lift it off the tablet. A pairs of boxes will appear. Move the cursor through the center of an imaginary line between the boxes. As you pass through each set of boxes, the next set will appear, until you pass through the top set of boxes. There will be eight sets per screen. In this experiment, being 100% accurate is more important than going quickly. As long as you pass the cursor within 1/2 inch of the actual center of the squares, it will be counted correct. Being more precise than that doesn't count. It doesn't matter what path the cursor takes from one set of boxes to the next, as long as you pass between the boxes in the center. The position gets recorded the first time through, you can't go back to fix an error.

As you do the task, you'll find that you are able to complete the screens more quickly. Go as quickly as you can, but be sure to keep 100% accuracy throughout the experiment. The clock starts as soon as the screen comes up and ends when you have passed through the last set of squares. The time to complete the screen will be displayed at the end of each trial. Now lets take a look at the screens and try a few for practice.

--- Experimenter performs one screen, subject performs four. ---

Now the screen will include dots that show you where the 1/2 inch target area is in the center of the squares. As long as the cursor passes between the dots, without touching them, the goal is met. Your score is the same as long as you are anywhere within the dots. If you do happen to make an error, still be accurate on the rest of the targets. Each target is evaluated independently.

--- Subject performs five screens. ---

In extremely rare instances the dots will not be in the center of the squares. If you see this, it is important that you still move the cursor through the midpoint of the squares, not the dots. If you see a set of dots that are not in the center of the squares, continue moving until you have passed through the top of the screen and your time is displayed. Press the space key to indicate that the dots were not centered. Be sure to press the key while the time is on the screen and before the next start box appears.

Every time you correctly say that a set of dots was offset you will receive $1.00. If you correctly say that the dots were not offset, you will earn 1¢. Every time you say there was no offset and you're wrong, you will lose 50¢ and every time you say there was an offset and there was not, you lose 10¢. See the table below. Remember, it is very rare that a set of dots will be offset.

```
Said

<table>
<thead>
<tr>
<th>Offset</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td>Offset</td>
<td>+$1.00</td>
</tr>
<tr>
<td>None</td>
<td>-50¢</td>
</tr>
<tr>
<td></td>
<td>-10¢</td>
</tr>
<tr>
<td></td>
<td>+1¢</td>
</tr>
</tbody>
</table>
```

176
Subject performs five screens with one offset on each.

These examples of the dots being offset were presented for demonstration purposes. In the actual experiment, such offsets will be extremely rare. Now, if you feel these instructions make sense and you're ready to begin, we'll start the data collection.
No Detection Trials

The last 50 trials will have no offset of the dots. This is done to give us a measure of how quickly you perform the task when you only have to do the cursor movement, without also having to check for the offset. Perform the trials just as you have been, but you do not have to look for the offset. None of the dots will be offset. The key press response will not be recorded, in fact the keyboard will be removed. Be sure to not hit any of the dots as you're going through them. Remember 100% accuracy is important.