Reliability Evaluation of an Expert System Diagnostic Aid for a Sleep and Respiration Experiment

by

Allen Atamer

Bachelor of Applied Science, Aerospace Engineering
University of Toronto, Toronto, Ontario, Canada, 1999

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

Masters of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

January 15, 2001

[February 2001]

Author .....
Department of Aeronautics and Astronautics
January 15, 2001

Certified by .....
Laurence R. Young, Sc.D
Apollo Program Professor of Aeronautics and Astronautics
Thesis Supervisor

Accepted by .....
Wallace E. Vander Velde
Professor of Aeronautics and Astronautics
Chair, Committee on Graduate Students

Reliability Evaluation of an expert system diagnostic aid for a Sleep and Respiration Experiment

by

Allen Atamer

Submitted to the Department of Aeronautics and Astronautics on January 15, 2001, in partial fulfillment of the requirements for the degree of Masters of Science in Aeronautics and Astronautics

Abstract

The expert system software Principal Investigator-in-a-box ([PI]) is designed to help astronauts in conducting space life science experiments outside their field of expertise. The current version of [PI] was applied to the Sleep and Respiration experiment which flew aboard STS-90 and STS-95. These are the results of a ground study to determine the efficacy of [PI] for helping astronaut surrogates with fault management of the sleep instrumentation are shown here. Thirty subjects (14 female, 16 male) were tested on two days, one with [PI] assistance and one without it.

Results: [PI] assistance was not found to improve the probability of a correct detection rate of subjects, but was found to improve the probability of finding a fault. The chance of a correct detection of [PI]'s indicator lights alone was on average lower than that of the subjects studied. By decoupling the software's reliability from the subjects', [PI] only needed to be 40% reliable for subjects to correctly detect anomalies at their best. A regression analysis of the time taken to find a fault showed that [PI] was helpful for diagnostic planning tasks on the first experimental day, and for execution of diagnostic tasks on the second day. This, along with [PI]'s ability to reduce the chance of an undetected anomaly, shows that [PI]'s intelligent interface tends to regulate fault management reliability. Characteristics of how each experimental group managed the faults, such as troubleshooting steps taken, and calibrations performed and are shown and compared.

Thesis Supervisor: Laurence R. Young
Title: Apollo Professor of Aeronautics and Astronautics

.

"For my mouth shall speak truth; and wickedness is an abomination to my lips. All the words of my mouth are in righteousness; there is nothing froward or perverse in them. They are all plain to him that understandeth, and right to them that find knowledge. Receive my instruction, and not silver; and knowledge rather than choice gold. For wisdom is better than rubies; and all the things that may be desired are not to be compared to it," Proverbs 8:7-11

"Rejoice, O young man, in thy youth; and let thy heart cheer thee in the days of thy youth, and walk in the ways of thine heart, and in the sight of thine eyes: but know thou, that for all these things God will bring thee into judgment," Ecclesiastes 11:9

Acknowledgements

This research was supported by the National Space Biomedical Research Institute, NASA Cooperative Agreement NCC 9-58, and the NASA Ames Research Center, grant number NCC 2-570.

Larry Young was also helpful in making major decisions about our experiment and what we hoped to achieve. I appreciate his confidence in me to conduct research of the highest quality and his guidance as an advisor in discussing my future.

Mom, you always tell me to stop dreaming that I will be an astronaut, but I know deep down inside you're more worried that the space-lasagna I'll eat won't be as good as yours. Dad, sunshine, I've seen the good, the bad and the ugly - it's my jump-shot, and I'm still working on it.

Hello... Jello... how could I forget Mindy, my favorite stress-ometer. Thanks for being supportive in my quest to shed my mama's boy image. What can I say, my voice is naturally loud. At least our discussions entertained all our labmates, so it was all good...

The people at Brigham and Women's were very helpful and generous with their time, patience, and spare SleepNets: especially Eymard, Ken, Jerzy, Rod, and Dr. Czeisler. Thanks for making us bona-fide sleep technicians.

Susanne, thanks so much for helping us get our feet off the ground with preparing for our experiment. I hope all the best for you and your family. Luca, the top tourguide in Texas, thanks for letting us party it up with you in Houston. Dennis, thanks for your help with the code - and promise you'll learn what snow is one day.

Will, our favorite volunteer "head." Thanks for being there to create faults (pun fully intended)! Good luck in the future with Aero/Astro and remember that the best engineers are physicists.

Joachim, I'm glad you were there to help us throughout the experiment. I'll probably bump into you at a jazz club somewhere someday. Alan, although we've had our disagreements, you still taught me to think twice about all important decisions. Thanks for translating my engineering-ese into the vernacular, and I'm sorry but pale is my color. Marsha, I'm glad that chair in your office can be devoted to me talking about mostly cheerful stuff. You have to realize by now that wasting time is the quiet way I deal with stress. If I meet a callipygious girl one day, I'll let you know. Thanks also go to Sung-Ho, my favorite buddy to waste time with because everybody else at MIT is too busy. Cheer up, you'll be a super-consultant in no time! Just make sure you don't attack me, the wimpy goalie.

Thanks to the MIT canucks for being Canadian, and for subliminally inducing me to hockey. Simon and Joe, I got both your numbers in net! Adam, thanks for lending me all the equipment, now if all else fails, at least I can try the NHL. Keep it real down here and always remember your roots...

Thanks to the MVL folks that kept things fun. Patricia, thanks for ignoring my effluvious hockey equipment in the lab. JJM, I need to get some mini-donuts from LaVerde's - come join me. Richards, the baller, shotcaller, it was fun. Joe, thanks for listening to my ideas. Carol, eyes open, eyes open, blink blink! Chris, Sylvie, Andy (and family), Kathy, Lisette, and everybody else: our lab doesn't have windows, but you all brought the sunshine in every day...

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The expert system, "Principal Investigator-in-a-box," or "[PI]," was designed to assist astronauts (or other untrained scientists) in performing experiments outside their expertise. Previous applications of [PI] were used in STS-40 and STS-58 (Rotating Dome experiment). The current version of [PI] helps detect and diagnose instrumentation problems for a Sleep and Respiration Experiment that flew on STS-90 (Neurolab) and STS-95. [PI] displays electrophysiological signals in real time, alerts astronauts via indicator lights when a poor signal quality is detected, and tells astronauts what to do to restore good signal quality. [PI]'s reasoning engine uses heuristic rules, developed with the help of a real sleep expert, to assess the quality of electrophysiological signals.

1.1 Artificial Intelligence and Expert Systems

Expert systems were one of the fundamental developments of the artificial intelligence (AI) community during the 1970s. The idea was that a particular problem-solving knowledge possessed by an expert could be encoded into a reasoning system of simple heuristic rules. It was particularly appealing because simple "if-then-else" rules are easy to understand, and correspond to Rasmussen's rule-based level of human behavior [13]. With this set of rules, an expert system software can interact with its domain in a way similar to the real expert. Many examples of successful applications of an expert system exist in industry, such as XCON [14]. Expert systems impacted many different aspects of industry, such as process control, electronics, and manufacturing.

1.2 Motivation

There are several driving factors behind the development of [PI]. Astronauts are highly intelligent individuals, but generally lack initial experience with a particular experiment. Their training on the apparatus, procedure and data collection on each experiment is a very small part of their overall training. Principal Investigators can rarely accompany their experiments into space; and ground-to-air contact with the astronauts performing the experiment is not always possible. Having [PI] alongside them as a coach can help astronauts answer questions they may have about the experiment, manage instrumentation problems, or interpret the data being collected.

This deficiency in astronauts' practical experience with an experiment is exacerbated by the long delays between their training and the experiment's execution. Astronauts also do not typically have direct contact with the real Principal Investigator on the ground because of limited communications bandwidth. Further, scientific objectives are sometimes lower priority than operational objectives of a mission, so astronauts may have even more limited resources to conduct experiments should contingencies need to be made. The time from training to launch may be nearly 6 months to one year for missions to the International Space Station (ISS). Stress from a long-duration mission, fatigue, and high mental workload can also increase the chances of an error. [PI] can cut down this risk by taking on some of the astronauts' workload for the experiment such as monitoring data collection, so that astronauts can focus better on tasks for which they are better suited.

Chapter 2

Background

2.1 Previous studies

A series of experiments were done by Rouse to see how humans diagnose faults, and how a computer decision-aiding system would help them in this task [17]. One experiment compared how subjects do with and without a decision aid, to see if there was any improvement. The diagnosis task was to debug a network of AND gates with known failure modes to see which component was faulty. The decision aid was a display which would rule out known good components by an optimization algorithm based on values probed by the human. It also had a design such that the training (from one day to the next) and transfer (switching from "Aid" to "no Aid") could be studied separately as effects across three different days. The results were that subjects did perform better with the decision aid available. Further, practicing (or the change across days) on the task was found not to be significant. Subjects who received the aid for Day 1 and Day 2 were found to maintain their performance without the aid on Day 3.

A second experiment assessed subjects' performance on the same fault diagnosis tasks with an explicit time limit set. Subjects would be asked to debug the circuit within 30, 60 and 90 seconds of it being displayed. Two strategies distinctly emerged from the data: one that was more "optimal" for the 90 seconds case, and one that was more "brute force" for the 30 and 60 second case. Further, the transfer effect found in the first experiment was not found here: those with the aid first were unable to repeat their performance for the "no aid" trial. The forced-paced strategies were different from the self-paced strategies of the first experiment. The amount of practice subjects had for each time limit was also a significant effect.

The experiments done were very context-free, since knowledge of a particular domain was not evident to perform the task. So Rouse studied some of the errors made in fault management by the crew of a supertanker [6]. After characterizing errors with the help of an experienced engineer, he found that

- 27% of errors were incomplete execution of procedures (such as omission of steps, out of sequence steps),
- 26% of errors were inappropriate identification of the failure, and
- 13% of errors were incomplete observation of the state of the system

He also found that there was a high correlation between the errors made in fault identification and with lack of knowledge. So assessing the "system state" after a failure, and failure identification could be a significant indicator of the progress made by the astronaut in a fault management task. In our experiment, we also characterize the fault management trajectory made by subjects throughout the trials.

Other studies which involved the evaluation of a software decision aid for a monitoring and control task were conducted for a satellite-management software called GT-MOCA, developed at Georgia Tech [8]. Results showed that the satellite mission control crew who used the software successfully offloaded much of their monitoring task to the software, and the chance they would make an error was significantly lowered compared to without having the aid.

2.2 The origin of [PI]

The expert system [PI] was designed to assist astronauts or other operators in performing experiments outside their field of expertise. The first version of [PI], also known as the Astronaut Science Advisor (ASA), is the first documented attempt to use a biomedical diagnostic expert system on a space mission. [PI] was used to assist astronauts in the per-

formance of the Rotating Dome Visual-Vestibular Interaction Experiment on the STS-58 Space Life Sciences 2 (SLS-2) Space Shuttle mission in 1993 [22]. This first version of [PI] provided data collection capabilities, as well as protocol assistance, scheduling, and protocol modification suggestions. An additional feature consisted of an "interesting data" filter, designed to perform quick-look data analysis and report any unexpected findings to the astronauts during the experiment. Although crew feedback on this demonstration was positive, no data was taken concerning the performance of [PI] or the correctness of the advisories that it issued.

2.3 The Sleep and Respiration Experiment

The sleep and respiration experiment was designed to gather electrophysiological and cardiorespiratory data on astronauts in space while they were sleeping. Aboard Neurolab (STS-90), the Sleep and Respiration Experiment was to assess sleep quality of astronauts, and the effects of melatonin as a countermeasure against fatigue and suboptimal performance due to circadian misalignment of astronauts during a mission. [PI]-Sleep was the data acquisition software used for the experiment. The intelligent features were the signal quality indicators and diagnostic messages, designed to assist the astronauts in setting up and debugging the instrumentation in the pre-sleep period.

2.4 Results from STS-90 and STS-95

[PI] flew with the Sleep and Respiration Experiment aboard STS-90 (Neurolab) and STS-95. Data was collected from the pre-sleep instrumentation session for the astronauts. The data files and [PI]'s signal quality indicators were analyzed and evaluated by trained sleep experts at MIT. The red indicator lights were found to be correct 81% of the time for both cardiorespiratory and electrophysiological signals that did not exhibit saturation, but were correct 55% of the time when saturated signals were included.On STS-95, the red indica-

tor lights were found to be correct 77% of the time for these signals.

2.5 Results from the Pilot Study and Phase I

Previous experiments were carried out as part of an NSBRI ground study to assess the utility of [PI] for managing faults on the Sleep and Respiration Experiment. The pilot study was an experiment with 12 subjects who were trained to detect and identify anomalous signals in a series of electrophysiological data. They were then asked to perform this task with and without [PI] help to see if [PI] would reduce detection time. Although no significant reduction in detection time was found when [PI] provided help, a training effect was identified and a significant reduction in the number of undetected anomalies was found with [PI] help.

The Phase I study was similar to the pilot study, except with more subjects. Two different stimulus files were used, one labeled "File A" and one "File B." In fact, a significant difference was found between subjects who received File A and subjects who received File B in the study. Though subjects found file A much harder to interpret than file B when [PI] was not active, this difference was reduced when [PI] was active. This result is encouraging. It suggests that any peculiar differences between stimuli that affect subject performance for correct identification of anomalies, are essentially nullified when [PI] is active. [PI] appears to be more effective in situations where subjects find detection to be difficult.

2.6 Fault-tree analysis for Phase II

The NASA troubleshooting flowcharts published for the on-board Sleep and Respiration Experiment had an exhaustive library of possible failure states, diagnostics lists, and repair procedures. The pilot study only tested subjects on anomalies that only affected one channel at a time. This is not the case in reality. Other failures are also possible. In fact, at

one point the astronauts aboard Neurolab improperly inserted the SleepNet connector into the DSR port, which prevented the proper collection of data.

We went through all the possible failures in the Sleep Experiment system we could envision to determine whether they could be used in the experiment, starting with an exhaustive list compiled for the Neurolab experiment [19]. Some errors were eliminated because they were too trivial to troubleshoot, such as not having the ThinkPad on. Other errors were eliminated because [PI] would not have a fair control against which to test its effectiveness. No errors were created on the EMG channels for this reason, since [PI] did not have any rules for assessing their signal quality.

Chapter 3

Experimental Apparatus

3.1 Hardware Overview

The ground study phase II was designed to evaluate [PI] in a controlled setting. The sleep experiment was a collaboration between MIT and Brigham and Women's Hospital (Harvard Medical School) and The University of California, San Diego. The basic components of the apparatus come from the Sleep Experiment which flew aboard Neurolab and STS-95. A schematic of the entire hardware setup is shown below.

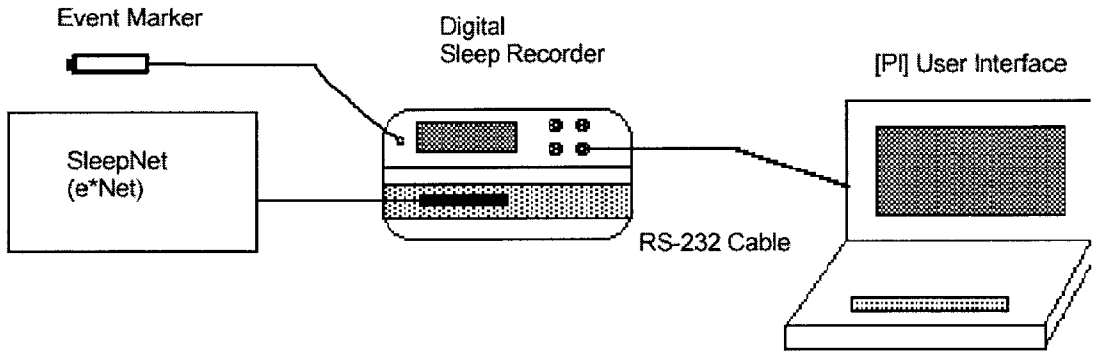


Figure 3.1: [PI] hardware schematic diagram

3.1.1 Electrode Net (e-Net)

The electrode-net, or SleepNet, is an elastic web-like cap worn by the sleep subjects during the experiment to record their EEG, EMG and EOG signals. It contains 13 electrode sockets designed to house hydrodots, which are disposable biosensors filled with a sticky, water-soluble gel that adheres to the skin to provide better contact. The e-Net is versatile since it locates the electrodes on the head in the same place, which makes it eas-

ier to align the electrodes in the standard locations. Both the SleepNet and hydrodots are manufactured by Physiometrix, Incorporated of North Billerica, Massachusetts.

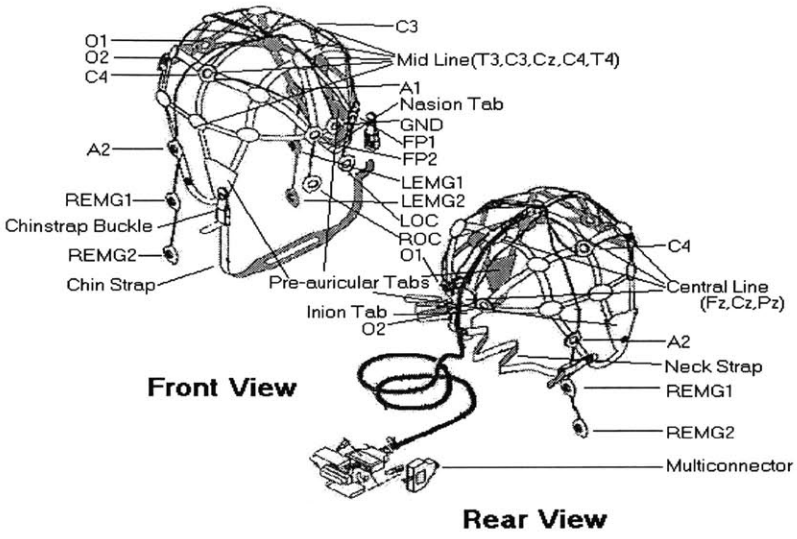


Figure 3.2: eNet electrode diagram

3.1.2 Digital Sleep Recorder

The digital sleep recorder (DSR) records signals coming from the SleepNet. The device converts the raw analog signals from the various electrodes and instrumentation to digital signals, which are then recorded onto a PCMCIA FlashRAM card. The DSR outputs to an IBM ThinkPad laptop via an RS-232 optically isolated serial cable. The DSR used in the study was the Vitaport2 recording system, made by TEMEC Incorporated.

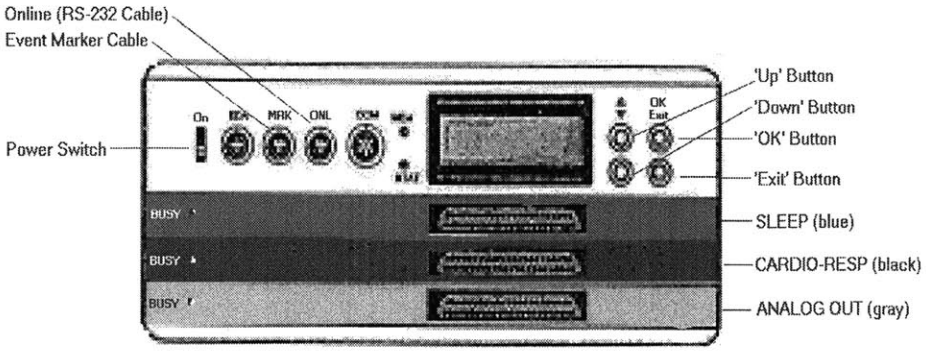


Figure 3.3: Digital Sleep Recorder

3.1.3 Thinkpad laptop

The computers used for displaying waveforms were IBM ThinkPad laptops - the same laptops which flew aboard the Shuttle. The ThinkPads were equipped with Pentium class processors running Windows 95 with [PI] software installed. The laptops displayed the [PI] interface to the subjects during the experiment.

3.1.4 Event marker

The event marker is a clicker used by the test subject to indicate when he begins the trial. It is attached to one of the Vitaport2 inputs.

3.2 The [PI] software and interface

[PI]'s knowledge base was developed at NASA Ames using the "C Language Integrated Production System" (CLIPS), a NASA-developed tool used for building expert systems. [PI]-Sleep version 4 flew aboard the STS-95 mission. [PI]-Sleep version 4.04 was used for this experiment. The raw electrophysiological data is displayed in real time to enable the subjects to view signals individually. Each large division on the display represents five seconds of data. The "state," or quality assessment of each signal is indicated using color-coded indicator lights that enable the subject to determine at a glance which signals require attention. A green indicator light indicates good signal quality; yellow indicates unknown or marginal quality; and red indicates poor quality. In addition, the system uses both text and graphic displays to alert the user to any problem with signal quality, in

the diagnostic messages window. These messages could appear automatically, or with a click on the indicator light. The [PI] interface is shown in Figure 3.3.

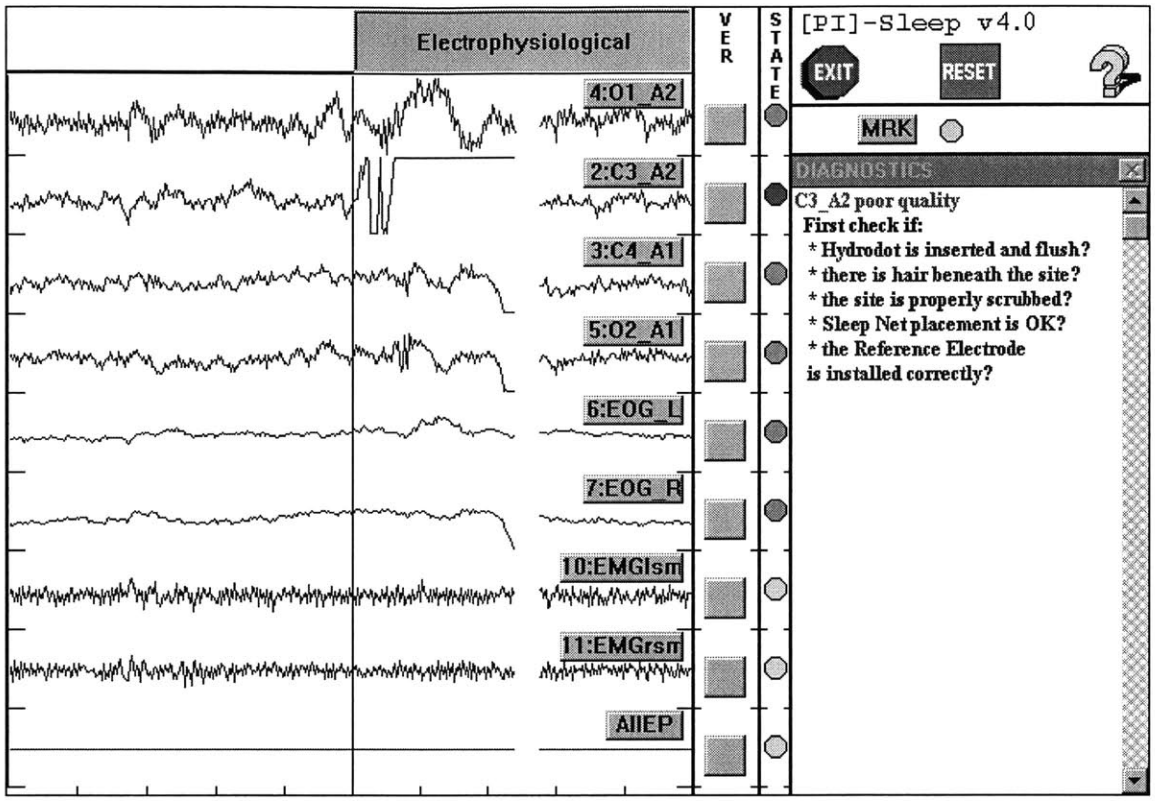


Figure 3.4: [PI] graphical user interface. indicator lights to the right of the waveforms indicate signal quality. Message box on the right displays diagnostic message for C3_A2 electrode.

The subject is then provided with suggestions and diagnostic procedures for eliminating the problem and returning the signals and system back to the nominal operating state. Changes to the software and interface from the flight version were:

- the amber indicator light was changed to yellow, so that it wouldn't be confused for red
- The cardiorespiratory (CR) section of the interface was disabled since it was irrelevant to the ground study
- The EKG checkbox and indicator light were replaced by a checkbox called "All EP" which was used to call up troubleshooting procedures for all signals.

• When subjects select the indicator lights to bring up the diagnostic messages, a little blue circle appears within the light to show that it was selected.

Also, we tried to create a semi-automatic indicator light, whereby [PI] would pop up messages automatically, and subjects could click the indicator light to keep the messages on the screen. But we decided the diagnostic messages should stay in manual mode.

3.2.1 [PI] logs and recordings

When the "Record data" option is set in [PI], it will record experiment events into a log file, whose name is specified in a configuration "DEF" file. [PI] can record clicks from the event marker, and both the subjects' and its own signal quality assessments.

Chapter 4

Signal overview and [PI] reasoning

4.1 Signals displayed by [PI]

The electrophysiological signals to be monitored consist of the electroencephalogram (EEG) and the electro-oculogram (EOG).

4.1.1 The electroencephalogram

The electroencephalogram (EEG) is the primary polysomnographic measure in evaluating and scoring sleep data. The four different stages of Non-Rapid Eye Movement (NREM) sleep can be distinguished based on the EEG signal characteristics alone. In general, a minimum of one central and one occipital EEG are recorded in sleep studies. As a fail-safe measure, two of each will be recorded for this experiment

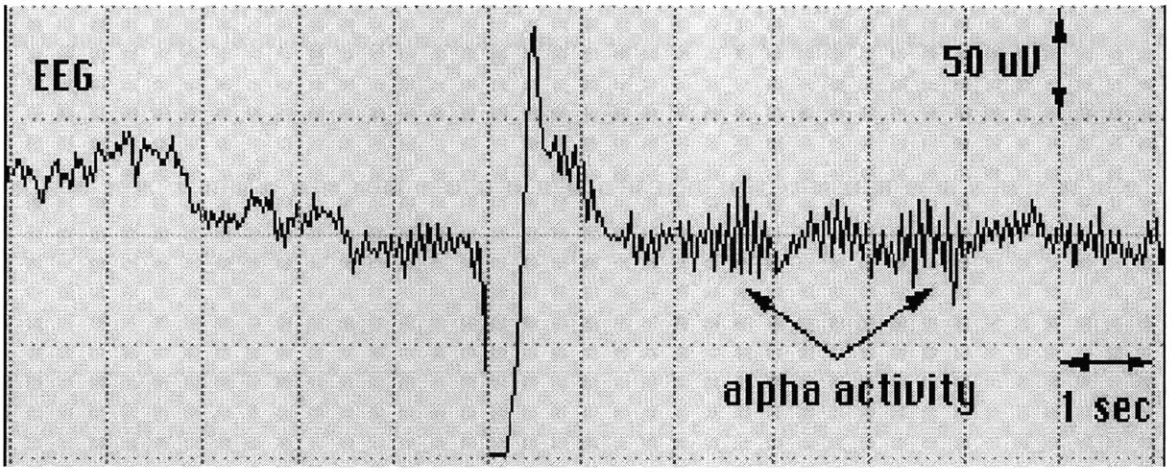


Figure 4.1: EEG waveform. Alpha activity is indicated.

While the subject is awake and relaxed with eyes open, the EEG is a mixed-frequency, low-amplitude signal. A standard amplitude range for the EEG signal is 10-200 mV. When the subject's eyes are closed, a rhythmic, higher frequency pattern becomes apparent in the EEG signal. This activity, known as alpha activity, is characterized by a frequency in the 8-12 Hz range. Alpha activity is generally most prominent in the occipital EEG, but is also

discernible in the central EEG. An example of EEG data, including alpha activity is illustrated in Figure 4.1

The EEG occipital and central electrodes are located on the head as indicated in Figure 4.2(a). The EEGs are referenced to either the A1 or A2 reference electrode, described later.

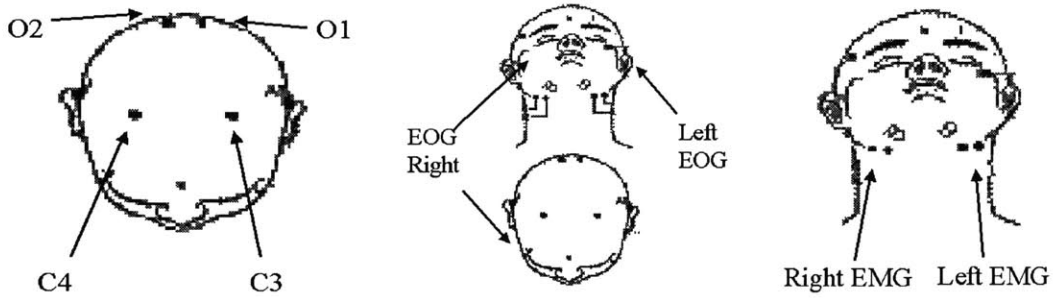


Figure 4.2: Electrode locations for (a) EEGs, (b) EOGs, (c) EMGs

4.1.2 The electro-oculogram

The electro-oculogram (EOG) consists of two waveforms, one for each eye. The EOG signals are extremely important for determining REM sleep, since any kind of eye movement is highly distinguishable on the EOG signals due to their shape. The EOG electrodes are located such that the signals from each eye are oppositely polarized, as shown in Figure 4.2(b). An example of an EOG signal is shown in Figure 4.3.

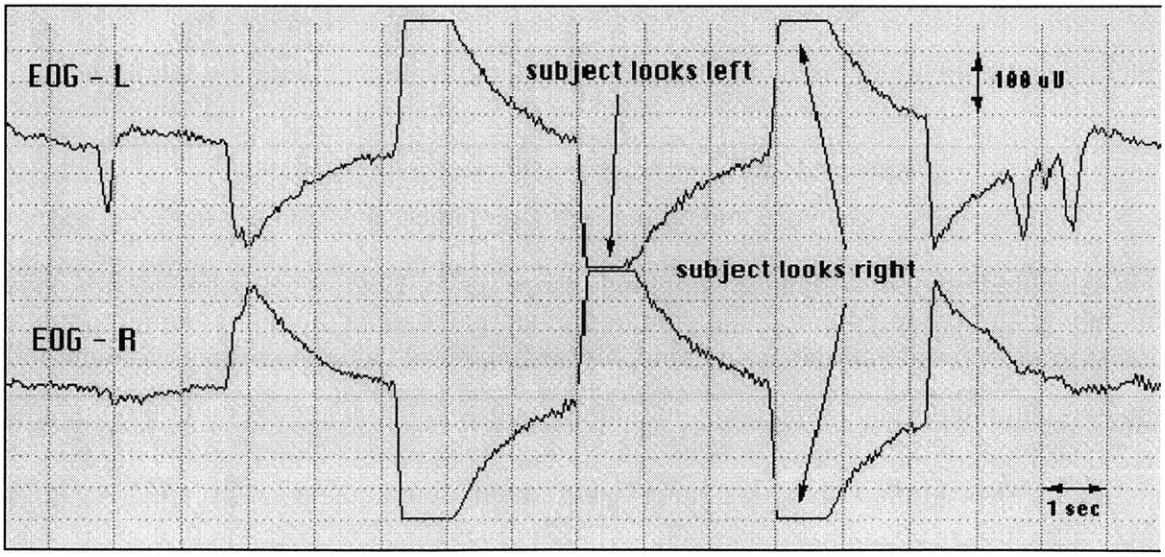


Figure 4.3: EOG signal. Deflections where subjects look left and right are shown

The EOG signals make use of the potential difference that exists across each eyeball. The amplitude of the EOG deflections can range from 20 to 500 microVolts. Since the EOG is AC-coupled, the signal decays to zero during a steady gaze.

4.1.3 The electromyogram

The third electrophysiological signal is the electromyogram, which records jaw muscle activity. Each EMG, as shown in Figure 4.2(c), measures the potential difference between two electrodes placed across the jaw muscle. The EMG helps to distinguish REM state, since one of the characteristics of rapid eye movement sleep is the loss of muscle tone. Generally, the EMG looks like a noisy, high frequency signal which dramatically increases in frequency and amplitude in the case of a muscle contraction (clenched jaw). The amplitude can vary between 20 and 300 microvolts. An example of an EMG is shown in Figure 4.4.

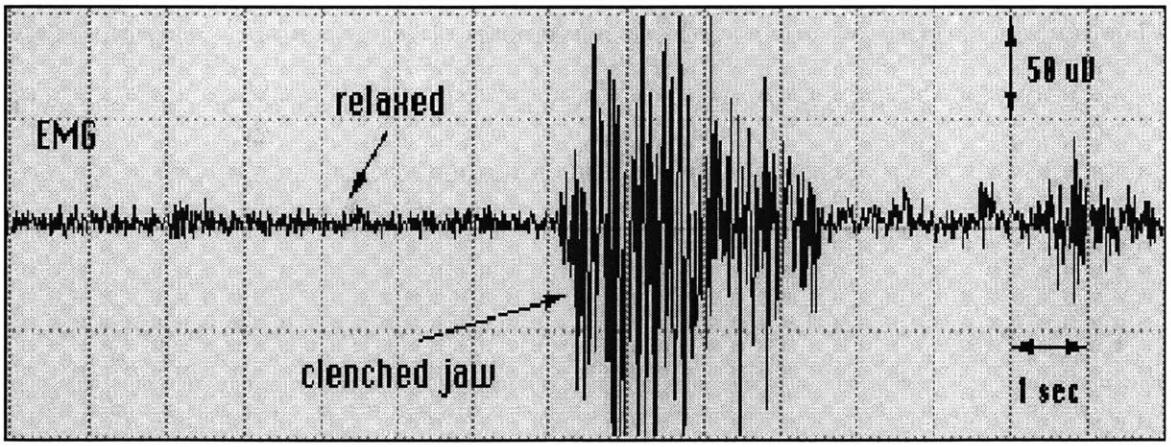


Figure 4.4: An EMG signal. Deflections where subjects clenched jaw are indicated.

4.2 Signal Calibrations

There are a number of calibrations which a trained sleep signal expert can perform to verify that the signal waveforms are working. Test subjects were trained how to carry out these calibrations and use them in the detection and diagnosis of failures.

4.2.1 Eye calibrations

Each EOG electrode is referenced to the corresponding reference electrode located behind the opposing ear. A reference electrode is different from the ground electrode, to which every signal is referenced. An eye movement across the horizontal plane produces a negative voltage in one eye and a positive voltage in the other. The deflection that these movements produce on the EOG signals are approximately equal in magnitude but opposite in polarity. If a sleep subject makes an eye movement to the right, it results in a negative (upward) deflection of the left EOG signal and a positive (downward) deflection of the right EOG signal. The exact opposite situation occurs for an eye movement to the left. Figure 4.3 shows a waveform of a left-right eye movement.

The right EOG is placed above the eye and the left EOG is placed below the eye to record vertical eye movement. An upward eye movement creates a negative (upward) deflection of the left EOG and a positive (downward) deflection on the right EOG signal. The reverse happens for downward eye movements. The deflection of the right EOG is more noticeable than the left because the right EOG is placed above the right eye, which also picks up movements of the upper eyelid. An example of such an up-down eye movement is shown in Figure 4.5.

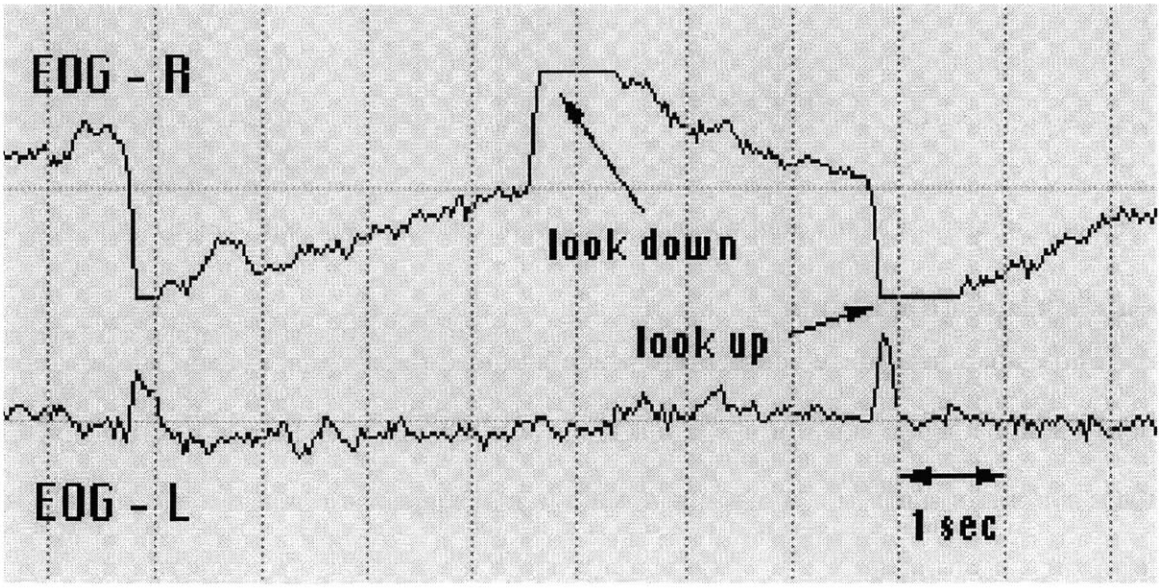


Figure 4.5: An EOG signal. Deflections where subjects looked up and down are indicated

4.2.2 Clenching the Jaw

Another easy calibration to perform is to have the sleep subject clench their jaw. The relaxed EMG waveform is of low amplitude; once the jaw muscles are contracted, the electrical activity will give a high amplitude EMG signal. An example of such a calibration is shown previously in Figure 4.4.

4.2.3 Other diagnostic techniques

Apart from the calibrations, test subjects may ask sleep subjects to do other things to help in their detection. Sometimes motion artifacts are picked up by the waveforms: sleep subjects moving their head, yawning, etc. Having the subject relax will settle these waveforms so that the EP signals can be analyzed. The test subject may also ask the sleep subject to create motion artifacts to see which signals are active; this is particularly helpful for finding a missing electrode signal - since all other signals would exhibit the motion artifact except for the "dead" channel.

4.3 Signal Artifacts on a single-channel

Signal deterioration can occur because there are motion artifacts, or because of several failures in the instrumentation. The indicator lights tell the test subject of [PI]'s assessment of the signal quality. However, sleep experts divide the anomalies further into three main categories, described here. The reason for categorizing these anomalies is that they can be traced back to possible diagnoses of faults in the system. Therefore, they are considered observables of instrumentation failures. With each anomaly description, a sequence of possible faults is described, in order of probability. These fault traces correspond to those used by sleep experts at BWH, and were taught to the astronaut surrogates during training for the experiment.

4.3.1 Popping artifacts

A "popping" signal occurs when there is poor or intermittent contact between the hydrodot and the scalp. Popping is very distinctive and easy to recognize. Popping signals are primarily caused by hydrodots not being inserted flush against the skin, hair beneath the hydrodot, poor electrode placement, or motion artifacts - in order of probability. Popping artifacts are similar to Figure 4.6.

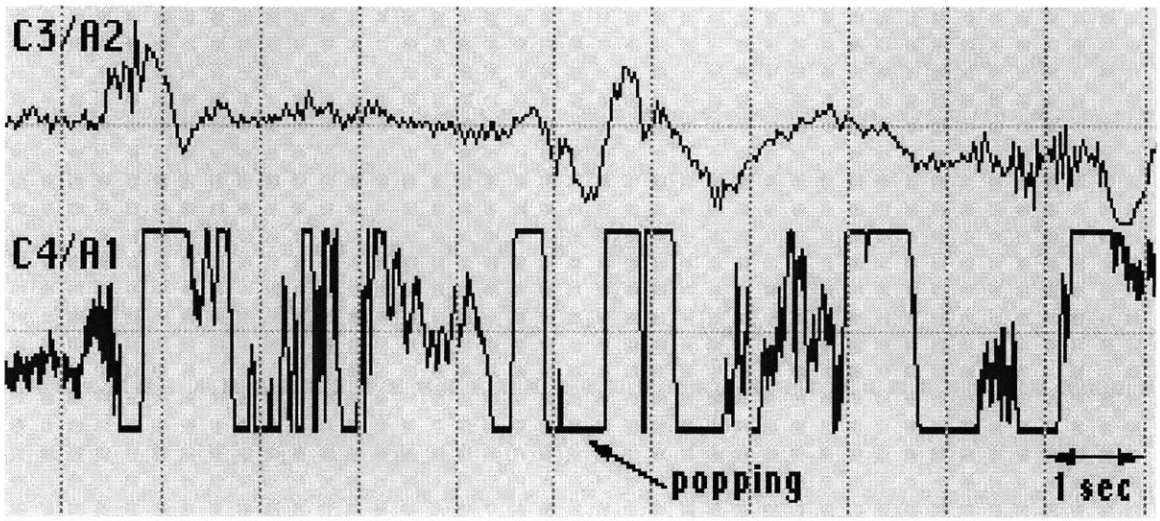


Figure 4.6: (above) a nominal signal, and (below) a popping signal

4.3.2 Flat artifacts

A flat signal occurs when an electrode is detached from the skin. Flat signals have a characteristic exponential decay as an electrode is pulled, and the potential difference will return to zero. A flat signal will exhibit no response when a calibration is performed on the

faulty channel, so it is an easy signal to diagnose for a sleep expert. Flat signals are caused by the absence of a hydrodot, or poor skin contact. Figure 4.7 shows a typical flat signal.

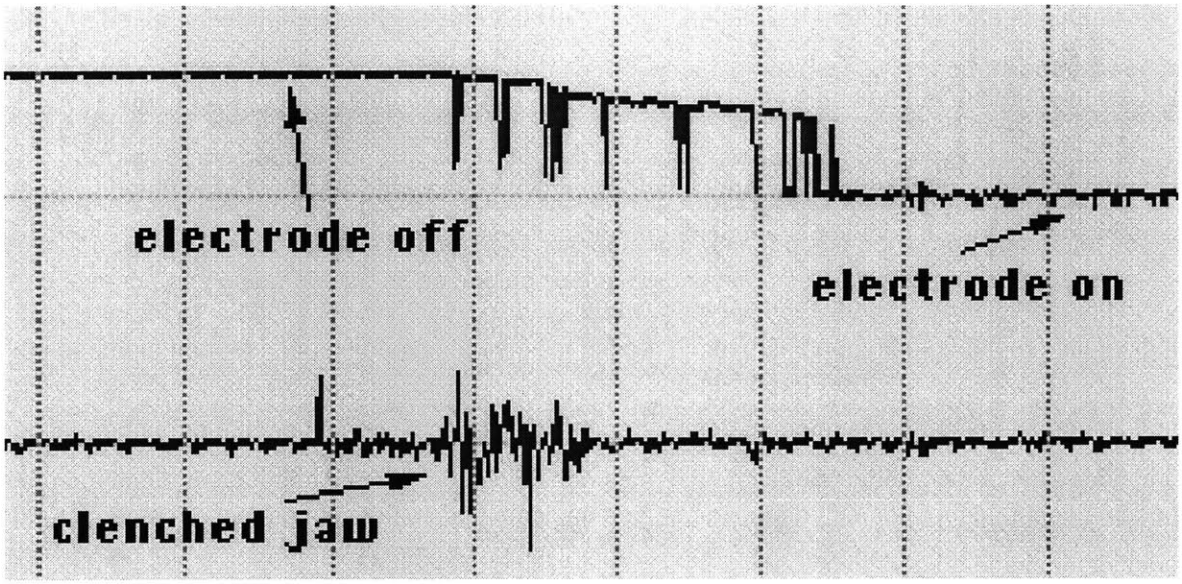


Figure 4.7: (above, left) a signal with electrode in place (above, right) a signal without electrode in place

4.3.3 Noisy signals

Noise usually appears as a random, high frequency signal. Noise appears when a high impedance exists between the electrode and the reference. The high impedance (over 10kOhms) will make the electrode a conductor of light radiation, and pick up the 60 Hz noise from the room lights. The causes may be a site not being properly scrubbed, or hair

between the hydrodot and skin (for an EEG). An example of a noisy signal is shown in Figure 4.8.

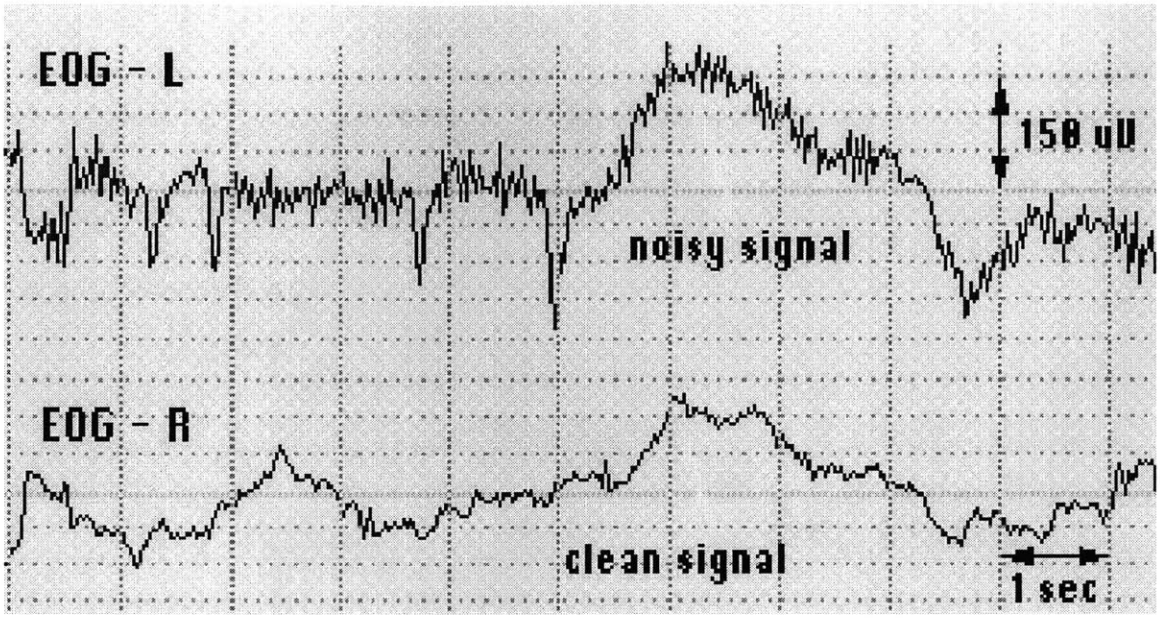


Figure 4.8: (above) a noisy signal, (below) a clean signal

4.3.4 Saturated signals

Saturation occurs when a signal goes outside the bounds displayed by the interface. Saturation can occur if an electrode is improperly placed, or there is a DC offset in the signal recorded by the DSR. Despite the fact that the signal quality cannot be interpreted by the display, [PI] can still assess the signal quality and display its indicator lights. Test subjects were trained to be wary of saturated signals as a potentially poor quality signal. An example of a saturated signal is shown in Figure 4.9

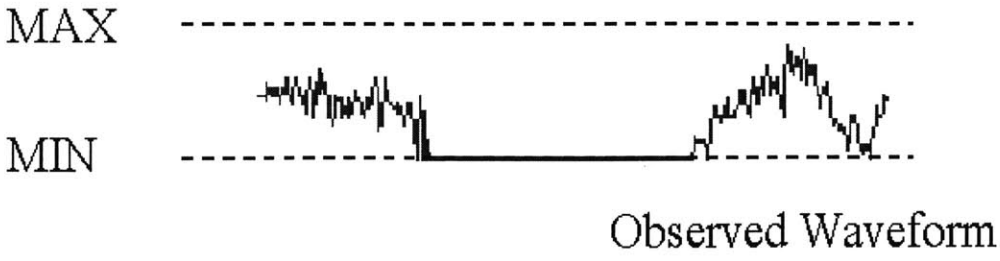


Figure 4.9: A saturated signal, with display boundaries indicated with a dotted line

4.3.5 Mixed anomalies

Sometimes two anomalies can appear at the same time in a waveform. Test subjects were trained to interpret the signal as best they could. However, a mixed anomaly is a more complicated signal from which to diagnose faults. Although mixed signals are not known to trace back to particular diagnoses, we still study the way subjects used them as observables to guide their troubleshooting questions. An example of a signal with popping and saturated behavior is given in Figure 4.10

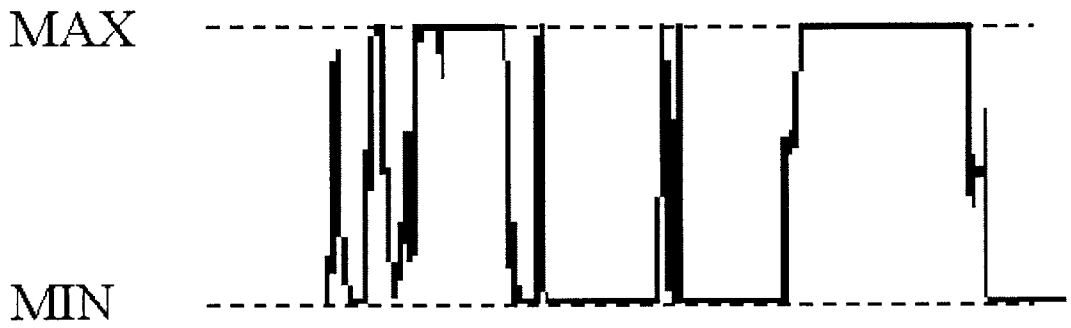


Figure 4.10: A mixed signal that is popping and is saturated at the same time

4.4 [PI] heuristics for signal presence and quality

4.4.1 Rules for single channel anomalies

The raw signals from the SleepNet are recorded by the DSR, and also sent to the ThinkPad through the serial cable. The signals are received by [PI] and stored in a buffer which holds data from the past 4 seconds of recording. [PI] uses statistical thresholds to assess signal quality. The rolling mean and variance of the 4-second buffer is calculated, and passed to the expert system reasoning engine. Upon activation, [PI] waits 20 seconds before analyzing signals, but checks each second for signals from the serial port.

Once the [PI] heuristic rules have established the presence of a signal, the signal quality is checked to be within the bounds specified. [PI] does not diagnose EMG signal quality because its "noisy" nature makes it difficult to assess its signal quality using thresholding methods.

4.4.2 Rules for assessing anomalies on all EP channels

[PI] also checks if EP signals are present with another set of rules [see 1]. If all EP signals are not being picked up, [PI] will automatically display diagnostics in the message window, but the "All EP" indicator light will not turn RED. Once all signals are present, [PI] will check if all signals are showing poor quality. Should all indicator lights turn red, such as when the ground electrode is not present, [PI] will show diagnostic messages. It takes time for all the indicator lights to turn red even when all the EP channels are poor quality. But when there are no EP signals present, [PI] can almost immediately detect it.

Signal	indicator status	Rules
EEG (microVolts)	Good	100 < variance < 4000
	Unknown	4000 < variance < 7500
	Poor	variance > 7500
EOG (microVolts)	Good	variance > 30 & -100 < mean < 100
	Unknown	variance <30 & -100 < mean < 100
	Poor	mean > 100 or mean < -100
EMG (microVolts)	Unknown	Currently no diagnostic rules
All EP signals	Good	at least one channel not poor quality
	Poor	All channels = Poor quality

Table 4.1: [PI] signal quality rules

Chapter 5

Experiment

5.1 Overview

The experiment on which this report is based ran through the month of January, 2000. It was conducted as phase II of a National Space Biomedical Research Institute (NSBRI) grant to study the efficacy of the [PI] expert system for space life science experiments - awarded to Professor Laurence Young in April, 1998. The objective of the experiment was to see if [PI] would successfully assist astronaut surrogates detect and diagnose signal anomalies in a life science experiment outside of their domain of expertise.

5.2 Experimental Design

We thought that two experimental days instead of three days as in the experiments conducted by Rouse would be more feasible to run and would be more realistic in terms of actual exposure to the experiment in space. Running subjects through the four groups for Day 2 and Day 3 would be the most instinctive design to choose. However, we expected the training effects of the subject to be negligible from Day 1 to Day 2. The experimental design selected consisted of the two "transfer groups," neglecting the possibility that going from aid to no aid would be the same as going from no aid to aid, i.e. a memory exists in the system. The consequence of the simpler design using two groups over only two days was that the training and transfer effects would be intertwined. In retrospect, the experimental design was oversimplified based on the effects observed.

The experiment used a balanced crossover design, as shown in Table 5.1.

Table 5.1: [PI] experimental design

	Day 1	Day 2
Group 1 (N=14)	[PI] on	[PI] off
Group 2 (N=16)	[PI] off	[PI] on

The two groups of subjects differed only in whether they began with or without [PI] assistance. Group 1 had [PI] assistance only on their Day 1 and Group 2 only on their Day 2 of testing. The assistants mistakenly did not administer [PI] to two subjects on Day 1, which is the reason the groups are not exactly balanced. We tested the subjects' ability to detect and subsequently troubleshoot faults in the electrophysiological instruments. A repeated measures approach allows each subject to be his own control.

Each subject was then tested in two thirty minute sessions on two separate days with the instrumentation. The pace of the experiment is somewhere in between self-paced and forced-paced – since there is a time limit on each fault. There may be a mixing of two or more fault management strategies for these trials.

5.3 Subjects

Subjects were students currently at MIT from various academic backgrounds, all between ages 17 and 34, with an average age of 22. One subject completed Day 1 but not Day 2 of the experiment. Thirty subjects, consisting of 14 females and 16 males, participated in the experiment on both days. Both males and females were allotted to each group, so the experiment was completely balanced in terms of gender effects. A sample of the consent form for test subjects is provided in Appendix . Throughout the study 16 subjects donned the instrumentation as sleep subjects. Sleep subjects were limited to participating in not more than 2 experimental sessions per day to prevent wear on the skin. A sample of their consent form is provided in Appendix .

When subjects arrived for training, they were asked to complete a questionnaire to

assess the uniformity of the subject pool. A copy of the pre-experiment questionnaire is also provided in Appendix C. Only two subjects were previously experienced with electrophysiological signals. Four subjects ran on two consecutive days, because of scheduling constraints and the timing of the experiment.

5.4 Training

Training was held on each Monday of a two week period, while each subject was scheduled for their two trial days within the same week. Each group received the same 3.5 hours of training on the sleep instrumentation and the [PI] interface. Two hours of training consisted of a slide presentation on the experiment, signal anomalies, and the experimental procedure. The remaining time was used to show a live instrumentation and brief demonstration of the [PI] interface. After the training session, subjects were given a quiz which was corrected in class. The quiz was mainly for didactic purposes for the subject.

5.5 Experimental Procedure

Each experiment involved the interaction of three people: an MIT research assistant trained by Brigham & Women's Hospital staff, a sleep subject who donned the instrumentation, and the test subject. First the assistant would begin each trial by creating a fault. Then, the test subject was asked to detect what fault, if any, existed, diagnose it, and instruct the assistant to correct the fault in real time. Interaction was limited: only the research assistants could handle the sleep instrumentation hardware.

Thirteen error trials were carried out on each day. These errors were broken down into three different fault types, null, single-channel and multi-channel faults. The null fault is different from the others because the detection task is to verify that all signals are okay rather than to detect and diagnose a particular fault. On Day 1, the order in which the errors were seen corresponds to that in Table 5.2, but on Day 2 the order was identically reversed, such that the last error on Day 1 was the first error on Day 2. The order in which

the errors appear was arranged such that different faults were randomly sorted. At the same time, the order was balanced according to the fault type. Faults in the system were

Error	Electrode	Fault	Desired Symptom	Fault Type
1	O2	hydrodot not flush with SleepNet	popping signal	single-channel
2	none	no fault introduced	all signals OK	null
3	C4	site not properly scrubbed	noisy signal	single-channel
4	All EP	RS-232 cable unplugged	display freezes	multi-channel
5	A2	reference electrode loose	3 flat signals	multi-channel
6	All EP	DSR stopped recording	display freezes	multi-channel
7	none	no fault introduced	all signals OK	null
8	O1	hydrodot not inserted	flat signal	single-channel
9	Ground	electrode not inserted	all signals poor quality	multi-channel
10	left EOG	hydrodot not flushed	flat signal	single-channel
11	All EP	SleepNet connector not plugged in	all signals poor quality	multi-channel
12	C3	hair beneath hydrodot	popping signal	single-channel
13	right EOG	site not properly scrubbed	noisy signal	single-channel

Table 5.2: A breakdown of errors introduced

created, detected, diagnosed, and fixed in real time. The experimental setup is shown in Figure 5.1. Test subjects were permitted to view the signal display while the assistant created a fault, but they wore headphones to prevent audio cues that might hint at the problem they were about to face. When a fault (or no fault) was created, the assistant alerted the

subject to begin his search for the problem. The subjects removed the headphones and clicked on an event marker, indicating in the data file that they had begun each scenario.

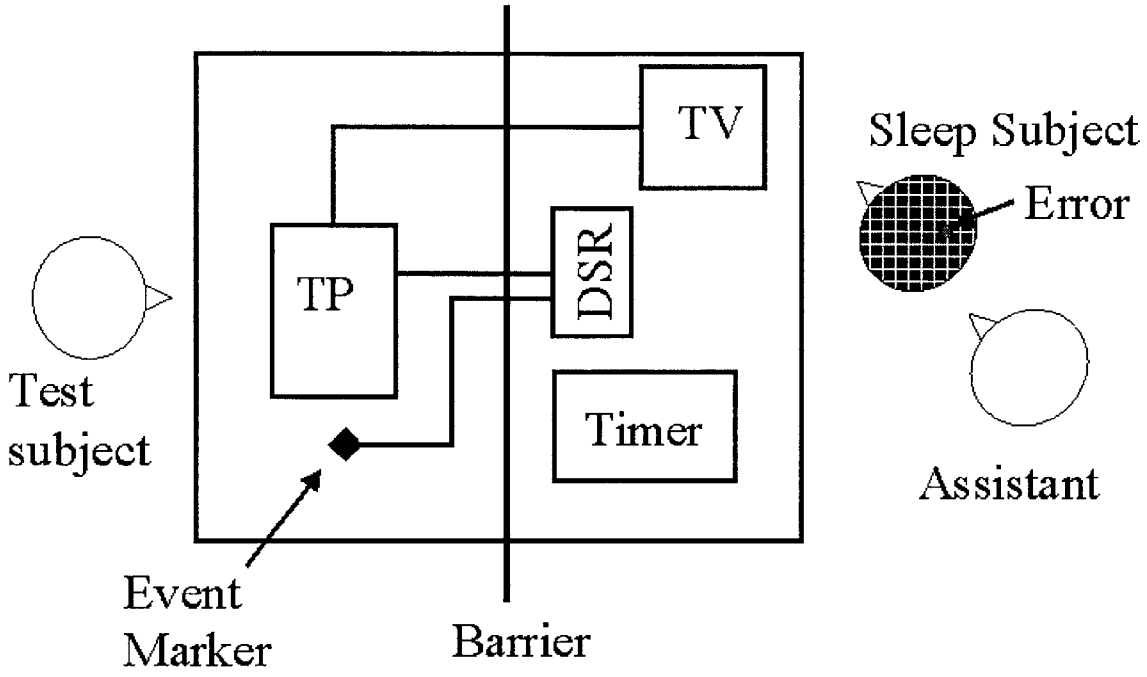


Figure 5.1: Experimental layout (top view)

They had 180 seconds to detect and diagnose a fault (if any) and analyze the signal waveforms and the color-coded indicator lights, if available, to determine if an error had been introduced into the system. Once a fault was known to exist, the subject would click on a gray checkbox and select a system state from a list of possible symptomatic states, listed in Table 5.3. The subjects were instructed to ask if there was "no error," if they felt

that all signals appeared to be good quality. Only after we confirmed this were they supposed to popup the checkbox to assess that all signals were OK.

State	Description
1	[PI] signals freeze
2	[PI] displays no signals
3	All EP signals are not present or poor quality
4	EEG signals not present or poor quality
5	EMG signals not present or poor quality
6	EOG signals not present or poor quality
7	System State OK - no problems
8	Other state

Table 5.3: Possible system states

After assessing the system state, the subject would isolate the fault by following the malfunction procedures outlined for the specific symptoms. The subject could use either [PI]'s diagnostic messages or the actual NASA troubleshooting guidelines given to the astronauts for use in-flight. The test subject could question the assistant to gain information about the instrumentation, but could not see the apparatus or interact with it physically. The subject was also trained to ask the sleep subject to make calibration movements (such as looking left and right) to verify signal presence and quality. Once the cause of the problem was determined, the test subject clicked on a gray checkbox to select from a list of possible problems associated with the symptoms recorded earlier. The subject would then ask the assistant to fix the specific fault they had diagnosed.

After the assistant confirmed that the fault was removed, the test subject would deselect any indicator lights, turn back to the first page of the NASA guideline, put on the headphones and prepare for the next scenario. Sometimes a fault was induced that was unintended by the assistants while they created the initial fault. These faults are called

"bonus errors." The assistant would let the test subject detect and diagnose both faults in order to keep the experiment going. However, in the analysis, only the primary errors will be analyzed. Following the experiment, the subjects completed a questionnaire and were debriefed by the assistants.

Chapter 6

Data collection, Analysis and Results

6.1 Overview

[PI] recorded data mainly for three things: one, the event marker that indicates the start of a trial; two, the subjects' state, problem and solution assessments for each trial; and three, the [PI] signal quality assessments. With these data, we can extract the detection time (t_d), troubleshooting time (t_{ts}), repair time (t_r), and the correctness of a state assessment. The response files of 61 test sessions were recorded; thirty subjects with two days of data each, and one subject with only Day 1 data.

Another data source was the troubleshooting logs, on which the assistants recorded the signal quality of the fault (e.g. "popping"), the diagnostic questions asked by the subjects, the calibrations performed by the subjects, the repair instructions, and other observations related to each trial. These data are used as secondary performance indicators to the time and reliability measurements. This data will help to break down the data into components related to each task necessary for this analysis.

The raw DSR recordings of the waveforms from the test session were downloaded and stored on the ThinkPad in a format known as VPD using software called Columbus. These VPD files are used as secondary data sources to analyze what signal quality the subjects witnessed during the instrumentation.

6.2 Data extraction and reduction

All recordings made by [PI] in the log file are accompanied by a time stamp of when the entry was recorded. Each type of data recorded by the log file had a specific formatting that distinguished itself from others. These formats are indicated in the following Table 6.1.

A data extraction program was developed to parse through the logs, and tabulate the data required. The program was written in MATLAB script language. The assistants went through each log file by hand after each experiment to verify each trial; all manual entries were preceded by a double slash "//". A breakdown of the syntax and formatting for the log files is provided in Appendix A.

The assistant logs were transferred to electronic tabular form after the experiment. The two data files were merged afterwards into one large file. Bonus errors were removed where possible, except when the primary error was not properly created (4 cases out of 793). Trials in which subjects did not click the event marker were included in the tabulated data. Assistants marked t_0 manually with the assumption that the onset of an error was about 30 seconds after the previous error was cleared. The tabulated data are stored in two formats: one which describes in detail the subjects' first attempt for each trial, an another which summarizes the subjects' actions (number of attempts, t_d, t_{ts}, etc.) for each trial.

6.3 Data Manipulation

Once the raw data was in tabular form for each trial, the reliability and time measurements

were derived as described below.

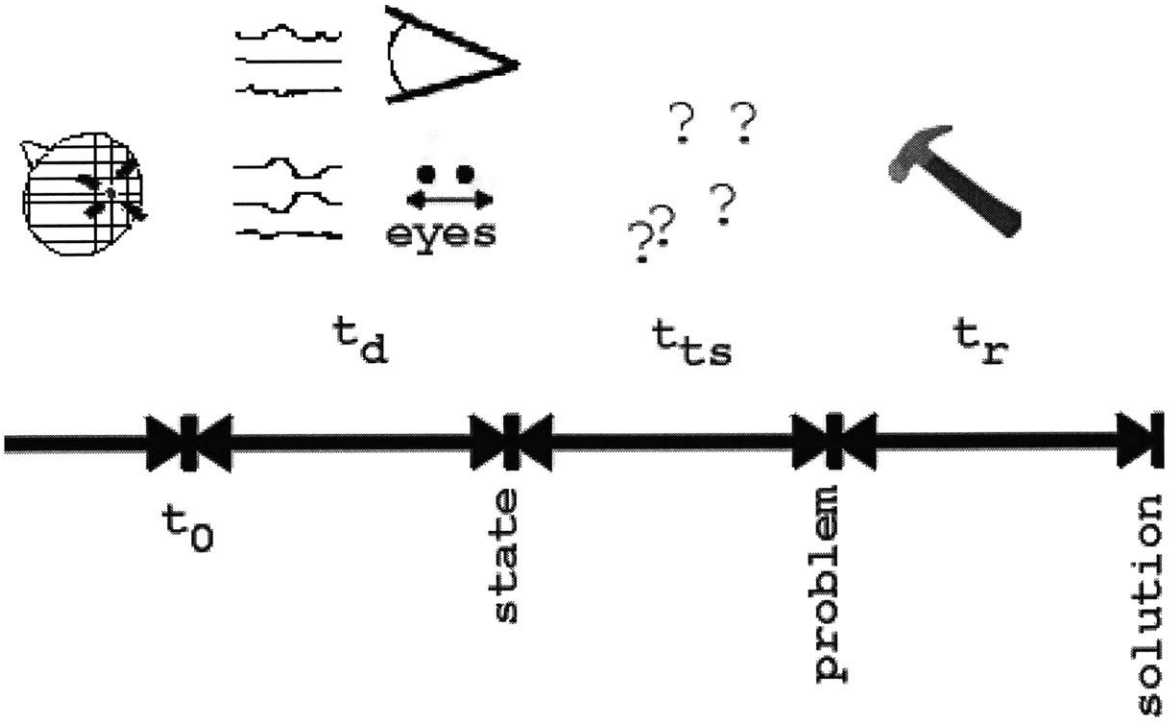


Figure 6.1: Timeline of an error. The four stages, along with pictorial representations of tasks are shown.

6.3.1 Time measurements

The subjects detection time t_d was defined to be the time from the trial onset to the time the state dialog pops up, on their first state assessment, as in Equation 6.1:

$$t_d \equiv t(\text{state dialog popup, try 1}) - t(\text{initial}) \tag{6.1}$$

The troubleshooting time is defined to be the time between when the state is selected and the time when the problem dialog pops up on their correct assessment of the problem, as in Equation 6.2:

$$t_{ts} \equiv t(\text{problem dialog popup, try i}) - t(\text{state dialog pop down, try 1}) \tag{6.2}$$

where i represents the number of tries the subject needed to find the fault. The sum of t_d and t_{ts} is effectively defined as the isolation time, $t_{isolation}$, because the repair time is

independent of the subject's diagnostic ability. $t_{isolation}$, which is less than or equal to 180 seconds, is defined in Equation 6.3

$$t_{isolation} \equiv t_d + t_{ts} \leq 180 \text{ seconds} \quad (6.3)$$

The detection time for [PI] of a given fault type is defined as the first time it displays a red or yellow indicator light for the channel corresponding to the fault on that trial. For a null fault, $p(t_d;[PI])$ is the first time [PI] displays a red or yellow indicator light for *any* channel. For multi-channel faults where the [PI] display is frozen, a $p(t_d;[PI])$ of zero seconds is assigned, since such a fault is detected at once by [PI].

6.4 Correct fault isolation

In trials where a fault was created in the system, we could compare the time taken in detecting and finding the fault for each subject with and without [PI] help. In trials where there was no fault to detect, we can compare the performance of subjects in establishing that there was indeed no fault in the system. This is a "null" fault. We have adopted the cumulative probability $p(t_d+t_{ts})$ of detecting a fault (or the absence of one) and isolating it within the time $(t_d + t_{ts})$ as a measure of performance.

Subjects with [PI] help had a higher $p(t_d+t_{ts})$ in a null trial than subjects without [PI] help ($p=0.053$, via paired Kolmogorov-Smirnov test), i.e. they detected the absence of a fault earlier, on the average. Moreover, the difference in cumulative probability fades

about 90 seconds after onset of a trial. By that time, the subjects who did not have [PI] also established their results.

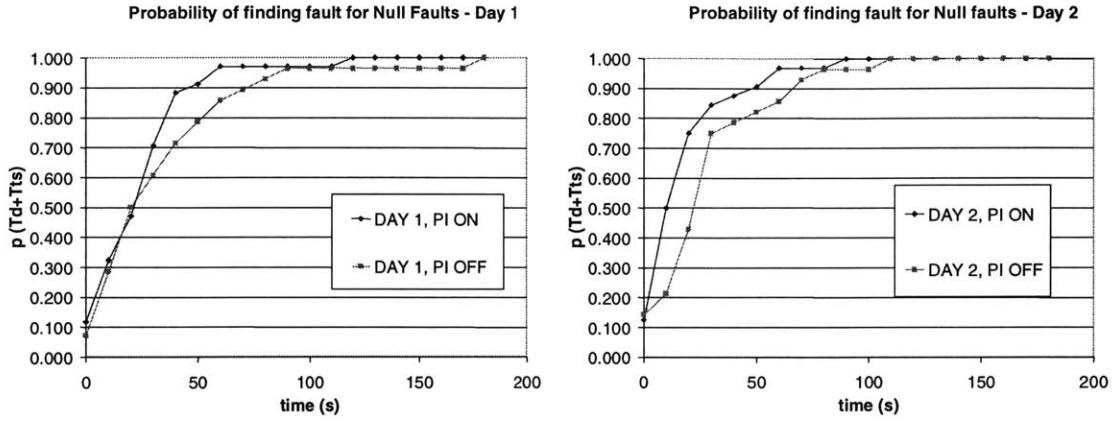


Figure 6.2: p (t_d+t_{ts}) for null faults (a) Day 1, and (b) Day 2

Group 2 which had [PI] assistance on the second day had a higher p(t_d+t_{ts}) than on Day 1 when they did not have [PI] assistance (p=0.041, via paired KS test). With [PI] help available, Group 2 had the highest cumulative probability of correct detection in the first 30 seconds among all cases. This could be because subjects in Group 2 learned more on Day 1 about the normal behavior of waveforms without [PI], when their counterparts were learning that and how to interpret the normal behavior of the [PI] indicator lights. Astronauts will need to learn how to assess the normal behavior of the indicator lights, and to reject false alarms when they turn red but no fault exists in the system.

A single-channel fault causes only one waveform to behave abnormally at a time. p (t_d+t_{ts}) was higher with [PI] than without it on Day 1 (p=0.021, via a paired KS test).

However, the help provided by [PI] did not improve the cumulative probability of finding a fault significantly on Day 2.

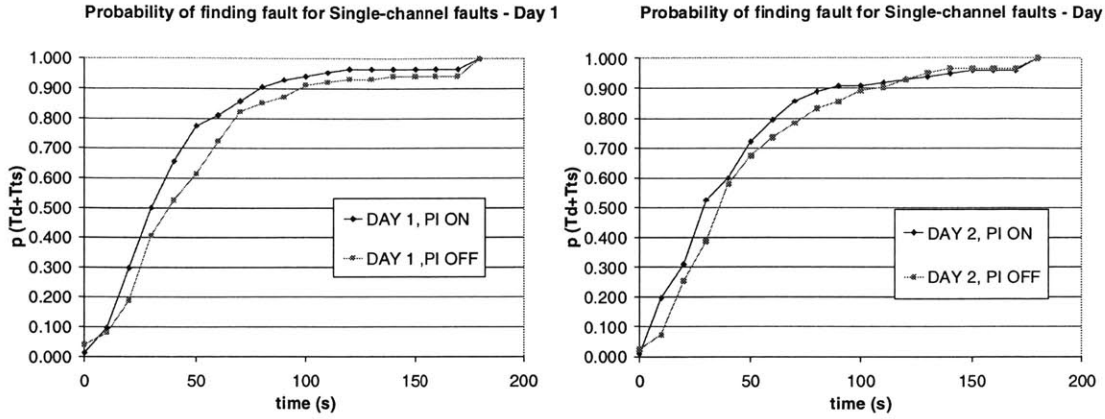


Figure 6.3: $p(t_d+t_{ts})$ for a single-channel fault on (a) Day 1, (b) Day 2

A multiple-channel fault causes either more than one waveform, or the entire system to behave abnormally. Subjects' $p(t_d+t_{ts})$ increased from Day 1 to Day 2 ($p=0.003$, via paired KS test) – improvement with practice. Those who had [PI] help also had a higher $p(t_d+t_{ts})$ than those without it ($p=0.0005$, via paired KS test). This improvement is most visible between 50 and 150 seconds after the onset of the trial. The skills learned with [PI] for troubleshooting multiple-channel faults for Group 1 were less effective if a fault took more than 30 seconds to find, because the [PI] interface merged troubleshooting steps from the NASA flowchart for two separate states into one list of troubleshooting messages displayed to the user. Subjects made many state assessment mistakes, and this discrepancy in the two troubleshooting lists could be the cause. Group 1 trained themselves during the experiment to following the "merged" checklist from the [PI] interface, and had difficulty looking up the correct pages in the NASA guideline. Beyond this impediment to Group 1,

subjects in Group 2 learned more about multi-channel faults on Day 1. This explains why they did very well with [PI] on Day 2.

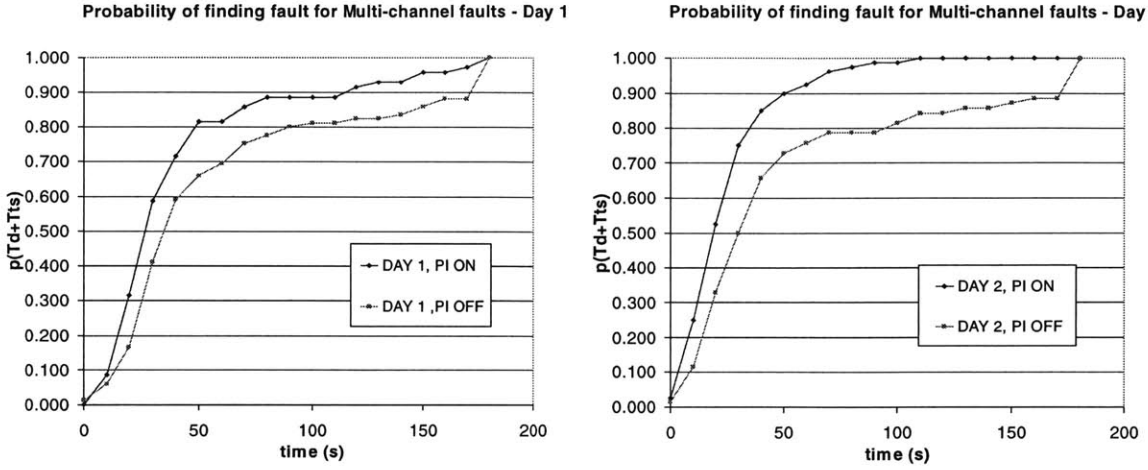


Figure 6.4: $p(t_d+t_{ts})$ for multi-channel faults on (a) Day 1, and (b) Day 2

Gender effects were not significant for either fault type (null, single-channel, and multi-channel fault p values were 0.998, 0.889, 0.896). This is different from the results obtained during the pilot study, where a gender effect was observed for detection time and reliability.

6.5 Correct detection on the first attempt

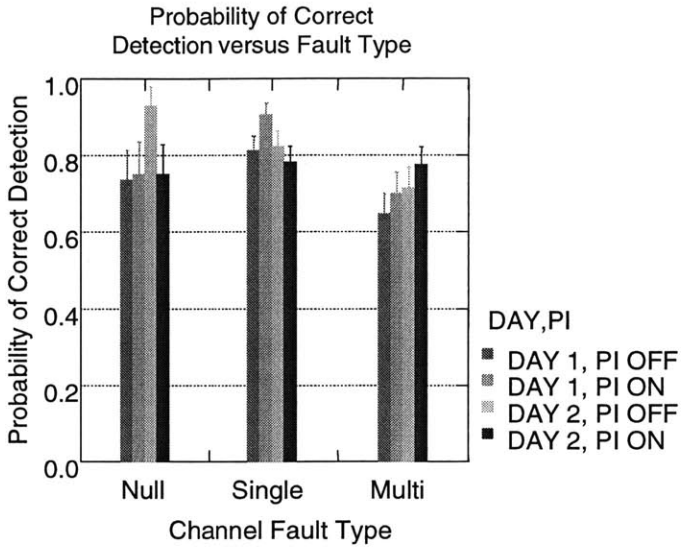


Figure 6.5: p (t_d; S) versus channel fault type

The cumulative probability $p(t_d+t_{ts})$ is a good aggregate measure of fault management since it takes into account the number of attempts the subject needed. It is also important to look at the correctness of the subject in detecting anomalous signals. The measure $p(t_d)$, the probability of detecting a fault correctly on the first attempt, is a pure measurement of the subjects' reliability in fault detection. This is because the subject's primary task during t_d is detection. During t_{ts}, subjects could detect signal quality and ask troubleshooting questions at the same time. The probability of correct detection was computed by averag-

ing over subjects' responses for each fault type. A detection of a single-, or multi-channel fault was considered correct if subjects' could find both the correct channel(s)

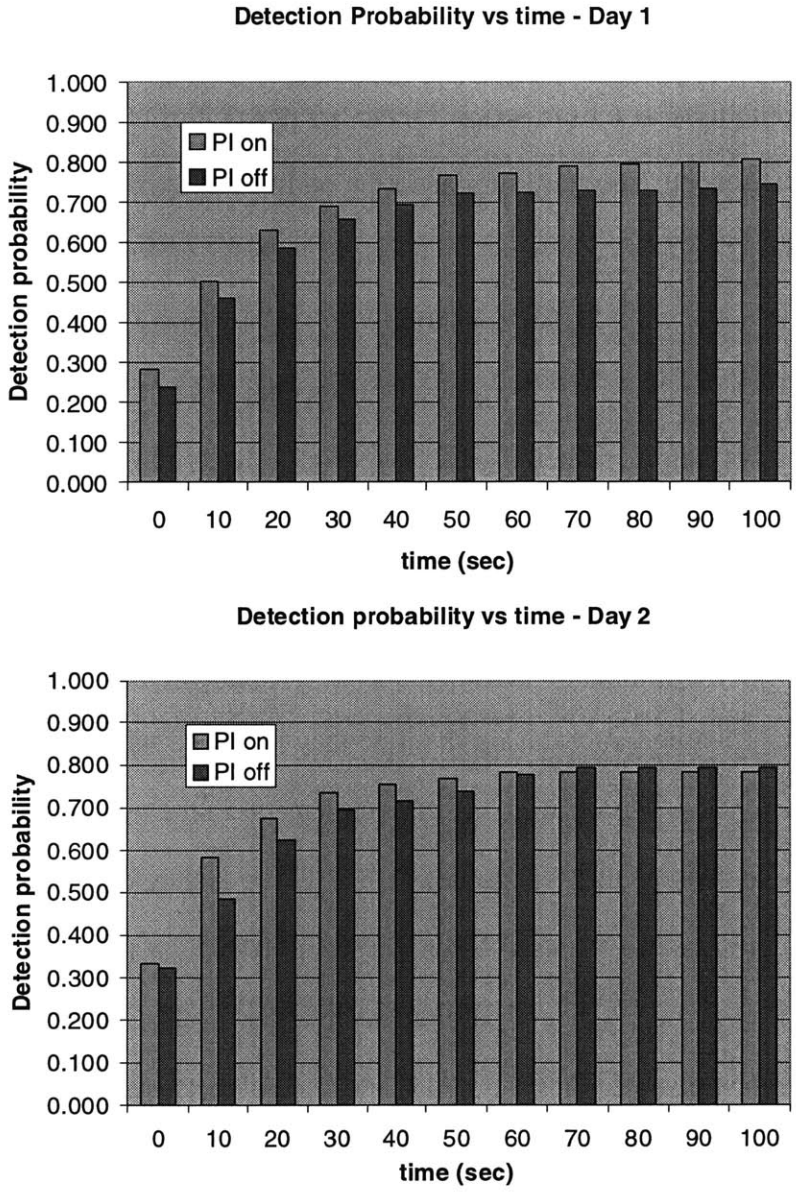


Figure 6.6: $p(t_d)$ versus time for Day 1 and Day 2

on which the fault was induced, and identify the correct system state which corresponded to the failure introduced in each trial. A null fault was considered correct if subjects found the system state to be OK - with no problems. The plots in Figure 6.5 show $p(t_d)$ for each fault type and for each Day, and [PI] setting. There does not seem to be any significant differences in percent correct detection between Day 1 and Day 2, or between

"[PI] on" and "[PI] off" settings. So on average, [PI] does not have any impact on reliability of detection.

However, since subjects took shorter or longer times to detect failures, it is important to look at the reliability as a time series. Figure 6.6 is a plot of $p(t_d)$ averaged over detection times in 10 second intervals, similar to the earlier $p(t_d+t_{ts})$ plots. The general trend shows that subjects could make their signal assessments faster with [PI] help than without it. The values of $p(t_d)$ were higher for those with [PI] on Day 1, but on Day 2, subjects with [PI] only had a higher p (t_d) for t_d about 70 seconds after fault onset. [PI] may afford different degrees of help for detecting each types of failure (null-, single-, or multi-channel).

[PI] does not increase $p(t_d)$ for a null fault, as seen in Figure 6.7. $p(t_d)$ increased for Group 1 from Day 1 with [PI] to Day 2 without it, which either means that [PI] is not beneficial or Group 1 learned a lot during the time they practiced with [PI]. In fact, Group 1 had a significantly higher probability of correct detection on Day 2 than Group 2 did on the same day (one-tail p=0.031, via z-test). As for Group 2, they benefitted on Day 2 with a higher $p(t_d)$ compared to themselves on Day 1 (p=0.041 via paired KS test). Further, from Day 1 to Day 2, the number of calibrations made by a subject decreased distinctly ($p < 0.005$, Kruskal-Wallis $\chi 2 = 7.85$, dof=1), because subjects became more experienced detecting a nominal waveform. What was expected was a large increase in reliability from Day 1 to Day 2 because of the additive effects of having one day's experience and [PI]

help at the same time. This shows that there is some effect here that cannot be explained by just [PI] and day effects.

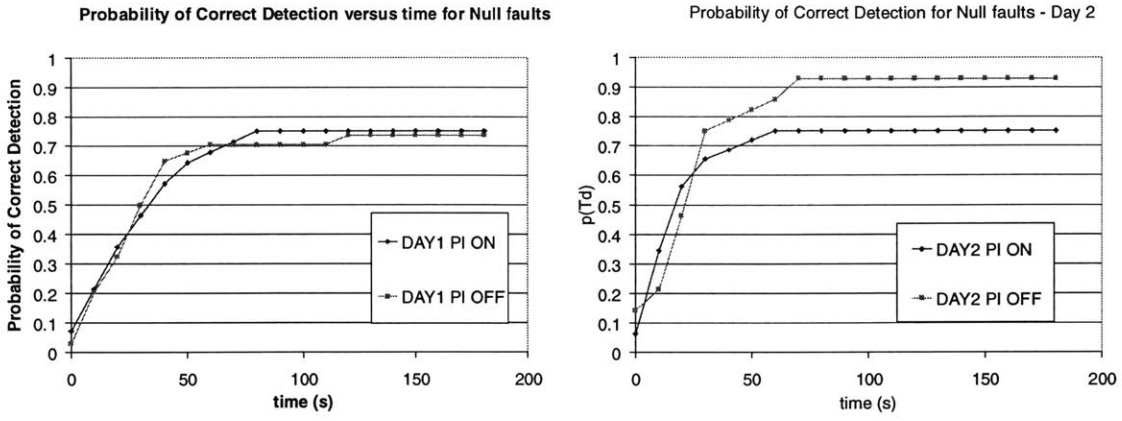


Figure 6.7: p (t_d) versus t_d for null faults on (a) Day 1, and (b) Day 2

For a single-channel fault, Group 1 was better on Day 1 than on Day 2 and Group 2 on either day, as seen in Figure 6.8. Group 1 was significantly better than Group 2 on Day 1 when detection time took longer than 70 seconds (one-tail p=0.026, via z-test). There was no significant difference overall between Day 1 and Day 2 performance for Group 2. It seems that Group 1 improved in interpreting signals and indicator lights together, but that was not helpful on Day 2 without [PI]. Moreover, Group 2 did not improve with [PI] on Day 2, because they did not adequately interpret the [PI] indicator lights. Group 2 made more calibrations than Group 1 ($p < 0.0005$, Kruskal-Wallis $\chi 2 = 18.189$, df=1), partly because it needed more information about the system than the waveforms alone provided, and partly because they assessed the signal quality of the EMGs for which [PI] help was unavailable. In fact, there were no intentional faults introduced in the EMGs throughout

the experiment. These differences contributed to the distinct behavioral patterns of the two Groups, which are outlined in detail later.

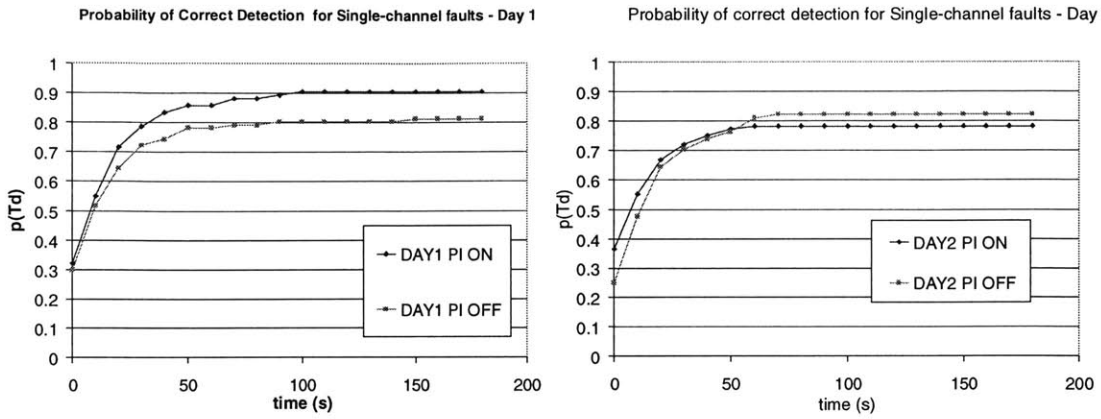
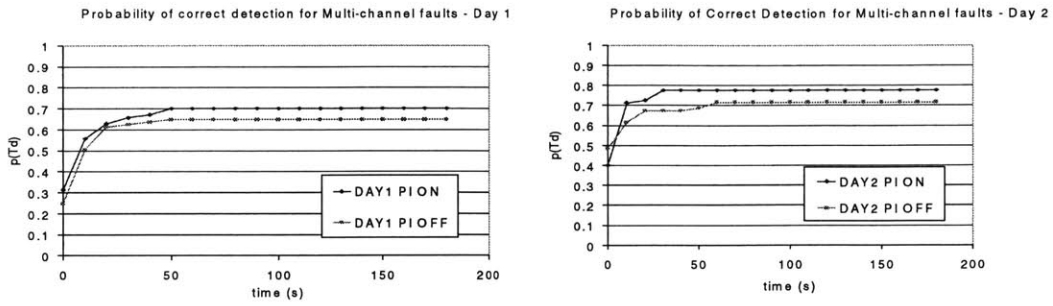


Figure 6.8: p (t_d) versus t_d for single-channel faults on (a) Day 1, and (b) Day 2

For a multi-channel fault, those with [PI] had a consistently higher p(t_d) than without it on Day 1. There is a small increase in p (t_d) from Day 1 to Day 2, but the main [PI] effect remains the same. The reliability of [PI] benefit is consistent from Day 1 to Day 2 because of the probability of correct detection for [PI] alone. It is higher for those with [PI] than without it, and on both days, meaning [PI] is just as easy to interpret for an experienced astronaut as for a novice one, for a multi-channel fault. This is because [PI] displayed diagnostic messages for these multi-channel faults without intervention from the subject - coercing him even more that something was wrong on all channels. On Day 2, subjects were worse with [PI] assistance for detection times less than 20 seconds. This can be attributed to an ambiguity between two of the three multi-channel fault states. Overall, those without [PI] generated hypotheses more slowly than those with [PI] ($p=0.0005$, KW $\chi 2 = 36.06$ df = 1). This is probably because those without [PI] needed more time to look

for the information they needed about the system before they could find the correct diagnosis.



6.6 Discriminability (d') of the system as a fault detector

Signal detection theory can be applied to the correctness of fault detection [18]. The basis of signal detection theory is to evaluate the discriminability of a diagnosis technique, or its ability to distinguish between both positive and negative stimuli. We analyzed only the responses for the subjects' first try, since this is the only time we are sure that subjects could spend their time analyzing waveforms. Table 6.1 summarizes the cases of both positive (single-, multi-channel faults) and negative stimuli (null faults).

We computed the positive probability, p (HIT), and the negative probability, p (FA), for

	Subject assesses state and channel correctly (**R+**)	Subject assesses state and channel incorrectly (**R-**)
single-, multi-channel(**S+**)	P(HIT), hit	P(MISS), miss
null(**S-**)	P(FA), false alarm	P(CR), correct rejection

Table 6.1: Stimulus-Response breakdown

detection of the three fault types. We found d' by subtracting z(p (HIT)) and z (p (FA)) where z is the Z-distribution parameter. Further discussion on calculating and interpreting d' can be found elsewhere [18, 23]. We computed a separate d' value for both the single-

channel faults and the multi-channel faults. These values are shown in Figure 6.9.

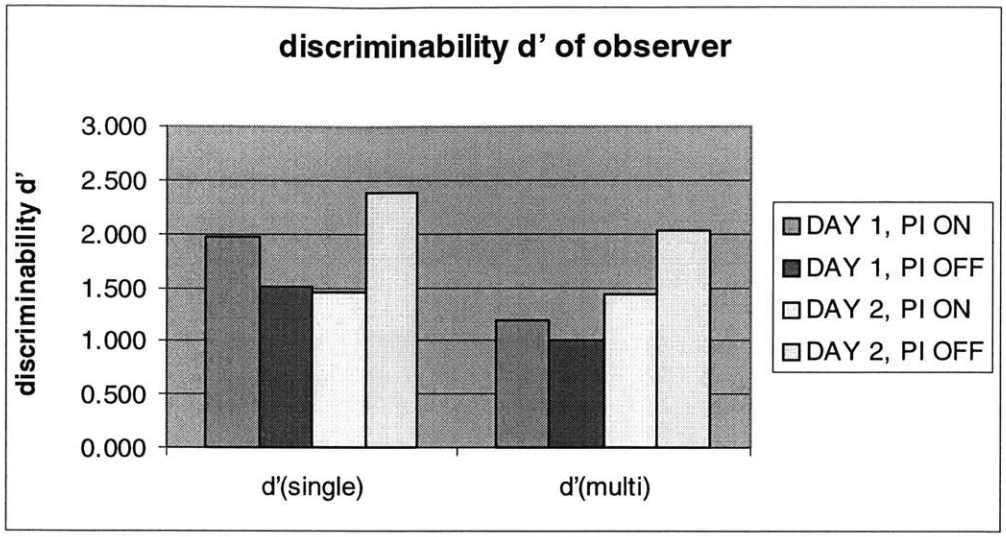


Figure 6.9: d' values for single- and multi-channel faults

On Day 1, the value of d', or ability to distinguish between a failure and a non-failure, was higher with [PI] than without [PI]. However, on Day 2, Group 2 had a lower d' than Group 1, despite having the aid. There are two effects interacting here: training allowed Group 1 to strengthen their discriminability from Day 1 to Day 2, but Group 2 was hampered by the need to learn how to interpret the [PI] indicator lights, so they did not improve as much as they could have. It is difficult to dissect these effects with only two experimental groups, especially since training was shown to be a significant factor for detection time [5].

There are some assumptions under which d' is valid. First, that the distributions of the p(HIT) and p(FA) are both normal and have equal variance. The number of data points for S+ exceeds that of S-, since there are only 2 null faults per day per subject. There may not be enough data for the S- case to make this assumption valid. Second, when plotting p(HIT) vs. p(FA) in 10-second intervals of t_d (see Figure 6.10) the area under the probability density function (pdf) for the false alarm was higher than the probability of a hit. Only after about 30 seconds, when the p(HIT) rises above p(FA), can the d' for each interval of time be meaningful. We neglected that and just took the probabilities for the entire time-

line.

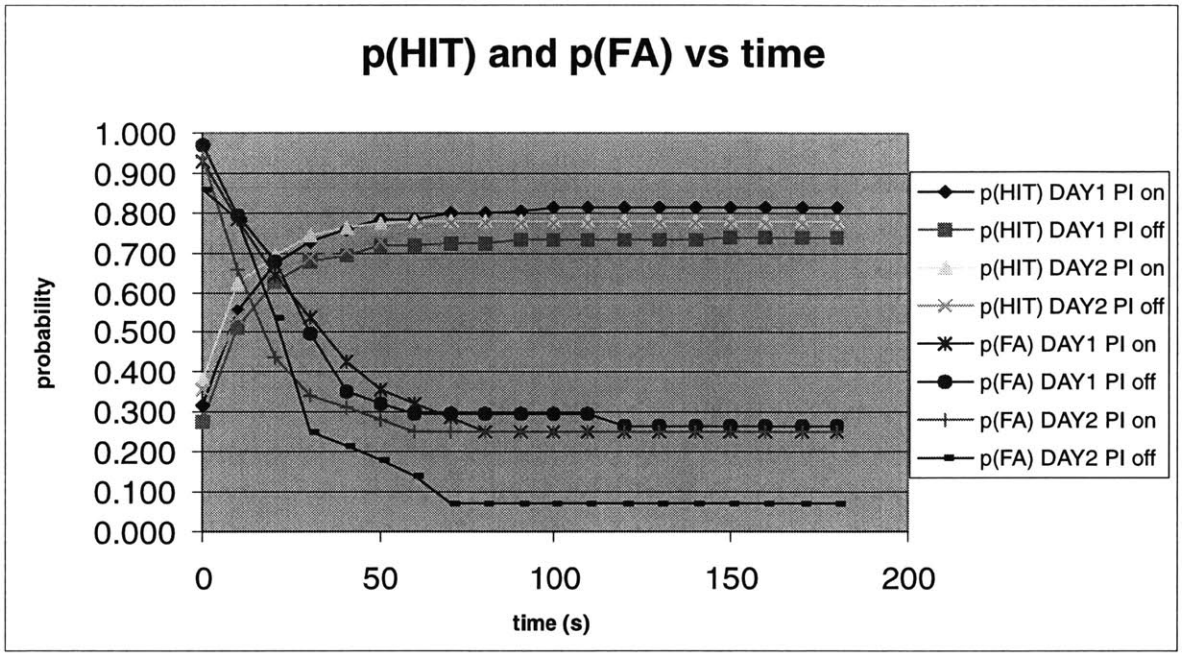


Figure 6.10: Probability of hit and false alarm rates across timeline

Third, these definitions of hit and false alarm are not correct in the strictest sense. For S+, if a subject correctly identified both the anomalous state and the correct channel, then it counted as a correct detection, or "HIT." A subject would have to explicitly say there was nothing wrong with the system on their first guess for it to count as a "miss," or an incorrect detection of S+. Too few subjects recorded misses, according to the above definition, for this probability to be meaningful.

6.7 [PI] indicator light reliability

Evaluating the reliability of [PI] for anomalous signal detection has been done in the past [2,1]. Within the controlled settings of this experiment, it can also be done to assess the effectiveness of the [PI] rules. The probability of a correct detection can be computed by counting a correct detection if [PI] fired a red indicator light at all during the detection time of each subject in each trial. This is a generous definition, since this discounts the

confusion that may arise from false alarms on other channels.

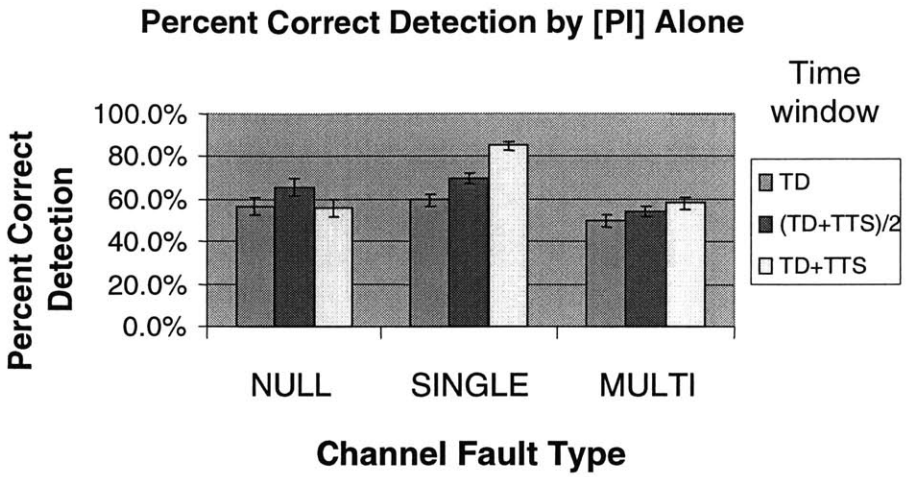


Figure 6.11: probability of correct detection by [PI] alone versus channel fault type

A plot of these reliabilities (see Figure 6.11) shows that overall the probability of [PI] correctly detecting a failure was about 60%, lower than that of the subject (see Figure 6.5) at the time they made their first detection t_d. But through the entire t_d+t_{ts} time interval, [PI] correctly detected a single channel fault 85% of the time. The reliability of [PI] for multi-channel fault detection was lowest because it required that all indicator lights turn red for it to detect such a fault, and under the "realistic" circumstances it was difficult to create these errors properly.

Figure 6.12: Subject percent correct versus anomaly type

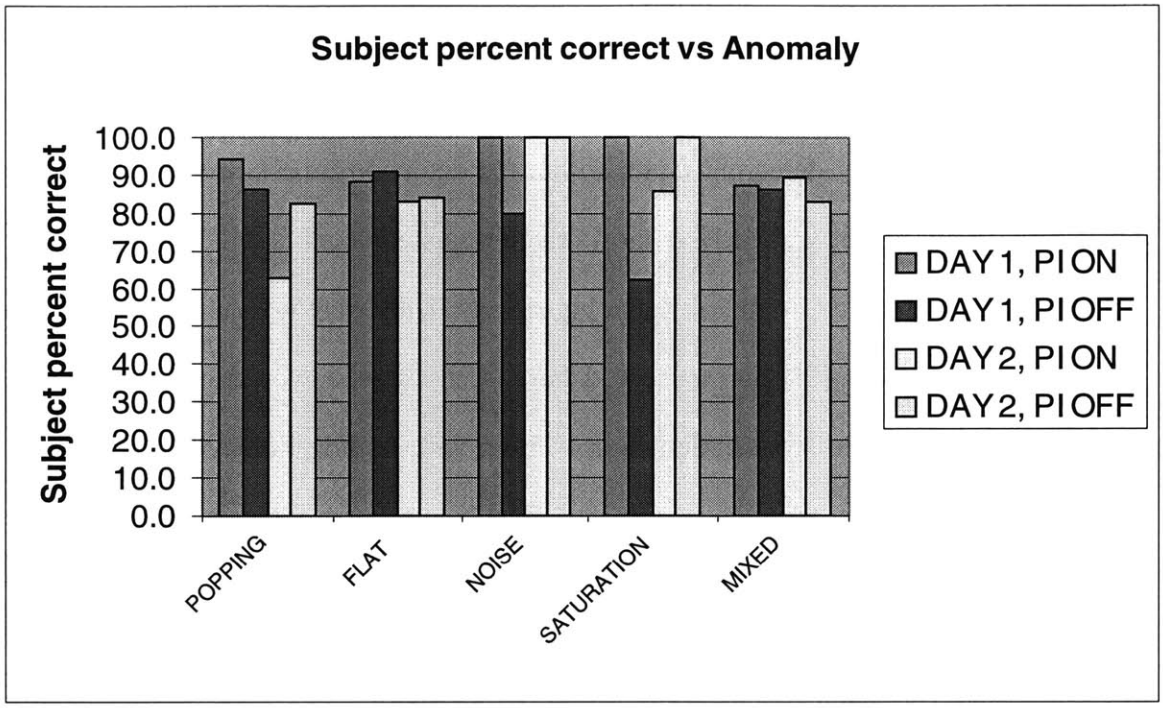


Figure 6.13: [PI] percent correct versus anomaly type

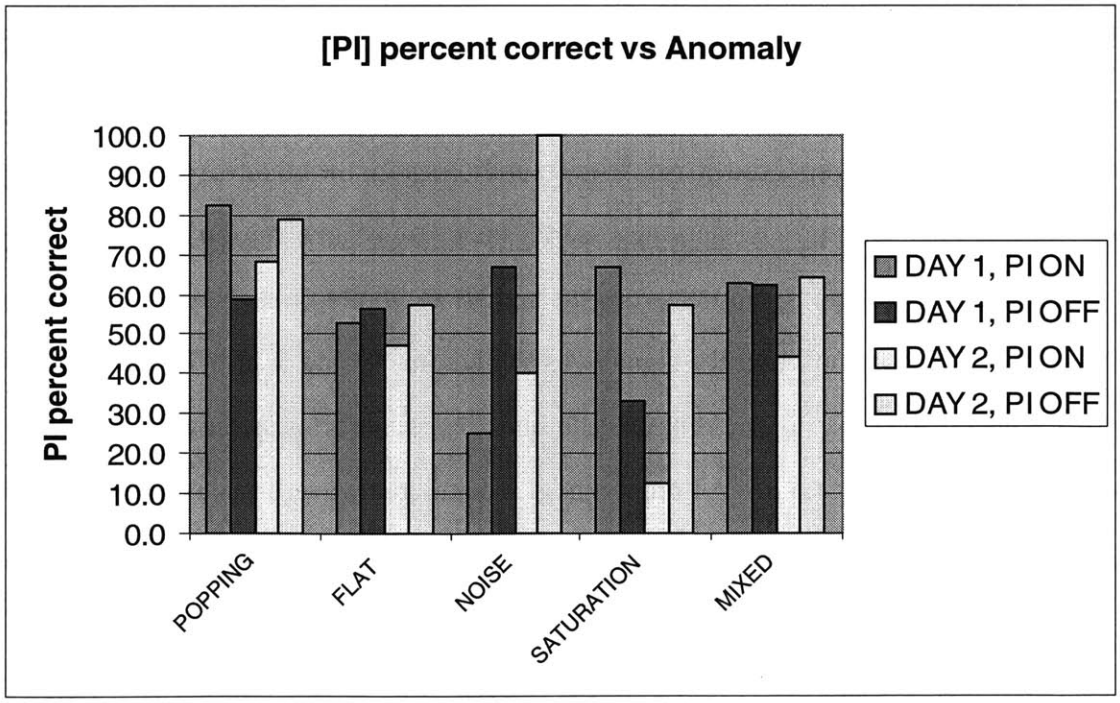


Figure 6.12 is a plot of the percentage correct based on $p(t_d)$ versus the anomaly type, and Figure 6.13 is a plot of the percent correct based on $p(t_d;[PI])$, for [PI] alone. On Day

1 subjects had a higher $p(t_d)$ for popping signal anomalies with [PI] than without it. On Day 2 those with [PI] had much lower $p(t_d)$ than without it. This is surprising, because subjects with [PI] performed worse than their controls on Day 2. A difference in the probability of [PI] correctly detecting faults within the detection time t_d explains this result. Since $p(t_d;[PI])$ is lower with [PI], it shows there was some confusion about the meaning of the indicator lights to Group 2 for popping signals. They were not getting as much reliable information from [PI] as they could have. Otherwise it would correspond to the higher [PI] probability of correct detection as those without [PI] on Day 2. This is further evidence that there is an asymmetric transfer effect between the two groups.

The probability of correct detection for subjects with and without [PI] was about the same for flat anomalies. Although there were not enough noise errors, the $p(t_d)$ was about the same for both days and both with and without [PI] settings for such anomalies. However, subjects had to overcome much confusion interpreting the indicator lights for noisy signals, since the $p(t_d;[PI])$ was very low with [PI]. Their high reliability with noise errors can be credited to their pre-experiment training on anomaly types.

On Day 1, Group 2 did worse with saturation anomalies than Group 1 because we emphasized that saturation was not necessarily the cause of an error during pre-experiment training. We taught them that since the signals could not be displayed on the screen, you could not be sure whether the signal was a good quality one or not. Therefore in some cases, subjects in Group 2 would ignore the saturated waveforms as they trained themselves to do since Day 1. Group 1 did not neglect saturated waveforms as much since they got used to interpreting them as the [PI] indicator lights turned red for some of the saturated signals. This effect is also mirrored in the data for the [PI] probability of correct detection, since Group 2 was unable to extract as much information from [PI] as Group 1 for a saturation signal.

There was little difference between days, and between groups for mixed anoma-

lies, but Group 2 with [PI] had slightly higher $p(t_d)$ than Group 1 without [PI]. This is in spite of the [PI] probability of a correct detection being lower with [PI] than without [PI] on Day 2. A mixed signal made up of two anomalies is too much for a trained observer to ignore for long, so subjects compensated for the lack of qualitative information in [PI]'s indicator lights with their own knowledge of the signals. Overall, Group 1 seemed to be better tuned to [PI] than Group 2, because $p(t_d; [PI])$ is higher with [PI] than without for Group 1.

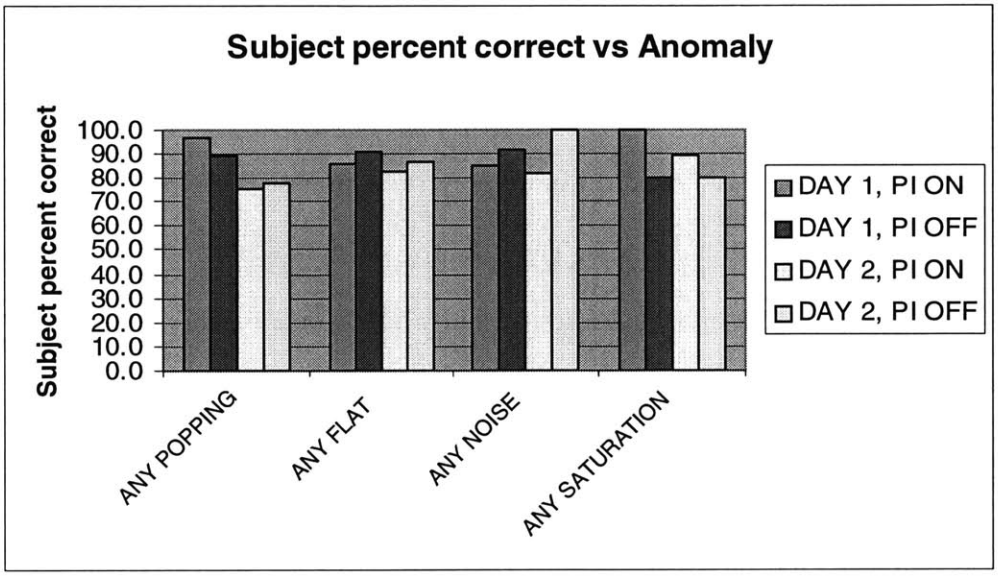


Figure 6.14: Subject percent correct for signals exhibiting particular anomalies

Figure 6.15: Percent correct for [PI] alone for signals exhibiting particular anomalies

Figure 6.14 shows the subjects' percent of correct responses for trials that exhibited at least one of the four main categories of anomaly. For instance, a waveform that was popping and was saturated counts in the "Any popping" and in the "Any saturation" category. Subjects had a higher $p(t_d)$ with [PI] than without for signals exhibiting any saturation – but this could be a consequence of the training given before the experiment. Otherwise, for signals exhibiting any noise or any flat, subjects with [PI] had a lower $p(t_d)$ than without [PI]. There is little difference in these anomaly categories except for the popping anomaly, in which subjects did worse on Day 2 than on Day 1. Figure 6.15 shows the

$p(t_d;[PI])$ for different anomaly types. $p(t_d; [PI])$ was a lot lower for noise anomalies without [PI] than for with the aid. Subjects were therefore very confused by the indicator lights when a noise anomaly was present. If a noise fault is combined with another anomaly, then the confusion is lessened. Subjects were not confused about popping or flat anomalies, as the $p(t_d;[PI])$ was higher with [PI] than without the aid. These observations are consistent with the test subjects' suggestions that [PI] should be a little more helpful with detecting poor quality signals.

6.7.1 [PI] indicator light Reliability: EEGs and EOGs

Waveforms displayed were of three forms of electrophysiological data; the EEG, the EOG, and the EMG. Since [PI] did not employ rules in assessing EMG signals, we will study the other two for effectiveness. The $p(t_d)$ for subjects was about the same for Day 1 between both groups, but on Day 2 Group 2 did a little worse than Group 1 despite having the [PI] available. On Day 1, subjects were better tuned to [PI] being on than those on Day 2. The $p(t_d, [PI])$ explains part of this difference, since it is no different with [PI] on than without it on Day 2. Subjects had a higher $p(t_d)$ for the EOG channels with [PI] for both days. This corresponds directly to them gaining information from [PI], since the $p(t_d;[PI])$ is higher with [PI] for both days. [PI] provided more useful information for the EOG to the

subject than the EEG. This may have to do with there being twice as many EEG channels as there are EOG channels.

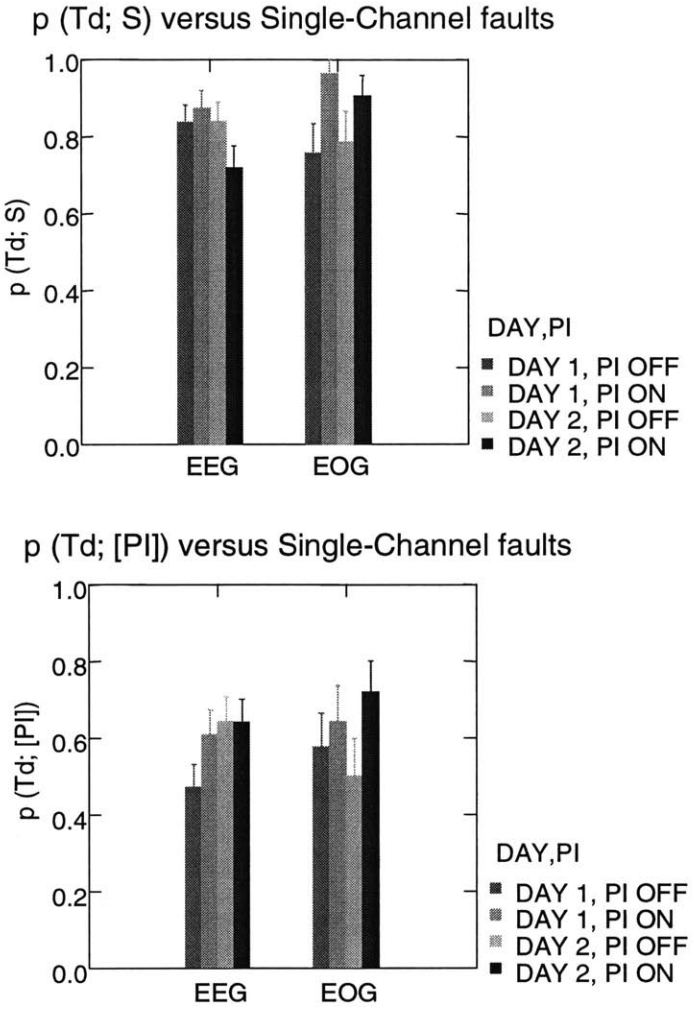


Figure 6.16: $p(t_d; S)$ and $p(t_d; [PI])$ for EEGs and EOGs

6.8 Subject correctness and [PI] Reliability Index

In their debriefing questionnaires, some subjects commented that [PI] gave false positives (i.e. displaying a red indicator light beside a signal which was not at fault) [5]. Some also reported misses, or a green indicator light displayed beside a faulty signal. For instance, when a signal went flat because an electrode had no hydrodot in place. The indicator light would be green, but as the waveform decayed into the bounds of a "poor signal", the indicator light would turn red. This is a defect in the [PI] system, since subjects who looked at the indicator lights at different points in time would get conflicting informa-

tion from [PI] on the same faulty waveform.

The data from the pilot study and the in-space recordings did not distinguish the benefit of [PI] to the subject when the indicator lights were very reliable, or when they were unreliable. We define t_{fire}, or the reliability index for the indicator lights, as the time that the indicator fires red for the faulty channel(s) divided by the sum of the time that the indicators go red for all channels. We only consider the time between the onset of a trial

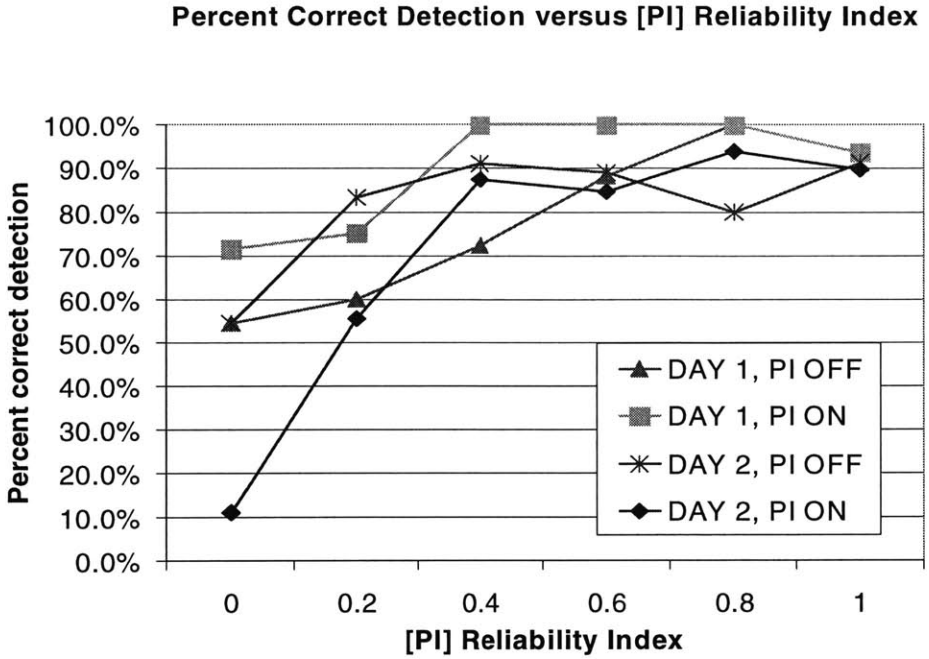
$$t_{fire} = \frac{t_{ch,\,correct}}{t_{ch,\,correct} + \sum_{i} t_{ch,\,incorrect}}$$

and the subject's first detection.

t_{fire}, the reliability index of the [PI] indicator lights (6.4)

Only single-channel faults were used in this analysis. The "[PI] off" setting is the natural response to anomalies as opposed to the "forced response" with [PI]. Remember that even though subjects would not see the indicator lights without [PI], it would still be recorded in the log file. As t_{fire} increases, so does the probability of correct detection, until above 0.4, where performance plateaus for all settings. This is good news, because it means that [PI] has the flexibility to not have to be 100% reliable for the astronaut to ben-

efit from the information it gives. Subjects with [PI] on Day 1 were more correct than



those without it on Day 1 for detection tasks in which [PI] was very unreliable. However, subjects with [PI] help on Day 2, who already had one day's experience with [PI], were hindered by [PI] because their correct detection rate was lower than subjects without [PI] on Day 2 for low reliabilities. These strong trends, although not statistically significant ($p=0.125$ via Sign test) show that it may take a long time for subjects to be trained with [PI] if they have had experience without [PI] help already. This along with the analysis of discriminability d' shows that subjects with one day's experience in detecting signals without the decision aid had a harder time training themselves with the [PI] signal quality indicators.

The aggregate measure of performance $p(t_d)$ is an average probability of a correct detection on the first try amongst all subjects studied. The subjects in Group 1 have a somewhat lower variability in $p(t_d; S)$ than those in Group 2. Both groups had similar variance between subjects for multi-channel fault $p(t_d; S)$, and there is not enough data to meaningfully assess null faults, but for single-channel faults, Group 2 has more variability for the [PI] settings for each subject than Group 1.

The larger variance in Group 2 can be explained by a trial-by-trial analysis. Figure 6.17 is a plot of $p(t_d; S)$ averaged over all subjects for each trial for single-channel faults. Group 1 exhibits a learning curve whereby $p(t_d; S)$ increases on Day 1, almost monotonically, and plateaus with the exposure to more trials of single-channel faults. Then on Day 2, Group 1's $p(t_d)$ plateaus, but at a lower value without [PI] than with it. On the other hand, Group 2's $p(t_d)$ is erratic on Day 1, and remains erratic on Day 2. [PI] was able to influence Group 1 positively, but most likely influenced Group 2 negatively, since subjects were less reliable in making a correct first detection on Day 2, despite having one day's experience detecting anomalies.

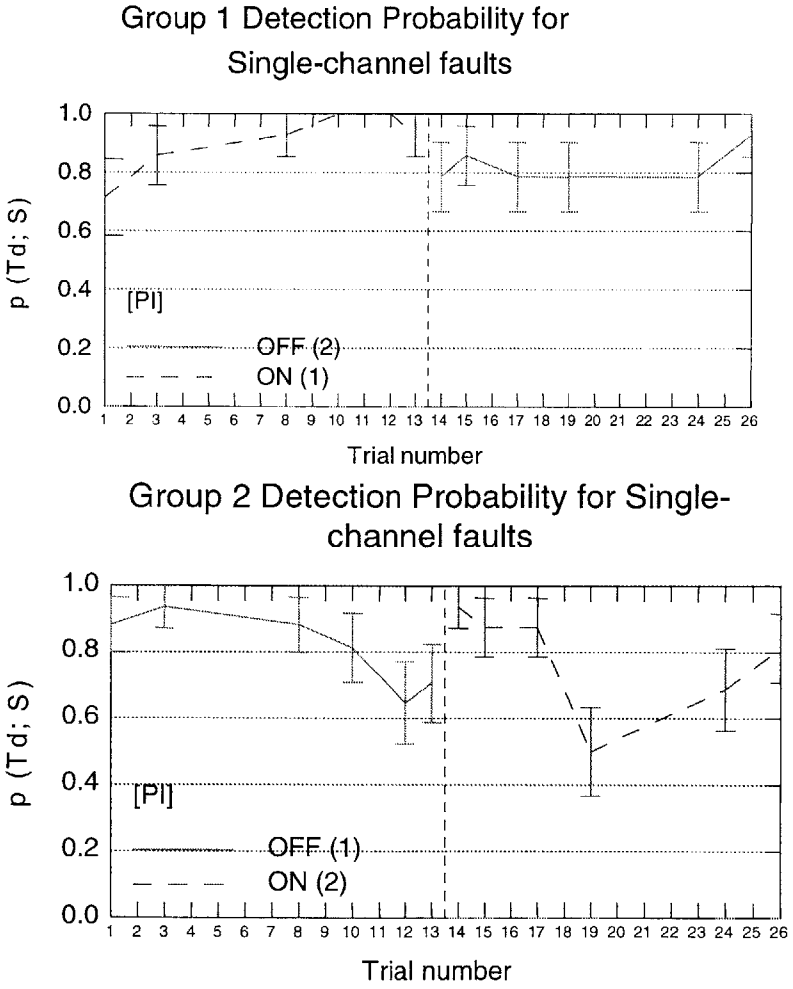


Figure 6.17: $p(t_d; S)$ versus trial number for each group. The day is indicated in brackets

On Day 1, Group 1 compensated for [PI]'s interpretations since their $p(t_d;[PI])$ showed an increase with the number of trials on Day 1 compared to the variances on Day 2. Group 2 did not exhibit compensation for the indicator lights until later in the series on Day 2, compared to the variances on Day 1. For the trials in each group where subjects showed compensation for [PI], i.e. increasing $p(t_d;[PI])$ with the number of trials seen, there is a subsequent increase in the subjects' $p(t_d)$.

6.9 Troubleshooting: the tradeoff between planning and execution

The total down time in fault management is attributable to an astronaut doing a number of tasks, such as probing the failed system, reading flowcharts, monitoring sensor data, etc. – essentially gathering information about the system. At the same time, the astronaut is planning; the next component to probe, the alternatives should their hypothesis be incorrect - interpreting the information provided from the system. Therefore, the astronaut is making a trade-off between planning a strategy, and executing it.

t_{ts} is a measurement of how an astronaut can perform both of these skills. It is a combination of planning tasks and execution. By regressing t_{ts} for each trial against the number of questions asked by the subject, we are left with how much time subjects spent planning the next course of action. A linear regression is justified since it takes approximately a constant time for the subject to ask a question, and a constant time for the assistant to answer the question. t_{ts} can also be regressed against the number of calibrations, since it takes the same time for a sleep subject to make eye movements, or clench their jaw, or relax. The subject's first, second, and subsequent attempts are summed into t_{ts}.

If subjects detected the wrong channel the first time, they tried to detect the failure after they realized that they were wrong in their initial assessment. A time is associated with each detection attempt which may be dependent on the number of previous failed attempts a subject made in finding a given fault. The thinking time taken to find a fault on the second attempt may decrease from the thinking time for the first attempt, since sub-

jects have a smaller diagnostic search space to look through after a failed search. It may also increase if the subject is responding more as a reflex to an anomaly and must take time to reset himself after the initial failure. Therefore, we regress t_{ts} against the number of diagnostic "tries" as a categorical, or dummy variable.

We regress t_{ts} stepwise on variables:

$$t_{ts} = \beta1*TRY1 + \beta1p*PI*TRY1 + \beta2*TRY2 + \beta2p*PI*TRY2 + \beta3*TRY3 + \beta3p*PI*TRY3 + \beta q*NUMQ + \beta qp*PI*NUMQ + \beta c*NUMCAL + \beta cp*PI*NUMCAL \quad (6.5)$$

where

NUMQ is the number of questions asked,
TRY1,TRY2,TRY3 are dummy variables (either 1,or 0) representing the number of tries (for instance, for two tries, TRY1=0, TRY2=1, TRY3=0),
PI is the presence of [PI] help, with 1 indicating [PI] was on, and 0 indicating [PI] was off, and
NUMCAL is the number of calibrations made

	Overall t_d+t_{ts} for DAY 1					
	R squared: 0.776		Standard error of estimate: 27.413			
Effect	Coefficient	Std Error	Std Coef	Tolerance	T	P (2 Tail)
NUMQ	6.36	1.118	0.301	0.238	5.69	**0.0005**
TRY1	35.779	3.055	0.586	0.266	11.71	**0.0005**
TRY2	59.915	6.015	0.319	0.649	9.962	**0.0005**
TRY3	37.425	14.308	0.093	0.524	2.616	**0.0090**
PI*TRY1	-7.107	3.16	-0.082	0.498	-2.249	**0.0250**
PI*TRY3	-76.344	23.163	-0.102	0.7	-3.296	**0.0010**

Table 6.2: t_d+t_{ts} regression analysis for Day 1

	Overall t_d+t_{ts} for DAY 2	
	R squared: 0.834	Standard error of estimate: 21.399

Table 6.3: t_d+t_{ts} regression results for Day 2

Effect	Coefficient	Std Error	Std Coef	Tolerance	T	P (2 Tail)
NUMQ	13.074	1.048	0.68	0.162	12.477	**0.0005**
TRY1	19.391	2.026	0.351	0.358	9.57	**0.0005**
TRY2	45.26	8.308	0.275	0.19	5.448	**0.0005**
PI*NUMQ	-6.332	1.03	-0.256	0.277	-6.146	**0.0005**
PI*TRY2	-16.786	9.277	-0.086	0.213	-1.809	0.0710
PI*TRY3	53.113	12.224	0.109	0.766	4.345	**0.0005**
PI*NUM-CAL	13.775	2.471	0.127	0.926	5.574	**0.0005**

Table 6.3: t_d+t_{ts} regression results for Day 2

We combined t_d+t_{ts} into $T_{isolation}$, measured for each fault type. We used $T_{isolation}$ for subjects who did not timeout (i.e. spend more than 180 seconds to isolate the fault), and who did not omit to click the event marker to record the onset of each trial, and faults where no bonus failures occurred.

On Day 1, the model R^2 is 0.776. The β_2 coefficient is higher than β_1. Subjects spent more time on Day 1 detecting faults on their second try than they did on the first try, confirming our hypothesis. The coefficients β_{1p}, β_{2p}, and β_{3p} can be subtracted from their counterparts β_1, β_2, and β_3 to derive the benefit attributed to [PI] for each case where subjects took 1, 2, and 3 tries respectively. With [PI], the down time decreased by about 7 seconds for those who took one attempt, but no significant impact on the second try. On the third try, some subjects found an enormous benefit with [PI], although statistics on the few subjects who took three attempts leads us to be cautious with this value.

On Day 2, the model R^2 is 0.834. Subjects trained themselves to plan their steps while asking questions, as indicated by the β_q coefficient increasing from Day 1 to Day 2. The time taken to ask each question decreased by 50% if the subject had [PI], as seen by the β_{qp} coefficient. In fact it took about the same time to ask a question on Day 1 as it did to ask a question on Day 2 with [PI]. This shows how [PI] is a "regulator" of performance

across each day, since there is no "overhead" thinking time in asking each question. [PI] does not impact any subject's first attempt on Day 2, since the β_{1p} coefficient did not come out significant. However, subjects with [PI] help spent 33% less time managing their faults than those without it if they took two attempts. This again shows the regulatory effects of [PI] by keeping the down time within reasonable bounds. Subjects needed [PI] more on their first tries on Day 1, and their second tries on Day 2. Nobody without [PI] took 3 attempts to complete any one trial, so β_{3p} is not an appreciable effect. Second order models (quadratic) were abandoned because negative curvatures were computed, which is not physically meaningful.

The regression results shows that even though [PI] had a significant impact on decreasing troubleshooting time for all subjects, subjects used [PI] differently on each day. It was helpful only for the subjects' first attempt on Day 1, but then was helpful for reducing the execution time, and (although not significant) thinking time for their second attempt on Day 2. The impact of [PI] was different on each day - first helping with planning time, then helping with executing a plan. [PI] helps to regulate fault management performance.

6.10 Calibrations

During training, test subjects were taught how to perform calibrations, as all scientists do with their equipment, to determine if a signal was operating properly. They were taught calibration with eye movements for EOGs and jaw movements for EMGs. They were also told that instructing the sleep subject to relax, which was helpful for EEG signals, or move

their head could also help detect and diagnose faults.

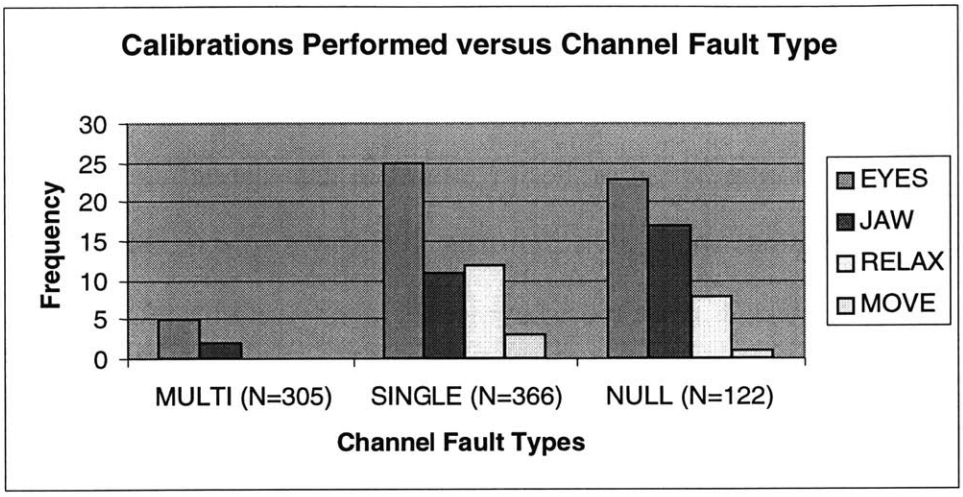


Figure 6.18: Calibrations performed for each fault type. Number of trials indicated in brackets.

Figure 6.18 shows the frequency of calibrations made for each channel fault type. By far the most frequent use of calibrations occurred with null faults, because subjects needed more information than just the static waveforms alone to determine if the system was okay. EMGs were calibrated (i.e. jaw movements) quite a few times even though no faults were created on them. Since the number of calibrations performed for eyes and jaw were comparable for null faults, one can conclude that calibrations are still a necessary task for ensuring signal quality even with [PI].

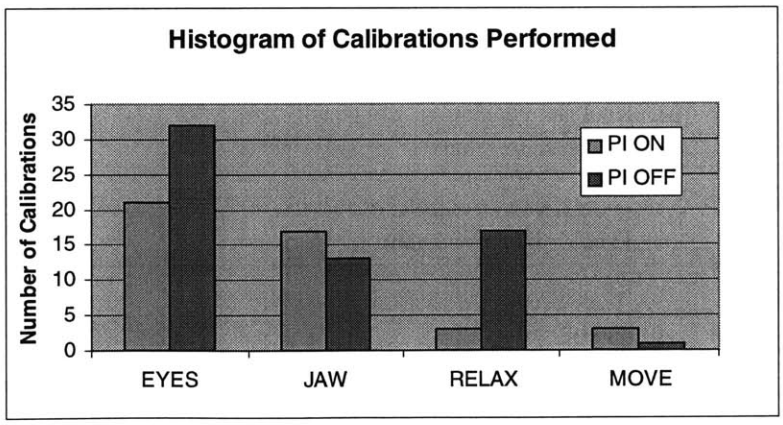


Figure 6.19: Calibrations performed with and without [PI]

Figure 6.19 shows the frequency of each calibration command both with and without [PI] available. Having [PI] reduced the number of calibrations performed ($p < 0.033$, Kruskal-Wallis $\chi 2=4.536$, df=1). Subjects said that they used the indicator lights as a "secondary source" of information to the waveforms in their debriefing questionnaires. The number of calibrations went down from Day 1 to Day 2 ($p < 0.0005$, K-W $\chi 2=14.705$, df=1) because subjects became more familiar with the waveforms after one day of experience. Group 2 performed more calibrations than Group 1 because they trained themselves to extract as much information as possible on Day 1 and continued this strategy to Day 2 ($p < 0.0005$, K-W $\chi 2=24.660$, df=1).

6.11 Probing strategy

Once a symptom is detected, subjects must identify a fault in order to repair it. The subject can use his training to deduce a fault from symptoms. Subjects can also ask questions to the assistant to learn if a particular fault exists. Each trial was categorized into a few groups which describe the subjects' trajectory for each trial. These categories are called "subject actions," and a breakdown of the various common actions for each trajectory are

Subject Action	Trajectory	# of Errors	% of Errors
STATE & CHANNEL	Correct state detection of state and channel on the first attempt	635	78.8%
STATE, <<, STATE	Correct state detection, went back, and chose the same state	12	1.5%
state, <<, STATE	Incorrect state detection, went back, and chose the correct state	43	5.4%
STATE & channel	Correct state detection, but on an incorrect channel	27	3.3%
state & CHANNEL	Incorrect state, but on the correct channel (mostly multiple channel faults)	56	6.9%

Table 6.4: Breakdown of subjects' fault management trajectory. Capitalization indicates correctness of assessments

STATE & CHANNEL, <<, T/O	Detected correct state and channel, went back, and timed out	20	2.5%
Other	Other trajectory	14	1.6%

Table 6.4: Breakdown of subjects' fault management trajectory. Capitalization indicates correctness of assessments

shown in Table 6.4

Figure 6.20 shows the frequency of each action as a percentage normalized by the number of trials under each condition of day and aid. The actions in capitals indicates that an action was the correct one. Group 2 seemed to be more sporadic than Group 1, since they had more second-guessing (STATE, <<, STATE), and chose the correct state but wrong channel more times (STATE & channel). Group 2's strategy involved asking questions rather than analyzing the waveforms. They would sometimes select a channel without fully interpreting the waveforms just to guess at some possible diagnoses. The percentage of subjects who corrected their initially incorrect state assessment (state, <<, STATE) decreased from Day 1 to Day 2, because subjects made fewer mistakes of this

form for the null faults in this interval.

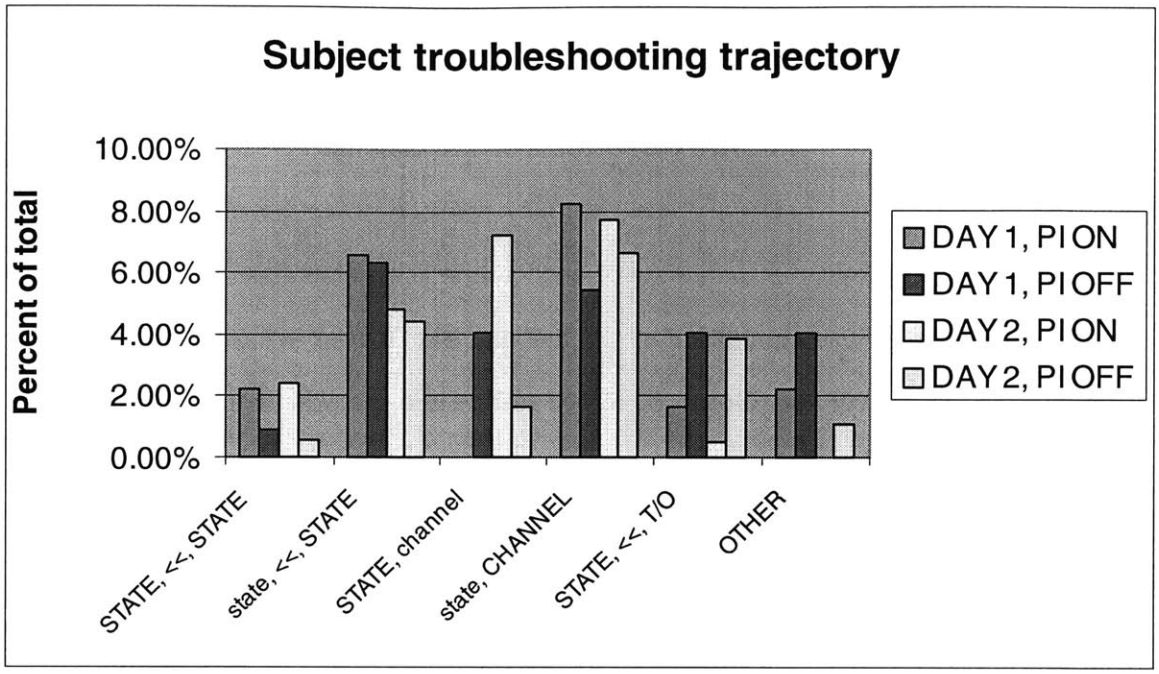


Figure 6.20: Breakdown of subjects' troubleshooting trajectory: percentages are out of total trials for each (DAY, PI) setting

But there are distinct patterns associated with [PI]. Subjects with [PI] were more susceptible to erring in their initial state assessments. Yet they tended to second-guess themselves and time out more than those without [PI], because the indicator lights reminded them what they were troubleshooting. The most noticeable pattern though is the number of times subjects could not diagnose the fault within the allotted 180 seconds, as seen in Table 6.5 ($p=0.007$, via Pearson $\chi^2 = 7.37$, $df=1$). Two of the 9 time outs with [PI] occurred for the ground electrode failure, and 12 out of 25 time-outs without [PI]. This also points to the effect of [PI] to regulate subjects, making their performance more consistent regardless of their experience level.

	Day 1	Day 2
[PI] ON	5	4
[PI] OFF	15	10

Table 6.5: Breakdown of undiagnosed faults (time-outs)

The strategies that Group 2 used in asking questions were very much different than Group 1's strategies. Group 2 had a greater tendency to deviate from the checklists and flowcharts than Group 1 (17.5% of trials versus 13.5%), although not statistically significant. Deviations are things such as not asking a troubleshooting question in the order specified, asking a question more than once in a trial, or not finishing a series of questions on a particular state before moving to a new one. Table 6.6 divides the deviations into several categories, described below:

- added steps are questions which are not part of the sequence in which the subject assessed the system state to be, such as asking if the O2 Hydrodot is in place if the subject selected an EOG signal to be poor quality,
- capture errors were made on the EOG faults, as subjects would ask the same sequence of questions for EEG troubleshooting, even though EOG is slightly different.
- out of turn steps are questions which were not asked in the sequence described by [PI]'s messages or by the NASA guideline,
- repeated questions are questions which were asked more than once in a trial
- subjects did not complete the sequence of questions on the channel
- skipped steps are questions omitted from the recommended sequence

	Fault Type			Group		[PI] help	
Category	Null	Single	Multi	1	2	ON	OFF
added step(s)	3	15	9	7	20	13	14
capture error	0	7	0	2	5	2	5
out of turn step(s)	2	8	4	10	4	9	5
repeated questions	0	2	3	1	4	2	3
did not complete sequence	3	3	4	5	5	2	8

Table 6.6: Breakdown of troubleshooting deviations (a) by Fault Type, (b) by Group, and (c) by [PI] setting

skipped step(s)	0	5	51	20	36	30	26
other	1	5	6	8	4	2	10
Total	9	45	77	53	78	60	71

Table 6.6: Breakdown of troubleshooting deviations (a) by Fault Type, (b) by Group, and (c) by [PI] setting

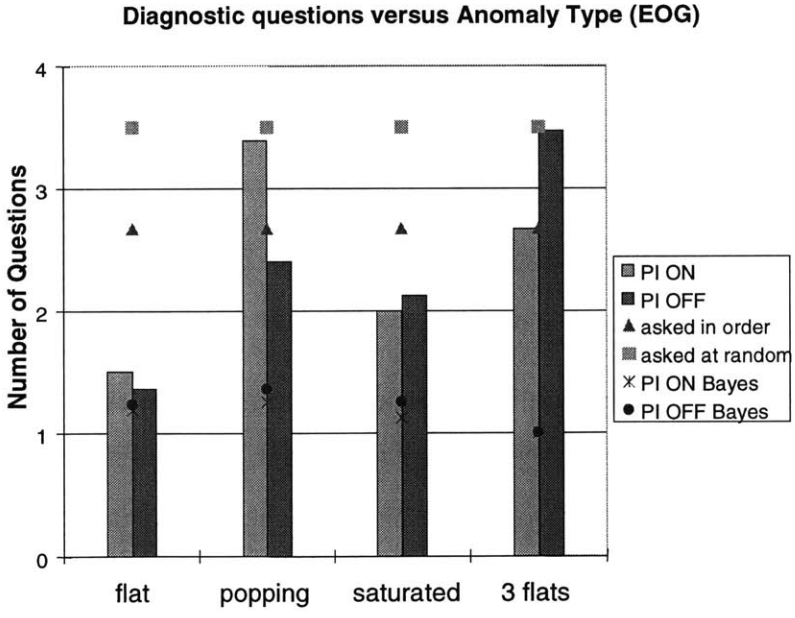
About 83% of the time (662 out of 793 trials), subjects followed the list of questions provided. However, many subjects did skip steps for the reference electrode failure and the ground electrode failure (37 and 9 out of the 51 path deviations for multi-channel failures, respectively). Many asked if one of the reference electrodes was not inserted on the multi-channel faults (34 out of 77 such trials), a diagnosis which could not explain all the channels having poor quality. Distinct differences appear between the two groups. Group 2 added steps to the sequence more times and skipped more steps than Group 1, a consequence of their tendency to ask questions more than analyze the waveforms. But Group 1 asked more steps out of turn than Group 2, since they were used to a checklist style of displaying possible faults. With checklists, it is easier to get back on track if you do ask steps out of turn, so this was a more favorable strategy than following the list step by step.

6.12 Fault diagnosis using qualitative knowledge of anomalies

The motivation to determine what strategies subjects used to diagnose faults came from previous studies of fault management [16]. In Rouse's study, the diagnostic search strategy of the subjects trained with the computer aid was found to mimic that of the computer. Extracting each group's strategy from the subjects' trajectory is difficult because some of the faults were not reliably created. The noise failure was very difficult to create because the makeup did not usually create the high impedance required. The ground electrode error, for which the system state was "All EP signals poor quality", was also troublesome to create because signals would sometimes not behave abnormally if the ground electrode was removed, since it is an auxiliary reference electrode. Therefore, applying a qualitative model to diagnose a fault directly from the symptoms is difficult.

The average number of questions asked by all subjects is plotted in Figure for EOG faults and Figure 6.22 for EEG faults against the observable anomaly, as characterized by an MIT sleep expert. The most likely fault to occur was the one with highest probability of p (fault | anomaly) calculated using Bayes' rule, and is shown on the graph as a lower bounds on the number of questions asked. A discussion of Bayesian inference and decision theory can be found in Sheridan and Ferrell [18]. The upper bounds shown were calculated based on the average number of questions asked if subjects asked questions in the order presented by the NASA guideline, or if they employed a random questioning strategy. We only used data for which subjects assessed the correct state and channel for the given fault.

Subjects were very good at detecting flat signal anomalies on the EOGs, since the difference between their response and the best Bayesian estimate were similar. Subjects asked fewer questions with [PI] only for the 3 flat signals, i.e. the reference electrode failure. [PI] did not help with the remaining anomaly types. In fact it was detrimental for diagnosis of the popping failure. This is a peculiar result, since [PI] was shown to reduce



troubleshooting time, and popping was a very common anomaly. Perhaps subjects could not use the "popping" diagnostic as effectively as a sign of a particular fault.

Figure 6.21: Average number of questions versus anomaly type for EOGs as asking questions at random.

On the EEG channels, subjects performed well for popping errors with respect to the Bayesian estimates, but [PI] provided no benefit to detecting this anomaly with respect to fault diagnosis. [PI] provided no help for detecting a flat signal on the EOGs, on which subjects without [PI] did well. Popping and noise, a mixed anomaly, was diagnosed well by all subjects. Actually, subjects without [PI] asked fewer questions on average than Bayesian estimates. This is a peculiar result, but it is because there is a slight variation in the posterior probabilities which makes the estimated number of questions so different from with to without [PI]. Only for a 3 flat signal anomaly did the number of questions asked decrease for subjects with [PI].

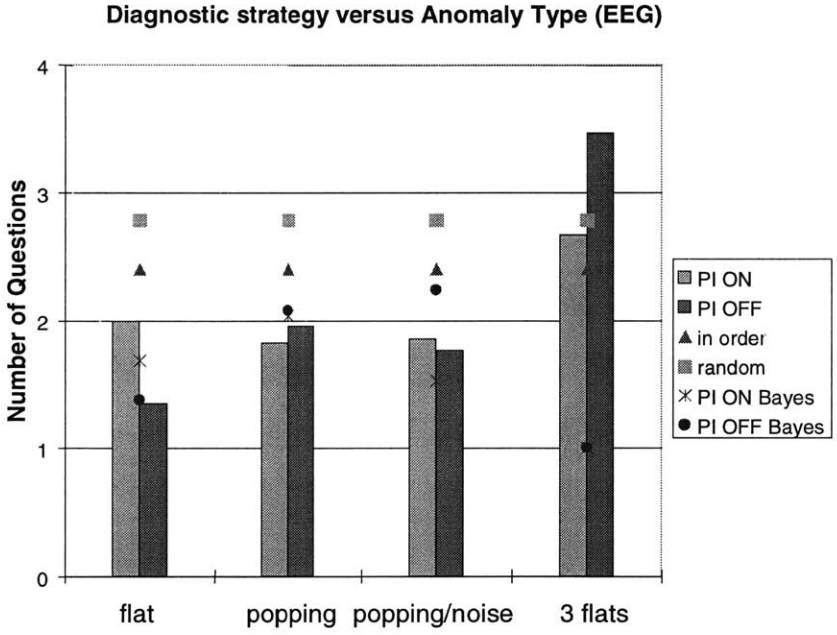


Figure 6.22: Average number of questions versus anomaly type for EEGs

[PI] did not seem help the astronaut in predicting faults using the observable anomalies that they were trained to recognize. This was expected, since there was no intelligent reasoning for diagnosing faults other than the priority ordering of the troubleshooting questions. In some way the analysis was dependent on the errors chosen for this study; subjects could learn during their trials errors that tended to point to particular diagnoses.

But to the credit of the [PI] development team, the questions were arranged so that subjects without any qualitative knowledge could ask the fewest number of questions to find a fault.

6.13 Discussion

Some benefits of using an on-board expert system for fault management are seen in this study; subjects discriminated signals (d') better with [PI] than without it for both single and multi-channel faults on Day 1. Subjects also had a higher probability of detecting faults on the first try with [PI] than without it. However, since each group reacted differently to the removal/inclusion of [PI] help, it is difficult to make a distinct conclusion for [PI]'s impact on Day 2. This effect was evident for d', and for the subject response to different reliability indices for [PI].

There were many distinctions made between the two groups in this study. Although the initial assumption was that there would be no group differences seen in the experiment, some subtle differences in style point to these groups as being different. In fact, subjects in Group 2 were shown to have significantly lower troubleshooting times than Group 1 [5]. Several distinctions between the two groups were discovered, both in their detection ability and their troubleshooting skills. Group 2 detected the absence of a fault on Day 2 less effectively than did Group 1, even with [PI] help. Group 2 also performed more calibrations than Group 1 ($p < 0.0005$, Kruskal-Wallis $\chi 2 = 24.66$, df=1). Group 1 waited longer before making an assessment than Group 2 (mean t_d of 36.5 seconds compared to 29.5), although not statistically significant. Group 2 subjects kept probing the system for more information with calibrations, and made more incorrect hypotheses than Group 1. Group 1 learned to use waveform information more wisely and were more patient than Group 2.

There is a parallel here between two fault management styles: the technician and the engineer [21]. A technician makes many observations, can find information regardless of the fault, and requires little intrinsic system knowledge. An engineer, on the other hand, makes few observations, and contemplates longer about the problem. There is no reason to decide which is better or worse, because the two approaches complement each other. In our experiment, we perceive Group 2 to behave more like a technician, and Group 1 to behave more like engineers in fault management. This may have developed because Group 2 was initially without a source of "knowledge" in real-time when faults appeared, and tended to probe for failures everywhere to gain more information about the system. Group 1 had [PI], a real-time signal anomaly interpreter, which taught subjects more about the system so that they could handle faults better on their own.

6.14 Comparison to previous studies

Data from this analysis supports the idea that [PI] regulates diagnostic performance of an astronaut. First, the [PI] reduced time spent on the first attempt on Day 1, and on the second attempts on Day 2. Second, the number of undetected faults was consistently less with [PI] than without it. Similar results were discovered by Jared Martin in his project, whereby [PI] seemed to nullify the differences encountered between two different stimulus files. Having [PI] regulate performance for potentially routine tasks such as fault management is a benefit.

In the pilot study, the definition of detection time was the difference between the time that a subject detected a signal anomaly and the time [PI] detected it. Only data which was not a bonus fault, and not a timeout, and for which [PI] detected an anomaly was used in this analysis. An analysis of variance (ANOVA) on the difference t_d-$t_{d,[PI]}$ showed that no significant effects existed.

Chapter 7

Conclusions

7.1 Summary

This experiment was the second phase of a ground study to assess the efficacy of a real-time expert system on a space life science experiment. Results show that use of the expert system increased the probability of correct diagnosis of multi-channel faults, and also correct detection of single-channel faults on Day 1. Further, it helped to regulate subject performance by reducing the thinking time on Day 1, and the execution time on Day 2 when subjects had more experience with the system. Its regulating ability also reduced the number of undetected anomalies (time outs).

False alarms on other channels that would potentially mislead subjects were shown to have little impact on reducing performance for a t_{fire} larger than 0.4. Group 2 on Day 2 was more likely to correctly detect a fault if [PI] fired a red indicator light with one day experience than Group 1 on the same day.

7.2 Suggestions for future experiments

The experiment was developed and tested within a short time span amid other constraints of time. In future experiments of this type, a different approach could be taken. In the design of this experiment, three factors were assessed with two unique treatments such that any two factors of training (i.e. Day), aid setting (with or without [PI]), and transfer (i.e. Group) will be confounded. A four-group design with two additional groups that either had [PI] on both days and no [PI] on both days would complete the design, would be more appropriate. The repeated measures approach, although beneficial for physiological experiments where large differences exist between subjects (e.g. heart rate, blood pressure) may not have been as important as originally thought in this experiment. Previous

experiments of these types of tasks found that training was a dominant effect compared to the transfer effect. For this experiment, it is indeterminate whether training or transfer was dominant, since the results are so much different from those expected and no explicit control exists for the Day effect.

We had a lot of "experimental noise" associated with our setup, since we created real faults on sleep subjects for our stimuli instead of a predefined file of waveforms. Some failures were not created as reliably as we hoped, along with occasional equipment failures. However, subjects actually liked interacting with both the assistant and the sleep subject in real time compared to pre-recorded stimulus files. Perhaps they would not be as enthusiastic had the experiment been simulated. Given the tradeoffs between realism, reproducibility, and time investment, a pre-recorded file may have been more appropriate.

7.3 Suggested Improvements for [PI]-Sleep

Due to the time constraints from the proposal [PI] could not improved to its fullest extent, especially the knowledge rulebase, because new rules would need time to be developed, and validated with the sleep experts at Brigham and Women's Hospital. The original [PI]-Sleep did not use spectral analysis rules because it was computationally expensive, but with faster computers this could be a useful characterization of expert signal detection skill. Other analysis techniques, such as Independent Component Analysis [9], show promise for detecting and removing signal artifacts. The rulebase should make use of all system information, such as conjunctive rules that can "AND" channels to test if the reference electrode is loose. Further, the signal quality can be displayed to the user, and the troubleshooting questions can be rearranged or modified based on the signal quality observed. So identifying a popping signal, and rearranging the diagnostic question sequence to reflect the signal quality will be more efficient. Developing rules for the

EMGs would also be important, because even though we did not plan to create one, problematic EMG signals appeared in 8 out of the 61 experiments run.

A calibration feature was to be added to [PI] for phase III of the study, but was not developed for use in this study. However, it would be a great benefit for subjects to use a calibration for fault detection and diagnosis. One approach could be to let [PI] cue the astronaut to ask the sleep subject to make eye movements, for example, and [PI] will analyze the deflections to verify the correct polarity, signal amplitude and signal quality. Using the calibration information, [PI] can better assess the signal quality and present its analysis to the astronaut.

The graphical user interface can be improved by adding a picture of a sleepnet with indicator lights located at the electrode sites to indicate the signal quality of the corresponding electrode. This feature was being developed initially for the STS-90 mission, but was abandoned. Some of the subjects preferred the NASA flowcharts to the [PI] diagnostic window, so an option of displaying either the NASA flowchart or the checklist style window in [PI] would make [PI] more flexible for the astronaut. The changes suggested here for [PI]-Sleep can be implemented for the ISS version of the sleep experiment should this experiment be carried out again.

7.4 The future of Artificial Intelligence in space

The [PI] interface could alleviate some of the human-system interface problems astronauts will encounter on the ISS, as described in the NASA critical path document [12]. Although highly intelligent, well-educated and versatile, most astronauts will inevitably face the need to execute an experiment outside their field of expertise. [PI] can offer the right knowledge for astronauts to use: diagnostic aids, an experimental planner/scheduler, and an interesting data filter. The ability of the human-autonomous agent team to carry out

expert decision-making in an isolated, confined environment will ensure that both scientific and mission objectives are reached.

Autonomous systems will be implemented in several applications in the ISS operation [3]. For instance, Node 3, to be launched in 2002, will be a connector for several of the U.S. modules. Vigilance monitoring for each of the 8 subsystems of the life support system will be prohibitively time consuming for station and ground crew. 3T, an autonomous software package developed at NASA, will run this life support system. The Remote Manipulator System Assistant (RMSA) will serve to automate the procedures relating to the shuttle's remote manipulator system. There is an equivalent autonomous system for the Space Station Remote Manipulator System (SSRMS) being designed as well. Moreover, the AerCAM, a soccer-ball-shaped free-flying assistant, is also being implemented in station operations. It is designed to inspect the modules for suspicious leaks or faults, and can be inserted in locations that may be risky for the astronaut. Autonomous systems will play an important role in the daily operation of the station.

To ensure the success and effectiveness of a mission, crew members must maintain a high level of cognitive performance and vigilance while operating and monitoring sophisticated instrumentation. Astronauts, however, commonly experience stress such as high workload, isolation, and sleep disruption during space flight. Moreover, astronauts aboard the International Space Station (ISS) will nominally have three- to six-month tours of duty. Because it is important for astronauts to maintain high levels of performance throughout long-duration space flight missions, there is a need to develop effective human-machine systems that can overcome these detriments. [PI] offers a promising way of addressing these problems.

This push for autonomy in complicated space systems may inspire the same autonomous support for the crew while running experiments on themselves and each other in

BIOPLEX. BIOPLEX provides a timely opportunity to develop and test software for interface between crew and autonomous agents in a stand-alone, closed habitat. BIOPLEX should seek to supply tools to enable the crew to maintain and repair the life support systems in a real lunar or Martian habitat. One of these tools should be the implementation of autonomous systems in the everyday operations of such a habitat, including conducting scientific experiments. [PI] was the seminal work for an investigation into implementing autonomous agents for scientific experiments. BIOPLEX provides an appropriate test bed for proving this autonomous agent technology.

7.5 [PI] for home sleep monitoring

The appeal of clinical sleep monitoring in home is growing. First, patients tend to sleep better in their own beds than in a hospital. Second, home sleep monitoring costs far less than monitoring done in a laboratory. With current systems of home sleep monitoring, technicians make house calls to instrument the patient with electrodes and setup the equipment. The sleep doctor can then monitor the patient's sleep pattern remotely, by downloading data from the home recording device. This system is generally reliable for home sleep monitoring, and there is tremendous interest in this from the private sector in terms of home health care.

However, there are problems with this scenario that the average patient or caregiver is not equipped to handle. Sometimes electrodes fall off after the technician leaves or during the night. As a result, data is lost or is of poor quality. [PI] as the home sleep monitoring software would detect anomalous signals and suggest ways the patient or caregiver might fix the problem. [PI]'s benefit can be extended from helping untrained astronauts to helping untrained sleep patients or caregivers fix problems with instrumentation. [PI] could be a cost-effective way of improving the reliability of the home sleep monitoring system.

Chapter 8

References

1. Callini, G. "Assessment of an Expert System for Space Life Sciences: a Preliminary Ground-Based Evaluation of PI-in-a-Box for the Neurolab Sleep and Respiration Experiment." Unpublished Master's Thesis, Massachusetts Institute of Technology, September 1998.

2. Callini, G., Essig, S.M., Heher, D., & Young, L.R. "Effectiveness of an expert system for astronaut assistance on a sleep experiment." Aviation Space and Environmental. Medicine, 71 (9): 1-10.

3. Chatterjee, Samprit (2000). "Regression analysis by example" 3rd ed.

4. Dorais, Gregory A., Bonasso, R. Peter, Kortenkamp, David, Pell Barney, and Schreckenghost, Debra. "Adjustable Autonomy for Human-Centered Autonomous Systems on Mars" In Proceedings of the First International Conference of the Mars Society, Aug/98.

5. Delaney, M. "Ground-Based Study of an Expert System for Human Assistance on the STS-95 Sleep and Respiration Experiment." Unpublished Master's Thesis, Massachusetts Institute of Technology, December 2000.

6. van Eekhout, J.M, and Rouse, W.B., "Human Errors in Detection, Diagnosis, and Compensation for Failures in the Engine Control Room of a Supertanker" IEEE Transactions on Systems, Man, and Cybernetics SMC-11 (12), December 1981

7. Franier RJ, Groleau N, Hazelton LR et al. "PI-in-a-Box: a Knowledge-based System for Space Science Experimentation." AI Magazine 1994; 15(1):39-5.

8. Jones, P.M., & Mitchell, C.M. (1995). "Human-computer cooperative problem solving: Theory, design, and evaluation of an intelligent associate system." IEEE Transactions on Systems, Man and Cybernetics, 25 (7).

9. Jung, T.P.; Makeig, S.; Humphries, C; Lee, T.W.; McKeown, M.J.; Iragui, V.; Sejnowski, T.J. (2000) "Removing electroencephalographic artifacts by blind source separation." Psychophysiology, 37 pp. 1-16.

10. Lafuse, S.A. (1991). "Development of an expert system for analysis of shuttle atmospheric revitalization and pressure control subsystem anomalies." Proceedings 21st International Conference on Environmental Systems.

11. Martin, J., "Ground based study and evaluation of Principal Investigator-in-a-box: Annual Project Report, Year 2." Unpublished report for the NSBRI, June 1999

12. NASA Critical Path Roadmap http://criticalpath.nasa.gov

13. Rasmussen, J. "New Technology and Human Error" pp. 53-61. 1987. Wiley.

14. Rauch-Hindin, W.B. "Artificial Intelligence in Business, Science and Industry. Volume II - Applications" pp.36-60.1995 Prentice-Hall.

15. Roth, E.M., Bennett, K., & Woods, D.D. "Human interactions with an 'intelligent' machine." International Journal of Man-Machine Studies, 27: 479-525.

16. Rouse, W.B., "A model of Human Decision-making in Fault Diagnosis Tasks that Include Feedback and Redundancy" IEEE Transactions on Systems, Man, and Cybernetics SMC-9 (4), April 1979

17. Rouse, W.B., "Human Problem Solving Performance in a Fault Diagnosis Task" IEEE Transactions on Systems, Man, and Cybernetics SMC-8 (4), April 1978

18. Sheridan, T.B., Ferrell, W.R. "Man-machine systems; information, control, and decision models of human performance" MIT Press 1974.

19. Smith, Robin. "Fault Tree Analysis and Diagnostics Development for PI-in-a-Box with the Neurolab Sleep and Respiration Experiment." Unpublished Master's Thesis, Massachusetts Institute of Technology, 1997.

20. Swartout, W.R. (1983). "XPLAIN: A system for creating and explaining expert consulting programs." Artificial Intelligence, 21: 285-325.

21. Vicente, K.J., "Cognitive Work Analysis: Towards Safe, Productive, and Healthy Computer-Based Work." published by Erlbaum, 1999

22. Young L.R. "PI-in-a-Box." Journal of the Society of Instrument and Control Engineers. (1994); 33(2):119-22.

23. Wickens, C.D., Gordon, S.E., Liu, Y. (1998) "An Introduction to Human Factors Engineering" published by Addison-Wesley

Appendix A

Data logging and extraction

Table A.1: Breakdown of syntax for log entries

Log entry format	Format	Description
[PI] = = = Event Marker = = =	Auto	subject records fault onset time (t0)
[PI] Comm failure with DSR	Auto	communication with DSR failed; used to mark t0 for Errors 4 and 6
//No event marker [description]	Manual	Event marker was not pressed by subject; used to mark t0**
Menu Popup	Auto	subject clicked checkbox to open dialog menu
Check X %chan_index	Auto	subject selected channel chan_index (once failure state is selected)
Check OFF %chan_index	Auto	subject de-selected checkbox chan_index (once failure is cleared by assistant)
OP chose State %state_index	Auto	subject selected state state_index from dialog menu
OP chose Problem %prob_index	Auto	subject selected problem prob_index from dialog menu
OP chose Solution %sol_index	Auto	subject selected solution sol_index from dialog menu
[PI] Warning %chan_desc %chan_index	Auto	[PI] detected anomalous behavior on chan_index described by chan_desc
[PI] Clear %chan_desc	Auto	[PI] detected nominal behavior from the corresponding channel, and cleared indicator light which showed chan_desc
User << State menu	Auto	Subject went back on his original state assessment, clearing checkbox
User << Problem menu	Auto	Subject went back on his original problem assessment
//Error %err_num [description]	Manual	Error number within trial
//Bonus error [description]	Manual	Bonus error encountered during trial

Appendix B

Informed Consent Form - Test subjects

INFORMED CONSENT FORM FOR TEST SUBJECTS

NSBRI PI-IN-A-BOX GROUND STUDY

Purpose

We would like permission to enroll you in a research study. The purpose of this study is to evaluate the efficacy of an expert system called PI-in-a-box in identifying the presence of artifacts in sleep data and suggesting corrective procedures to eliminate these artifacts. A version of PI-in-a-box has already been developed to assist astronauts in performing a sleep experiment in space. This experiment is designed to quantify the effectiveness of an expert system in a laboratory environment in terms of both time and accuracy.

Participation in this study is voluntary and you are free to withdraw your consent and discontinue participation in the experiment at any time without prejudice.

Procedures

You will be given approximately five to seven hours of training on a "training day". It is intended to provide an overview of the equipment used in sleep recordings, and the characteristics of each signal recorded. You will be trained in the possible problems that can arise with the instrumentation of a Sleep*Net, a web-like cap used to record electrophysiological signals. Another volunteer will be wearing the Sleep*Net on their heads. You will learn how to detect, troubleshoot and, correct problems which will occur in the instrumentation session. You will also be trained on the use of a computer decision aid called PI-in-a-box, which runs on a laptop computer. It will display the signals and use color-coded lights to indicate the quality of each signal. In addition, PI-in-a-box displays a "diagnostics" window which contains procedures for correcting poor quality signals.

Testing will take place over the course of three to four days, with one "training day" of five to seven hours, and two test days which will involve no more than one and a half hours on each day. Total testing time will be between 8-10 hours. During the test sessions, your task will be o detect problems in the sleep signal system, troubleshoot the problem to find the cause of these problems, and instruct a sleep technician to fix the problem and restore the quality of the sleep signals. One test day will be performed with the decision aid, and one without the aid.

Risks and Discomforts

No known risks associated with this component of the experiment.

Benefits

A prorated payment of $7.00 per hour will be provided to participants.

In the unlikely event of a physical injury resulting from participation in this research, I understand that med-

ical treatment will be available from the MIT Medical Department, including first aid emergency treatment and follow-up care as needed, and that my insurance carrier may be billed for the cost of such treatment. However, no compensation can be provided for medical apart from the foregoing. I further understand that making such medical treatment available; or providing it, does not imply that such injury is the Investigator's fault. I also understand that by my participation in this study, I am not waiving any of my legal rights.

I understand that I may also contact the Chairman of the Committee on the use of Humans as Experimental Subjects, MIT 253-6787, if I feel I have been treated unfairly as a subject.

Signature

I have been fully informed as to the procedures to be followed, including those which are investigational, and have been given a description of the attendant discomforts, risks, and benefits to be expected. In signing this consent form, I agree to participate in the project and I understand that I am free to withdraw my consent and have this study discontinued at any time. I understand also that if I have any questions at any time, they will be answered.

_____ _____

Subject's Signature Date

Appendix C

Informed consent form - Sleep subjects

INFORMED CONSENT FORM FOR SLEEP SUBJECTS

NSBRI PI-IN-A-BOX GROUND STUDY

Purpose

We would like permission to enroll you in a research study. The purpose of this study is to evaluate the efficacy of an expert system called PI-in-a-box in identifying the presence of artifacts in sleep data and suggesting corrective procedures to eliminate these artifacts. A version of PI-in-a-box has already been developed to assist astronauts in performing a sleep experiment in space. This experiment is designed to quantify the effectiveness of an expert system in a laboratory environment in terms of both time and accuracy.

Participation in this study is voluntary and you are free to withdraw your consent and discontinue participation in the experiment at any time without prejudice.

Procedures

You will act as a sleep subject in two instrumenting test sessions per day in a span of two to three days. You will don a Sleep*Net, a web-like cap used to record electrophysiological signals such as EEG, EOG and EMG signals. A mildly abrasive cream will be used to scrub each electrode site prior to electrode application. Small adhesive discs will be used to apply the facial electrodes. During the test session, the technician will loosen or remove electrodes in the setup to deliberately introduce problems in the sleep signals.

Risks and Discomforts

A mild, abrasive cream will be used to scrub each electrode site prior to applying the electrodes. Minor irritation may result from this cleansing process.

Facial electrodes will be applied to the skin using small adhesive discs. The glue on these adhesives may cause minor discomfort or skin irritation.

Discomfort may be experienced as electrodes are poked out and put back into the Sleep*Net, but there will be breaks between test sessions, and you will not be required to wear the Sleep*Net for more than one hour at a time.

Benefits

A prorated payment of $7.00 per hour will be provided to participants.

In the unlikely event of a physical injury resulting from participation in this research, I understand that medical treatment will be available from the MIT Medical Department, including first aid emergency treatment and follow-up care as needed, and that my insurance carrier may be billed for the cost of such treatment. However, no compensation can be provided for medical apart from the foregoing. I further understand that

making such medical treatment available; or providing it, does not imply that such injury is the Investigator's fault. I also understand that by my participation in this study, I am not waiving any of my legal rights.

I understand that I may also contact the Chairman of the Committee on the use of Humans as Experimental Subjects, MIT 253-6787, if I feel I have been treated unfairly as a subject.

Signature

I have been fully informed as to the procedures to be followed, including those which are investigational, and have been given a description of the attendant discomforts, risks, and benefits to be expected. In signing this consent form, I agree to participate in the project and I understand that I am free to withdraw my consent and have this study discontinued at any time. I understand also that if I have any questions at any time, they will be answered.

_____ _____

Subject's SignatureDate

Appendix D

Pre-experiment questionnaire

Subject Number: _____

Date: _____

Principal-Investigator-in-a-Box
Ground Based Evaluation Study Year 2

January 2000

Subject Name:_____

Home Phone Number:_____ MIT ID#: _____

E-Mail Address:_____

Please fill out the following for payment purposes:

Social Security Number (for payment):_____

Address: _____

_____

Country of Citizenship: _____

Age (as of today):_____Gender: Male Female

Year in School (freshman= 1): 1 2 3 4 Graduate

Field of Study:_____

How many hours per week do you use a personal computer or workstation? _____

Have you ever seen or worked on experiments involving electrophysiological signals such as EEGs, EKGs, etc ? YES NO

(If YES, please elaborate.) _____

_____

Do you have any experience as a repair or support technician? YES NO

(If YES, please elaborate.) _____

_____

Are you color blind? YES NO

Do you wear corrective lenses? YES NO

If YES, are you currently wearing GLASSES or CONTACTS (circle one)

Are you right or left-handed? LEFT RIGHT

Appendix E

Data Collection sheet for Assistant

Data Sheet

Name:_____ Expt. Day: 2 Order: O2-EOG Head Name: _____

Time: _____ PI: Y/N Sex: M/F Date:_____ Training Session:_____ SleepNet:_____

T/S Q #	Description	Fix Cmd	Description	Slpr Cmd	description
1	Is RS 232 Cable connected?	1	Plug in RS 232 Cable	BLINK	Blink eyes
2	Is Sleep Net plugged in?	2	Plug Sleep net into blue slice	JAW	Clench jaw muscles
3	Is Sleep Net placement OK?	3	Adjust Sleep Net placement	RELAX	Relax
4	Is DSR recording?	4	Start DSR recording	EYES	Look left, look right, etc
5	Is DSR on?	5	Turn DSR on. Start DSR recording	HEAD	Move head around
6	Are Hydrodots properly inserted?	6	Insert and flush Hydrodots		
7	Is the ground electrode properly installed?	7	Replace ground electrode	Special case: Null error T/S Question	
8	Is Hydrodot properly inserted?	8	Insert Hydrodot until flush	NULL	Is there no error?
9	Is there hair beneath Hydrodot?	9	Remove hair		
10	Is the site scrubbed?	10	Rescrub site		
11	Is reference electrode properly installed?	11	Install reference electrode		
12	Is the thin side of socket applied to skin?	12	Apply thin side of socket to skin		
13	Is electrode placement good?	13	Place electrode in right spot		

Error	Description	T/S Q #	Elect #	Fix cmd #	Elect #	Sleeper Cmd
1	O2 Hydrodot not flush with sleepnet					
2	Null error					
3	C4 EEG site not properly scrubbed					
4	RS 232 Cable not connected					
5	A2 Reference Electrode loose					

Data Sheet

Name:_____ Expt. Day: 2 Order: O2-EOG Head Name: _____

Time: _____ PI: Y/N Sex: M/F Date:_____ Training Session:_____ SleepNet:_____

T/S Q #	Description	Fix Cmd	Description	Slpr Cmd	description
1	Is RS 232 Cable connected?	1	Plug in RS 232 Cable	BLINK	Blink eyes
2	Is Sleep Net plugged in?	2	Plug Sleep net into blue slice	JAW	Clench jaw muscles
3	Is Sleep Net placement OK?	3	Adjust Sleep Net placement	RELAX	Relax
4	Is DSR recording?	4	Start DSR recording	EYES	Look left, look right, etc
5	Is DSR on?	5	Turn DSR on. Start DSR recording	HEAD	Move head around
6	Are Hydrodots properly inserted?	6	Insert and flush Hydrodots		
7	Is the ground electrode properly installed?	7	Replace ground electrode	Special case: Null error T/S Question	
8	Is Hydrodot properly inserted?	8	Insert Hydrodot until flush	NULL	Is there no error?
9	Is there hair beneath Hydrodot?	9	Remove hair		
10	Is the site scrubbed?	10	Rescrub site		
11	Is reference electrode properly installed?	11	Install reference electrode		
12	Is the thin side of socket applied to skin?	12	Apply thin side of socket to skin		
13	Is electrode placement good?	13	Place electrode in right spot		

Error	Description	T/S Q #	Elect #	Fix cmd #	Elect #	Sleeper Cmd
6	DSR stopped recording					
7	Null error					
8	O1 EEG Hydrodot not inserted					
9	Ground Electrode missing					
10	EOG-L Hydrodot Not inserted					

Data Sheet

me:_____ Expt. Day: 2 Order: O2-EOG Head Name: _____

ne: _____ PI: Y/N Sex: M/F Date:_____ Training Session:_____ SleepNet:_____

S Q #	Description	Fix Cmd	Description	Slpr Cmd	description
	Is RS 232 Cable connected?	1	Plug in RS 232 Cable	BLINK	Blink eyes
	Is Sleep Net plugged in?	2	Plug Sleep net into blue slice	JAW	Clench jaw muscles
	Is Sleep Net placement OK?	3	Adjust Sleep Net placement	RELAX	Relax
	Is DSR recording?	4	Start DSR recording	EYES	Look left, look right, etc
	Is DSR on?	5	Turn DSR on. Start DSR recording	HEAD	Move head around
	Are Hydrodots properly inserted?	6	Insert and flush Hydrodots		
	Is the ground electrode properly installed?	7	Replace ground electrode	Special case: Null error T/S Question	
	Is Hydrodot properly inserted?	8	Insert Hydrodot until flush	NULL	Is there no error?
	Is there hair beneath Hydrodot?	9	Remove hair		
	Is the site scrubbed?	10	Rescrub site		
	Is reference electrode properly installed?	11	Install reference electrode		
	Is the thin side of socket applied to skin?	12	Apply thin side of socket to skin		
	Is electrode placement good?	13	Place electrode in right spot		

or	Description	T/S Q #	Elect #	Fix cmd #	Elect #	Sleeper Cmd
	Sleep net not plugged in					
	C3 EEG Hair beneath site					
	EOG-R site not properly scrubbed					

Appendix F

Subject Debriefing Questionnaire

Please circle the level of rating you think best represents the following statements

STATEMENT	RATING						
The level of accuracy of monitoring signal quality:	Poor			Satisfactory			Good
Q1. Using [PI] assessments (colored lights)	1	2	3	4	5	6	7
Q2. Observing the signal waveforms	1	2	3	4	5	6	7
Q3. Using both [PI] assessments and observing signal waveforms	1	2	3	4	5	6	7
The effectiveness of figuring out the cause of the problem:							
Q4. Using [PI] diagnostics messages	1	2	3	4	5	6	7
Q5. Using the troubleshooting procedures manual	1	2	3	4	5	6	7
Q6. Using both [PI] messages and the troubleshooting procedures	1	2	3	4	5	6	7
Q7. How well did you understand the troubleshooting directions given in the procedures?	1	2	3	4	5	6	7
Q8. How closely did you follow the troubleshooting directions given in the NASA guideline?	1	2	3	4	5	6	7
Q9. How helpful were the [PI] diagnostics instructions in determining how to correct the problem?	1	2	3	4	5	6	7
Q10. How effective was the training session in preparing you for the experiment?	1	2	3	4	5	6	7
Q11. Describe the usefulness of [PI] as a completely autonomous decision-making tool	1	2	3	4	5	6	7
Q12. Describe the usefulness of [PI] as a troubleshooting advisory tool	1	2	3	4	5	6	7

Were you able to rectify the problem using the procedures?

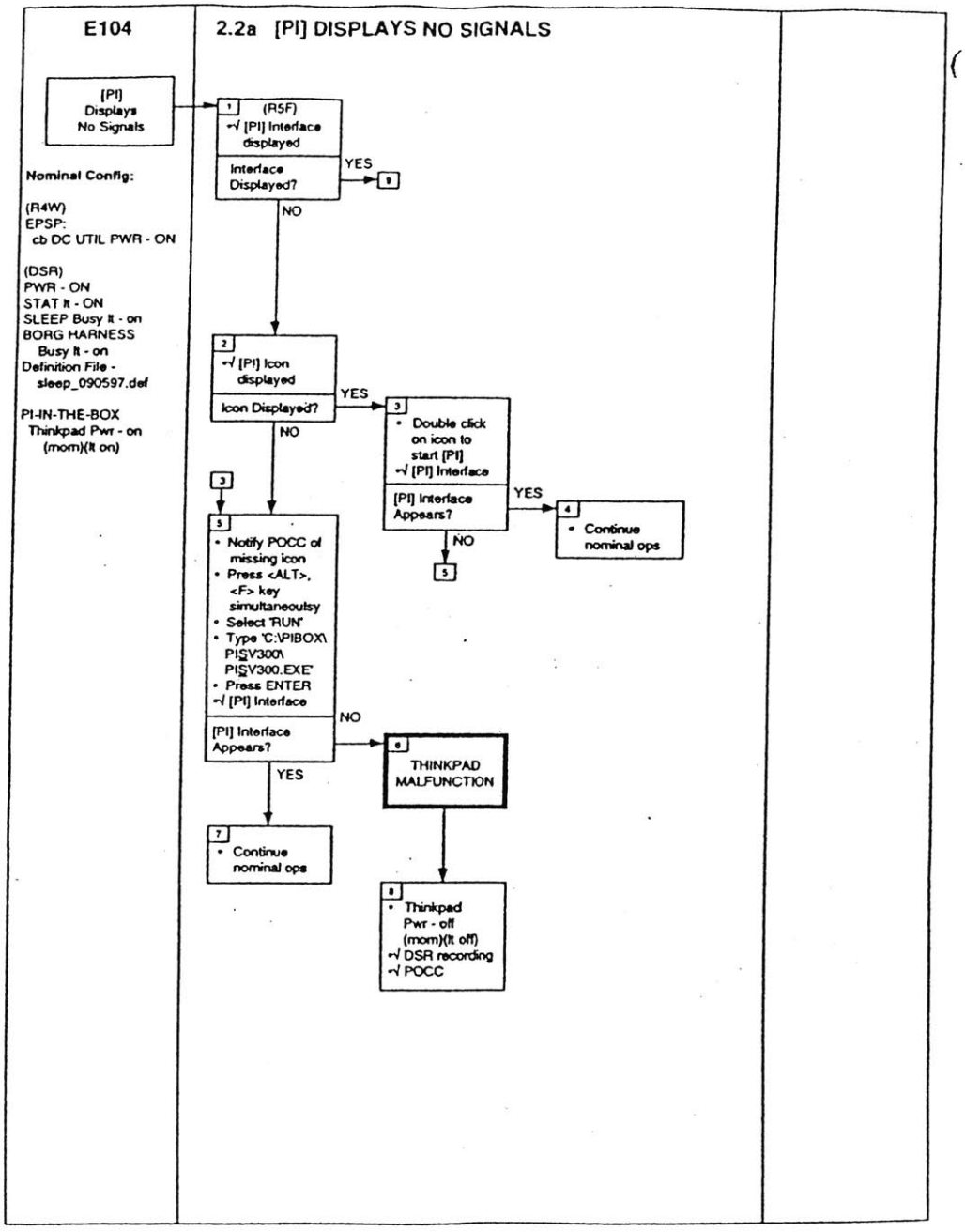
If NO, please explain:

Please feel free to include any comments you may have about our experiment

Thank you for your time ...

Appendix G

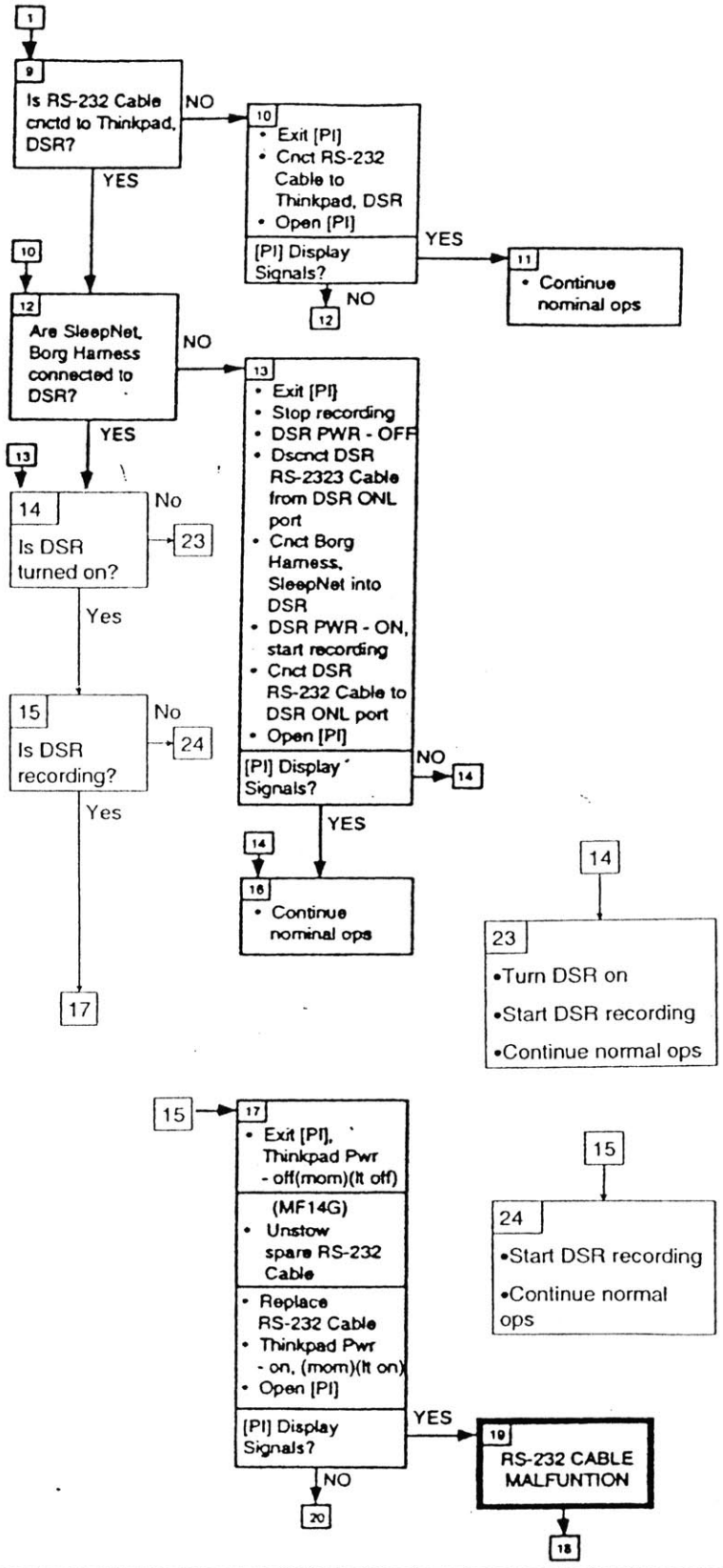
NASA Troubleshooting Guideline



2-55 SLEEP PROC/NL/FINAL, REV A

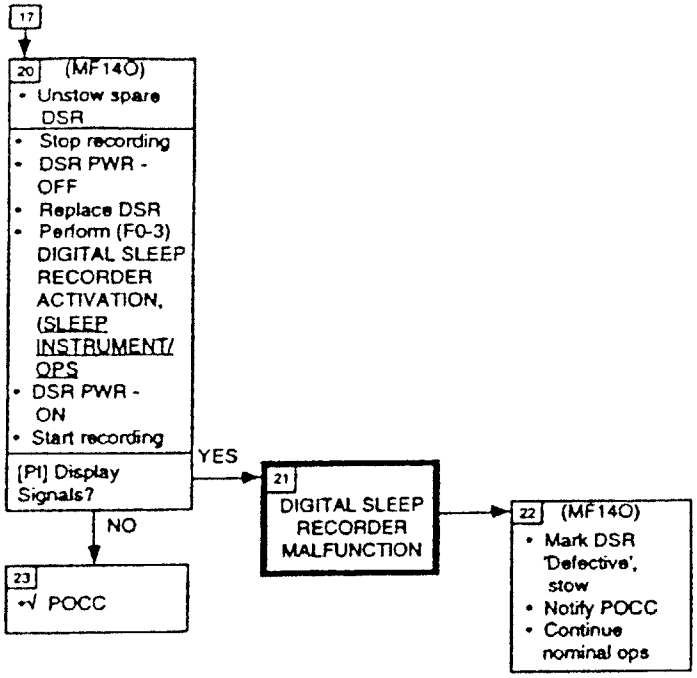
E104

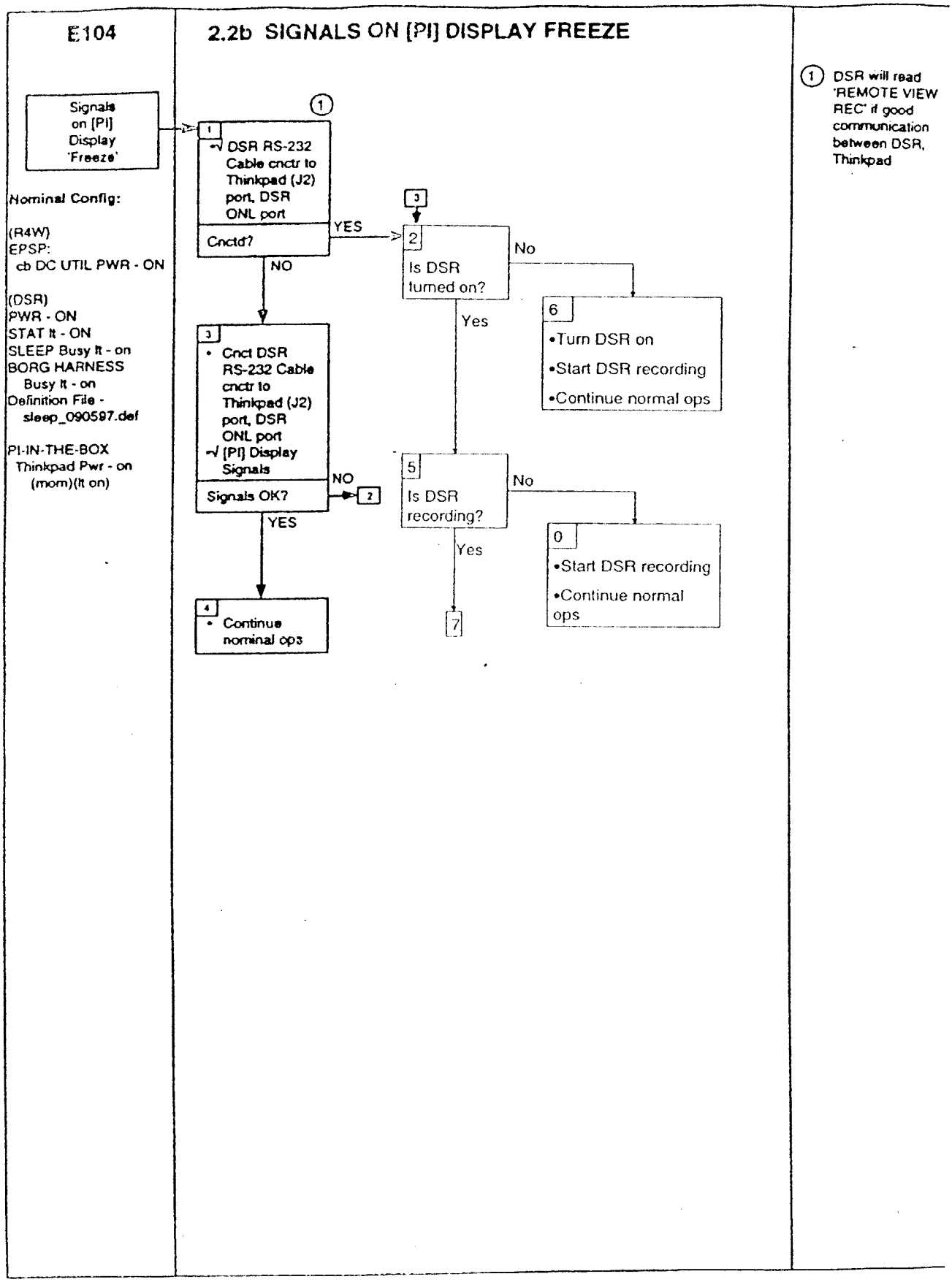
2.2a [PI] DISPLAYS NO SIGNALS (CONT'D)



E104

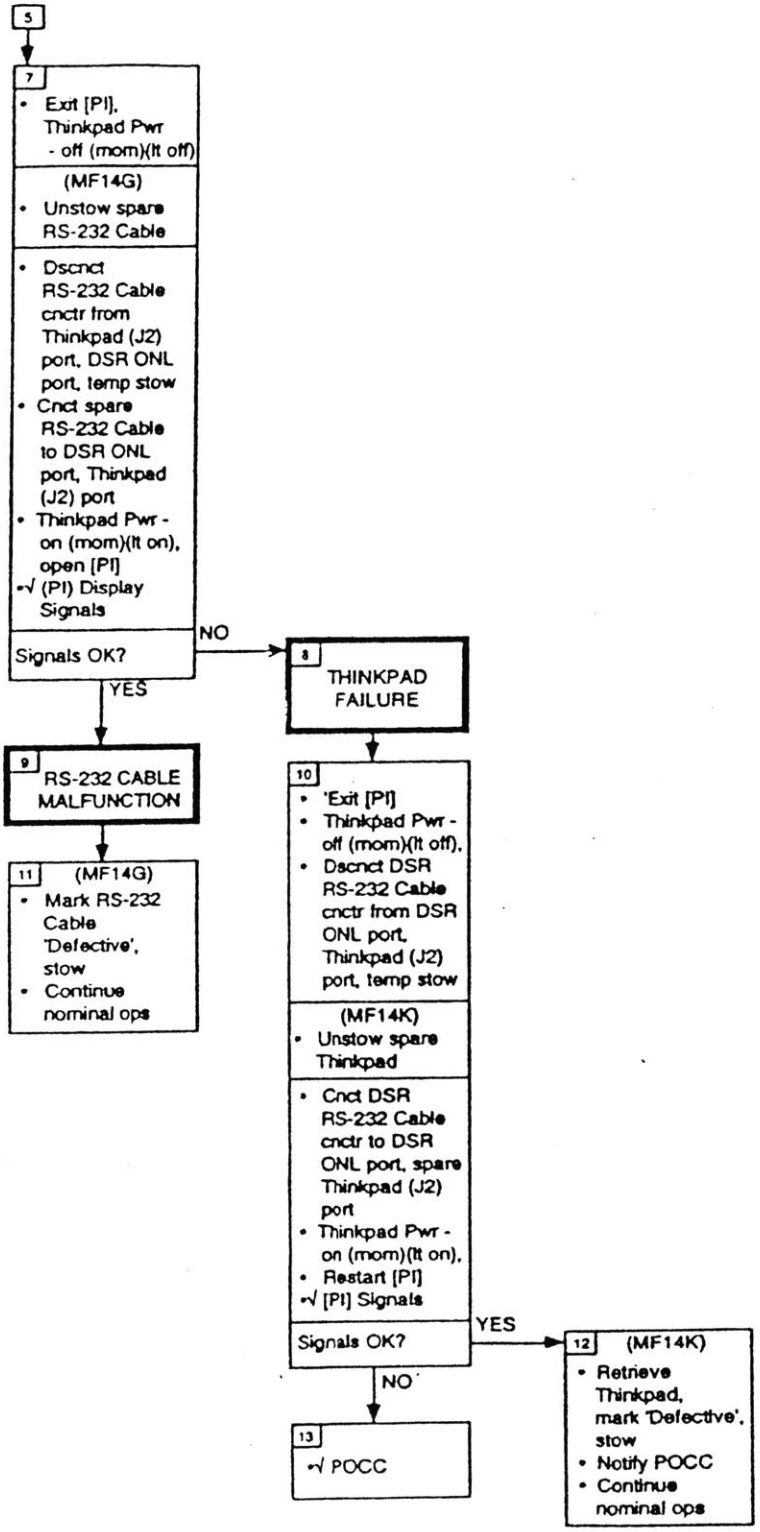
2.2a [PI] DISPLAYS NO SIGNALS (CONT'D)

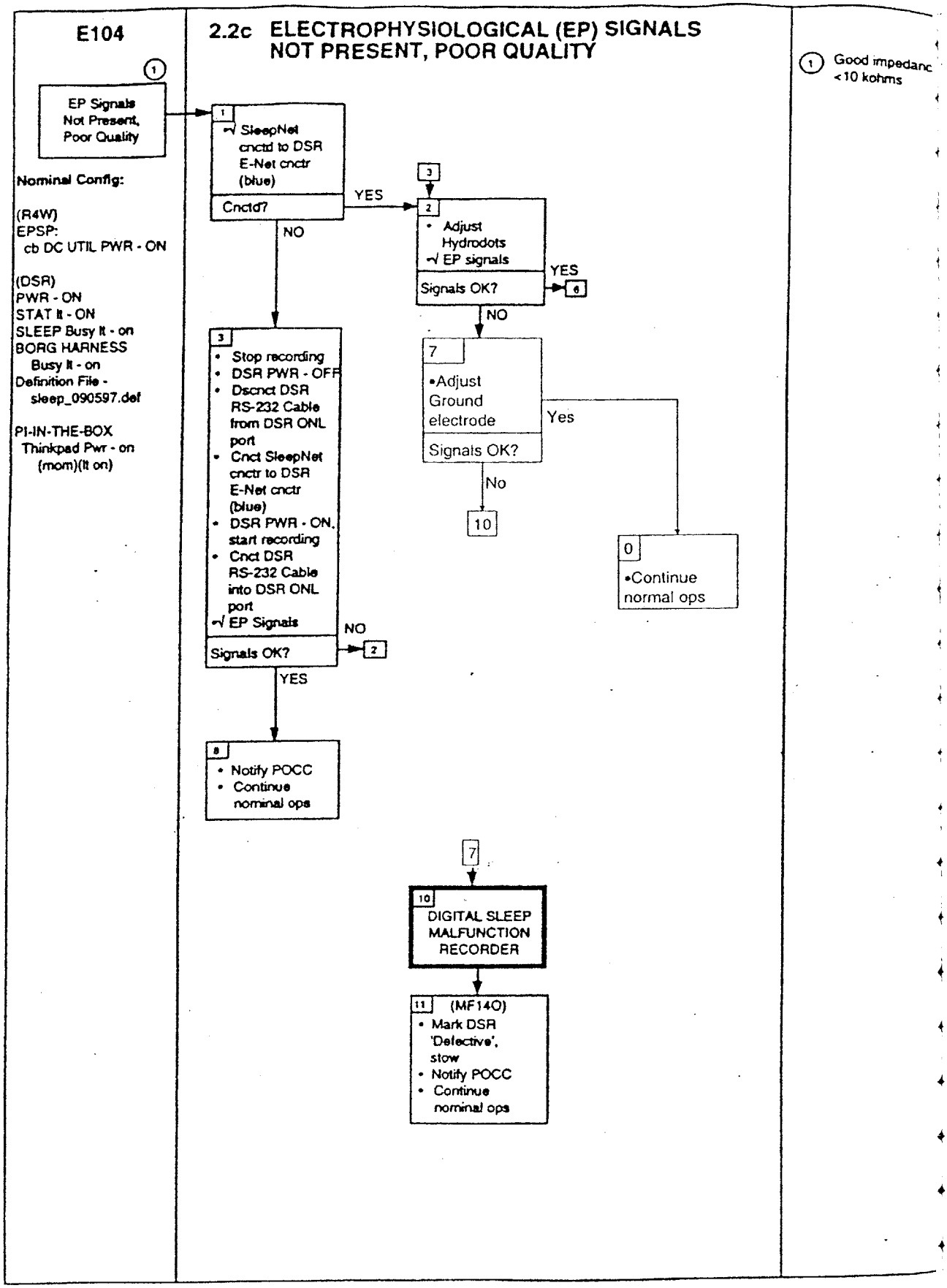


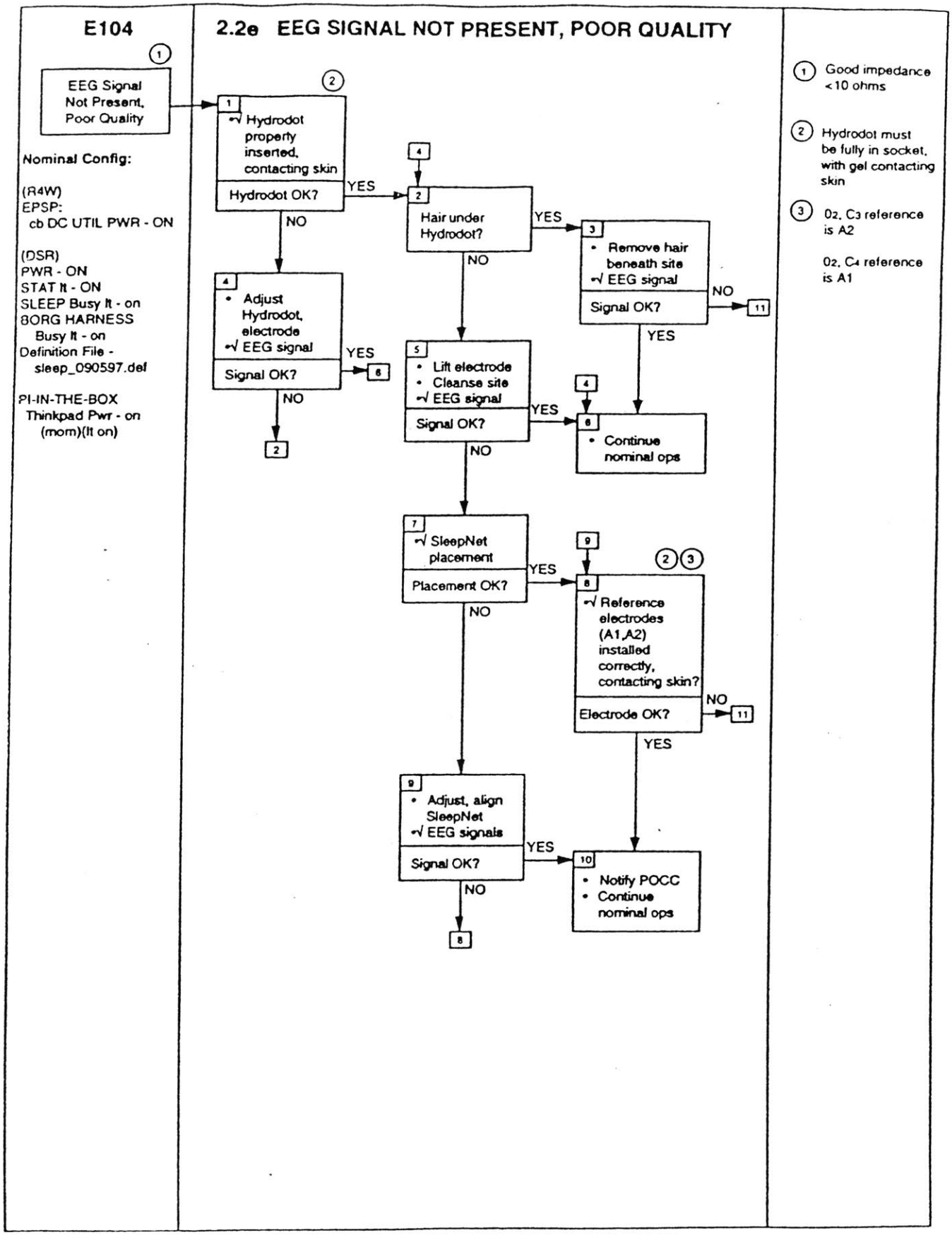

E104
2.2b SIGNALS ON [PI] DISPLAY FREEZE
Signals on [PI] Display 'Freeze'
Nominal Config:
(R4W)
EPSP:
cb DC UTIL PWR - ON
(DSR)
PWR - ON
STAT lt - ON
SLEEP Busy lt - on
BORG HARNESS
Busy lt - on
Definition File -
sleep_090597.def
PI-IN-THE-BOX
Thinkpad Pwr - on
(mom)(lt on)
1
-√ DSR RS-232 Cable cnctr to Thinkpad (J2) port, DSR ONL port
Cnctd?
YES
NO
2
Is DSR turned on?
No
Yes
6
•Turn DSR on
•Start DSR recording
•Continue normal ops
3
• Cnct DSR RS-232 Cable cnctr to Thinkpad (J2) port, DSR ONL port
-√ [PI] Display Signals
Signals OK?
NO
YES
5
Is DSR recording?
No
Yes
0
•Start DSR recording
•Continue normal ops
4
• Continue nominal ops
7
1 DSR will read 'REMOTE VIEW REC' if good communication between DSR, Thinkpad

E104

2.2b SIGNALS ON [PI] DISPLAY FREEZE (CONT'D)

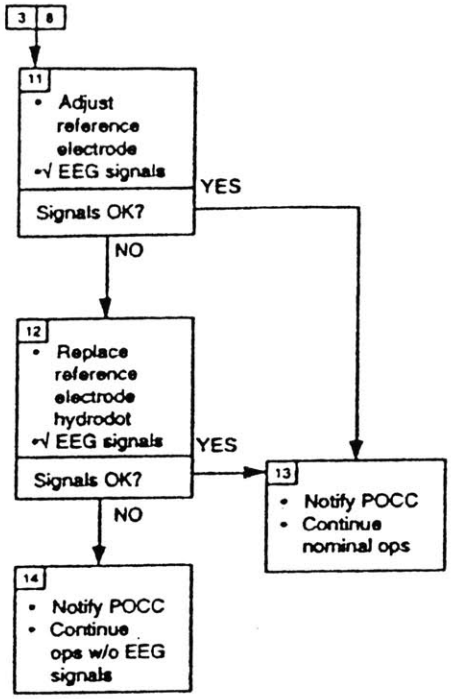


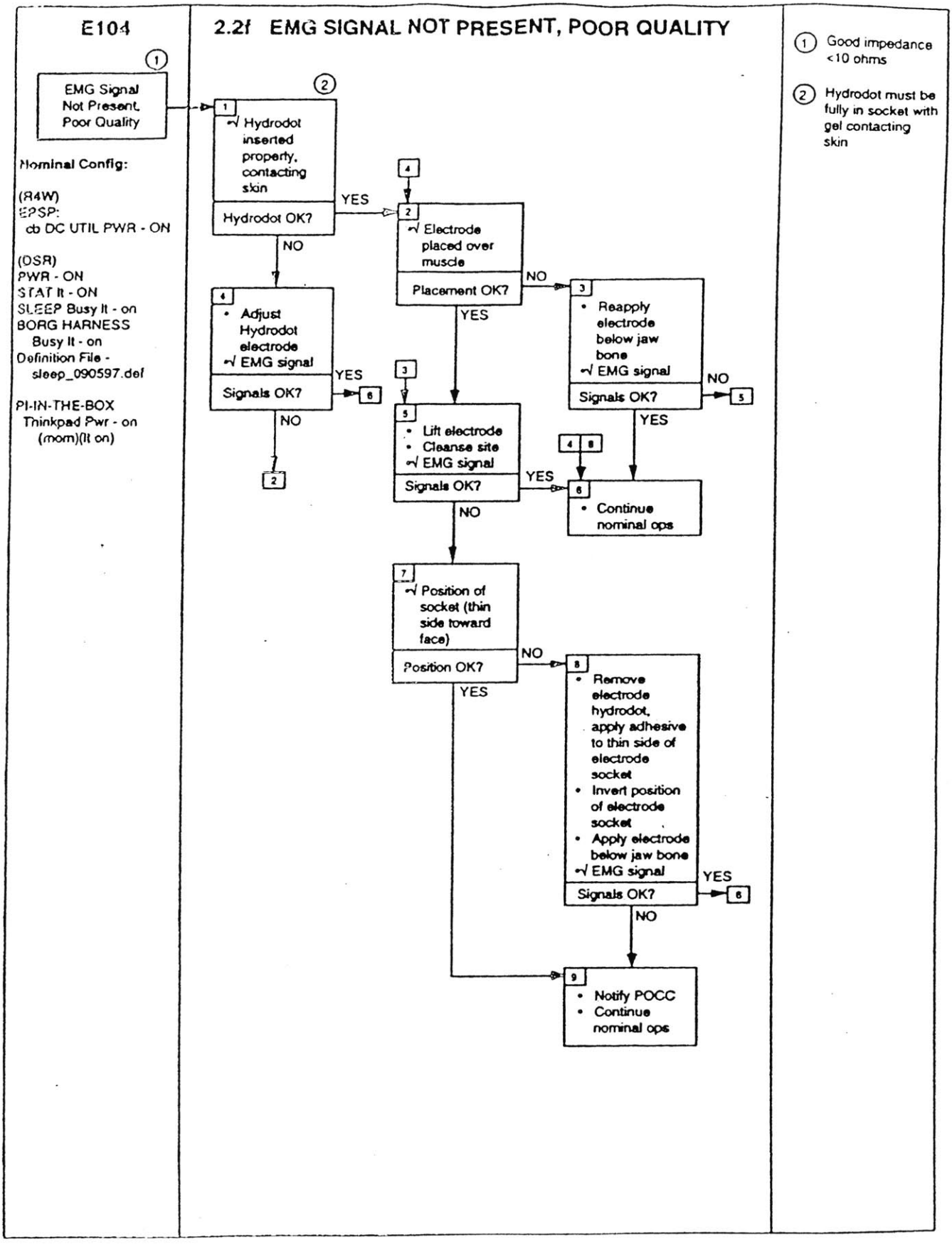

E104
2.2c ELECTROPHYSIOLOGICAL (EP) SIGNALS NOT PRESENT, POOR QUALITY
1 Good impedanc <10 kohms
1
EP Signals Not Present, Poor Quality
Nominal Config:
(R4W)
EPSP:
cb DC UTIL PWR - ON
(DSR)
PWR - ON
STAT lt - ON
SLEEP Busy lt - on
BORG HARNESS Busy lt - on
Definition File - sleep_090597.def
PI-IN-THE-BOX
Thinkpad Pwr - on (mom)(lt on)
1
√ SleepNet cnctd to DSR E-Net cnctr (blue)
Cnctd?
YES
NO
3
2
• Adjust Hydrodots
√ EP signals
Signals OK?
YES
6
NO
3
• Stop recording
• DSR PWR - OFF
• Dscnct DSR RS-232 Cable from DSR ONL port
• Cnct SleepNet cnctr to DSR E-Net cnctr (blue)
• DSR PWR - ON, start recording
• Cnct DSR RS-232 Cable into DSR ONL port
√ EP Signals
Signals OK?
NO
2
YES
8
• Notify POCC
• Continue nominal ops
7
•Adjust Ground electrode
Signals OK?
Yes
No
10
0
•Continue normal ops
7
10
DIGITAL SLEEP MALFUNCTION RECORDER
11
(MF140)
• Mark DSR 'Defective', stow
• Notify POCC
• Continue nominal ops

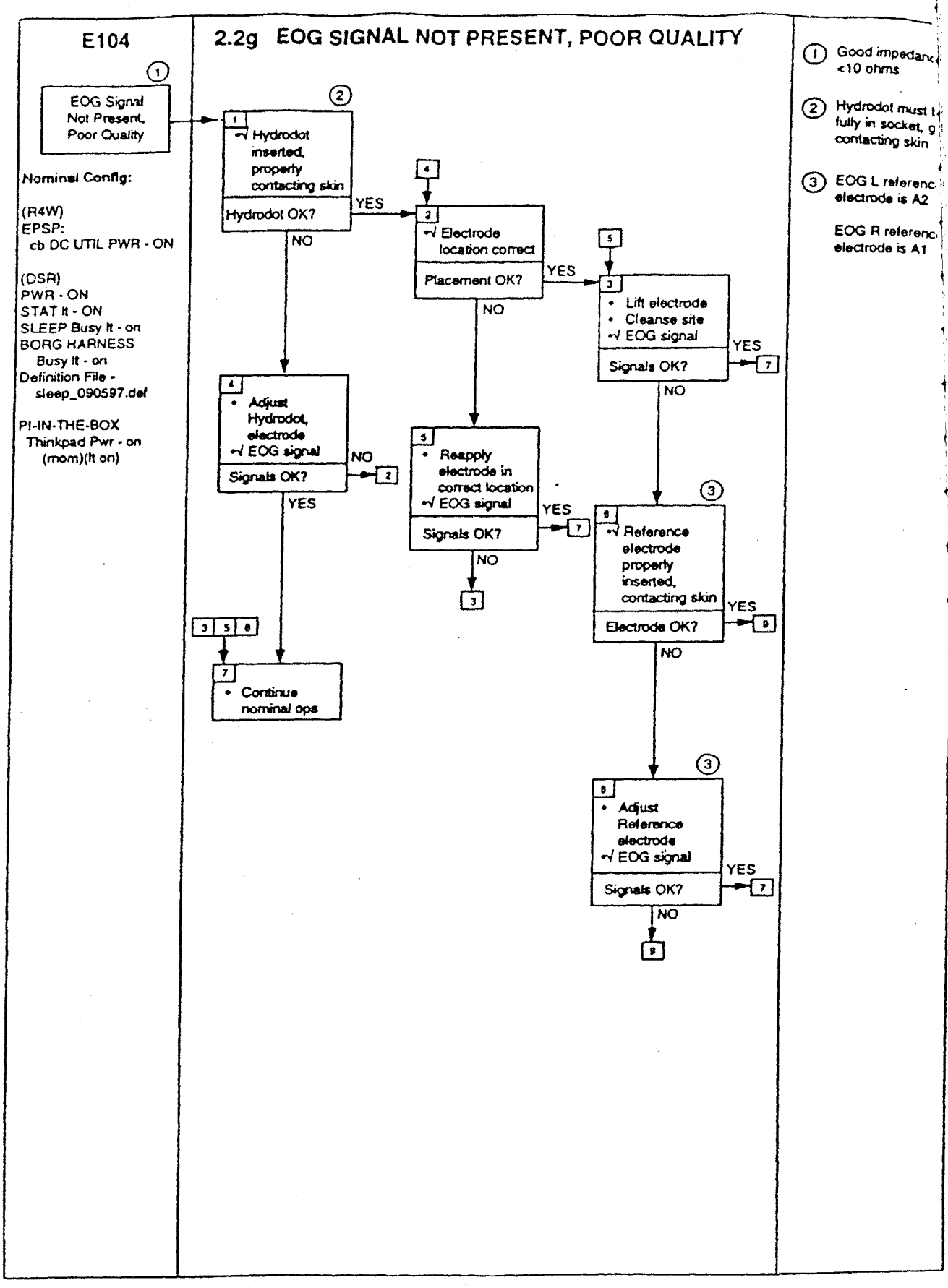

E104
1
EEG Signal Not Present, Poor Quality
Nominal Config:
(R4W)
EPSP:
cb DC UTIL PWR - ON
(DSR)
PWR - ON
STAT lt - ON
SLEEP Busy lt - on
8ORG HARNESS
Busy lt - on
Definition File -
sleep_090597.def
PI-IN-THE-BOX
Thinkpad Pwr - on
(mom)(lt on)
2.2e EEG SIGNAL NOT PRESENT, POOR QUALITY
2
1
√ Hydrodot properly inserted, contacting skin
Hydrodot OK?
YES
NO
4
2
Hair under Hydrodot?
YES
NO
3
• Remove hair beneath site
√ EEG signal
Signal OK?
NO
11
YES
4
• Adjust Hydrodot, electrode
√ EEG signal
Signal OK?
YES
6
NO
2
5
• Lift electrode
• Cleanse site
√ EEG signal
Signal OK?
YES
4
6
• Continue nominal ops
NO
7
√ SleepNet placement
Placement OK?
YES
9
2 3
8
√ Reference electrodes (A1,A2) installed correctly, contacting skin?
Electrode OK?
NO
11
YES
NO
9
• Adjust, align SleepNet
√ EEG signals
Signal OK?
YES
10
• Notify POCC
• Continue nominal ops
NO
8
1 Good impedance <10 ohms
2 Hydrodot must be fully in socket, with gel contacting skin
3 O2, C3 reference is A2
O2, C4 reference is A1

E104

2.2e EEG SIGNAL NOT PRESENT, POOR QUALITY (CONT'D)

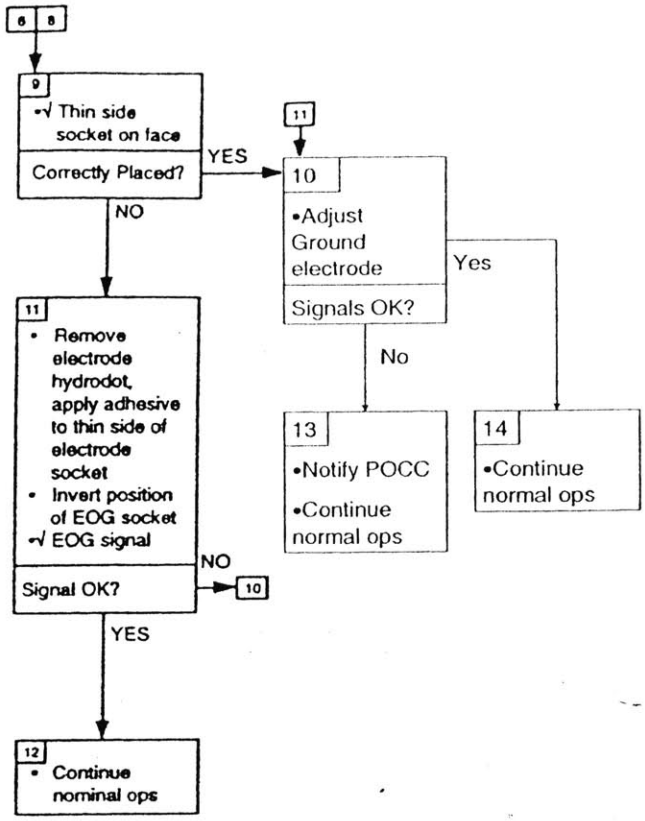



E104
2.2f EMG SIGNAL NOT PRESENT, POOR QUALITY
1 Good impedance <10 ohms
2 Hydrodot must be fully in socket with gel contacting skin
1
EMG Signal Not Present, Poor Quality
Nominal Config:
(R4W)
EPSP:
cb DC UTIL PWR - ON
(DSR)
PWR - ON
STAT lt - ON
SLEEP Busy lt - on
BORG HARNESS
Busy lt - on
Definition File -
sleep_090597.def
PI-IN-THE-BOX
Thinkpad Pwr - on
(mom)(lt on)
2
1
√ Hydrodot inserted property, contacting skin
Hydrodot OK?
YES
NO
4
2
√ Electrode placed over muscle
Placement OK?
NO
YES
3
• Reapply electrode below jaw bone
√ EMG signal
Signals OK?
NO
YES
5
4
• Adjust Hydrodot electrode
√ EMG signal
Signals OK?
YES
6
NO
2
3
5
• Lift electrode
• Cleanse site
√ EMG signal
Signals OK?
YES
4 8
6
• Continue nominal ops
NO
7
√ Position of socket (thin side toward face)
Position OK?
NO
YES
8
• Remove electrode hydrodot, apply adhesive to thin side of electrode socket
• Invert position of electrode socket
• Apply electrode below jaw bone
√ EMG signal
Signals OK?
YES
6
NO
9
• Notify POCC
• Continue nominal ops


E104
2.2g EOG SIGNAL NOT PRESENT, POOR QUALITY
EOG Signal Not Present, Poor Quality
Nominal Config:
(R4W)
EPSP:
cb DC UTIL PWR - ON
(DSR)
PWR - ON
STAT lt - ON
SLEEP Busy lt - on
BORG HARNESS
Busy lt - on
Definition File -
sleep_090597.def
PI-IN-THE-BOX
Thinkpad Pwr - on
(mom)(lt on)
Hydrodot inserted, properly contacting skin
Hydrodot OK?
YES
NO
Electrode location correct
Placement OK?
Lift electrode
Cleanse site
EOG signal
Signals OK?
Adjust Hydrodot, electrode
EOG signal
Reapply electrode in correct location
Reference electrode properly inserted, contacting skin
Electrode OK?
Adjust Reference electrode
Continue nominal ops
Good impedanc
<10 ohms
Hydrodot must b
fully in socket, g
contacting skin
EOG L referenc
electrode is A2
EOG R referenc
electrode is A1

E104

2.2g EOG SIGNAL NOT PRESENT, POOR QUALITY (CONT'D)



Appendix H

Experimental Procedure

Instrumenting Head:
Clean face with alcohol swab.
Place sleep net on head and line it up.
Adjust chin and rear straps for proper fit.
Scrub under chin profusely to prepare the EMG site.
Apply nasion pad and insert ground electrode.
Apply EEG electrodes by removing hair scrubbing site and inserting Hydrodots.
Insert reference electrodes on mastoid by same procedure.
Insert the EOG electrodes in proper locations EOGR above and EOGL below.
Insert EMG electrodes under jawbone.
Make sure to use pads on EOG and EMG electrodes pad goes yellow side against the electrode white side against skin.

Turning on software:
Plug in monitor out.
For the PI computers (Win95) :

Turn on the think pad. (If using Orion Laptop change screen resolution to 800x640). Attach RS 232 cable to back of think pad.
Turn on DSR. Plug sleep net into DSR. Have DSR start recording.
Plug RS 232 cable into DSR.
Start the correct version of PI: select icon for PI with diagnostics, PI diagnosticless
Check to verify that all signals are working properly.

For the NT computers:

Turn on the ThinkPad.
Start the Matlab timer script.

Reinstrument any electrodes, which appear to need reinstrumentation.
Do calibrations to assure that PI is working properly.

Running Experiment with the order O2-EOG:

For ALL problems:

Before creating the error, start the create timer script.
After creating the error,

Start the problem timer script.
tap the subject on the shoulder.
Ask test subject to:

(a) remove headphones, and

(b) press the event marker to begin. Check to see if the event marker light turned "green"

During the diagnostic part of each error:

(a) Answer test subject's questions about the setup. Do not answer questions until they have any checkbox marked with an "X"

(a) Note on the recorder sheet the channel number and problem code for each question. Watch the monitor to see what the test subject is doing, and make notes of any observations.

After the correct fix has been made and before the next problem is created, ask test subject to:

(a) Click on the solution dialog to remove the "X" from a problem checkbox

(a) turn troubleshooting guide back to page 1, and

(a) place headphones on

If more than one problem still exists, ask the test subjects to use the remaining problem time to diagnose the other errors. When their time is up, make sure these other problems are fixed before moving on.

If not, then

(a) Ask the test subject to ignore those channels, and

(a) Change to the channel OPPOSITE it if there are any other problems involving that electrode

(a) Make a note on recording sheet of the bad channel

Error 1: Hydrodot not flush with sleep net

- Creating: Wait 0 seconds before tapping test subject. Pull out the O2 Hydrodot so it is not flush with the sleep net and no longer contacting the scalp. Try to make the signal a *popping* signal.
- Ask sleeper to: Tap electrode so that the signal appears to be popping.

Error 2: No error

- Creating: Wait 30 seconds before tapping test subject.

Error 3: EEG site not properly scrubbed

- Creating: Wait 0 seconds. Try to make the signal appear *noisy.* Apply make up to C4 electrode so it is completely brown. Signal should be different from other three EEG signals.
- If this error cannot be created, produce any error to call the attention of the test subject, and record the discrepancy.
- Hydrodot will probably need to be replaced after this error.

Error 4: RS –232 Cable not connected

- Creating: Wait 28 seconds. Unplug the RS-232 cable from DSR from the LEMO end. PI signals should *freeze.*

Error 5: Reference Electrode Loose

- Creating: Wait 25 seconds. Pull A2 electrode away from skin. There should be three *flat* signals.
- Ask sleeper to: hold reference electrode off skin, but not to pull too hard.

Error 6: DSR stopped recording.

- Creating: Wait 28 seconds. Stop DSR recording. PI signals should *freeze.*
- After problem is solved, make sure to select Append to Data when the DSR prompts you that the card is not empty.

Error 7: Null error

- Creating: Wait 30 seconds before tapping test subject.

Error 8: EEG Hydrodot not inserted.

- Creating: Wait 10 seconds. Remove O1 electrode. Signal should be *flat.*

Error 9: Ground Electrode missing.

- Creating: Wait 15 seconds. Remove ground electrode carefully be sure not to unstick sleep net from forehead pad or will need to replace forehead pad. *All signals* should have *poor quality.*
- Ask sleeper to: insert finger into Ground electrode if All signals don't look bad.

Error 10: EOG Hydrodot not inserted.

- Creating: Wait 15 seconds. Remove Hydrodot from EOGL. Signal should be *flat* and unresponsive to eye movements.

Error 11: Sleep Net not plugged in.

- Creating: Wait 28 seconds. Unplug Sleep Net from DSR blue slice. *All signals* should be *poor quality.*

Error 12: Hair beneath EEG

- Creating: Wait 15 seconds. Lift C3 electrode and put hair under it. Signal should be either *flat* or *popping*, depending on the hair.
- If this error cannot be created, create a popping error by asking the sleeper to tap the electrode, and make note of the discrepancy.

Error 13: EOG site not properly scrubbed.

- Creating. Wait 0 seconds. Cover the EOGR Hydrodot with makeup so it is completely brown. Signal should be *noisy* and have poor response to eye movements.

- If this error cannot be created, produce any error to call the attention of the test subject, and record the discrepancy.
- Hydrodot will probably need to be replaced after this error.

After the experiment:

Exit PI

Find the logfile of the experimental session, make sure it is from the correct directory.

Filename: **Pis_recd.txt**

Rename file to **<subjectname>_dd_mm_yy.txt**

<subjectname> is the first name of the test subject

And dd,mm,yy refer to date,month, and year respectively

Move the file to directory **C:\PI-in-a-box\Jan\logs**

Stop DSR recording.

Turn DSR off.

Remove FlashRAM card by pressing down on the center of it first and then pulling it out.

Insert FlashRAM into ThinkPad PCMCIA port.

Wait for Thinkpad to recognize the card (A window will popup showing drive E)

Copy the **vpdata.raw** file to **C:\PI-in-a-box\Jan\vpdata**

Rename the file to **<subjectname>_dd_mm_yy.txt**

<subjectname> is the first name of the test subject

And dd,mm,yy refer to date,month, and year respectively

Remove FlashRAM card from PCMCIA port, and reinsert into DSR.

Emergency heads:

Will Fournier

Andy Liu

Joe Saleh

Heiko Hecht

Richard Delaney

Jason Richards