U.S. Department
of Transportation
**National Highway
Traffic Safety
Administration**

DOT HS 808 247

December 1994

Final Report

Research on Vehicle-Based Driver Status/Performance Monitoring; Development, Validation, and Refinement of Algorithms For Detection of Driver Drowsiness

This document is available to the public from the National Technical Information Service, Springfield, Virginia 22161.

This publication is distributed by the U.S. Department of Transportation, National Highway Traffic Safety Administration, in the interest of information exchange. The opinions, findings and conclusions expressed in this publication are those of the author(s) and not necessarily those of the Department of Transportation or the National Highway Traffic Safety Administration. The United States Government assumes no liability for its contents or use thereof. If trade or manufacturers' name or products are mentioned, it is because they are considered essential to the object of the publication and should not be construed as an endorsement. The United States Government does not endorse products or manufacturers.

Technical Report Documentation Page

1. Report No. DOT HS 808 247	2. Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle

RESEARCH ON VEHICLE-BASED DRIVER STATUS/PERFORMANCE MONITORING; DEVELOPMENT, VALIDATION, AND REFINEMENT OF ALGORITHMS FOR DETECTION OF DRIVER DROWSINESS

5. Report Date

December 23, 1994

6. Performing Organization Code

VRISU ISE

7. Author's! W. W. Wierwille L. A. Ellsworth
S. S. Wreggit R. J. Fairbanks
C. L. Kirn

8. Performing Organization Report No.

VPISU ISE 94-04

9. Performing Organization Name and Address

Vehicle Analysis and Simulation Laboratory
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061-0118

10. Work Unit No. (TRAIS)

TRAIS

11. Contract or Grant No.

DTNH-22091-Y-07266

12. Sponsoring Agency Name and Address

Office of Crash Avoidance Research
National Highway Traffic Safety Administration
400 Seventh Street, SW
Washington, DC 20590

13. Type of Report and Period Covered

Three-year Report
September 24, 1991 to
December 23, 1994

14. Sponsoring Agency Code

15. 5upplementary Notes

16. Abstract

This report summarizes the results of a 3-year research project to develop reliable algorithms for the detection of motor vehicle driver impairment due to drowsiness. These algorithms are based on driving performance measures that can potentially be computed on-board a vehicle during highway driving, such as measures of steering wheel movements and lane tracking. A principal objective of such algorithms is that they correlate highly with, and thus are indicative of, psychophysiological measures of driver alertness/fatigue. Additional objectives are that developed algorithms produce low false alarm rates, that there should be minimal encumbering of (interference with) the driver, and that the algorithms should be suitable for later field testing. This report describes driving simulation and other studies performed to develop, validate, and refine such algorithms.

17. Key Words

Drowsy Driver Detection, Driver Performance Monitoring, Asleep-at-the-Wheel, Crash Avoidance Countermeasures, Driver Fatigue, Driving Simulation

18. Distribution Statement

This document is available to the public through the National Technical Information Service, Springfield, VA 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages 219	22. Price

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

Page

ACKNOWLEDGMENTS

The authors wish to thank Dr. Ron Knipling of NHTSA, who served as the contract technical representative. Dr. Knipling contributed many suggestions throughout the project and also provided results from his accident data base studies for inclusion in the first semi-annual research report and in experiment planning. In addition, Dr. Knipling handled contractual matters promptly, efficiently, and constructively.

The authors also wish to thank Michael Perel, Dr. Michael Goodman, Robert Clarke, and William Leasure of NHTSA for helpful suggestions throughout the project. Mr. Leasure additionally made it possible for the authors to present the results of the research to large technical audiences both in Europe and the United States.

EXECUTIVE SUMMARY

The purpose of this report is to summarize the results of a three-year study in which the main objective was to develop reliable algorithms for the detection of driver impairment due to drowsiness. More specifically, the goal of the research was to develop the best possible algorithms for detection of drowsiness, based on measures that could be computed on-board in a vehicle. Additional objectives were that developed algorithms would produce low false alarm rates, that there should be minimal encumbering of (interference with) the driver, and that the algorithms should be-suitable for later field testing. This report describes the various studies that were performed to develop, validate, and refine such algorithms. Included are chapter summaries of the six preceding semi-annual research periods, summaries of additional supplemental research, and remarks concerning future research in regard to implementation of a full-scale driver-drowsiness detection and alerting system. Because of the large amount of research and documentation generated during the three year period, this report necessarily represents an overview. Ordinarily, this document would have been called a final report. However, the project has been extended, Therefore, to avoid confusion, this report is called a "Three-Year Report."

This report is comprised of eight chapters. The first six correspond to summaries of the six semi-annual research periods, as shown in Figure 1. The seventh chapter describes three additional analyses performed late in the project that were directed at further refinement of algorithms and gathering of additional information about their effectiveness. The eighth chapter consists of a final summary of findings and recommendations. The remainder of this executive summary provides a brief description of each chapter of this three-year report.

Chapter One (First Semi-Annual Research Period; Wierwille, Wreggit, and Mitchell, 1992)

This chapter contains a brief review of motor vehicle accident data bases for characteristics of drowsy driver accident scenarios and a review of the drowsiness related literature. There were three specific objectives in performing this review. The first was to provide information about scenarios most likely to lead to drowsiness-related accidents. The

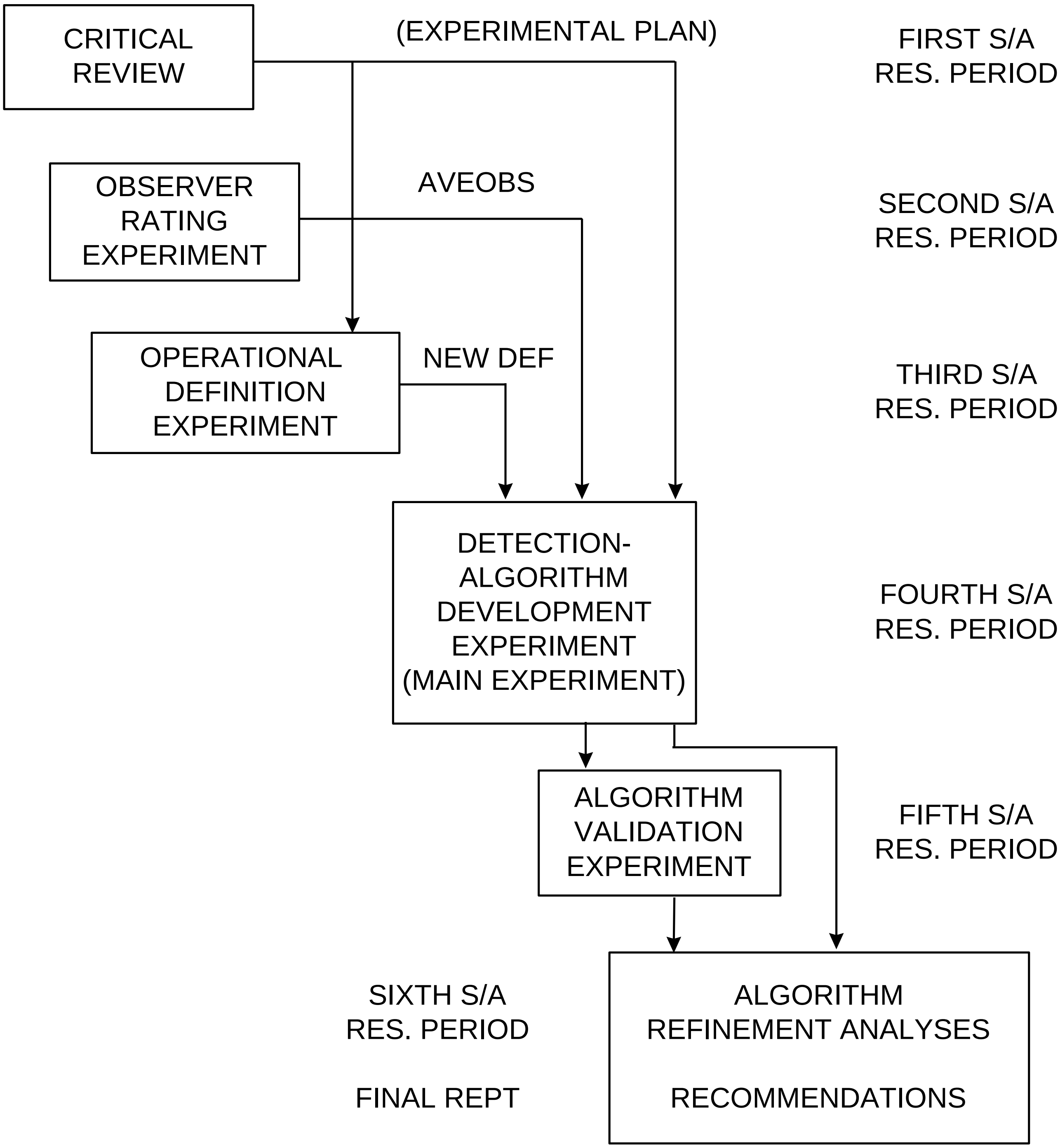


Figure 1: Project Overview

second was to determine which operationally-obtainable measures are believed to covary with the level of drowsiness, and the third was to determine how drowsiness level should be defined. All of these information gathering tasks were directed at developing the best experimental plan for drowsiness detection algorithm development.

Chapter Two (Second Semi-Annual Research Period; Wierwille and Ellsworth, 1992)

One of the findings of the literature review was that insufficient information existed on defining the level of drowsiness of drivers in a practical way. Therefore, efforts were directed toward the development of an operational definition based on ratings by informed observers (persons familiar with the behavior of drowsy individuals). For this study, informed observers were to rate the drowsiness level of drivers based on videotaped facial images. Such videotapes already existed from previous experiments and could be used for this preliminary definitional study. The experiment used six behaviorally trained raters who received a “drowsiness definition” statement and a rating scale. After reading the drowsiness definition statement they performed ratings of48 segments of drivers in various stages of drowsiness.

Factors considered in the data analysis were test-retest reliability, inter-rater reliability, intra-rater reliability and sensitivity. It was determined that agreement between raters and within raters was quite good and reliability levels were quite high. Thus, observer rating or averages of observer ratings can be used to define the level of drowsiness of drivers on a minute-by-minute basis.

Chapter Three (Third Semi-Annual Research Period; Ellsworth, Wreggit, and Wierwille, 1993)

This study also focused on defining the level of drowsiness of drivers. While its main purpose was the development of one or more additional definitional measures, it also allowed the preliminary check of hardware needed for the detection algorithm development experiment that was to follow.

The fundamental concept of this study was to attempt to use a “blend” of eye-closure and other physiological measures to predict performance in cognitive tasks. If such a blend could predict performance under a variety of apparent drowsiness levels, then it could serve as an alternative definition of drowsiness level.

Eight sleep deprived subjects performed two interleaved (non-driving) tasks, one being a lower level cognitive task (simple visual search task), the other being a higher level cognitive task (mental arithmetic). By exposing the subjects to these two tasks, it was possible to determine performance decrements in both lower-level functions and higher-level functions. All subjects had been awake for at least 17 hours before the experiment began.

Results of the experiment showed that a good definition of drowsiness level (as defined by performance decrements on the tasks) could be obtained by combining eye closure, electroencephalogram (EEG), and heart rate measures in a linear multiple regression model. These measures would not be difficult to obtain in an automobile simulation since they can be obtained without discomfort or intrusion. Therefore, a regression model could be used as an alternate definition of drowsiness level.

Chapter Four (Fourth Semi-Annual Research Period; Wreggit, Kim, and Wierwille, 1993)

This study was directed at the central objective of the project, namely, at developing a wide variety of usable algorithms for detection of driver drowsiness. The dependent measures in this study were definitional measures of drowsiness that were not considered to be operationally obtainable in an actual vehicle. The independent measures in this study were operational measures that would be obtainable in an actual vehicle. The objective was to find optimum combinations of independent measures that would best predict levels of drowsiness.

The independent measures collected during this study included *driving-related* measures, *driver-related* measures (determined by Ellsworth, Wreggit, and Wierwille, 1993), and *secondary task* performance measures. The various measures were used to create algorithms for the detection of drowsiness while driving. The detection algorithms were developed through the use of multiple regression analyses.

The dependent measures collected during the study included two eyelid-closure measures, the average observer rating developed by Wierwille and Ellsworth (1992), and an operational definition of drowsiness developed by Ellsworth, Wreggit, and Wierwille (1993). A measure that was comprised of the standardized sum of the above dependent measures was also used as an operational definition of drowsiness.

Twelve sleep deprived subjects drove an automobile simulator from approximately 12:30 A.M. to 3:00 A.M. Four subjects performed a secondary task, four subjects manipulated dash controls, and four subjects simply drove.

The secondary task ("A/O task") consisted of a task that involved an auditory presentation of simple words every fifteen seconds. If the presented word contained an "A" or "O" the subject was to press the button labeled "YES" located on the steering wheel. If the presented word did not include an "A" or "O" the subject was to press the button labeled "NO" located on the steering wheel. This task demonstrated performance of low cognitive-load tasks.

The task of manipulating various controls on the instrument panel involved following auditory commands to adjust radio controls, push buttons, and vertical slide controls every eight to ten minutes. The manipulation of the controls by some of the subjects was for the reason of introducing factors that may be experienced by drivers on an actual roadway.

Performance measures, behavioral measures, and physiological measures were collected and analyzed through the use of multiple regression-and discriminant analysis. These measures consisted of the five dependent measures (definitions of drowsiness) and thirty-three independent variables. Multiple regression analyses were undertaken to determine which independent variables were significant predictors of drowsiness. Discriminant analyses employed the sets of independent variables that were found through multiple regression to be significant predictors of drowsiness. The results showed that multiple regression was as accurate as discriminant analysis in classifying levels of drowsiness. Since multiple regression analysis does have some inherent advantages over discriminant analysis when dealing with detection algorithm development and use, it was decided that all algorithms would be developed using multiple regression techniques. Analyses were then undertaken to determine which independent variables were significant predictors of drowsiness.

Typical algorithms contained four to seven measures and corresponding coefficients (weightings). A typical good algorithm consisting of steering-related, lateral accelerometer-related, and lane-related measures produced an R value of 0.87. (See Figure A8, Appendix A.) The R value indicates the correlation between the actual drowsiness measure and the algorithm output (predicted measure). An accuracy rate of 79% was attained when *all* misclassifications were considered (i.e. observed alert, questionable, and drowsy segments being classified in any erroneous category) and a 98% accuracy rate was attained when only *large* misclassifications were considered (i.e. observed alert segment being erroneously classified as drowsy or vice versa). Thus, large misclassification error rates of 2% to 3% are likely to occur. Classification accuracy rates attained during the algorithm development study and subsequent algorithm validation study reflect accuracy rates for algorithms when applied to partially sleep deprived driver-subjects. It is believed that false alarm rates would be lower for alert drivers because, if the level of actual drowsiness is very low (drivers are alert), the detection-algorithm output will not (in a great majority of cases) erroneously exceed the predetermined “impairment threshold.” In other words, the actual level of drowsiness in an alert driver will be so far from threshold that it is unlikely that a misclassification would occur.

To further minimize false alarms, a two-stage detection system could be used. The first stage would detect probable drowsiness based on driver and vehicle related measures, and the second stage would further discriminate using a secondary task.

Chapter Five (Fifth Semi-Annual Research-Period; Wreggit, Kim, and Wierwille, 1994)

This experiment was conducted with the primary purpose of algorithm validation, that is, determining algorithm classification accuracy for data from a new set of driver-subjects. While estimates of algorithm accuracy were obtained along with the algorithms when they were developed, it was not certain that such estimates could be relied upon for new groups of drivers operating under similar conditions. Therefore, it was deemed necessary to apply typical developed algorithms to a new data set for the purpose of obtaining a "validated" estimate of accuracy.

In this experiment, twelve driver-subjects drove the moving-base, computer-controlled simulator at Virginia Tech while measures used in the previously derived algorithms were gathered. Subjects were kept awake until approximately 12:15 A.M. when they were placed in the simulator. They drove the simulator until about 3:00 A.M.

The conditions used in this experiment were similar but not identical to those of the algorithm development (earlier) experiment. The reasons for using a slightly modified design in the validation experiment were:

1. to determine algorithm accuracy under similar, but not identical conditions, thereby "simulating" the likely conditions of an application, and
2. to use the data for additional purposes, such as determining the effects of cruise control on algorithm detection accuracy.

Typical previously-developed algorithms were selected and then tested for detection accuracy on the measures. The accuracy analysis was divided into two major categories. The first category was for algorithms based solely on driver-vehicle performance measures, and the second was for algorithms including A/O task performance measures. The second category contained half the data of the first category since subjects performed the A/O task only in two of the four quartiles during the data collection run.

The results of the driver-vehicle performance category showed that on the average there was no appreciable degradation in algorithm accuracy when the algorithms were applied to the new data. Error classification matrices similarly showed no degradation. When the various quartiles were compared, it was found that some minor variations in algorithm accuracy did occur.

The results of the A/O task performance category showed that there was a reduction in algorithm accuracy, as evidenced by R values, when the algorithms were applied to the new data. On the average, R values were reduced from 0.81 to 0.61. Surprisingly, however, classification matrices did not exhibit a corresponding reduction in accuracy. In other words, the ability of the algorithms to classify correctly remained reasonably high.

The results of the validation study make it possible to draw several important conclusions about drowsy driver detection. They are as follows:

- There was no degradation in detection accuracy for previously developed driver-vehicle-performance algorithms when they were applied to new data.
  - Both R values and classification matrix accuracies maintained their values.
  - The use of twelve representative subjects is therefore probably sufficient to characterize algorithms for general use.
- There was a degradation in R values for previously developed algorithms that included A/O (secondary) task measures, when the algorithms were applied to new data. The drop in value averaged 0.2. However; classification matrices did not exhibit a correspondingly large decrease in accuracy. Instead, their reduction in accuracy was small.
  - The reduction in R values is probably a result of using only four subjects to develop the algorithms, or possibly a result of the limited number of bouts of drowsiness in the new data.
  - R values for validation results may underestimate the capabilities of algorithms to classify correctly, when a small subject sample is used to develop the algorithms.
- This experiment has shown that appreciable losses in accuracy do not occur when appropriately-developed drowsy-driver detection algorithms are applied to similar new data. Therefore, the algorithms retain their ability to detect drowsiness.
- Any future algorithm development work should be based on twelve or more representative subjects, and both R values and classification matrices should be evaluated.
- The results of this experiment generally support the feasibility of drowsy-driver detection.

Chapter Six (Sixth Semi-Annual Research Period; Kim, Wreggit, and Wierwille, 1994)

This study employed the same set of data collected during the validation phase of the main study. As indicated previously, the validation study was planned in such a way that additional analyses could be conducted. Three additional issues were examined using this set of data, including: 1) an analysis of how forward-velocity measures covaried with level of drowsiness, and whether or not they could be used to improve algorithm detection accuracy, 2) an investigation of whether the performance of a secondary-task had an alerting effect on drivers and, 3) determination of whether or not cruise-control increased levels of driver drowsiness.

Use of Forward velocity-related measures. Velocity-related measures were used from the non-cruise control segments of each driver' s data run. The results of correlations between these measures and the five definitional measures of drowsiness gave an indication of the relationship between longitudinal measures and drowsiness. Results suggest that the relationship between velocity-related measures and drowsiness is moderately strong only when drivers are not stimulated by a secondary task. In other words, under the dullest of driving conditions, there is a moderately strong relationship. Otherwise, the relationship is weak.

To determine whether velocity-related measures contributed significantly to algorithm detection accuracy, algorithms were developed both with and without velocity-related measures so that direct comparisons could be made. Results showed that velocity-related measures contributed only minimally. Similarly, error classification matrices showed only the slightest improvements when velocity-related measures were included in the algorithms. It must thus be concluded that velocity-related measures do not provide a substantial increase in drowsiness detection accuracy.

Effects of Cruise Control, Secondary Task, and Velocity-Related Measures on Driver control. Results suggest that there is no strong alerting effect of the A/O task on level of drowsiness, and similarly, there is no strong drowsiness inducing effect of cruise control usage on level of drowsiness. Although not conclusively demonstrated by the present experiment, there are indications that the very dullest of conditions (no A/O task and cruise engaged) caused increases in drowsiness level and decreases in lane-keeping performance. If this hypothesis is indeed correct, then stimulating the driver by any means should be helpful to a degree in maintaining alertness.

Chapter Seven; Part One (Fairbanks and Wierwille, 1994)

This study focused on the effects of using higher-order (non-linear) algorithms on drowsy-driver detection accuracy. Measures from the development phase (Wreggit, Kim, and Wierwille, 1993) and validation phase (Wreggit, Kirn, and Wierwille, 1994) were squared or multiplied with each other to obtain cross products. The second order terms, combined with first order terms, were used to calculate predictive algorithms using data from the development phase. The developed algorithms were then applied to the validation data.

The results of this study suggest that the use of second-order terms in driver-drowsiness-detection algorithms does not result in detection accuracy improvement. Although not conclusively proven by the present study, the results do support the hypothesis that higher-order algorithms produce more and larger outliers when applied to new data than do linear algorithms. When outliers resulting from algorithm outputs (predictions) were limited to the maximum and minimum values of the observed scores (in other words, the outliers were "clipped" from the data and set to a value equal to the largest and smallest observed data) the R values increased on average from 0.612 to 0.800.

Chapter Seven: Part Two (Wreggit and Wierwille, 1994a)

Various A/O task algorithms were developed and validated in previous phases of this study (Wreggit, Kim, and Wierwille, 1993 and Wreggit, Kim, and Wierwille, 1994). It was found that the R values in the validation phase (using new A/O data) were significantly lower than the R values obtained in the development phase (using original A/O data). When the results of the Fairbanks and Wierwille (1994) study became available, it was felt that the significant decrease in R values from the development phase (using A/O data) to the validation phase (using A/O data) in the main study could have been due to the effects of prediction outliers.

The purpose of this follow-up study was to examine the possibility of potential improvement in multiple-R values by limiting the output (prediction) values obtained from previously developed and validated A/O task algorithms. The algorithm output values were limited to the minimum and maximum values of the corresponding observed data. Thus, no outliers were present when the subsequent correlation analyses were run. A comparison of R ‘values obtained from analyses of original data (main study: development phase), new data (main study: validation phase), and “clipped” data were examined.

The outliers present in the prediction data were very limited in number and in magnitude. It was concluded from the results of this study that the significant decrease in A/O task algorithm R values in the validation phase was not a result of outliers. Instead, it is most likely that the use of only four subjects in A/O task algorithm development limited the prediction capabilities somewhat.

Chapter Seven: Part Three (Wreggit and Wierwille, 1994b)

In developing and validating algorithms for drowsiness detection, drivers were purposely subjected to partial sleep deprivation and driving in the early morning hours. False alarm rates obtained would thus reflect those corresponding to such drivers, and not drivers who are fully alert. Therefore an additional analysis was performed to assess the alert-driver false-alarm rate. To accomplish this task, data from the algorithm development phase were screened for segments in which drivers were fully alert (that is, alert based on the definitional measures). These segments were extracted and then new R values and classification matrices were computed (for the extracted data). The results showed large decreases in R values and large improvements in classification accuracy.

It was concluded that if drivers are fully alert, 1) R values will not accurately reflect the detection capability of a given algorithm, and 2) there will be substantially fewer false alarms than earlier classification matrices would indicate.

Chapter Eight

This chapter includes a summary of findings and recommendations for future research. The reader is referred directly to this chapter for a summary.

Appendix A

This appendix contains regression summaries and classification matrices for selected algorithms. The coeffkients (B-values) in the regression summaries specify the weighting that should be used when algorithms are employed in an application.

Chapter One

Literature Review

(This chapter is based on material drawn from the First Semiannual Research Report, dated April 10, 1992 and referred to as Wierwille Wreggit, and Mitchell, 1992. The material presented here has been updated based on recent findings in the literature)

INTRODUCTION

This chapter contains 1) a brief review of motor vehicle accident data bases for characteristics of drowsy driver accident scenarios and 2) a review of the drowsiness related literature.

The purpose of examining accident data bases was to provide information about scenarios most likely to lead to drowsiness-related accidents. The purpose of reviewing a wide variety of literature pertaining to past driver-drowsiness and general-drowsiness research was to determine 1) which operationally-obtainable measures are believed to covary with the level of drowsiness and 2) how drowsiness level should be operationally defined. These information gathering tasks were directed at developing the best experimental plan for drowsiness detection algorithm development.

ACCIDENT SCENARIOS

One of the leading causes of single- and multiple-car accidents is driver impairment due to drowsiness (Office of Crash Avoidance Research, 1991). Unfortunately, drowsiness while driving may be perceived as less of a problem than it actually is because of the difficulty of attributing drowsiness as a cause of an accident. However, as more research is being conducted concerning drowsy drivers, it is becoming indisputable that a major problem does exist. Also, there may be many more incidents in which the initial cause of the loss of control of a vehicle is drowsiness yet is reported to be caused by something other than drowsiness.

A study conducted in 1973 at Duke University by Tilley, Erwin, and Gianmrco provides evidence that drowsiness while driving is an all too common occurrence. In this study two experimenters, stationed in the Durham, North Carolina Department of Motor Vehicles, administered questionnaires pertaining to driving habits and behaviors to 1500 individuals who were successful in renewing their driver's license. Of those 1500 people, 64% responded that they had, at one time or another, become drowsy while driving. Also, over 7% responded that they had gone to sleep for short periods while driving. Of those who answered that they had had trouble with drowsiness while driving, 3 1.2% responded that they had become drowsy before they were aware of their condition. Of those who did experience drowsiness while driving, approximately 10% reported that they had been in one or more accidents due to drowsiness or falling asleep at the wheel. Another 10% responded that they had been in a near accident due to drowsiness.

A survey completed in 1980 by the Kanagawa Prefectural Police, based on questionnaires collected near the Tokyo-Nagoya Expressway, shows that approximately 75% of the drivers admitted to being sleepy while driving (Seko, 1984). Unfortunately, Seko does not give an indication in his article concerning the extent of the drowsiness experienced by the polled drivers. Seko also cites a survey of the causes of accidents on the Tokyo-Nagoya Expressway since 1969. He found that most of the rear-end collisions at night were attributable to drowsiness. A survey cited by Seko (1984) which was carried out by the

Shizuoko Prefectural Police states that in 1973 nine percent of all traffic accidents were caused by drowsiness and 45% of all deaths were due to drowsiness. Planque, Chaput, Petit, Tarriere, and Chabanon (1991) report that fatigue is the cause of 26% of the fatal accidents occurring on the highways in France.

A recent NHTSA Research Note (Knipling and Wang, 1994) summarized available national statistics for the years 1989-93 based on General Estimates System (GES) data representative of all crashes and Fatal Accident Reporting System (FARS) statistics on fatal crashes. A summary of their findings is as follows:

- There were an average of 56,000 police-reported crashes in which driver drowsiness/fatigue was cited (0.9 percent of all crashes).

- An annual average of 1,542 fatalities were associated with these crashes (3.6 percent of all fatalities).

- These crashes resulted in an estimated 40,000 non-fatal injuries (all non-fatal severity levels).

- Due to underreporting, all of the above statistics are regarded as conservative.

Statistics on crash characteristics of "drowsiness-cited" cases indicated the following:

- Drowsy driver crashes peak in the early a.m. hours with a second smaller peak in the afternoon. Fifty five (55) percent occurred between midnight and 7:59am, and another 18 percent occurred between 1:00 and 4:59am.

- Most occurred in non-urban areas, generally on roadways with 55-65 mph speed limits.

- Eighty (80) percent were single-vehicle crashes or collisions- with parked vehicles. An additional 6.6 percent were subject vehicle-striking rear-end crashes.

- In 76 percent of crashes the driver was the only occupant of the subject vehicle.

- Fifteen (15) percent of drowsiness/fatigue, crashes also involve alcohol.

- Involvement is strongly related to both driver sex and driver age. For the 1989-93 period, 76 percent of subject drivers were male, and 59 percent were under the age of 30.

- In addition to young male drivers, commercial (i.e. long-haul truck) drivers are at risk, primarily due to their high mileage exposure.

REVIEW OF THE DROWSINESS LITERATURE

This section examines previous studies that have focused on physiological measures, driver-performance measures, and behavioral measures. The overall purpose of this section is to present those operational indicators or drowsiness which have shown promise in detecting driver drowsiness and are either currently obtainable or may soon be obtainable on-the-road. This information will be used to formulate a set of measures which should be incorporated into the simulator testing and the development of the detection algorithms.

Many measures have been examined as predictors of driver impairment and can be divided into two basic categories: objective measures and subjective measures. In general, objective measures have demonstrated greater promise as predictors of driver impairment than subjective measures (Dingus, Hardee, and Wierwille, 1985). A subclassification exists within this objective measure category and is comprised of physiological measures and performance measures. In large part, performance measures have shown potential both in terms of driver impairment prediction, as well as practicality for on-the-road implementation (Dingus et al., 1985). Conversely, physiological measures typically cannot be obtained on-the-road in a manner that is feasible but are of interest since they can be predictive in nature concerning the onset of drowsiness.

A study conducted in 1984 by Skipper, Wierwille, and Hardee found results that indicated that it was possible to detect the onset of driver drowsiness by observing drivers' reactions to steering wheel torque and front wheel disturbances produced by the automobile simulator. However, while subjects were involved in a normal driving scenario the experimenters found that it was also possible to predict the onset of drowsiness. Several variables were examined, but eyelid closure was the most consequential.

Dingus, Hardee, and Wierwille (1985) performed a study that examined the effects of drowsiness on driver performance. Dingus, et al. employed both sleep deprived subjects and a control group consisting of the same subjects in a rested condition. The sleep-deprived runs took place from 2:00 a.m. to 3:30 a.m. The initial analyses of the collected data were correlation analyses between the eyelid closure measures and lane position measures. The lane position measures were indicators of driver impairment while the eyelid closure measures were indicators of drowsiness. Eyelid closure was recommended by Erwin (1976) since it has been found that eyelid closure is a very stable physiological indicator of drowsiness. It was found that a relatively high correlation between eyelid and lane position measures was present, as seen in Table 1.

Dingus et al. (1985) ran a second set of correlation analyses between the indicators of driver impairment, which included eyelid closure and lane position measures, and other dependent measures. Dingus et al. state that any measure that demonstrated reasonably consistent correlations across the impairment indicators of approximately 0.25 or greater was considered promising. The potentially reliable impairment detectors based on correlation analyses run by Dingus et al. concerning drowsiness can be seen in Table 2. Table 3 shows drowsiness impairment indicators and associated classification matrix for six-minute interval data from the Dingus et al. study. The six-minute interval data were found to provide slightly better discrimination of drowsiness-induced impairment.

It was found through stepwise discriminant analyses that YAWMEAN, YAWVAR, STEXEED, SEATMOV, and LANDEVSQ contained significant independent detection information. By employing eyelid closure as a definition of drowsiness it was possible for Dingus et al. to create several models of driver impairment based upon driving performance. This was an important development since the performance measures could be unobtrusively implemented using an in-car drowsiness detection system. Performance measures will be discussed in more detail in the *Driving Performance Measures* section of this chapter.

In the study conducted by Dingus, Hardee, and Wierwille (1985), EYEMEAS seems to be affected by degree of sleep deprivation as well as time on task. However, when compared to EYEMEAS data found in the Hardee, Dingus, and Wierwille (1985) study it becomes apparent that time on task may be as important as the degree of sleep deprivation.

Table 1: Eye Measure vs. Lane Measure Correlations. (From Dingus, Hardee and Wierwille, 1985)

	EYEMEAN	EYEMEAS	PERCLOS
LANEX	.47	.54	.62
LANEDEVV	.50	.55	.60
LANEDEVSQ	.55	.59	.60
LANEDEV4	.36	.40	.40

- EYEMEAN: Mean eyelid closure (zero = wide open)
- EYEMEAS: The mean-square percentage of the eyelid closure signal.
- PERCLOS: Proportion of time that the eyes are 80% to 100% closed.
- LANEX: Count of the number of samples taken while the simulated vehicle was out of the lane.
- LANEDEVV: Lane position variance.
- LANEDEVSQ: Weighted lane deviation. Heavier weighting away from the center of the lane by a squared function.
- LANEDEV4: Heavily weighted lane deviation. Heavier weighting away from the center of the lane by a fourth power function,

Table 2: Impairment Detectors Based on Correlation Analysis (From Dingus, Hardee and Wierwille, 1985)

YAW-VAR
STEXEED
STVELVAR
LGREV
SEATMOV
HRTRTM
HRTRTV

- YAWVAR: Yaw deviation variance.
- STEXEED: Count of steering velocity occurrences over 150 degrees per second.
- STVELVAR: Steering velocity variance.
- LGREV: The number of times the steering wheel position increment exceeds 5 degrees (after steering wheel velocity passed through zero).
- SEATMOV: Seat movement counter.
- HRTRTM: Heart rate mean.
- HRTRTV: Heart rate variance.

Table 3: Drowsiness Impairment discriminant Analysis. Six-Minute Interval Data -- Best Results (From Dingus, Hardee and Wierwille, 1985)

		Predicted		
		Impaired	Not Impaired	
Actual	Impaired	20	8 (28.57%)	28
	Not Impaired	4 (2.63%)	148	152
		24	156	180

APER = 6.7%

Model Variables:

YAWVAR	.7692
SEATMOV	.6218
LANDEVSQ	.4152
YAWMEAN	.2460
STEXEED	-.0292

In Figure 2 a comparison of the EYEMEAS data from both studies with starting times aligned is presented.

Stages of Sleep

To understand the terminology concerning drowsiness and sleep a discussion of the stages of wakefulness will be presented. An understanding of the various stages of sleep is important when carrying out studies that examine the physiology, psychology, or behavior of sleep deprived subjects. It should be noted however, that while driving an automobile, a person will most likely be at one of two stages of wakefulness -- either stage W or stage 1 sleep. Below is a summary explanation of the various stages of sleep. The descriptions of the stages is taken, in part, from Carskadon (1980).

- Stage W sleep. Stage W does not actually describe a sleep state but rather a state of wakefulness. This stage is usually accompanied by a relatively high tonic EMG. Rapid eye movements and eye blinks are present in this stage.

- Stage 1 sleep. This is the stage that intervenes between wakefulness and other sleep stages. In most subjects, the duration of stage 1 sleep usually is not longer than several minutes. Stage 1 sleep may occur after large movements of the body which are caused by the relaxation of the muscles in the body of a person. entering this stage of wakefulness. Stage 1 sleep following wakefulness is often accompanied by slow eye movements. Each slow eye movement may be several seconds in duration. Rapid eye movements are absent at this stage. Tonic EMG levels are commonly below EMG signals of individuals in a relaxed but wakeful state.

- Stage 2 sleep. No eye movements are usually seen in stage 2 sleep. Stage 2 sleep can last as long as one hour and become interspersed with periods of REM sleep.

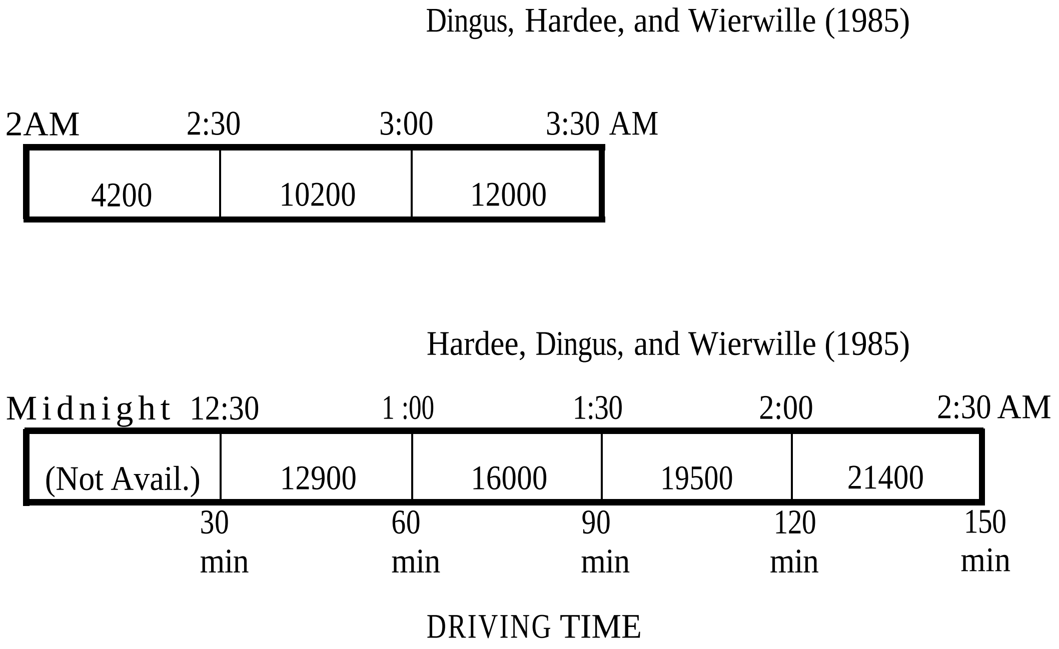


Figure 2: Comparison of EYEMEAS Values for Two-Experiments with Start-Times Aligned.

- Stage 3 sleep. Stage 3 sleep is a transitional stage between stage 2 and stage 4 sleep early in the night. Sometimes stage 3 sleep will not be followed by stage 4 if it occurs in the NREM portion of the sleep cycle.

- Stage 4 sleep. Stage 4 sleep usually occurs during the first third of the night. This stage is characterized by a predominance of high amplitude slow brain waves.

- REM sleep. This stage usually occurs within the first 100 minutes after sleep onset. REM sleep is characterized by low voltage, mixed frequency EEG, bursts of rapid eye movement (REM) and low amplitude EMG.

Promising Physiological Measures Found in the Literature

The purpose of this section is to discuss measures that may lead to a more refined, operational definition of drowsiness and the onset of stage 1 sleep or drowsiness. A review of the sleep and drowsiness literature has been conducted and the most likely measures to be successfully employed in the refinement of the operational definition of drowsiness, that may eventually aid in the detection of the onset of drowsiness in various applications, are discussed below.

Eyelid closure. Eyelid closure has been found to be a very reliable predictor of the onset of sleep (Erwin, 1976) and degraded task performance (Dingus, Hardee, and Wierwille, 1985; Hardee, Dingus, and Wierwille. 1985; Skipper, Wierwille, and Hardee, 1984). Erwin examined various measures to determine whether they were predictive of sleep onset, including plethysmography, respiration rate, electroencephalography (EEG), skin electrical characteristics, electromyography (EMG), heart rate variability, and eyeiid closure. It was found that eyelid closure was the most reliable predictor of the onset of sleep among the dependent measures examined. Eyelid closure is indicative of sleep onset and undoubtedly the cause of poor performance in visual tasks, especially tracking tasks such as driving. It seems quite obvious that if a driver's eyelids are closed, the ability to operate a vehicle would be greatly hampered.

Skipper, Wierwille, and Hardee (1984) examined the ability of sleep deprived drivers to perform a one and one half hour driving task. Various disturbances were purposely input into the steering system of the driving simulator to mimic on-the-road conditions. It was found that performance measures such as lane deviation, yaw deviation, and steering velocity were highly correlated with eyelid closures.

The apparatus used to capture eyelid closures in the Dingus et al, Hardee et al., and Skipper et al. studies was a low-light level camera. A linear potentiometer was used by an experimenter to track the eyelid movement of the subjects manually.

Eye movement. There are two general reasons that one may desire to record eye movements during sleep or before sleep. First, a principal sign of REM sleep is the phasic burst of rapid eye movements. Second, the onset of sleep in most subjects is heralded by or accompanied by slow, rolling eye movements (Carskadon, 1980).

Slow, rolling eye movements may accompany the onset of sleep or are precursors of sleep onset. This phenomenon also occurs with the transition to stage 1 sleep during the night. The characteristics of human eye movements change greatly with alertness level. Slow eye movements (SEMs) prove to be one of the most characteristic signs of the phase of transition between wakefulness and sleep (Planque, Chaput, Petit, and Tarriere, 1991). Dement (1975) states that the SEM event is a crucial occurrence in the sleep onset process.

Slow, lateral eye movements are quite different than eye movements typically seen in a person who is fully awake. A completely awake individual can be observed as having quick eye movements. As subjects become drowsy their eyes move in a pendular motion from left to right (Hiroshige, and Niyata, 1990) and the number of quick, voluntary movements of the eyes begins to lessen. Endo, Inomata, and Sugiyama (1978) found that attentiveness begins to disappear in conjunction with drowsiness due to the lessened number of lateral voluntary eye movements that would normally be used in a driving situation to check the rear view mirrors, side windows, etc. In other words, as alertness decreases, attentiveness also decreases. Numerous SEMs are detected during stage 1 sleep, but they also appear during the long period separating waking from sleep (Hiroshige, and Niyata, 1990). Torsvall and Akerstedt (1988) noted that the proportion of SEMs increased sharply with the advent of drowsiness on train drivers making long trips. Convergence of the eyes is also possible when a person becomes drowsy.

Electrooculography (EOG) involves the measuring of eye movements via electrodes in contact with the skin surrounding the eyes. The process of measuring eye movements with EOG is quite simple due to the electrical nature of the human body. In the eyeball, there is a small electropotential difference from the front to the back. The front (cornea) of the eye is positive with respect to the back (retina) of the eye.

Before a certain point in a person's awake but drowsy state, SEMs do not exist. However, after a particular moment in the onset of sleep, slow, rolling, lateral, ocular movements create sinusoidal activity in the EOG (Dement, 1975). On the EOG signal, the SEMs are translated by slow deflections lasting more than a second. It is likely that amplitudes of at least 100 microvolts will be seen (Torsvall and Akerstedt, 1988). The EOG waves that are normally observed are moderate in amplitude initially, but increase with the degree of drowsiness (Santamaria and Chiappa, 1987).

Planque et al. (1991) found that after several minutes of driving only blinking and glances at simulator instrumentation were recorded. Approximately 30 minutes into the study deterioration of deliberate eye movement was seen. Planque et al. state that by analyzing the EOG, it is possible to follow clearly the deterioration of alertness.

<u>Muscle activity</u>. Sleep onset may be accompanied by the reduction of muscle activity, or muscle tonicity, especially in the facial muscles. However, Erwin (1976) states that measures of muscle activity offer essentially no predictive information pertaining to sleep onset and that significant sleep can occur for several minutes prior to any significant change in muscle tone. Unfortunately, it is not clear as to which muscle groups he examined.

A study conducted by Yabuta, Iizuka, Yanagishima, Kataoka, and Seno (1985) demonstrates that facial expression is effected by drowsiness. Yabuta et al. state that special attention was focused on subjects' facial expression, among other physiological measures, because facial expression is known to vary according to the alertness level of the subject. Observing the muscle activity which causes the changes in facial expression with drowsiness is one method of quantifying this measure.

Electromyography (EMG) is a common method used for recording muscle activity. Often times, EMGs are used to evaluate various sleep and muscle disorders. EMG measures of facial muscles may be an efficient method of quantifying facial expression, or more specifically, facial muscle tone.

Hauri (1982) demonstrates that EMG recorded on the chin steadily, though not dramatically, decrease as a person nears stage 1 sleep. Even when a subject is totally relaxed, small muscle potentials will be seen (Carskadon, 1980). This is due to the fact that every muscle is composed of many contractile fibers that are innervated by nerves. When a muscle fiber is activated through nerve innervation, a change in the electrical potential is seen. When the muscle is relaxed, fewer nerves discharge, thus a smaller EMG potential is recorded.

Brain wave activity. Sleep produces distinctive alterations in the amplitude and frequency of the signals from the brain. Erwin (1976) states that there is no reliable alteration in background brain activity prior to eyelid closure. Upon eyelid closure Erwin found that a very rapid shift in brain wave patterns takes place. This shift is identifiable as stage 1 sleep. However, Planque et al. (1991) states that sharp changes in the frequency content of brain wave activity are observed during the crossing from alertness to a stage of hypoalertness. then to drowsiness, and finally to sleep. A slowdown of the cerebral activity, in general, an increase in the percentage of alpha waves and, in turn, a decrease in the percentage of beta waves, is observed at the same time that a decline in performance is seen. Seko (1984) reports that alpha waves appear during decreased alertness such as absentmindedness or "cloudy consciousness." Seko cites the work of Kuroki, Kitakawa, and

Oe (1974) wherein alpha waves were hardly detected at the beginning of a driving session but as the driving session continued and the level of driver/subject alertness decreased, high-amplitude alpha waves occurred frequently. Planque et al. (199 1) suggest that analysis of the beta, alpha, and theta frequencies are the most appropriate for examining/detecting the onset of sleep.

Planque et al. state that automatic processing of the EEG signal has proved very difficult to implement. Presently, various phases of sleep (stage 1, stage 2, REM, etc.) are identifiable via automated methods, however an examination of drowsiness and sleep onset is distinguished by much less distinctive physiological events. Therefore, Planque et al. suggest the manual method for analysis of EEG as well as EOG which was discussed previously.

Skin potential level. The SPL measures the potential difference between the outermost layer of skin (stratum corneum) and the layer immediately below it (stratum lucidurn). In a study by Erwin, Hartwell, Volow, and Alberti (1976) it was found that a correlation exists between changes in skin potential level (SPL) and stages of arousal. In all cases, EEG-defined-sleep occurred only after a shift in skin potential level (Erwin et al., 1976). It was also found that significant shifts in skin potential level preceded not only stage 1sleep but also the transition that occurs prior to stage 1 sleep. In the several minutes following the SPL shift, subjects oftentimes became drowsy as evidenced by decreased performance, frequent eyelid closures of more than one second, and occasionally, EEG manifestations of sleep (Erwin, 1976). Although decreased skin potential negativity was shown to be a prerequisite of sleep onset, decreased potential values preceding sleep onset varied in lengths of time. This fact may indicate that SPL is by no means the only deterministic factor of arousal level (Erwin et al., 1976).

Erwin et al. (1976) discounted the hypothesis that electrodermal shifts are simply a function of time from initial arousal. This was done by observing that spontaneous and evoked EEG arousal was accompanied by a return to waking skin potential levels.

Some obstacles do exist, however, when it comes to applying SPL as a measure of drowsiness. First, recordings of some subjects may give indications of shift changes in SPL without showing drowsy behavior or sleep onset and with no performance decrements seen. As stated earlier, in all cases, EEG-defined sleep occurred only after a shift in skin potential level. From this, it can be hypothesized that SPL shifts must occur for a person to drift into sleep although a shift in SPL is not always followed by sleep onset. Second, there is a considerable variation in baseline values of SPL. This variation can be seen between and within subjects. SPL is susceptible to alterations in subjects' mood, activity level, and temperature.

Heart rate variability. Heart beat interval variability has been found to correlate with drivers' fatigue level (Wierwille and Muto, 1981). As cited in the literature by Wierwille et al., (198 1) Sugarman and Cozad (1972) and Riemersma, Sanders, Widervanck, and Gaillard (1977) found, even greater amounts of variability in heart rate with fatigue. On the other hand, Volow and Erwin (1973) found no correlation between heart rate variability and sleep onset. However, Volow states that in real (or simulated) driving situations there may be sufficient motoric demands on the driver such that the interaction of driving activity may produce significant variations.

Pupil aperture size variability. The pupil serves as a window into central nervous system activity. Spontaneous pupillary movement in darkness in the normal awake individual has been described as reflecting "tiredness," "fatigue," and "sleepiness" (Lowenstein and Loewenfeld, 1963; Lowenstein and Loewenfeld, 1964). The state of the autonomic nervous system has been thought to reflect fatigue and wakefulness for quite some time. For instance, over 200 years ago, pupillary constriction was believed to be associated with sleep (Fontana, 1765). Marked changes in pupillary stability and extent of oscillations have been consistently shown to occur in normal "tired" subjects (Lowenstein and Loewenfeld, 1951; Lowenstein and Loewenfeld, 1963; Lowenstein and Loewenfeld, 1964). Pupillary behavior in individuals suggests that the actions of the pupil do reflect autonomic events and that it is consequently an indirect but accurate indicator of sleepiness or arousal level.

Secondary Task Measures

A study was conducted by Hardee, Dingus, and Wierwille (1985) which employed secondary tasks in a simulator study using sleep deprived subjects. This experiment was run starting at 12:00 a.m. instead of 2:00 a.m. as in the Dingus, Hardee, and Wierwille (1985) study. Hardee et al. found that auditory or visual secondary tasks, along with heart rate variability, predicted quite well whether a subject was impaired or unimpaired due to drowsiness. However, it was found that secondary tasks did not keep the subjects from becoming drowsy.

Subjective Ratings

Observer ratings. Most of the studies that have been carried out rely on the subjective evaluation of drowsiness by the subjects themselves. One study that does, investigate observer rating of drowsiness was carried out by Carroll, Blisewise, and Dement (1989). The results of this study show a high interrater reliability for observations of the sleep-wake cycle of 39 nursing home residents.

Driving Performance Measures as Indicators of Driver Drowsiness

Driving performance measures that can be used to predict the onset of drowsiness are important since it has been shown that sleep loss produces decrements in driving skills (Hulbert, 1972). Driving performance measures include lane-related measures, steering-related measures, and heading- and lateral acceleration-related measures. These measures are obviously important since drivers must maintain proper lane position to avoid vehicles in nearby lanes and objects located on the side of the roadway. The purpose of this section is to discuss various measures used in the past to evaluate driver drowsiness while a subject is actually behind the wheel of an automobile (either simulated or on the road). Measures of performance have potential for driver impairment prediction and are, in some cases, relatively easy to install in an on-the-road vehicle. An overview of performance measures as indicators of driver drowsiness has been addressed by Wierwille, Wreggit, and Mitchell (1992) and will be presented, in summary, below.

Lane-related measures. Several studies have found lateral control measures to be closely related to prolonged driving. Dureman and Boden (1972) found that lane tracking ability degrades as time on task increases over a four-hour period. Several other researchers (Mast, Jones, and Heimstra, 1966; Sussman, Sugarman, and Knight, 197 1) found similar results in that lane position errors increased over a four-hour period. Several lane-related measures have been found to be accurate and reliable measures for the detection of drowsiness, all of which are feasible for on-the-road use. The names of the measures described below are simply the variable names used in previous studies.

- LANEDEVM: Lane deviations which were heavily weighted for lane exceedences were found to be highly correlated with eye closure and were influenced by sleep deprivation and time on task (Skipper, Wierwille, and Hardee, 1984).

- LANESTD: The standard deviation of the lane’position was found to be highly correlated with eye closure and was influenced by sleep deprivation and time on task (Skipper, Wierwille, and Hardee, 1984).

- LANEDEV: the global maximum lane deviation was found to be highly correlated with eye closure and was influenced by sleep deprivation and time on task (Skipper, Wierwille, and Hardee, 1984).

- LANEDEVSQ: The mean square of the lane deviation has been found to contain a significant amount of independent information. The measure is considered to be an accurate and reliable measure for the detection of drowsiness (Dingus, Hardee, and Wierwille, 1985).

- LATPOSM: The mean square of the high pass lateral position (heavily weighted for rapid changes in lateral position) shows potential as a drowsiness indicator (Dingus, Hardee, and Wierwille, 1985).

Steering-related measures. The frequency and type of steering reversal is related to lane tracking. This relation is seen since drivers who are impaired due to drowsiness are typically inattentive to the driving task. As a result, the number of "micro-wheel adjustments" may decrease. Ryder, Malin, and Kinsley (198 1) found that steering reversals decreased in frequency with time on task and Hulbert (1963), cited in Dingus, Hardee, and Wierwille (1985) found that sleep deprived drivers have a lower frequency of steering reversals than rested drivers. Sugarman and Cozad (1972) found that steering magnitude increased with time. Other researchers such as Dureman and Boden (1972), cited in Haworth, Vulcan, Triggs, and Fildes (1989) and Mast, Jones, and Heimstra (1966), cited in Haworth et al. (1989) have found that there is a deterioration of steering performance with drowsiness. Erwin (1976) has also found a reduction of "micro-wheel adjustments" during drowsiness. However, Erwin states that the wheel adjustment measure may not be predictive since EEG signals that indicate the onset of drowsiness precede the change in steering wheel adjustment behavior. Several steering-related measures have been found to be accurate and reliable measures for the detection of drowsiness, all of which are feasible for on-the-road use. The names of the measures described below are simply the variable names used in previous studies.

- STVELM: The steering velocity weighted heavily for fast maneuvers has been found to be highly correlated with eye closure and was influenced by sleep deprivation and time on task (Skipper, Wierwille, and Hardee, 1984).

- STEXEED: The number of times steering velocity exceeded a criterion (150 degrees/second over a three minute interval) was found to contain a significant amount of independent information. This measure is considered to show some potential as a drowsiness indicator (Dingus, Hardee, and Wierwille, 1985).

- STVELV: Steering velocity variance (calculated over a three-minute interval) was found to show potential as a drowsiness indicator (Dingus, Hardee, and Wierwille, 1985).

. LGREV: The number of times the steering wheel position increment exceeded 5 degrees (after steering wheel velocity passed through zero) was found to show potential as a drowsiness indicator (Dingus, Hardee, and Wierwille, 1985).

Heading/head rate/lateral acceleration related measures. Heading errors can quickly become a major problem when driving at high speeds. For example, if heading changes by 1 degree from straight ahead at 60 miles per hour, the lateral velocity will be approximately 1.5 feet per second. It is easy to see that heading is closely related to lane maintenance and steering-related measures. It is no surprise then that changes in heading and heading rate may also be possible measures that could be employed to detect drowsiness. Several heading- and heading rate-related measures have been found to be accurate and reliable measures for the detection of drowsiness, all of which are feasible for one-the-road use. The names of the measures described below are simply the variable names used in previous studies.

- YAWDEV: The global maximum yaw deviation was found to be highly correlated with eye closure and was influenced by sleep deprivation and time on task (Skipper, Wierwille, and Hardee, 1984).

- YAWVAR: The yaw deviation variance (calculated over a three-minute period) was found to contain a significant amount of independent information. This measure is considered to be an accurate and reliable measure for the detection of drowsiness (Dingus, Hardee, and Wierwille, 1985).

- YAWMEAN The mean yaw deviation (calculated over a three-minute period) was found to contain a significant amount of independent information. This measure is considered to be an accurate and reliable measure for the detection of drowsiness (Dingus, Hardee, and Wierwille, 1985).

Braking and acceleration measures. The ability of a driver to apply brakes and accelerator adequately so as to maintain consistent driving speed is of obvious importance. Erratic driving or slowed braking responses may be a factor that could contribute to an accident.

Hulbert (1963) found that sleep deprivation contributes to the slowing of accelerator behavior. Safford and Rockwell (1967) found that accelerator pedal reversals were highly correlated with time during a twenty-four hour driving study. However, a literature review conducted by Hardee, Dingus, and Wierwille (1985) reported little evidence that accelerator behavior was related to time on task or drowsiness. Several other studies confirm the findings by Hardee et al (Brown, 1965; Brown, 1966; Brown, Simmonds, and Tickner, 1967; Huntley and Centybear, 1974). It was also found by Huntley and Centybear that brake usage did not significantly change with sleep deprivation. Several other studies also confirm these findings (Brown, 1965; Brown, 1966: Brown, et al., 1967).

Related to braking and acceleration behavior is speed-related behavior. Speed variability, including longitudinal acculturation and velocity maintenance, have not shown consistent results with regard to performance degradation in sleep deprived subjects. Mast, Jones, and Heimstra (1966) found significant differences between subjects' abilities to maintain constant velocity during the first and last hours of both four- and six-hour simulated driving sessions. Riemersma, Sanders, Wildervanck, and Gaillard (1977) found that speed variability significantly increased during night driving. However, three studies (Brown, 1965; Brown, 1966; Brown, et al., 1967) did not find a significant change in velocity maintenance ability in both eight- and twelve-hour driving tasks. Safford and Rockwell (1967) found no increases in speed variability during a 24 hour driving test.

The ability to follow a lead car at a consistent and safe distance is quite important while driving at high rates of speed. It was found by Muto and Wierwille (1982) that subjects' reaction times to an emergency situation involving the sudden deceleration of a lead car in a simulated car-following task were significantly greater after driving for 30, 60, and 150 minutes when compared to baseline runs. Muto and Wierwille state, however, that repeated response trials may not provide valid indications of fatigue-induced decrements in performance.

Chapter Two

Evaluation of Driver Drowsiness by Trained Raters:

Development of AVEOBS Operational Definition of Drowsiness

(This chapter represents an extended summary of work reported in the Second Semiannual Research Report, dated October 15, 1992, and referred to as Wierwille and Ellsworth, 1992)

INTRODUCTION

One of the findings of the literature review was that insufficient information existed on defining the level of drowsiness of drivers in a practical way. Specifically, the human factors research literature contains very few reports of studies which have used observers to rate the level of drowsiness exhibited by an individual. Most of the existing literature addresses studies that employed subjects to perform subjective self ratings. However, a study carried out by Carroll, Blisewise, and Dement (1989) in several nursing homes investigated the ability of observers to rate levels of drowsiness. The results of the study suggested that the use of observer ratings is a valid approach to studying drowsiness.

Therefore, efforts were directed toward the development of operational definitions of drowsiness based on observer rating. The primary objective of this study was to determine if an accurate operational definition of drowsiness could be developed by rating video taped segments of drivers at various levels of alertness. This study employed trained raters (subjects of the study in this case) who were familiar with the behavior of drowsy individuals. The raters were trained to look for behaviors oftentimes exhibited by drowsy individuals. Specifically, as an individual becomes drowsy, behaviors such as rubbing of the face or eyes, scratching, facial contortions, and moving restlessly in the seat may be exhibited. These actions are thought of as countermeasures to drowsiness. They occur during the intermediate stages of drowsiness. As an individual becomes very drowsy eyelid closures may exceed two or three seconds. These slow eyelid closures may be accompanied by a upwards or sideways rolling movement of the eyes themselves. A drowsy individual may also appear not to be focusing the eyes properly, or may exhibit a cross-eyed (lack of proper vergence) look. Facial tone will probably have decreased. Very drowsy drivers may also exhibit a lack of apparent activity and there may be large isolated (or punctuating) movements, such as providing a large correction to steering or reorienting the head from a leaning or tilting position.

A determination was made as to whether trained raters were able to consistently and accurately rate levels of drowsiness through the observation of the video taped faces of drowsy drivers. Consistency within raters and consistency between raters were evaluated. The specific objectives of this study were as follows:

Objective 1: Evaluate intrarrater reliability. (To determine if a rater assigns scores consistently.)

Objective 2: Evaluate test-retest reliability. (To determine whether a rater will score similarly on the same measure at two different points in time.)

Objective 3: Evaluate inter-rater reliability. (To determine if different raters assign similar scores using the same instrument under the same conditions.)

Objective 4: Create a new operational measure of drowsiness based on observer rating of drowsy drivers for later use in the development of drowsiness-detection algorithms.

If an observer rating method could be devised thatcould predict performance under a variety of apparent drowsiness levels, then it could serve as an alternative operational definition of drowsiness in the algorithm development study.

METHOD

Raters

Six individuals (three males and three females) volunteered to participate in this study. All participants were graduate students in the Human Factors Engineering program at Virginia Polytechnic Institute and State University. Human Factors students were chosen because of their familiarity with subjective rating procedures and human factors methodology (It was assumed that persons performing drowsiness evaluations in research or in applications would have received behavioral training). Each individual participated in two sessions, each lasting approximately two hours.

Apparatus

Previous experiments involving drowsy drivers had been performed in the Vehicle Analysis and Simulation Laboratory, in which low-light level video recordings of the drivers' faces had been made. The videotapes were retained for archival purposes and were available for use in the present study. The tapes showed drivers driving a computer controlled, moving-base driving simulator, and contained episodes of a variety of levels of apparent drowsiness. Thus, segments of the tapes could be transferred to new master tapes for use in the present experiment.

The subjective ratings of the video segments were performed in the Vehicle Analysis and Simulation Laboratory using a JVC Super VHS stereo video cassette recorder and a 20-inch Sony Trinitron color monitor. This system was used to playback segment recordings of different drivers at various levels of drowsiness. The segments were dubbed onto two separate tapes. The segments to be dubbed were located using a Panasonic VHS stereo video cassette recorder and a 20-inch General Electric color monitor. Once located, the segments were transferred from tape to tape using a JVC Super VHS camcorder and the JVC Super VHS stereo video cassette recorder mentioned above. After all segments were transferred from one tape to the other. the tapes were audio dubbed using the JVC Super VHS stereo video cassette recorder.

The scale used to perform the rating task was a form of the Likert Scale known as a Descriptive Graphics Scale. The continuous scale consisted of five descriptors: Not Drowsy, Slightly Drowsy, Moderately Drowsy, Very Drowsy, and Extremely Drowsy. There was one scale for each segment totaling 48 scales (24 scales for each session). However, for the experiment, each scale was on a separate slip of paper, approximately 22 cm wide and 7 cm high to avoid influences from previous scores.

A Macintosh II personal computer was used to analyze the data from this experiment. SuperANOVA 1.11 and Microsoft Excel 3.0 were used to perform statistical analyses of the resulting data.

Experimental Design

The experimental design used in this study was a single factor within-subject complete factorial design. The single factor was rater. This main factor (with six levels) was treated as the independent variable. By treating rater as the independent variable, each cell of the experimental design contained 48 replications of the rating task. In this experimental design, Subject was a within-rating-task variable rather than rating-task being a within-subject variable. The dependent variables were the raw-error-rating-scores. In both cases, errors were defined as differences from the mean across raters. There were 48 scores per experimental ceil giving a total of 288 data points.

The 48 segments to be rated were divided into two groups and then recorded onto two video tapes (24 segments per tape). The segments represented various levels of alertness/drowsiness and were assigned to a location on the tape. One tape was presented during the first session and the other tape was presented during the second session, which occurred approximately one week after the first session. A counterbalanced design was used in which half the raters received Tape A first followed by Tape B, while the other half of the raters received Tape B first followed by Tape A. Raters 1, 2 and 3 (two males and one female) received Tape A then Tape B while raters 4, 5 and 6 (two females and one male) received Tape B then Tape A. The reason for having two video tapes separated by one week was to allow the determination of test-retest reliability.

Intrarrater reliability. On each of the tapes, three of the segments were repeated on that same tape. On Tape A, Segments 3, 8, and 10 were repeated as Segments 23, 18, and 19 respectively. Likewise, on Tape B Segments 3, 8, and 10 were repeated as Segments 23, 18, and 19, respectively. During a particular session, the rater was exposed to the three segments twice. The repetition of segments gave six sets (pairs) of scores for each rater which were used to determine intrarrater reliability.

Test-retest reliability. In addition to having three of the segments repeated within each session, three different segments from the first session were repeated in the second session. Segments 5, 12, and 20 from Tape A were repeated as Segments 20, 12, and 5, respectively, on Tape B. Therefore, each rater was exposed to these three segments a second time during Session 2. This procedure of repeating segments gave three pairs of scores per subject for use in determining test-retest reliability.

Interrater reliability. To determine interrater reliability, all repeated segments and the first segment (the practice segment) were temporarily deleted from the data. Only those segments that were not repeated were used in the statistical analysis. After deleting the repeated segments, there remained 28 segments per rater (14 segments from each session).

Procedure

On the first day of the experiment, the rater was asked to read the general instructions for the experiment. These instructions described the nature of the experiment, the tasks to be performed, and the approximate length and timing of the two sessions. The instructions made it clear that the rating scale was a continuous one and that the rater could place a rating anywhere on the scale, not just at one of the descriptors. Once the instructions were read, the rater was asked to read the informed consent form and sign the form if he or she agreed to the conditions of the study. Any questions concerning the instructions, the informed consent form, or the experiment in general were answered. The rater was then seated in front of the video recorder and monitor. At this point the experimenter reviewed the instructions and gave additional instructions. These additional instructions included showing the rater the rating forms, giving examples of how to correctly mark the scales, and providing the rater with the "Description of Drowsiness Continuum" form. This form contained a description of the various levels of drowsiness and gave an idea of the characteristics to look for when rating the segments, The rater read the description form before the experiment began and was also allowed to refer back to the description form during the experiment. Once all questions had been answered, the first experimental session began.

When the rater returned after approximately one week for the second session, the written instructions were offered for review. Once the instructions were reviewed the experimenter asked the rater to review the Description of Drowsiness Continuum form.

Experimental task procedures. The rating task consisted of viewing 24 segments of different drivers at various levels of drowsiness for each session and subjectively rating each segment on its corresponding rating scale form. When the experimental session began, the first videotaped image appeared on the screen. A short time thereafter, a recorded voice instructed the rater to begin the evaluation for that segment. For example, at the beginning of Segment 1 , the rater heard "Begin, Segment 1." This command informed the rater that the evaluation period for segment 1 had begun. The rater observed the videotaped driver until a second voice command, "End, Segment 1" was given. (The length of the evaluation period was one minute.) The "End" command informed the rater that the evaluation period was over and that a rating on the scale was to be provided. The rater observed the beginning of the videotaped image prior to the "Begin, Segment __" command, but was instructed to only rate the interval between the "Begin" and "End" commands.

After the "End" command was given, the segment continued for 15 seconds, but the rater did not evaluate this section. Once the 15 seconds had elapsed, the screen went blank for 10 seconds before the next segment appeared. The rater used the last 15 seconds of the segment and the 10 seconds of blank screen in between the segments (totaling 25 seconds) to provide a rating.- If this amount of time was insufficient, the rater asked the experimenter (who sat behind the rater) to pause the tape until the evaluation was completed. This pausing technique allowed the rater to refer to the Description of Drowsiness Continuum sheet. Once the rating was accomplished, the experimenter restarted the tape. The rater could also change an answer if desired, but only if the rating was changed before the next segment started. (Only the current segment rating could be changed.) The rater was not permitted to go back to a previous segment to change a rating. The sequence continued until all 24 segments had been rated.

RESULTS

The first step in analyzing the data was to convert the subjective scores on the rating scale to a numerical value. This task was accomplished by converting the scale to a hundred point scale and then measuring the location of the given rating. The second step was to pair the repeated segments to perform the correlations on these data. The third step was to eliminate the paired data points and the practice segment from the original data set. The final step was to convert the original data scores into raw error rating scores and absolute error rating scores. The mean of each segment was determined and ranked from low to high. Each subject's raw data were then plotted against this ranking, as shown in Figure 3. The graph suggests that there was little, if any, error of central tendency in the experiment. Error of central tendency refers to placing ratings in the middle of the scale and avoiding extreme positions. It can be seen from the graph that subjects rated at both the low ends and the high ends of the scale as well as near the middle.

The raw error scores were obtained by subtracting each segment's mean score from the score given by each rater. Because there is no numerical (or objective) definition of drowsiness, an independent variable did not exist for this experiment. Therefore the raters' scores were compared to the mean segment score to determine consistency of the scores. The absolute error scores were obtained by taking the absolute value of the raw error scores. Thus, 28 raw error scores and 28 absolute error scores were derived for each subject resulting in a total of 168 raw and absolute error scores across all six subjects. Because of the way the error scores were calculated, a positive error score indicated that the segment was overrated compared to the segment mean rating and a negative error score indicated that the segment was underrated compared to the mean rating for that segment.

Four different correlations and four paired t-tests were performed on the data. The criterion for acceptability for the correlations was 0.80. The first correlation compared first exposure to second exposure for the segments that were repeated within a session. As

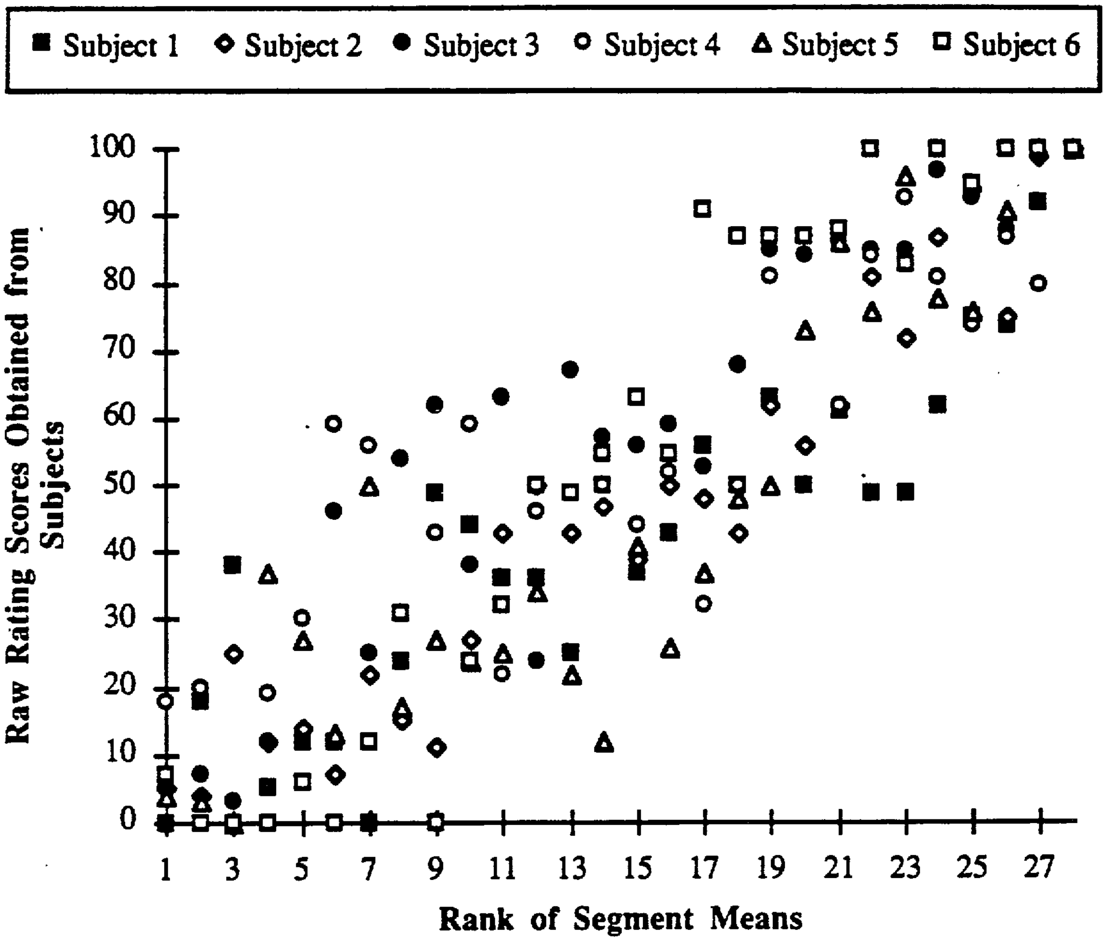


Figure 3: Raw Rating Scores as a Function of Segment Mean Rank.

mentioned, three segments were repeated in Session 1 and three segments were repeated in Session 2. Therefore, six pairs of data points per subject existed, giving a total of 36 data pairs with which to perform the correlation. This correlation was used to determine if raters tend to be consistent within themselves when scoring segments during the same time period. The measurement is an indication of intrarrater reliability. A paired t-test was also performed.

The second correlation was performed to determine the relationship between first exposure ratings and second exposure ratings from session one only. There were 3 data pairs per rater (a total of 18 pairs) for this correlation. The third correlation was also used to determine the relationship between first exposure ratings and second exposure ratings, but these data came from session two. Again, three data pairs per rater, giving a total of 18 pairs of data, were used to calculate the correlation. Both of these correlations are indications of intrarrater reliability, but the sessions were analyzed separately to determine if a fatigue/learning effect existed. Once computed, the correlations were compared to one another to determine if a significant difference existed. Separate t-tests were performed on each of the two sets of data.

The fourth correlation was performed to determine test-retest reliability. The three segments from session one were paired with the corresponding repeated segments from session two. The three pairs per rater gave a total of 18 pairs with which to determine if raters consistently rated segments at two different points in time (i.e., over a week' s period). A fourth r-test was performed on the data to determine if the difference between pairs was significantly different from zero.

An additional correlation analysis was performed as part of the interrater reliability analysis. It involved correlating the raw ratings of each rater with every other rater, as a means of quantitatively assessing consistency.

Two Analyses of Variance (ANOVAs) were conducted on the data. The first ANOVA compared the raw error rating scores in each experimental cell to determine if there were any biases in subject-ratings. Positive biases are indicative of a tendency to overrate segments as compared to the mean rating for that segment, while negative biases are indicative of a tendency to underrate segments as compared to the mean rating for that segment. In short, this analysis was used to determine if the subjects rated the segments consistently from one observer to the next. The analysis gave an indication of inter-rater reliability.

The second ANOVA compared the absolute values of the raw error rating scores in each cell to the mean of the segment. This measure made it possible to determine each subject's score deviation from the mean segment rating. Clearly, absolute error rating scores that are close (or equal) to zero indicate accuracy of rating with respect to the mean, whereas absolute error rating scores that are greater than zero signify less-than-accurate ratings with respect to the mean. Thus, the ANOVA indicated whether a difference existed in score deviations from the mean when comparing subjects.

Post-hoc analyses of significant main effects were performed using the Newman-Keuls pairwise comparison technique. This procedure was used to determine exactly which observers were significantly different from one another on the rating task.

Analysis of Intrarrater Reliability

The Pearson r correlation procedure gave a correlation value of 0.88 (t = 10.92, d.f. = 34) for intrarrater reliability. This value was significant ($p < 0.001$). This result indicates that raters consistently rated the level of drowsiness when asked to rate the same segment twice. The paired t-test gave a value t = 0.032 (d.f. = 35); $p > 0.20$.

Session 1 Correlation versus Session 2 Correlation

The comparison between the two correlations did not show a significant difference. The correlation value for Session 1 was 0.93 and the correlation value for Session 2 was 0.85. Both these correlations are significant ($p < 0.001$; t = 9.98, d.f. = 16 for Session 1 and t = 6.35. d.f. = 16 for Session 2). Comparison of the two values indicated that they did not differ significantly from one another (p = 0.1335).This comparison suggests there was no learning effect from Session 1 to Session 2. The t-tests performed on these data were not significant (p > 0.20; t = -0.46, d.f. = 17 for Session 1 and t = 0.33, d.f. = 17 for Session 2). The results of the t-tests indicate that the differences within data pairs in both sets of data were not significantly different from zero.

Analysis of Test-Retest Reliability

The correlation value for test-retest reliability as determined by the Pearson r correlation procedure was 0.81 (t = 5.45, d.f. = 16). This value was significant (p < 0.001) and indicates that raters consistently rated the level of drowsiness when asked to rate the same segment twice with a given period of time (i.e., one week) separating the two exposures. The paired t-test yielded the value t = 0.66 (d.f. = 17) which is not significant (p > 0.20), indicating that the differences within pairs of data were not significantly different from zero.

Analysis of Interrater Reliability

The ANOVA performed on the raw error scores revealed a significant main effect of rater (F = 5.159, p = 0.00 1). This effect indicates that raters demonstrate differential biases when rating the level of alertness/drowsiness. Raters 1, 2, and 5 tended to underrate the level of drowsiness with respect to the mean and raters 3, 4, and 6 tended to overrate the level of drowsiness (Figure 4). Post-hoc analysis using the Newman-Keuls technique (a = 0.05) revealed which raters were significantly different from one another. Rater 3 (mean = 8.17) rated significantly different from raters 1 (mean = -6.90), 2 (mean = -3.94) and 5 (mean = -4.65). Raters 3 tended to overrate as compared to the mean while raters 1, 2, and 5 tended to underrate. Rater 4 (mean = 3.78) rated significantly higher as compared to the mean than rater 1 (mean = -6.90). Rater 6 (mean = 3.53) rated significantly higher than raters 1 (mean = -6.90) and 2 (mean = -3.94). Surprisingly, the Newman-Keuls post-hoc test did not show a significant difference between raters 6 and 5. The test also did not show a significant difference of rater 4 and raters 2 or 5. The differences in the means of these raters are greater than other differences in means which are significant, and seem to be an

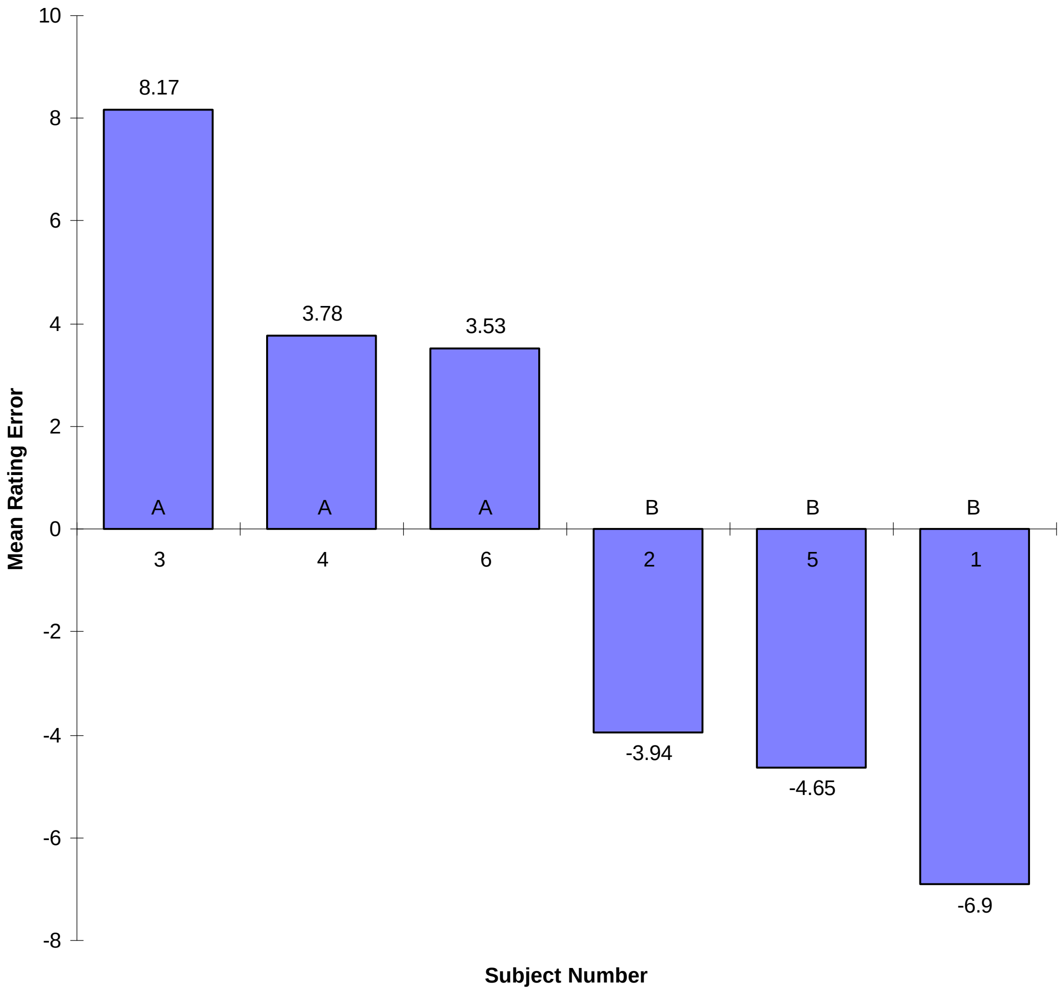


Figure 4: Mean Rating Error as a Function of Observer. (Mean ratings having common letters do not differ significantly, a = 0.05)

artifact of the Newman-Keuls post-hoc test. If one rater is removed and the test is readministered using only 5 raters, then the above mentioned non-significant differences become significant. For example, if rater 2 is removed from the data, then rater 4 is significantly different from rater 5. If rater 5 is removed, rater 4 becomes significantly different from rater 2. And, if the test is performed after removing rater 4, rater 6 is significantly different from rater 5. These observations indicate that the above mentioned subjects should be considered as significantly different from one another. Accordingly, the raters then fall into two groups that are significantly different from one another.

Descriptively speaking, the means ranged from -6.90 to 8.17. The average standard deviation across all raters' scores was 12.5. This value is an indication of the spread of ratings that can be expected anytime an observer performs a rating on the level of drowsiness.

Analysis of Absolute Error Scores

The ANOVA performed on the absolute error scores revealed no significant effect of Subject, F(5,135) = 1.537, p = 0.1929. This result suggests that individuals tend to display the same accuracy when it comes to rating the level of drowsiness. The mean absolute rating error for each subject is depicted graphically in Figure 5. (The reader is cautioned however that differences are not statistically significant.) The average mean absolute rating error across all raters was 10.85. This value is an indication of the expected magnitude of error (from the mean) for a subject rating the level of drowsiness.

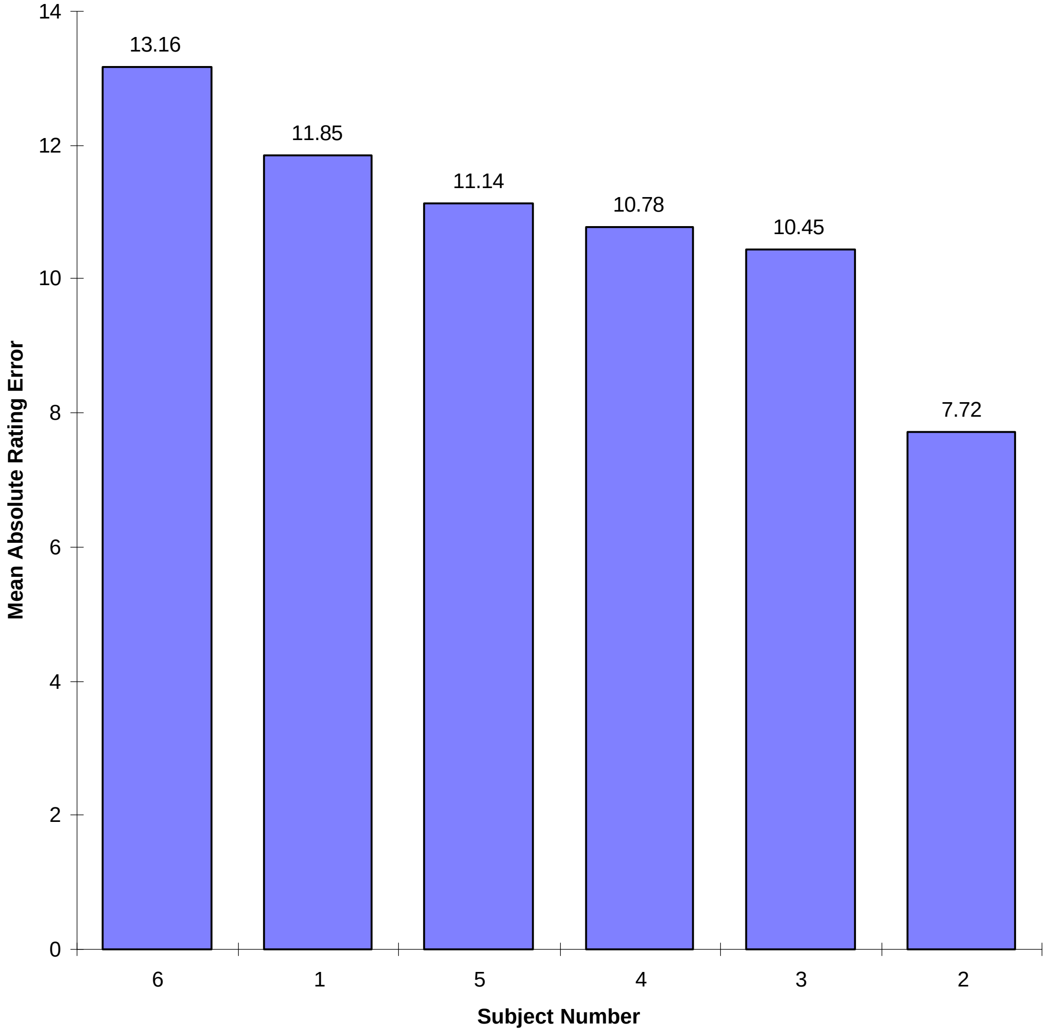


Figure 5: Mean Absolute Rating Error as a Function of Observer. (Differences are not statistically significant. a = 0.05.)

DISCUSSION ANDCONCLUSIONS

Interpretation of Results

The correlation values for intrarrater reliability and for test-retest reliability were greater than 0.80 indicating that raters tended to be consistent within themselves. Intrarrater reliability correlation (0.88) was slightly higher than test-retest reliability correlation (0.81) and suggests that the raters may lose a small amount of consistency over time. However, according to the statistical test to determine if a learning/fatigue effect existed between Session 1 and Session 2, the two correlations were not significantly different.

It is not surprising to find that there was a significant rater main effect in the ANOVA that was performed on the raw error rating scores. However, the previously mentioned study by Carroll et al. (1989) indicated that interrater reliability was high for observers studying the disturbances of the sleep-wake cycle. The rater main effect indicates that raters demonstrate differential biases when rating the level of drowsiness. Variability between raters can most likely be attributed to differences in individual definitions of drowsiness. Even though each rater was provided with the same Description of Drowsiness Continuum form, the interpretations of these descriptions may vary across raters.

According to the Analysis of Variance performed on the absolute rating error scores, raters' absolute error rating scores were not significantly different from one another. This result suggests that informed raters tend to display the same accuracy when it comes to rating the level of drowsiness.

Conclusions of the Study

The results from this study indicate that there is a good degree of consistency among and within raters when rating the level of drowsiness using videotaped segments of drivers' faces. The intrarrater reliability and the test-retest reliability indicate that raters are consistent within themselves. Even though the ANOVA of the raw error rating scores showed a significant effect of rater suggesting that inconsistent biases in ratings exist between raters, one must look at the spread of the means of raw error rating scores as compared to the scale used. The means ranged from -6.9 to 8.17 giving a spread of approximately 15 points. Very small increments were used for the divisions on the scale used to convert the ratings to numerical values. The distance between any two descriptors on, the scale was 25 points. The 15 point spread of means constitutes only 3/5 of the distance between one descriptor and the next. Furthermore, the ANOVA performed on absolute ratings as a function of rater was not significant. Therefore, a good degree of consistency is present between raters when rating the level of drowsiness in this study.

Finally, it is clear that the raters in this study were willing to use the entire scale. They ascribed widely different values to what they observed in the various videotaped segments. These findings, along with the reliability findings, suggest that ratings of drowsiness by informed raters do consistently discriminate between presented conditions.

Indications of Validity

The previously described experiment shows that there is consistency and reliability in the ratings produced. However, the experiment does not and *cannot* indicate the extent to which the raters are rating the "'true drowsiness level," since drowsiness is not a precisely or numerically defined quantity. It will be recalled that individual rating errors had to be defined in terms of deviations from the mean of all the raters. because there is no universal definition of drowsiness or drowsiness level. How, then, does one determine the validity of a drowsiness assessment procedure, such as that obtained from the rating process described in this paper? Or in short, how does one establish validity?

There are several approaches to validity. One approach is to apply the rating procedure to an actual or operational situation and determine whether the procedure "measures what it is supposed to measure" (Ghiselli, 1964). This is an application-oriented approach. Another possible approach is to compare the rating procedure to other supposed indicators of drowsiness in a controlled experiment. Such indicators might be physiological, performance based, or subjective. If it can be shown that the candidate assessment method provides results that covary with a variety of other known indicators, then the new method reflects changes associated with the common independent variables.

To provide answers to questions about validity, an additional, new experiment was conducted. Briefly, the experiment involved having sleep-deprived subjects perform alternating letter search and arithmetic tasks on a computer screen while a variety of measures were taken (Ellsworth, Wreggit, and Wierwille, 1993). The various measures were then correlated with informed-rater drowsiness ratings using a procedure identical to that described in this paper.

Typical results are shown in Table 4. As can be seen, correlations of rater ratings with eye closure and subject ratings are high, and correlations with physiological and performance measures are moderate. The results are for eight subjects and four raters. Three of the eight subjects did not exhibit any signs of drowsiness whatsoever. When they were eliminated from the data analysis, correlations values increased. These results, taken together, support the validity of rater assessment of drowsiness, and suggest that rater assessment is a viable method of drowsiness assessment when a video image of the vehicle operator is obtainable.

Table 4: Correlations of Rater Drowsiness Ratings with Other Indicators

PERCLOS	AVECLOS	EYEMEAS	SUBRATE	RTMTHCOR	RTLTCOR
0.711	0.911	0.875	0.833	0.322	0.547
MNALPHA	MNTHETA	ABRATIO	THREOG	MNHRT	MNSQHRT
0.568	0.567	0.468	0.483	-0.547	-0.525

Indicator Measures

PERCLOS: percent time that the eyes were more than 80 percent closed

AVECLOS: mean percent eye closure

EYEMEAS: mean square of percent eye closure

SUBRATE: subject on-line rating of drowsiness using an adjustable bar-knob control

RTMTHCOR: mean time to correct response in the math task

RTLTCOR: mean time to correct response in the letter search task

MNALPHA: mean amplitude of the EEG alpha wave, measures at the occipital lobe.

MNTHETA: mean amplitude of the EEG theta wave, measures at the occipital lobe.

ABRATIO: ratio of MNALPHA to mean amplitude of the EEG beta wave, measured at the occipital lobe

THREOG: percent time that the electrooculogram was above a set threshold (indicating eye blink or eye roll or both)

MNHRT: mean of instantaneous pulse rate

MNSQHRT: mean square of instantaneous pulse rate

Chapter Three

Initial Drowsiness Definition Experiment:

The Development of NEWDEF

(This chapter represents an extended summary of work reported in the Third Semiannual Research Report, data April 10, 1993 and referred to as Ellsworth, Wreggit, and Wierwille, 1993)

INTRODUCTION

This study focused on the development of an operational definition of drowsiness based on a combination of slow eyelid closure and other physiological measures. Although slow eyelid closure is a very accurate operational definitional of drowsiness, more accuracy may be gained if other measures are "blended" with slow eyelid closure measures. The primary limitation of the slow eyelid closure measures is that drivers may not exhibit this behavior until they are severely drowsy and/or impaired. Therefore, the purpose of this study was to determine if other physiological measures, used in conjunction with slow eyelid closures, could be used to create an enhanced definition of drowsiness. If such a blend could predict performance under a variety of apparent drowsiness levels, then it could serve as an alternative definition of drowsiness in the algorithm development study.

The Virginia Tech driving simulator was used for the experiment, however, the subjects in the experiment did not drive. Instead, the subjects viewed the simulator display and performed two types of tasks which were presented on the display. Push-buttons on the steering wheel were used by the subjects to respond to the two tasks.

A detection task (low-level cognitive task) consisted of a group of random-letter characters being displayed on the screen. If the subject detected one of the two target characters, the subject pressed the "yes" push-button located on the steering wheel. If none of the target letters were present in the field, the subject pressed the "no" push-button.

An arithmetic task (high-level cognitive task) consisted of mathematical problems that had numerical integers for answers. The subject was instructed to press the "even" push-button located on the steering wheel if the answer to the problem was even and to press the "odd" push-button located on the steering wheel if the answer to the problem was odd.

Correlation analyses and multiple regression were performed on the collected data. The purpose of the correlation analyses was to determine which measures could reliably detect performance impairment.

The purpose of the multiple regression analyses was to determine linear combinations of the impairment detection measures that would best predict impairment resulting from drowsiness. Multiple regression analyses were conducted on the physiological measures most highly correlated with performance measures to determine linear relationships between measures to predict performance impairment (due to drowsiness). Upon completion of the multiple regression analyses, various algorithms had been developed that contained a combination of slow eyelid closure measures and other physiological measures.

METHOD

Subjects

Eight subjects (four males and four females) volunteered to participate in this study. All potential subjects filled out a questionnaire regarding driving habits and sleeping habits before the experiment was run. Individuals who were not prone to drowsiness (found through use of the questionnaire) and those exhibiting pathological sleep disorders were not used in the experiment. In addition, potential subjects who were heavy smokers (more than three cigarettes per day) were not considered. The decision to exclude heavy smokers was made on the basis that these individuals would not be permitted to smoke for a substantial period of time (from approximately 7 P.M. to 3 A.M.). All subjects were required to have 20/30 corrected vision.

Apparatus

Figure 6 shows the equipment arrangement for the drowsiness experiment. Items of equipment are discussed separately in the following sections.

Simulator. The simulator used for the proposed research was a computer-controlled, moving-base automobile driving simulator located in the Vehicle Analysis and Simulation Laboratory. However, subjects did not drive this driving simulator. The simulator remained static during the entire experiment. The subjects viewed the simulator display with the lens system removed the usual display generation equipment disconnected. Specific tasks that the subjects were required to perform were exhibited on the simulator display. The steering wheel of the simulator had two push-buttons the subjects were to use to complete the tasks. Signals from the steering wheel push-buttons were sent to an IBM-PC for data collection.

IBM-PC Computer and Metrabyte PIO-2 Logic Interface Card. An IBM-PC generated the tasks on the simulator display. A new task appeared every 10 seconds. Task generation was accomplished using a program written in the BASICA programming language. The tasks were simultaneously displayed on the PC display and the simulator

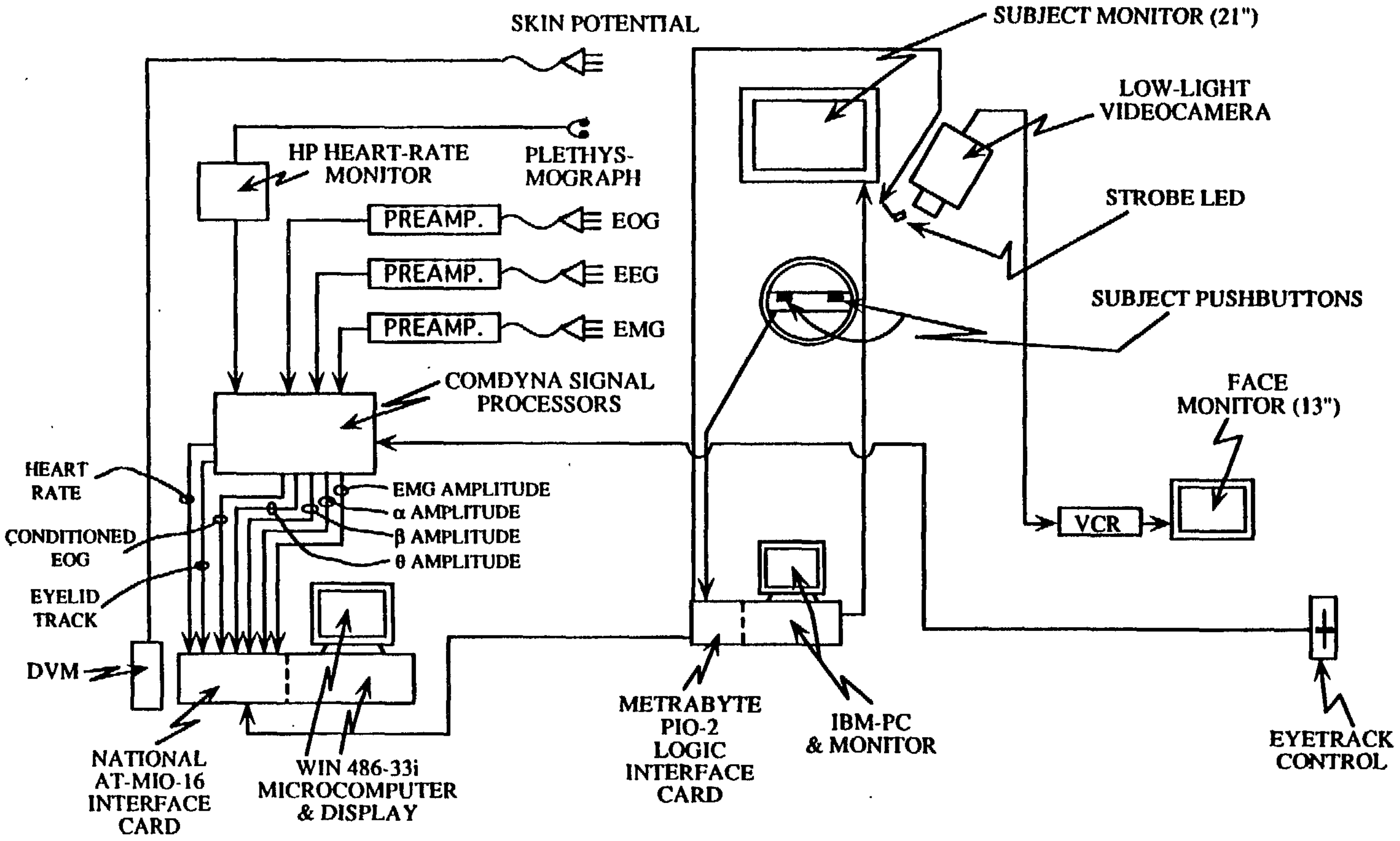


Figure 6: Equipment Arrangement for Definitional Experiment

display. Two types of tasks (a mathematical task and a letter detection search task) were generated alternately by the BASICA program and displayed on the monitors.

The PC was also used to collect the subject responses to the tasks. When a subject used the push-buttons located on the steering wheel of the simulator to complete the displayed task, a signal was sent to the IBM-PC via a Metrabyte PIO-2 8-channel logic interface card. The computer was programmed to recognize the button presses and record whether the responses were correct or incorrect. After every one-minute period, the PC stopped collecting data and calculated the number of correct responses, incorrect responses, and no responses for the mathematical task, the number of hits, misses, correct rejections, false alarms, and no responses for the detection task, the average response time for both types of tasks regardless of whether the response was correct or incorrect, and the average response time of correct responses for both task types. For the detection task, the average response time for correct response was split into average response time of correct response and average response time of correct rejection. These two averages were calculated separately.

Another function that the IBM-PC performed was to send signals to a WIN 486-33i, microcomputer telling the WIN when to start and when to stop collecting data. A high logic signal telling the WTN to start collecting data was sent as soon as the PC program started. At the end of each minute, the PC sent a low signal to the WIN telling it to stop collecting data and to complete calculations, store the results, and clear. At the beginning of the next minute, the PC sent a high signal telling the WIN computer to once again begin data collection. This sequence continued for the entire study.

The fourth function that the IBM-PC performed was to send signals to an LED mounted in front of the low light level camera shooting the subject’s face. A low signal was sent at the end of each one-minute segment. When the low signal was sent, the LED turned on and indicated on the videotape that a one-minute interval had ended. This low signal was the same signal that told the WIN to stop collecting data, to complete calculations, store the results and clear. The LED stayed on until the calculations had been completed. Once the calculations had been completed, a high signal was sent which turned the LED off. This high signal was the same signal that told the WIN to once again start the data collection process for the next one-minute duration.

Win 486-33i Microcomputer and National Instruments AT-MIO-16 Interface Card

The WIN 486-33i microcomputer collected data via an AT-MIO-16 interface card on the following physiological measures - heart rate, eyeball roll, muscle tension, alpha, beta, and theta wave amplitudes, and eye closure measures. Physiological data were received either from preamplifiers which were connected to electrodes placed on the subject, or from a direct line from a closed circuit television for the eye closure measure. Before data collection began, the WIN received a signal sent from the IBM-PC via the AT-MIO-16 interface card. The WIN was programmed using a QuickBASIC software package to recognize the signal from the PC as an indication to start collecting data. After a one-minute interval, another signal was received from the PC and recognized by the WIN computer as an indication to stop collecting data, compute calculations, store data, and clear registers to prepare for a new data collection interval. The WIN computer continued to receive the signals mentioned (from the PC) at every one-minute interval throughout the experiment.

The calculations that the WIN program performed at the end of a one-minute interval were mean heart rate, mean squared-heart rate, heart rate standard deviation, heart rate variance, the proportion of the time that the eyes were 80% or more closed, the average percent that the eyes were closed during the one-minute period, the squared value of the eye closure, the proportion of time the eyerolls were outside a threshold value, the number of times the eyerolls fell outside the threshold value, mean alpha wave signal, mean beta wave signal, mean theta wave signal, the ratio of alpha to beta, the ratio of theta to beta, the ratio of alpha plus theta to beta, mean EMG signal, the mean subjective rating of drowsiness, the mean square subjective rating, the subjective rating standard deviation and the subjective rating variance.

Skin electrodes. Biopotential skin electrodes were placed in various locations on the subject to gather physiological data such as eye roll, muscle tension, alpha, beta and theta wave amplitudes and skin potential. All signals from the electrodes except those for the skin potential passed through a GRASS high performance preamplifier system before reaching the COMDYNA signal processor. The preamplifiers feature high gain, adjustable filters, low noise and an output that can be interfaced with computers.

Eye roll measures were obtained using electrooculography (EOG) via the skin electrodes placed around the eyes. Because of the constant electropotential difference between the front and the back of the eyeball, movement of the eyes is easily measured using electrodes placed on the skin surrounding the eyes. Two electrodes were used to collect the data. One electrode was placed on the right outer canthus (ROC) area (the temple area of the right eye) and the other electrode was placed about two centimeters inside the left outer canthus area. The electrodes were offset from the horizontal midline of the eye by approximately one centimeter with the ROC electrode being above the horizontal midline and the LOC electrode being below the horizontal midline. Positioning of the electrodes in this manner allowed the detection of both horizontal and vertical eyeball movements.

A measure of muscle tension was obtained from electromyography (EMG) through electrodes placed on the chin and jaw area. Specifically, one electrode was placed on the chin and one electrode was placed under the jaw near the platysma muscle. This latter electrode was offset from the vertical midline of the face by approximately three centimeters and was located on the left side of the jaw. These electrodes were intended to detect a lack of muscle activity as the facial muscles became relaxed.

Alpha, beta. and theta wave amplitude measures were obtained by passing EEG signals (from the GRASS preamplifiers) through bandpass filters and detectors programmed on the COMDYNA processors. Two electrodes were applied to the occipital region of the scalp to record the brain wave activity. To obtain acceptable connections, the electrodes were placed and taped (using adhesive pads) to the subject’s scalp after the hair had been parted in the occipital region. A headband fitted from the back of the head , over the ears and around the forehead was used to supplement the adhesive pads in holding the electrodes in place. (This procedure was followed so that it would be unnecessary to shave small patches of hair from subjects.)

The skin potential was measured using two electrodes — one was placed on the left forearm closer to the inside of the elbow and the other was placed on the left forearm closer to the wrist. Readings were made using a MICRONTA digital multimeter which allows direct measurement of DC skin potential level without the use of DC preamplifiers and amplifiers. This potential level was not collected by the WIN computer. Instead, one of the experimenters read and recorded the skin potential level 30 seconds into each one-minute interval. (Skin potential is a slowly varying voltage.)

A “common” electrode was located on the subject’s forehead just below the hair line. All electrode wires were taped, drawn to the back of the head and bundled in a pony-tail like fashion behind the head. The wires were kept out of the subject’s view and hopefully, were fairly unobtrusive.

Ear plethysmograph Measures of heart rate and heart rate variability were collected using an ear plethysmograph and commercial heart rate monitor (Hewlett-Packard 7807C). This form of measurement was easy to implement and was unobtrusive to the subjects. The data collected from the plethysmograph was passed through the COMDYNA processors for signal level amplification before reaching the WIN via the AT-MIO- 16 interface card.

Closed circuit television A low-light level closed-circuit television camera (RCA TC 1004-U01) was used to continuously monitor the eye closures of the subjects. This video camera shot the subject’s face and eyes and was placed in such a location that the subject’s view was not obstructed. The image after passing through a VCR appeared on a Sanyo VM 45 12A monitor so that one of the experimenters could manually track the eyelid closures using a linear potentiometer. The track signal from the linear potentiometer was sent to the WIN via the AT-MIO-16 interface card after processing by the COMDYNA processor.

A General Electric Hi-fi Audio HD VHS video cassette recorder recorded the image of the subject's face for later analysis by behaviorally trained raters. Because the camera and recording required no additional lighting and was placed in an inconspicuous position, the described setup resulted in an unobtrusive way of measuring the subject's eye closure.

Subjective rating device. A continuous, rotational control was used to collect the subjects' feelings of drowsiness. It was located to the right of the subjects' right leg, in the horizontal plane. The continuous control was labeled "drowsiness," and had "Max," "Mod," and "Min" markings. In addition, there was a single marker line between each of the above settings. The subjects rated themselves during the experiment in terms of the drowsiness level that they felt. The subjects were asked to change the rating device setting any time they felt the level of drowsiness had changed. This signal was sent to the WIN via the AT-MIO-16 interface card after processing by the COMDYNA processors.

Experimental Design

The experimental design used in this study was a single factor within-subject complete factorial design. The main factor was time-on-task (four levels). The first level was the first 30 minutes, the second level was the second 30 minutes, and so on. During each of the four 30 minute segments, subjects were exposed to a different sequence of alternating mathematical and search tasks. There were four separate sequences of tasks for the four levels of time-on-task. Each of the eight subjects received all levels of the main factor.

A counterbalanced design was used to control for order effects of the four task sequences. Four subjects were randomly assigned to the presentation order of the four sequences of tasks using a Latin square design. The remaining four subjects were randomly assigned to the conditions of a second Latin square.

Twenty-one dependent variables were collected during the experiment. These variables consisted of the following:

1. Heart rate
2. Eyelid closures

3. EOG readings
4. EEG readings- alpha waves, beta waves and theta waves
5. EMG readings
6. Skin potential readings
7. Subject rating — having the subjects rate themselves on the level of drowsiness they were experiencing
8. Rater rating — having a behaviorally trained raters analyze the subjects in terms of the level of drowsiness (post experiment, using videotapes)
9. Number of correct responses for the mathematical task
10. Number of incorrect responses for the mathematical task
11. Number of no responses for the mathematical task
12. The average response time for the mathematical task — regardless of whether the response was correct or incorrect
13. The average response time of correct responses for the mathematical task
14. Number of hits for the search task
15. Number of misses for the search task
16. Number of correct rejections for the search task
17. Number of false alarms for the search task
18. Number of no responses for the search task
19. The average response time for the search task - regardless of whether the response was correct or incorrect
20. The average response time of correct responses for the search task
21. The average response time of correct rejection responses for the search task

Procedure

Subject procedure. Each subject who passed the screening tests was asked to read the general instructions for the experiment and read and signed an informed consent form. Any questions concerning the instructions, the informed consent form, or the experiment in general were answered (both prior to and following signing the form).

Each subject participated in one session which lasted about nine hours. The subjects arose at 7 A.M. or before on the established experiment day and went through their normal daytime activities without resting or napping. At 6 P.M., a member of the experimental team picked the subject up at his or her residence. The team member took the subject to a fast food restaurant for dinner. The beverages consumed were limited to non-caffinated and non-sugared drinks. The subject was permitted to smoke during or immediately after dinner after which time the subject was taken to the Vehicle Analysis and Simulation Laboratory. The subject was allowed to read, study, watch TV, or listen to a personal headset stereo. The subject was not allowed to take naps, eat, smoke, or drink caffmated or sugared beverages. A research team member remained in the lab to ensure that the subject remained awake throughout the evening. Just before midnight, the subject was given the instruction sheet and informed consent form to reread.

At midnight, the experimental session began. Two new experimenters placed the subject in the simulator and verbal instructions were given. The subject was then given a ten minute practice session. Once the practice session was completed, the physiological monitoring equipment was fitted to the subject, the lights were dimmed, and the data collection began. When the two experimenters felt the equipment was ready to go, the lights were dimmed and the experiment was begun. The subject performed the tasks on the screen for the entire experiment, which took approximately 130 minutes. If the subject fell asleep during the data collection period, one of the experimenters woke the subject and asked him or her to continue with the experiment.

At the end of the experiment, the physiological monitoring equipment was removed from the subject. The subject was debriefed. paid, and then driven home by one of the experimenters.

Experimental task procedures. The experimental tasks that the subjects were required to perform consisted of two alternating tasks — a mathematical task and a letter search task. The mathematical task was considered a higher level cognitive task and the letter search task. was a simpler, lower level cognitive task.

The mathematical tasks consisted of addition, subtraction, multiplication and division problems displayed on the simulator screen. Examples of these types of problems are 7x8 = __, 25/5 = ___, 10x___ = 30, 5+___ = 20, 25-8 = ___. Each problem had a numerical integer for an answer. The subject was required to solve the problem, decide whether the answer was odd or even, and then press the corresponding push-button on the simulator steering wheel. The task, as it appeared on the simulator screen, is shown in Figure 7.

The letter-search-task consisted of a group of letter characters displayed on the simulator screen. The letters were randomly selected and were placed in random locations on the screen. The subject was required to detect one of two target letters (A or B). If either target letter was detected, the subject was to press the "yes" push-button on the simulator steering wheel. If no target letter was detected, the subject was to press the "no" push-button. Figure 8 shows an example of the letter search task as it would appear on the simulator screen.

When the experimental session began, the two tasks were alternately displayed on the simulator screen. Each task took approximately 10 seconds before the next task appeared. If the subject did not respond to the displayed task, the task remained on the screen for the entire 10 second period. If the subject responded to the task, a feedback remark was displayed on the screen for the remaining time in the 10 second period. Examples of remarks that were used for correct responses were "GOOD JOB," "EXCELLENT" and "WAY TO GO". For incorrect responses, examples of feedback remarks used were "WRONG," "YOU MISSED A LETTER" and "TRY AGAIN". This procedure continued throughout the entire experiment.

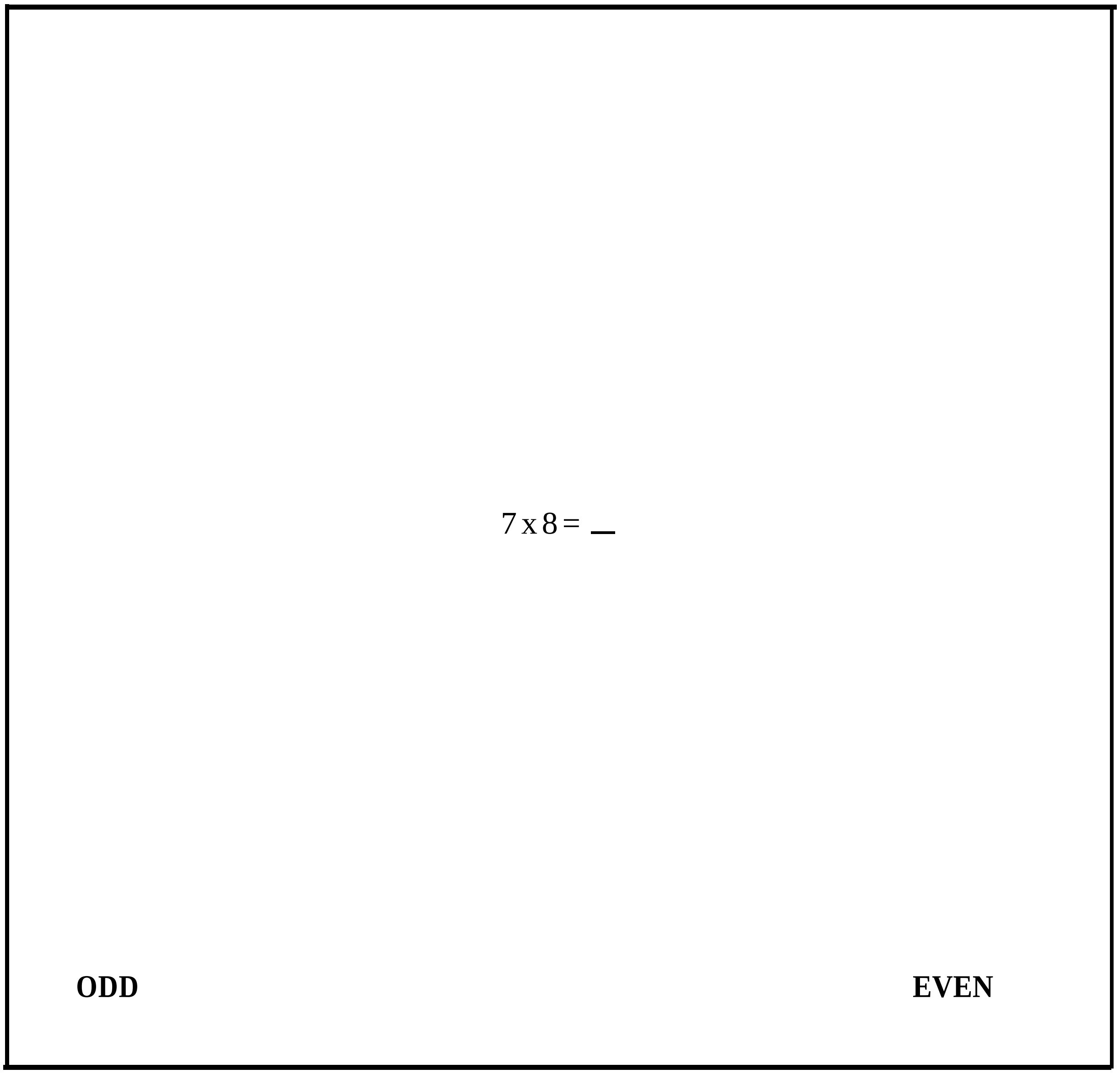


Figure 7: Example of a Mathematical Task (as it appeared on the simulator screen).

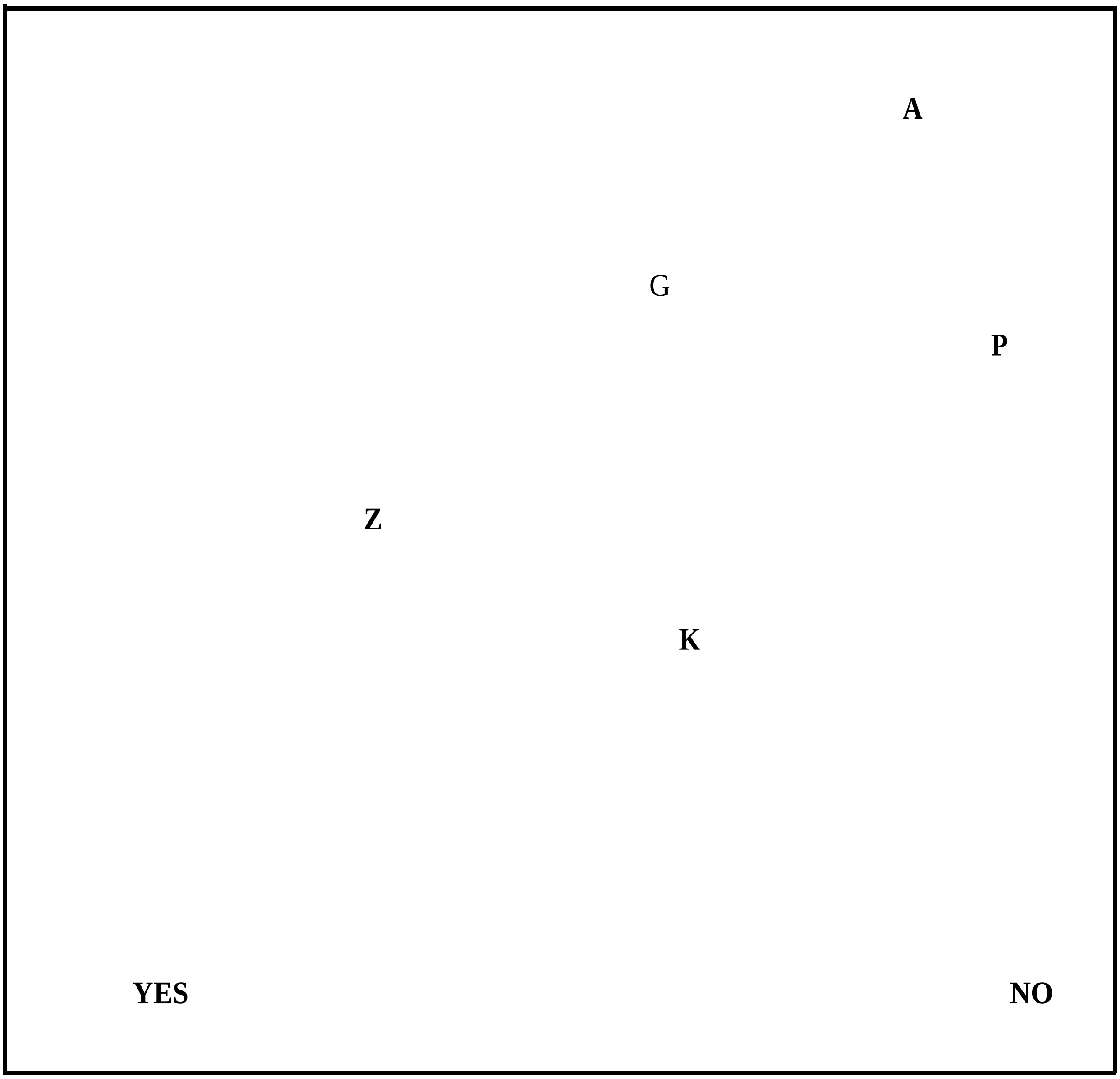


Figure 8: Example of a Letter Search Task (as it appeared on the simulator screen).

Data Manipulation

Several data manipulation steps were performed before the data were analyzed. The first step in manipulating the data was to eliminate or combine those measures that gave the same information as another measure. For example, the sum of the math errors is a combination of the sum of wrong math responses and the sum of no math responses. Therefore, the latter two measures can be eliminated and the sum of math errors measure can be used as one of the performance measures. When this procedure was completed, the following measures remained for the data analysis:

. PERCLOS: The percentage of time that the eyes were 80% to 100% closed over a one-minute interval.

• AVECLOSE: The average percent that the eyes were closed over a one-minute interval.

. EYEMEAS: The mean-square of the eyelid closure signal sampled over a one-minute interval. (EYEMEAS is more heavily weighted as eye closure increases.)

• AVEOBS: The average drowsiness rating of four informed observers for a one-minute interval.

• MNSUBRAT: The mean subjective rating over a one-minute interval (The subject moved a continuous “drowsiness” control marked with settings of maximum, moderate, and minimum. There were additional scale markers between minimum and moderate, and between moderate and maximum).

• SUMTHERR: The total number of math task errors over a one-minute interval (the number of wrong math responses and the number of no math responses).

• SUMLTERR: The total number of letter task errors over a one-minute interval (the number of wrong letter responses and the number of no letter responses).

• RTMTHCOR: The average response time to a correct math response over a one-minute interval. In situations where subjects gave an incorrect response or did not respond, a value of 10 seconds was inputted for the response time. This value is the minimum amount of time in which a subject could have responded correctly.

- RTLTCOR: The average response time to a correct letter response over a one-minute interval. In situations where subjects gave an incorrect response or did not respond, a value of 10 seconds was inputted for the response time. This value is the minimum amount of time in which a subject could have responded correctly.
- GLOBAL: Sum of SUMTHERR, RTMTHCOR, SUMLTERR, and RTLTCOR (data were non-baselined).
- MNALPHA: The mean alpha EEG amplitude over a one-minute interval. (The alpha wave was defined as including those frequencies between 8 and 12 Hz.)

. MNBETA: The mean beta EEG amplitude over a one-minute interval. (The beta wave was defined as including those frequencies between 12 and 24 Hz.)

. MNTHETA: The mean theta EEG amplitude over a one-minute interval. (The theta wave was defined as including those frequencies between 4 and 8 Hz.)

. ABRATIO: The ratio of mean alpha wave to mean beta wave amplitudes.

- TBRATIO: The ratio of mean theta wave to mean beta wave amplitudes.

. ATBRATIO: The ratio of mean alpha wave plus mean theta wave to mean beta wave amplitudes.

- MNEMG: The mean EMG amplitude over a one-minute interval. (The EMG was collected between 3 and 3000 Hz.)
- THREOG: The proportion of time that the eye-rolls go above threshold over a one-minute interval. (Threshold was set so that substantial eye-rolls would be detected).
- NUMRLBLK: The number of times the eye-rolls or blinks exceeded threshold over a one-minute interval.

. MNHRT: The instantaneous heart rate signal in beats per minute averaged over a one-minute interval.

. MNSQHRT: The mean-square of the heart rate signal sampled over a one-minute interval.

. VARHRT: The variance of the instantaneous heart rate signal calculated for a one-minute interval.

. SKINPOT: The skin potential voltage reading which was sampled 30 seconds into every one-minute interval.

The second step was to calculate two-minute intervals, four-minute intervals and six minute intervals. The two-minute interval data were calculated by taking an average of the data for two one-minute intervals for each variable. For example, one-minute intervals one and two were averaged to give two-minute interval one; one-minute intervals three and four were averaged together to give two-minute interval two; and so on. The four-minute interval data and six-minute interval data were calculated likewise except that an appropriate number of segments were used for averaging.

The third step in the data manipulation process was to delete some of the intervals from the data set. Even though the subjects had a ten minute practice session before the actual experiment started, two subjects missed two out of three math problems in the first minute and several other subjects missed at least one math problem in the first or second minute interval. Clearly, these mistakes were due to the subjects settling into the experiment and not due to drowsiness. For the one-minute interval data, the first two minutes of data were removed. For the two-, four-, and six-minute interval data the first interval was discarded. Note that the first interval for two-minute interval data consisted of the average of the first two minutes of the one-minute interval data, the first interval for four-minute interval data consisted of the average of the first four minutes of the one-minute interval data and the first interval for six-minute interval data consisted of the average of the first six minutes of the one-minute interval data. Therefore, two minutes were discarded for the two-minute interval data, four minutes were discarded for the four-minute interval data and six minutes were discarded for the six-minute interval data.

Data Analysis Overview

The data analysis for this research was composed of two major parts. The first part of the analysis consisted of correlation analyses of all the data. The purpose of these analyses was to determine which of the dependent variables could reliably detect impairment due to drowsiness. The second part of the analysis consisted of linear multiple regression analyses. The purpose of these regression analyses was to find one or more optimized linear combinations of variables that would best predict impairment resulting from drowsiness.

Various physiological and performance measures were collected and computations were made on line using the WIN 486-33i microcomputer and the IBM-PC. In addition, one of the experimenters manually collected the skin potential level for every one-minute interval, and a trained experimenter tracked the level of eyelid closure over each one-minute interval.

Correlation analyses. Correlations were performed between the collected physiological measures and the collected performance measures. For example, mean heart rate and heart rate standard deviation were correlated with each collected performance measure (i.e., the number of correct responses, incorrect responses and no responses for the mathematical task). In addition, eye closure measures were correlated with all other physiological measures.

Correlations were perforrned for one-minute intervals of data (118 data pairs), two-minute intervals of data (59 data pairs), four-minute intervals of data (29 data pairs), and six-minute intervals of data (19 data pairs). These intervals were calculated to determine whether data collected over longer intervals provided greater reliability for drowsiness detection. The studies by Dingus et al., 1985 and Hardee et al., 1985 found that in fact longer intervals (six-minute intervals versus three-minute intervals) were superior in the detection and prediction of impairment.

Several correlations were performed with the data in different configurations, including: all subjects/all data for one-minute intervals, two-minute intervals, four-minute intervals, and six-minute intervals. "Selecting" subjects/all data (data from subjects demonstrating performance decrements for a specific performance measure to determine the averaged correlation matrix) for one-, two-, four- and six-minute intervals; "selecting" subjects/pick & choose data (this method consists of using data from each subject and categorizing that data into high performance decrement, medium performance decrement, and low performance decrement categories) for one-minute interval data.

Multiple regression analysis. Once the correlations were obtained, those showing highest values were used to help construct possible "definitions" of drowsiness. Multiple linear regression analysis was employed for definition development. The purpose of multiple regression analysis is to produce an equation which can be used for prediction of a given measure at a future time.

Twenty-one performance and behavioral measures were collected and analyzed. Sixteen of these measures were used to predict the degree of performance impairment. These predictors of impairment were measures of eyelid closures, subjective self-ratings, observer ratings, and task performance. Eight of the collected measures were evaluated for reliability of performance impairment detection. These impairment detectors included heart rate, eyelid closures, EOG, EEG, EMG, skin potential, subjective ratings, and observer ratings.

RESULTS

Correlation Anaylses Results

It was found that the six-minute interval data were more reliable for drowsiness detection than the one-, two- or four-minute interval data. For two types of data configurations (all subject/all data method and the "selecting" subject/all data method), a trend towards increasing correlations existed towards the longer averages. Table 5 contains a summary of the six-minute correlation results. See Ellsworth, Wreggit, and Wierwille (1993) for a complete set of results.

In most cases, when the two data configurations are compared, the "selecting" subjects/all data procedure produces some improvement in the correlations.

The pick and choose method was used in an attempt to balance a design between drowsiness and non-drowsiness. The averaged "selecting" subjects/pick & choose correlations using only those subjects showing performance decrements produced better results than either the all subject/all data method or the "selecting" subject/all data method.

The pooled pick & choose method consisted of combining all individual subjects' pick & choose data together and running a correlation. In general, this procedure produced poorer results as compared to the previous methods.

Regression Analyses Results

Regression models for the global performance measure showed multiple correlation values (R) ranging from 0.76 to 0.86. The use of the GLOBAL measure allowed for the development of an overall linear regression model to predict drowsiness.

The final regression equation recommended for use in further studies to predict performance impairment due to drowsiness is:

$$\begin{aligned}\text{NEWDEF} = {} & 6.9\ 1500 + 18.45722(\text{PERCLOS}) - 0.01569(\text{MNALPHA}) + \\ & 0.020173(\text{MNTHETA}) - 0.00549(\text{MNBETA}) + \\ & 0.000698(\text{MNSQHRT}).\end{aligned}$$

The corresponding regression summary is presented in Table 6.

Table 5: Summary of the Six-Minute Correlation Results

	PERCLOS	AVECLOSE	EYEMEAS	AVEOBS	MNSUBRAT	SUMTHERR	SUMLTERR	RTMTHCOR	RTLTCOR
PERCLOS	1.0000								
AVECLOSE	0.8506	1.0000							
EYEMEAS	0.9147	0.9844	1.0000						
AVEOBS	0.7111	0.9108	0.8746	1.0000					
MNSUBRAT	0.6749	0.8027	0.7928	0.8334	1.0000				
SUMTHERR						1.0000			
SUMLTERR	0.3109	0.3337	0.3478	0.3372	0.2940		1.0000		
RTMTHCOR	0.3253	0.2939	0.3149	0.3201	0.2975	0.7634		1.0000	
RTLTCOR	0.5192	0.5546	0.5710	0.5471	0.4538		0.5629	0.4330	1.0000
MNALPHA	0.5442	0.5739	0.5678	0.5682	0.5271				
MNBETA									
MNTHETA	0.5363	0.5660	0.5642	0.5677	0.4755		0.3313	0.3554	0.4276
ABRATIO	0.6105	0.5078	0.5508	0.4683	0.4997		0.2522		0.3032
TBRATIO	0.3943	0.3810	0.3972	0.3660	0.3273	0.3258	0.3065	0.4144	0.4061
ATBRATIO	0.5858	0.5061	0.3440	0.4758	0.4887		0.3065	0.3264	0.3780
MNFMG									
THREOG		0.3632	0.3236	0.4827	0.4247				
NUMRLBLK	-0.3281								
MNHRT	-0.5780	-0.6122	-0.6195	-0.5471	-0.4314		-0.2807	-0.2772	-0.4780
MNSQHRT	-0.5660	-0.5951	-0.6032	-0.5252	-0.4137		-0.2718	-0.2696	-0.4630
VARHRT				0.2992	0.2558				
SKINPOT									

Table 6: NEWDEF Regression Table

Regression Statistics

Multiple R	0.821551
R Square	0.674946
Adjusted R Square	0.663814
Standard Error	1.863721
Observations	152

Analysis of Variance

	df	Sum of Squares	Mean Square	F	Significance F
Regression	5	1052.998	210.5996	60.63114	6.36E-34
Residual	146	507.1247	3.473457		
Total	151	1560.123			

	Coefficients	Standard Error	t Statistic	P-value	Lower 95%	Upper 95%
Intercept	6.915003	0.18183	38.02996	3.09E-79	6.555643	7.274363
PERCLOS	18.45722	1.584909	11.64561	9.06E-23	15.3249	21.58955
MNALPHA	-0.01569	0.00243 1	-6.4535 1	1.40E-09	-0.02049	-0.01088
MNTHETA	.0.020 173	0.004716	4.277585	3.34E-05	0.010852	0.029493
MNBETA	-0.00549	0.001599	-3.43482	0.000766	-0.00865	-0.00233
M-NSOHRT	0.000698	0.000296	2.353869	0.019867	0.000112	0.001283

(See pages 59, 60, and 61 for short definitions of independent measures. See page 82 for ft.111 definitions.)

The regression results are best understood by studying Table 7. This table shows the value of R and the number of terms for both non-baselined and baselined data. The "averages" column shows that the use of all significant terms results in nearly equal R values for both non-baselined and baselined data (R = 0.77 in each case). However, the baselining method used an average of 5.0 independent variables to achieve the regression model while the non-baselining method used 7.8 independent variables. This suggests that the baselining method can produce similar values of R with fewer independent variables and is, therefore, more efficient than the non-baselining method.

When attempts were made to reduce the number of terms in the regression model while retaining nearly the same value of R as the full model, non-baselined data show greater vulnerability. For example, using an average of 4.6 terms in the reduced models, the non-baselined R value is 0.71 while the baselined R value is 0.74. These results suggest that the baselining method provides small improvements in the models and is, therefore, an advantageous procedure. Accordingly, emphasis was placed on baselined regression in the subsequent analyses.

It was found that the two variables associated with the letter task (RTLTCOR and SUMLTERR) are more easily predicted than those associated with the math task (RTMTHCOR and SUMTHERR). Reasons for this difference are unclear, however, the letter task was considered to be much easier than the math task. Therefore, errors may have been made in the math task that were not drowsiness related. If so, drowsiness related independent variables would not have been able to predict performance as accurately.

When an attempt was made to predict the GLOBAL measure of performance, the prediction becomes better (R = 0.85). This is believed to be an excellent fit, considering the variability of the data. The table shows that a measure of eye closure, two measures of EEG and two measures of heart rate are sufficient to provide a good predictor of overall task performance.

Table 7: Summary of Regression Analyses (From Ellsworth, Wreggit, and Wierwille, 1993)
Note: All table numbers in Table 7 refer to Ellsworth, Wreggit, and Wierwille, 1993

	Measure	RTLTCOR	RTMTHCOR	SUMLTERR	SUMTHERR	GLOBAL	AVERAGES
Non- ***Baselined***	All Significant Measures (a = 0.01)	0.85 (9) Tahle 3 I	0.77 (9) Table 32	0.82 (7) Table 33	0.58 (6) Table 34	0.84 (8) Table 35	0.77 (7.8)
	Reduced Set Of Measures	0.79 (4) Table 36	0.66 (4) Tahles 37.4 I	0.78 (5) Table 38	0.56 (5) Table 39	0.76 (5) Table 40	0.71 (4.6)
Baselined	All Significant Measures (cl = 0.01)	0.87 (6) Table 42	0.75 (4) Table 43	0.83 (6) Table 44	0.54 (3) Table 45	0.86 (6) Table 46	0.77(5.0)
	Reduced Set Of Measures	084 (5) Table 47	0.71 (4) Table 48	0.81 (5) Table 49	0.51 (4) Table SO	0.85 (5) Table 5 I	0.74 (4.6)

- **RTLTCOR:** The average response time to a correct letter response over a one-minute interval. In situations where subjects gave an incorrect response or did not respond, a value of 10 seconds was inputted for the response time. **This value is the minimum amount of time in which a subject could have responded correctly.**

- **RTMTHCOR:** **The average response time to a correct math response over a one-minute interval. In situations where subjects gave an incorrect response or** did not respond, a value of **10 seconds was** inputted for the response time. This value is the minimum amount of time in which a subject could have **responded correctly.**

- **SUMLTERR:** The total number of letter task errors over a one-minute interval **(the** number of wrong letter responses and the number of no letter responses).

- **SUMTHERR:** The total number of math task errors over a one-minute interval **(the** number of wrong math responses and the number of no math **responses).**

- **GLOBAL:** **Sum of SUMTHERR, RTMTHCOR, SUMLTERR, and RTLTCOR (data were non-baselined).**

DISCUSSION

The objective of this research has been to provide a "definition" of drowsiness, using candidate measures that the research literature suggests should be sensitive indicators of drowsiness. The regression analyses performed in this study represent an attempt to relate these candidate drowsiness indicators to measurable task performance indicators. Thus, the independent measures and the ways these measures are combined in the regression analyses are in fact candidate "definitions" of drowsiness.

Four measures of performance were studied, both individually and in combination (GLOBAL measure). Final results appear in Ellsworth, Wreggit, and Wierwille (1993). These results show that with as many as five and as few as two independent variables, it is possible to achieve a relatively large R value. The independent variables are not difficult to obtain, and are limited to eye closure, simple EEG and simple heart rate measures. Thus, all measures can be obtained without overly encumbering subjects with electrodes.

Chapter Four

Development of Driver-Drowsiness Detection Algorithms

(This chapter represents an extended summary of work reported in the Fourth Semiannual Research Report, dated October 15, 1993 and referred to as Wreggit, Kirn, and Wierwille, 1993)

INTRODUCTION

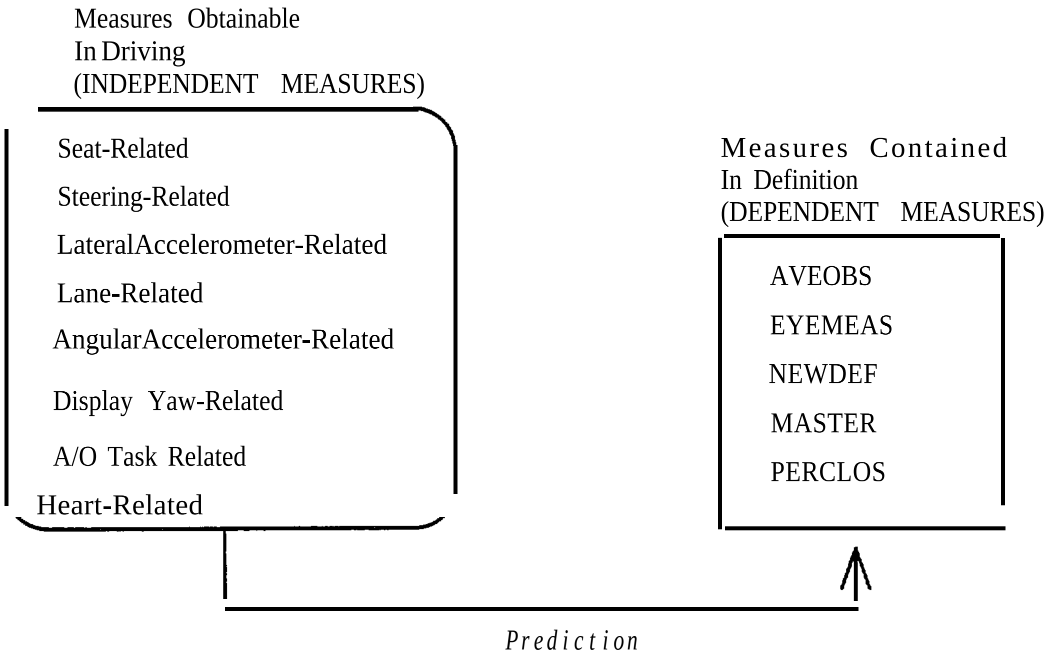
This study focused on two tasks, including: 1) determining the best statistical procedure for drowsiness-detection algorithm development and 2) developing a wide variety of usable algorithms for detection of driver drowsiness.

The dependent measures in this study were definitional measures of drowsiness that were not considered to be operationally obtainable in an actual vehicle. These measures included two eyelid-closure measures, the average observer rating measure (developed by Wierwille and Ellsworth, 1992 -- Chapter Two), an operational definition of drowsiness developed by Ellsworth, Wreggit, and Wierwille, 1993 -- Chapter Three), and a measure that was comprised of the standardized sum of the above dependent measures.

The independent measures in this study were operational measures that would be obtainable in an on-the-road vehicle. The independent measures collected during this study included *driving-related* measures, *driver-related* measures (determined by Ellsworth, Wreggit, and Wierwille, 1993), and *secondary task* performance measures. The various measures were used to create algorithms for the detection of drowsiness while driving.

Multiple regression and discriminant analyses were performed on the collected data to determine the best predictors of drowsiness. A pictorial representation of the multiple regression/discriminant analysis objective is given in Figure 9. The results from the two statistical procedures were compared for classification accuracy.

It should be noted here that in regression and discriminant analysis the definitions of independent/dependent variables are different than the definitions of independent/dependent variables in traditional experimental design. In traditional experimental design the variable being manipulated by an experimenter is the independent variable and the dependent variable is the measure affected by the independent variable. However, in regression and similar statistical techniques, the term independent variable refers to predictor variables and the term dependent variable refers to the variable that is being predicted.



NEWDEF score is from Ellsworth et al. (1993)

Figure 9: Multiple Regression/Discriminant Analysis Objective.

Multiple regression analyses were initially undertaken to determine optimum combinations of independent measures that would best predict levels of drowsiness. Discriminant analyses employed the same sets .of independent variables that were developed through the use of multiple regression. Classification matrices were then constructed for both multiple regression output and discriminant analysis output. The results showed that multiple regression was as accurate as discriminant analysis in classifying levels of drowsiness. Since multiple regression analysis does have some inherent advantages over discriminant analysis when dealing with detection algorithm development and use, it was decided that all algorithms would be developed using multiple regression techniques.

After determining that multiple regression analysis was best suited for the development of driver-drowsiness detection algorithms, further examination and analysis of the algorithms could be undertaken. Numerous algorithms were developed using various classes of measures. The classes of measures included lane-related, steering-related, lateral accelerometer-related, and secondary task-related measures, among others. By employing different combinations of measures, a step-up, step-down procedure could be achieved. Some detection algorithms employ steering and lateral accelerometer measures and other sets of detection algorithms employ steering, lateral accelerometer, and lane-related measures, for example. Therefore, loss of a lane-related measure does not cause failure of the detection system. Rather, the system simply "steps-down" to a model that does not contain lane-related measures.

Multiple algorithms containing various combinations of classes were obtained in order to 1) allow the use of a "step-up, step-down" procedure and 2) allow the use of different operational definitions of drowsiness.

METHOD

Subjects

Twelve volunteer subjects (six male and six female) were used in this study. All subjects lived in the Blacksburg, Virginia area. As part of a screening procedure all potential subjects were asked various questions over the phone concerning their driving habits, sleeping habits, and other relevant questions. Subjects who had atypical sleeping patterns, sleeping disorders, or were not prone to drowsiness were not used in the study. Potential subjects who smoked more than three cigarettes per day were not employed as subjects. This decision was made because subjects would not be allowed to smoke from approximately 7 P. M. to 3 A. M. It was felt that if heavy smokers either smoked during those hours or not did not smoke during those hours the subjects' arousal level may have been affected.

Subjects' ages ranged from 18 to 40 years. This age range was chosen because most accidents due to drowsiness that occur at night involve drivers of this age group. All subjects were given a Landholt C vision exam and had to demonstrate that they had corrected vision of at least 20/30. All subjects were required to have a valid driver's license.

During the data collection one subject drove the automobile simulator in an unrealistic and inconsistent manner. In particular, this subject consistently drove on the shoulder for extended periods of time. Another subject seemed highly stimulated and exhibited no signs of drowsiness. It was suspected by the experimenters that the latter subject either napped during the day or surreptitiously ingested a stimulant before the data gathering run took place. These subjects were run through the entire experiment and paid for their time. However, the data collected from these two subjects were not used. Two other volunteer subjects were used as replacements.

Apparatus

A pictorial representation of the peripheral equipment used for algorithm development is seen in Figure 10.

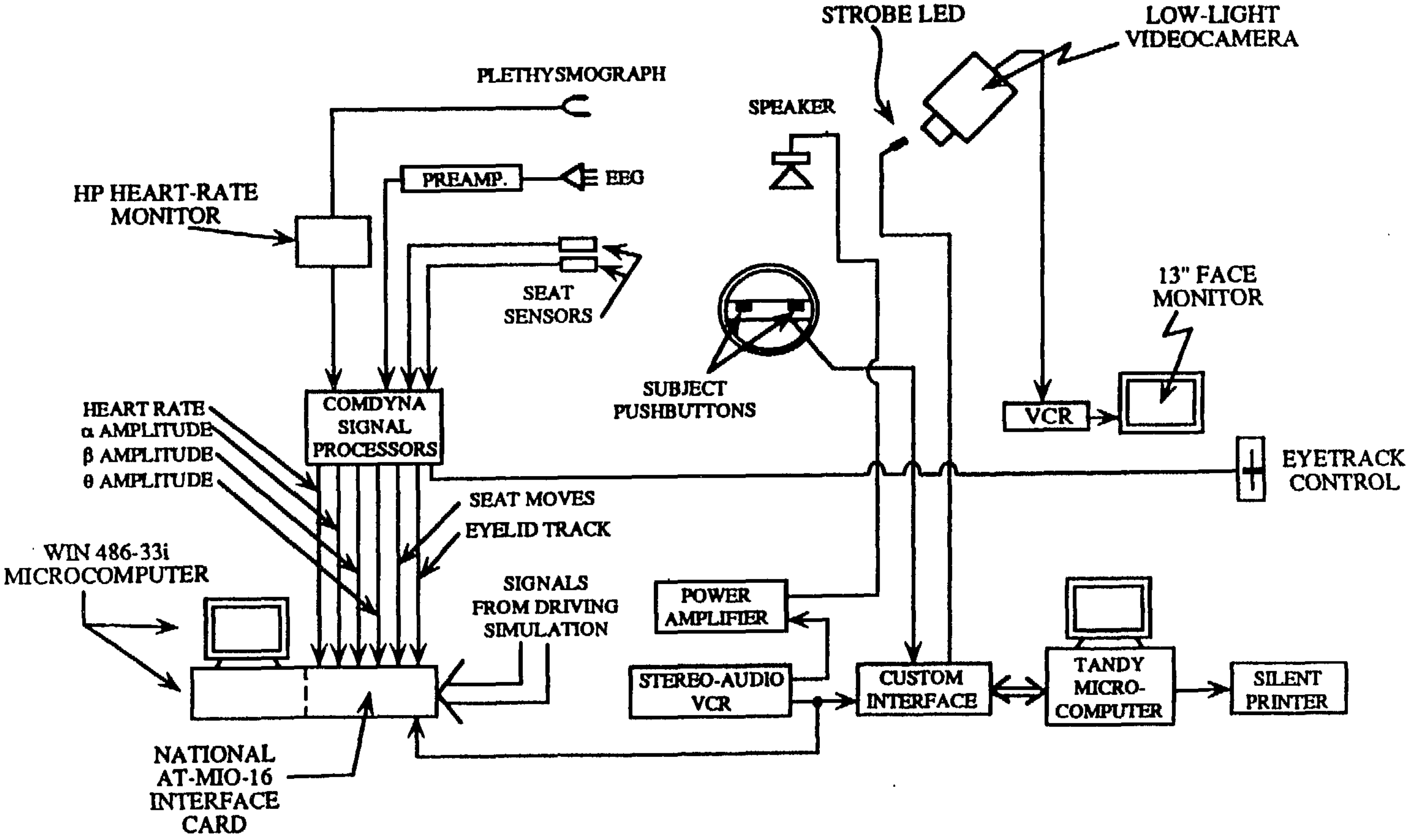


Figure 10: Peripheral Equipment Used for Algorithm Development

Simulator. The simulator used in the study was an automobile simulator that handles like a midsize vehicle. The simulator had been validated by Leonard and Wierwille (1975) with regard to driver-vehicle performance measures by comparing it with an actual automobile. It had also been validated in regard to visual glance times for in-vehicle tasks (Kurokawa and Wierwille, 1990).

The simulator was computer controlled and had a hydraulically powered moving-base with four degrees of freedom. The physical motions included pitch, yaw, lateral movement, . and longitudinal movement. The moving base was also capable of mimicking roadway vibration. Time delays inherent in the motion platform over and above normal vehicle delays were estimated to be 25 milliseconds (Dingus, Hardee, and Wierwille, 1985) and were compensated for in the vehicle dynamics.

The roadway imaging system of the simulator provided an image of a two-lane roadway with a center strip and side markings. Additionally, horizontal lines were displayed to give the driver a feeling of looking at a roadway that was embedded in the horizontal plane. This was important to further the impression that the simulated roadway continued into the distance. A monochrome CRT was used to present the roadway image to the driver. The CRT was viewed through a Fresnel lens. When the driver’s eyes were focused on the simulated roadway a majority of their peripheral vision was used to view the screen. Also present in the subject’s view was a simulated automobile hood that appeared at the correct distance and was of the correct size.

An audio system was included in the design of the automobile simulator to provide additional realism. Simulated sounds included tire noise, engine/drive train noise, tire screech on severe braking, and tire squeal on severe cornering (Dingus, Hardee, and Wierwille, 1985).

Video recording equipment. A low light level camera (RCA TC1004-UOl) was used to continuously monitor a subject’s entire face, including eye movements. Since the camera could operate at very low light levels, it was unobtrusive. The video signal was passed through a VCR and was then viewed by an experimenter on a Sanyo VM 45 12A monitor.

After all subjects completed the study the research team viewed the recorded images of the subjects so that further analyses could be performed.

Linear potentiometer. An experimenter manually tracked the subject’ s eyes by means of a linear potentiometer. As the subject’ s eyes closed, the potentiometer was pushed down so as to track the movement of the eyelids. If the subject’ s eyes were 100% closed the potentiometer was moved to the bottom of its range. If the subject’ s eyes were 0% closed the potentiometer was moved to the top.

Steering wheel controls for subsidiary task.The simulator steering wheel had been altered so that it included two push buttons on the cross member (thick spokes). One button was located on the left and the other on the right. The right button was labeled “YES” and the left button was labeled “NO”. The subjects responded to subsidiary task stimuli presented to them by pressing either the “YES” or “NO” button. The responses were interfaced to a microcomputer for storage and analyses. This microcomputer was dedicated to subsidiary-task response scoring and timing.

Win 486-33i microcomputer and analog-digital interface card. A majority of the data gathering for this experiment was performed by another microcomputer (Win 486-33i) and interface card. The interface card used was a National Instruments AT-MIO-16 card. This allowed for the collection of analog data which was converted to digital format for compatibility with a microcomputer.

Yet another computer/processor was used to collect data for several physiological measures including heart rate, eye closure, and EEG. The EEG measures included alpha, beta, and theta waves. The signals that were received by the processor included signals generated from two electrodes placed over the occipital lobe (EEG measures), an earplethysmograph (heart rate), a linear potentiometer (eye closure), and various signals from the automobile simulator. Output from the processor was then routed to the WIN 486-33i microcomputer, which was programmed in QuickBASIC to collect and store the appropriate data as well as to perform on-line calculations.

The on-line calculations that were performed by the WIN microcomputer on the collected data took place over each one-minute segment. These calculations resulted in the proportion of the time that a subject' s eyes were closed 80% or more (PERCLOS), the mean-square of the eyelid closure signal (EYEMEAS), mean alpha amplitude (MNALPHA), mean beta amplitude (MNBETA), mean theta amplitude (MNTHETA), mean heart rate (MNHRT), and squared mean heart rate (MNSQHRT), among others.

Electrodes and plethysmograph. Biopotential skin electrodes were placed over the occipital lobe and the lead wires were secured behind the driver/subject so that they could not be seen by the subject. An athletic headband was placed around the subject' s head so that the electrodes were held-securely to the subject' s head. As mentioned above, the EEG signals passed through a GRASS high performance preamplifier and then to the processor. Once the signals passed through the processor they were sampled by the AT-MIO-16 analog to digital card. After this stage the signal was ready for measures computation.

The plethysmograph sensor was placed on the antihelix of the subject' s ear for collection of heart rate data. The plethysmograph' s lead wire was secured behind the subject so that it was unobtrusive. To keep the sensor and lead wire in place the same athletic headband that was used for the electrodes was used to hold them to the subject' s head. The signal obtained from the plethysmograph passed through a Hewlett-Packard 7807C heart rate monitor. The signal then passed through the signal processor and on to the AT-MIO-16 card before reaching the microcomputer.

Experimental Design

The experimental design involved a regression- discriminant analysis approach to data analysis. Of the twelve subjects employed in the study, four were asked to simply drive the simulator, four subjects carried out a secondary task every fifteen seconds while driving, and four subjects interacted with the dashboard controls approximately every eight to ten minutes while driving. In each group there were two randomly assigned females and two randomly assigned males.

Several categories of measures were gathered during the study. Within each category were various measures. Below is a list of the 33 collected measures grouped within the appropriate category. (Each measure was initially calculated over each one-minute interval.)

Seat movement-related measures:

- NMRMOVS: The number of times the seatback sensor signals exceeded the threshold value (corresponding to the number of times the driver went from a static position to a moving position in the seat.)
- THRESMVS: The proportion of total time that the seat sensor signals exceeded the threshold value (corresponding to the proportion of total time that the driver was moving in the seat.)

Steering-related measures:

- NMRHOLD: The number of times the hold circuit output on the steering wheel exceeded a threshold value (corresponding to holding the steering wheel still for 0.4 second or longer). (Each time the steering wheel was held still for 0.4 second or longer, the count was increased by one.)
- THRSHLD: The proportion of total time that the hold circuit on the steering wheel exceeded a threshold value. (This proportion would begin to increase after 0.4 second of hold and would continue until the steering wheel was moved.)
- STVELV: The variance of steering velocity, where velocity was measured in degrees per second.
- LGREV: The number of times that steering excursions exceeded 15 degrees after steering velocity passed through zero.

- MDREV: The number of times that steering excursions exceeded 5 degrees (but less than 15 degrees) after steering velocity passed through zero.
- SMREV: The number of times that steering excursions exceeded 1 degree (but less than 5 degrees) after steering velocity passed through zero.
- STEXED: The proportion of time that steering velocity exceeded 125 degrees per second.

Lane-related measures:

- LANDEV: The standard deviation of lateral position relative to the lane, where lane position was measured in feet.
- LANVAR: The variance of the lateral position relative to the lane (square of LANDEV).
- LNMNSQ: The mean square of lane position in feet. (The “zero” position was defined as that position occurring when the vehicle was centered in the lane.)
- LNRTDEV: The standard deviation of the time derivative of lane position (relative to the lane) in feet per second.
- LNRTVAR: The variance of the time derivative of lane position (square of LNRTDEV).
- LANEX: The proportion of time that any part of the vehicle exceeded either lane boundary.
- LNERRSQ: The mean square of the horizontal difference (in feet) between the outside edge of the vehicle and the lane edge when the vehicle exceeded the lane. When the vehicle did not exceed the lane, the contribution to the measure was zero.

Accelerometer-related measures:

- ACCDEV: The standard deviation of the smoothed output of a simulated lateral accelerometer, where the output was first converted to feet per second-squared. (Smoothing was accomplished with a single-pole low-pass filter having a comer frequency at 7.25 Hz.)

- ACCVAR: The variance of the smoothed output of the accelerometer. (square of ACCDEV)
- INTACDEV: The standard deviation of the lateral velocity of the vehicle. (This signal was obtained by passing the smoothed accelerometer signal through an additional single-pole low-pass filter (leaking integrator) with a comer frequency of 0.004 Hz. The unit of output was volts, in which one unit (volt) corresponds to a smoothed lateral velocity of 73.34 feet per second.)
- INTACVAR: The variance of the lateral velocity of the vehicle (square of INTACDEV).
- ACEXEED: The proportion of time that the magnitude of lateral acceleration exceeded a threshold of 0.3 g (9.66 ft/second2).

Heading-related measures:

- HPHDGDEV: The standard deviation of the high-pass heading signal, in degrees. (The heading signal was passed through a single-pole high-pass filter with a comer frequency of 0.0 16Hz.)
- HPHDGVAR: The variance of the high-pass heading signal (square of HPHDGDEV).
- DSYAWDEV: The standard deviation of the display yaw signal in degrees. (This signal was the angular difference between vehicle heading and instantaneous roadway tangent.)
- DSYAWVAR: The variance of the display yaw signal (square of DSYAWDEV).

Subsidiary (A/O) task-related measures: (Obtained from four of the driver subjects.)

- AOTIME: Mean response time to a correct response. Incorrect responses and no-responses were specified as 12 seconds.
- NMWRONG: Mean number of incorrect responses. (Those instances in which there was no response were not included in this measure.)
- NUMNR: Mean number of stimuli for which there was no response.

Brain wave activity:

- MNALPHA: Mean alpha amplitude. (The detected amplitude of the output of a bandpass filter of the EEG having a passband from 8 to 12 Hz. The filter had a single complex pole pair with $\zeta = 0.1$ and $\omega n = 61.6$ rad. per second.)
- MNBETA: Mean beta amplitude. (The detected amplitude of the output of a bandpass filter of the EEG having a passband from 12 to 24 Hz. The filter had a single complex pole pair with $\zeta = 0.1$ and $\omega n = 109$ rad. per second.)
- MNTHETA: Mean theta amplitude. (The detected amplitude of the output of a bandpass filter of the EEG having a passband from 4 to 8 Hz. The filter had a single complex pole pair with $\zeta = 0.1$ and $\omega n = 35.8$ rad. per second.)

Heart rate measures:

- MNHRT: Mean heart rate. (The mean of the instantaneous output of the heart rate monitor in pulses per minute.)
- MNSQHRT: Mean-square heart rate. (The mean square of the instantaneous output of the heart rate monitor in pulses per minute.)

The data that were collected during this study were used to compute several definitional measures of drowsiness. The drowsiness measures were:

- EYEMEAS: The mean square of the percentage of the subject' s eye closure. (Eyes wide open represented zero percent and eyes closed represented 100 percent.)
- PERCLOS: The proportion of the time that a subject' s eyes were closed 80% or more. (Again, eyes wide open represented zero percent and eyes closed represented 100 percent.)
- AVEOBS: The average drowsiness rating of three observers for each one-minute interval. (Scale extremes were zero for " not drowsy" and 100 for " extremely drowsy" . This measure was obtained after the experimental runs by viewing the videotapes of the subjects' faces. A rating was obtained for each minute.)

. NEWDEF: Definition developed by Ellsworth, Wreggit, and Wierwille (1993): NEWDEF = 18.45722(PERCLOS) - 0.01569(MNALPHA) + 0.020173(MNTHETA) - 0.00549(MNBETA) + 0.000698(MNSQHRT). (See page 66.)

. MASTER: The sum of the standardized values of AVEOBS, EYEMEAS, NEWDEF, AND PERCLOS. (Standardization was performed after data gathering and included all 6-minute average values of the given measure (e.g., PERCLOS) from all subjects.

Procedure

Subject procedure. All subjects were involved in two sessions. The first was a screening process that took place over the telephone. During the screening session all potential subjects were asked questions in regard to driving habits, smoking habits, work schedules, and health.

Subjects that passed the screening and were chosen for the study were told to carry out their normal activities during the day on which the study was scheduled. It was mandatory that all subjects awoke at approximately 7:00 A.M. Individuals who slept during the day were not allowed to participate. At 6:00 P.M. a member of the experimenter team met the subject at the subject's residence. The experimenter took the subject to dinner at a fast-food restaurant. Subjects were not allowed to intake sugar, caffeine, alcohol or any other stimulant or depressant after 6:00 P.M. Subjects were allowed to smoke during or immediately following dinner. By coincidence, no smokers participated in the study. After eating dinner, subjects were driven to the Vehicle Analysis and Simulation Lab.

The subject was given a Landholt C vision exam upon arrival at the laboratory. Each subject was required to demonstrate corrected vision of at least 20/30. Once a subject passed the vision test he or she was given an instruction sheet that gave further details concerning the experiment. After reading the instructions the subject was asked if there were any questions concerning the study. Once questions were answered by the experimenter the subject was asked to sign an informed consent form. While subjects waited for the study to begin they were allowed to watch television, read, study, etc. An experimenter stayed with the subject at all times except restroom breaks.

The experiment was run from approximately midnight to 3:00 A.M. At midnight, two rested experimenters arrived to relieve the first member of the team. At that time the subject was again asked if there were any questions concerning the study. After any questions were answered, an experimenter placed the subject in the simulator and the laboratory lights were dimmed. The subject practiced driving the simulator for approximately five minutes. Once the five-minute practice session was complete the laboratory lights were turned on and the subject was allowed to get out of the simulator for a short time before beginning the experiment. This procedure was used to acclimatize the subject to the simulator.

The subjects in the group that were to interact with’the dashboard controls were shown the various controls and displays that they would have to use. Several practice commands were given to the subjects to familiarize them with the controls. The subjects that were in the auditory-search task group were given several practice commands as well. Any questions that the subjects had at this time were answered by the experimenter.

The experimenters began applying physiological monitoring equipment to the subject at approximately 12:15 A.M. Various equipment was turned on and the laboratory lights were dimmed. Thereafter the subject was told to begin driving the simulator and accelerate to 60 miles per hour. At the beginning of the driving session several more practice tasks were given to the subjects who were to manipulate the dash board controls or perform the secondary task. Several minutes after the subject began to drive and the experimenters felt that the driver was maintaining 60 m.p.h. in a consistent manner, data collection was initiated. The driving session in which data were collected lasted 2 l/2 hours.

After completion of the study the physiological monitoring equipment was removed from the subject by an experimenter. The subject was assisted out of the simulator, paid for time spent, and debriefed. The subject was then driven home by one of the experimenters.

Experimental task. The subject drove the simulated automobile as if it were an actual car. The subject attempted to stay within the side markings of the simulated roadway and in the appropriate lane. However, since this was a simulated roadway and vehicle, the driver was not harmed if the “vehicle” left the roadway or went into the wrong lane.

Four of the twelve subjects were asked to perform a secondary task. This task involved an auditory presentation of various words. If the presented word contained an “A” or “O” the subject was to press the button labeled “YES” located on the steering wheel. If the presented word did not include an “A” or “O” the subject was to press the button labeled “NO” located on the steering wheel. A new word was presented verbally every 15 seconds by means of an audio track on a pre-recorded videotape. The letters “A” and “O” were chosen as target letters because words could be found that include the letters “A” and “O” and are easily distinguishable from other words.

Four of the twelve subjects were asked to manipulate various controls on the dash board. These tasks involved following auditory commands to adjust radio controls, push buttons, and operate vertical slide controls. One auditory command was given approximately every eight to ten minutes. This dash board manipulation task was used simply to distract the driver from the driving task as would happen in an actual on-the-. road setting. This was important because the data would then include small amounts of “noise” that would actually be seen in an automobile. The commands were fairly infrequent so that the task of manipulating the controls by the subject would not create too much of an arousal effect. Also, the frequency of control manipulation would be similar to a person’s activity while driving on the road.

Data Analysis Overview

All measures were first computed over one-minute intervals. Data manipulation procedures were then undertaken to prepare data for statistical analyses'. Initially, the first two minutes from all measures was deleted. This was done so that the data to be analyzed did not include the time when subjects were suspected of "settling in" to the driving task. Even though all subjects were given a practice driving session it was thought that in the first two minutes of driving some subjects demonstrated inconsistencies concerning their driving 'behavior, reactions, and physiological measures.

All independent measures were baselined using the average of the first ten minutes (after the actual first two minutes had been deleted) of data. The average of the first ten minutes of data was then subtracted from every subsequent data point within that measure. Each data point consisted of a one-minute average of data. Aftercompletion of the baselining procedure the data were averaged across six-minute intervals. The first two intervals were five-minute averages to compensate for the earlier deletion of the first two minutes of data. Six-minute averages had been shown previously to have higher correlation values than either one-minute, two-minute, or four-minute averages (Ellsworth, Wreggit, and Wierwille, 1993). See Figure 11 for a pictorial overview of the data manipulation procedure. After data manipulation, multiple regression and discriminant analyses were performed on the collected data to determine the best predictors of drowsiness (as previously defined).

The difference between multiple regression and discriminant analysis can be seen in the methods used to choose the coefficients. In multiple regression the coefficients are selected to minimize the sum of the squared differences between a person's predicted and actual criterion score. In discriminant analysis the coefficients are selected to maximize correct classification. Also, the criterion variable for discriminant analysis is discrete rather than continuous as with multiple regression. The main purpose of the multiple regression and discriminant analyses was to find optimized combinations of variables that would best

2 minutes: The first 2 minutes are deleted from all data	
5 minutes: 5 minute average calculated	The first 10 minutes of data used for baselining*
5 minutes: 5 minute average calculated	
6 minutes: 6 minute average calculated	
6 minutes: 6 minute average calculated	
6 minutes: 6 minute average calculated	
6 minutes: 6 minute average calculated	
6-minute averages continued until the entire set of one-minute segments (150 minutes) was manipulated. 148 total minutes were used due to the deletion of the first two minutes of data. Therefore, 25 data points were created including two data points of 5-minute averages and twenty-three data points of 6-minute averages.	

* Baselining is a procedure in which the initial ten minutes of data are averaged and then subtracted from all subsequent one-minute segments. Baselining was carried out so that data relative to the subject's initial data values could be obtained.

Figure 11: Pre-Analysis Data Manipulation Procedures.

predict “drowsiness” during driving sessions.

Multiple regression analyses were initially used for several reasons. First, it was possible to track any portion of the data using multiple regression. Another important consideration was the fact that the threshold value could be changed to any level after application in the future. In other words, it would be possible to change the “sensitivity” of an onboard detection system if algorithms developed through the use of multiple regression were employed. Also, by using multiple regression, the experimenters were able to gain valuable insight into which measures contributed consistently to the prediction of drowsiness. For example, it was found that seat movement measures (NMRMOVS and THRESMVS) did not significantly contribute to the prediction of drowsiness and therefore they were dropped from further analyses. Finally, multiple regression was also used to determine which measures would be used in the discriminant analyses.

Multiple regression was performed on all twelve subjects and separately on the four subjects involved with the A/O auditory task. A block diagram of the algorithm development procedure is shown in Figure 12. When performing the multiple regression analyses the B weights of the various measures were first examined. This allowed for the removal of measures that were linearly related. Measures that contained large, offsetting coefficients were eliminated one at a time. (The equal and opposite coefficients demonstrated that the measures contained approximately the same predictive information. Therefore one had to be removed from the analysis.) Once any large, offsetting coefficients had been taken care of, the elimination of nonsignificant measures ($p > 0.05$) began, starting with the measure having the smallest F-ratio. Once the set of measures was reduced to four or five measures (sometimes more or less), substitution of various measures back into the set began. From this backward stepwise approach to multiple regression the best set of results were found.

Once the best multiple regression results were found for each dependent variable, MNHRT and MNSQHRT were added to the final set of independent measures. The purpose

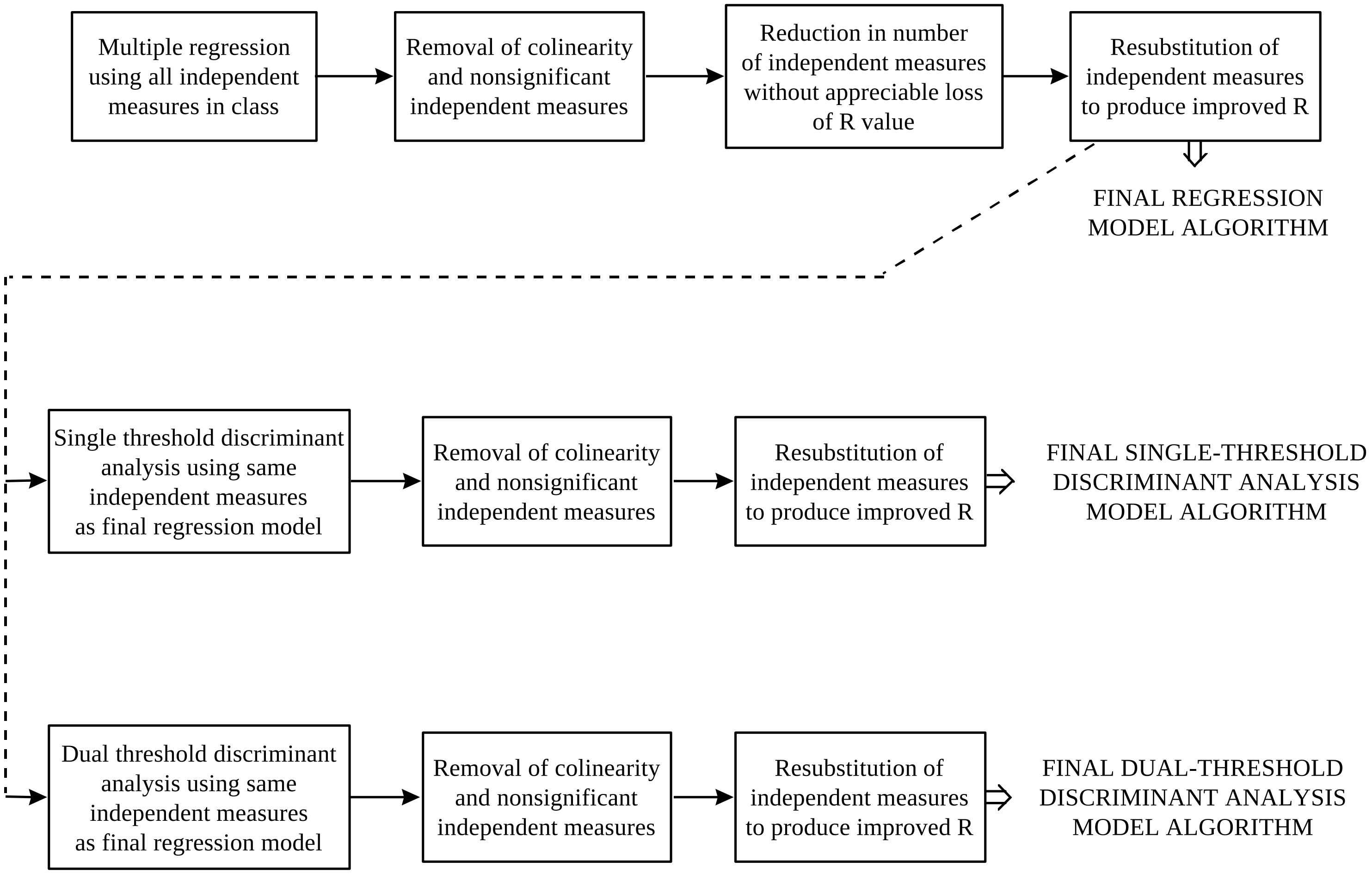

Multiple regression using all independent measures in class
Removal of colinearity and nonsignificant independent measures
Reduction in number of independent measures without appreciable loss of R value
Resubstitution of independent measures to produce improved R
FINAL REGRESSION MODEL ALGORITHM
Single threshold discriminant analysis using same independent measures as final regression model
Removal of colinearity and nonsignificant independent measures
Resubstitution of independent measures to produce improved R
FINAL SINGLE-THRESHOLD DISCRIMINANT ANALYSIS MODEL ALGORITHM
Dual threshold discriminant analysis using same independent measures as final regression model
Removal of colinearity and nonsignificant independent measures
Resubstitution of independent measures to produce improved R
FINAL DUAL-THRESHOLD DISCRIMINANT ANALYSIS MODEL ALGORITHM

Figure 12: Block Diagram of the Main Steps in the Algorithm Development Procedure

of this procedure was to examine whether heart rate would increase the accuracy of drowsiness prediction. After adding the two heart rate variables it was found that some of the significant measures found previously would become nonsignificant due to the inclusion of the heart rate variables.

The measures that were used in the discriminant analyses were based on the measures found to have the most predictive power in the multiple regression analyses. By using the measures found to be significant predictors of drowsiness in multiple regression it was felt that the discriminant analyses would begin with a strong foundation of measures. Since multiple regression attempts to fit predicted and observed data as closely as possible, it was hypothesized that these variables would contain integrity in a similar setting (other subjects carrying out similar activities). In using the measures found to be significant in multiple regression an attempt was made to bolster the future accuracy of the algorithms developed with discriminant analyses.

The discriminant analyses that were carried out examined the predictability of two distinct categories of wakefulness (awake and drowsy) and three distinct categories of wakefulness (awake, questionable, and drowsy). As seen in Figure 13 the dependent (definitional) variable PERCLOS has been graphed for each subject with threshold lines drawn in. In this graphs, the first 25 points on the abscissa correspond to subject 1, the next 25 points correspond to subject 2, and so on. The upper and lower threshold levels that were chosen for the three category discriminant analyses were based upon visual examination of the five dependent variables in conjunction with the known driving performance of each subject. For example, the experimenters rated subjects 5, 7, 9 and 10 as “alert” or “moderately alert” and these subjects performed adequately while driving. Therefore, the criterion line between “awake” and “questionable” was drawn so as to avoid the inclusion of a great majority of these subjects’ data. The spikes in the dependent variable data that extend into the “drowsy” category of the graphs correspond with poor driving performance. Therefore, the criterion line between “questionable” and “drowsy” was drawn to include the

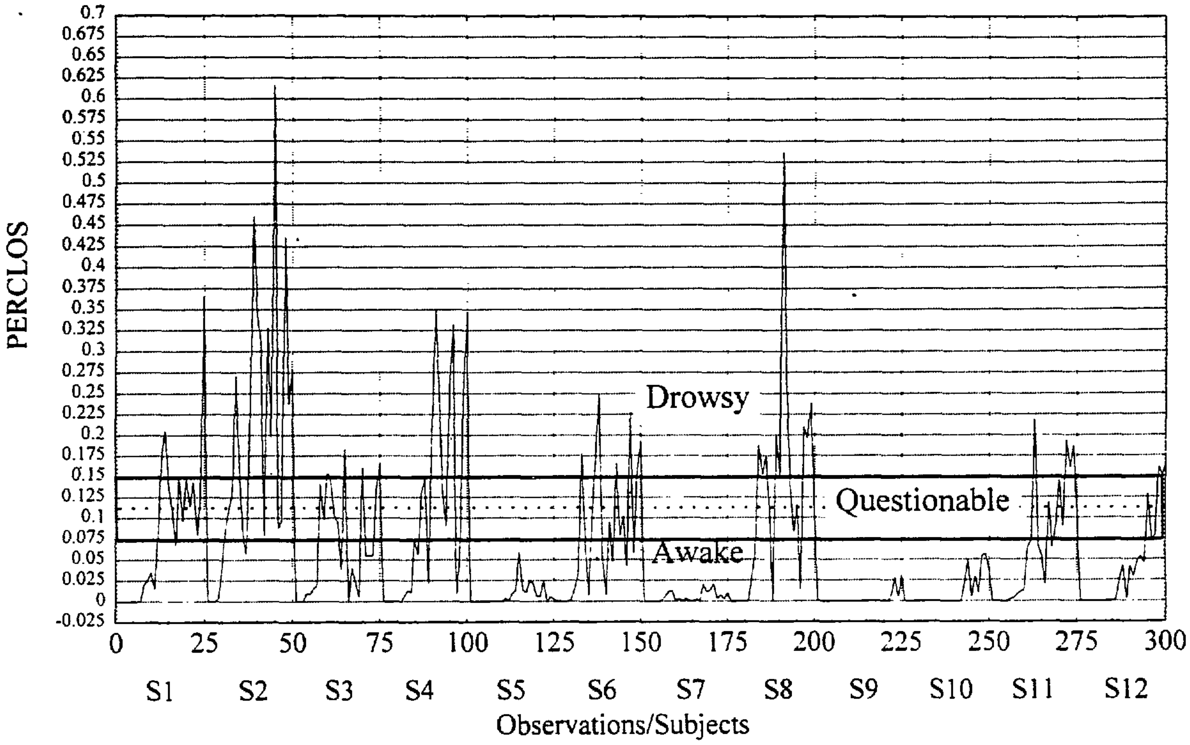


Figure 13: PERCLOS Data With Upper and Lower Criterion Lines for Three Categories and Single Criterion Line for Two Categories.

spikes in the data that corresponded with poor driving performance. The placement of the criterion line for the two category discriminant analyses was calculated by taking the average of the upper and lower thresholds of the three category analyses. In other words, the threshold is at the center of the "questionable" band.

Various drowsiness-detection algorithms were developed for possible implementation in an on-board detection system. Each set of algorithms used a slightly different set of measures so that loss of any measure does not mean failure of the detection system. The concept of using several algorithms for the detection of drowsiness employs a "step-up" and "step-down" approach. For example, if all signals are valid, the best available algorithm for drowsiness detection would be used. However, if one of the sensors necessary for the best algorithm is not providing a valid signal, the next best algorithm that does not require the invalid signal would be used. This procedure uses the "step-down" approach. A "step-up" procedure involves the use of newly validated signals. Table 8 shows the different sets of measures that were used in the multiple regression analyses and the discriminant analyses that make it possible to use the "step-up" and "step-down" process. (In the table, "accelerometer" refers to lateral accelerometer.)

Table 8: Sets of Measures Used in Multiple Regression and Discriminant Analyses for Each Dependent Measure.

Independent Measures	Dependent Measures				
	AVEOBS	EYEMEAS	NEWDEF	PERCLOS	MASTER
Steering and Accelerometer					
Steering, Accelerometer, & HPHDGDEV/VAR					
Steering, Accelerometer, LANDEV/VAR, LNMNSQ, LANEX, & LNERRSQ					
Steering, Accelerometer, & all lane measures (includes LNRTDEV/VAR)					
Steering, Accelerometer. all lane measures, & DSYAWDEV/VAR					
A/O Task Measures Only					
A/O Task, Steering, & Accelerometer					
A/O Task, Steering, Accelerometer, & HPHDGDEV/VAR					
A/O Task, LANDEV/VAR, LNMNSQ, LANEX & LNERRSQ					
A/O Task. Steering, Accelerometer, LANDEV/VAR. LNMNSQ, LANEX & LNERRSQ					
Heart, Steering, & Accelerometer					
Heart, Steering, Accelerometer, & HPHDGDEV/VAR					
Heart, Steering, Accelerometer, LANDEV/VAR. LNMNSQ. LANEX. & LNERRSQ					
Heart. Steering. Accelerometer. & all lane measures (includes LNRTDEV/VAR)					
Heart. Steering. Accelerometer, all lane measures. & DSYAWDEV/VAR					
A/O Task and Heart					
A/O Task. Heart. Steering, & Accelerometer					
A/O Task. Heart, Steering, Accelerometer. & HPHDGDEV/VAR					
A/O Task. Heart. LANDEV/VAR, LNMNSQ. LANEX. & LNERRSQ					
A/O, Heart. Steering, Accelerometer. LANDEV/VAR. LNMNSQ. LANEX, & LNERRSQ					

RESULTS

Two groups of data were analyzed. As explained earlier, the two groups included the A/O auditory-task group that consisted of four subjects, and a group including all subjects. It was found through the use of multiple regression and discriminant analyses that the use of only four subjects resulted in higher R values and lower Wilk's Lambda scores than when using data from twelve subjects. These results occur because an increase in the number of subjects causes greater difficulty in fitting predicted and observed data. It must be noted here that this was expected and must be kept in mind when reviewing the results of this study.

Multiple Regression

Table 9 is a summary of results that were attained from the multiple regression analyses. Multiple regression tables and classification matrices associated with the bolded cells in Table 9 are presented in Appendix A. The algorithms in Appendix A were chosen because they represent typical algorithms that may be employed in a full-scale on-the-road study. See Wreggit, Kim, and Wierwille (1993) for a complete set of results.

An examination of the average R scores across all sets of independent variables for each of the five dependent variables gives a good idea of the relative predictive strengths of the dependent variables. The results of the average R-score analysis seen below were obtained by averaging the R values contained within each column of Table 9.

1. MASTER: Average R = 0.8775 across 11 sets.
2. PERCLOS: Average R = 0.8563 across 12 sets
3. AVEOBS: Average R = 0.8303 across 16 sets
4. EYEMEAS: Average R = 0.8154 across 9 sets
5. NEWDEF: Average R = 0.7523 across 16 sets

The number of sets used to calculate each average was determined by the number of independent variable sets used to independently predict each dependent variable.

Table 9: Summary Table of Multiple Regression Analyses Results Showing R Values.

Independent Measures		Dependent Measures				
		AVEOBS	EYEMEAS	NEWDEF	PERCLOS	MASTER
D	Steering and Accelerometer	0.747	0.764	0.677	0.789	0.801
E	Steering, Accelerometer, & HPHDGDEV/VAR	0.793	0.809	0.700	0.847	0.852
F	Steering, Accelerometer, LANDEV/VAR, LNMNSQ, LANEX, & LNERRSQ	0.826	0.837	0.731	0.872	0.886
G	Steering, Accelerometer, & all lane measures (includes LNRTDEV/VAR)	✳ 0.824	▲	0.757	0.872	▲
H	Steering, Accelerometer, all lane measures, & DSYAWDEV/VAR	0.826	0.836	✳ 0.751	▲	▲
I	A/O Task Measures Only	0.761	0.768	0.660	0.810	0.822
J	A/O Task, Steering, & Accelerometer	Accel. 0.824	0.824	Accel. 0.740	0.836	0.876
K	A/O Task, Steering, Accelerometer, & HPHDGDEV/VAR.	0.917	Steering 0.855	▲	Accel. 0.868	0.903
L	A/O Task, LANDEV/VAR, LNMNSQ, LANEX, & LNERRSQ	Δ	0.874	0.768	0.875	0.903
M	A/O Task, Steering, Accelerometer, LANDEV/VAR, LNMNSQ, LANEX, & LNERRSQ	0.922	Δ	Δ	0.902	0.936
N	Heart, Steering, & Accelerometer	0.785	0.772	0.711	❐	❐
O	Heart, Steering, Accelerometer, & HPHDGDEV/VAR	0.813	❐	0.761	0.851	0.854
P	Heart, Steering, Accelerometer, LANDEV/VAR, LNMNSQ, LANEX, & LNERRSQ	0.838	❐	0.774	0.874	❐
Q	Heart, Steering, Accelerometer, & all lane measures (includes LNRTDEV/VAR)	✳ 0.816	❐	0.802	❐	❐
R	Heart, Steering, Accelerometer, all lane measures, & DSYAWDEV/VAR	Lane Rate 0.837	❐	steer/LnRT 0.797	▲	❐
S	A/O Task and Heart	❐	❐	0.774	❐	❐
T	A/O Task, Heart, Steering, & Accelerometer	Accel. 0.837	❐	Accel. 0.810	❐	❐
U	A/O Task, Heart, Steering, Accelerometer, & HPHDGDEV/VAR	0.918	❐	▲	Accel. 0.880	0.909
V	A/O Task, Heart, LANDEV/VAR, LNMNSQ, LANEX, & LNERRSQ	❐	❐	0.823	❐	0.910
	A/O, Heart, Steering, Accelerometer, LANDEV/VAR, LNMNSQ, LANEX, & LNERRSQ	❐	Δ	Δ	❐	❐

KEY:

▲ In regression analyses introduction of variable did not improve R value. See entry directly above .for model with same R and fewer terms.

❐ Heart Measures did not improve regression as compared with non-heart equivalent. See corresponding non-heart entry for model with same R value and fewer terms.

Δ A/O task measures did not improve regression as compared with non-A/O task measure equivalent. See corresponding non-A/O task measure entry for model with same R value and fewer terms.

✳ A cell designated with an asterisk could have been given the ▲ symbol. However, the asterisk denotes a substantially changed algorithm in term of measures used.

NOTES:
- Any measure specified in a cell was deleted because of nonsignificance.
- Letter in left-hand column corresponds to appendix (Wreggit, Kim, and Wierwille, 1993) in which analysis is presented.
- Multiple regression tables and classification matrices associated with bolded cells can be seen in Appendix A of this report.

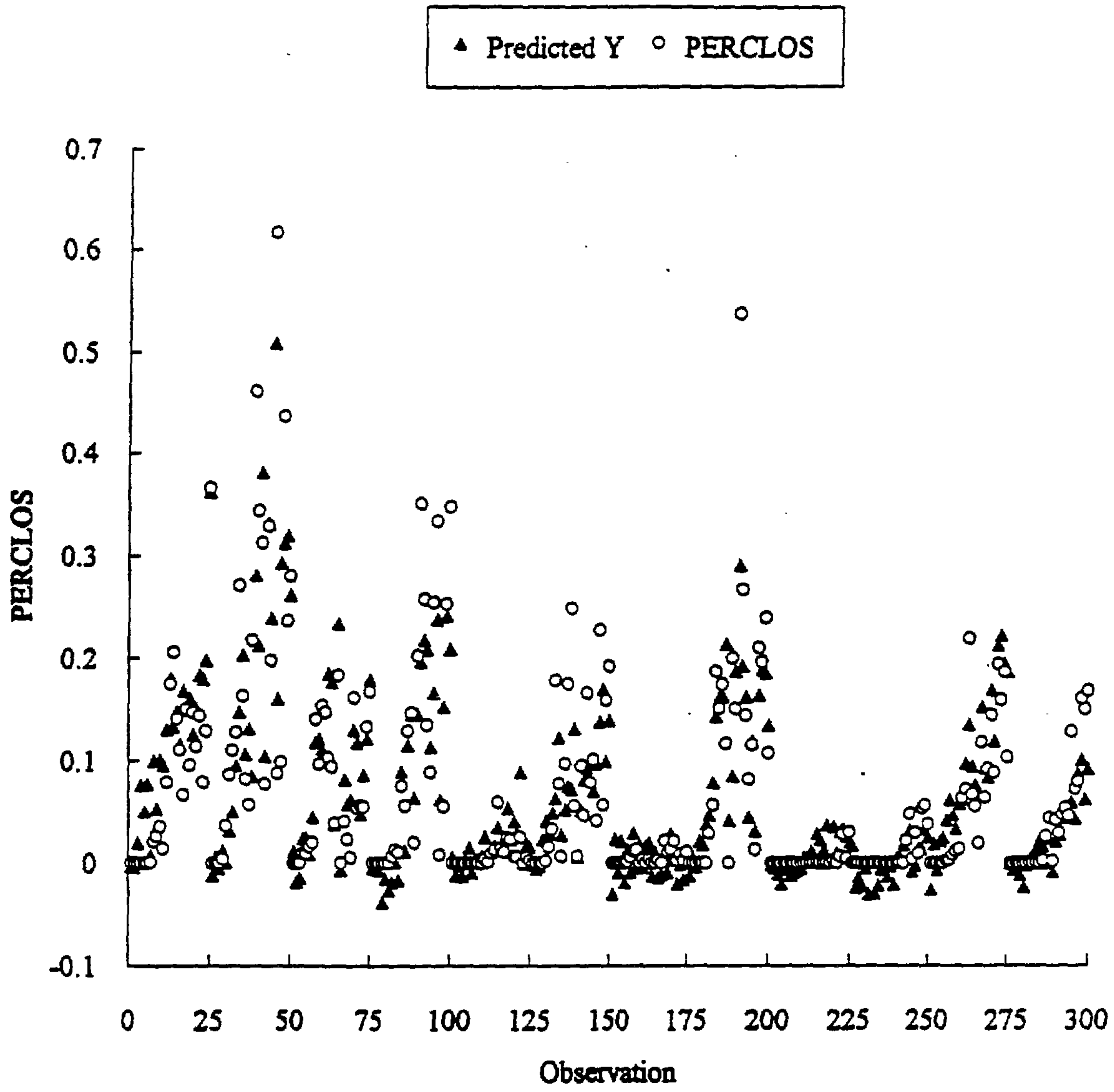
The multiple regression procedure was carried out in several steps. Mean heart rate and mean square heart rate measures were added to the best multiple regression sets to determine whether the heart rate variables contributed to the prediction of drowsiness. A general increase in R scores was seen with the addition of heart rate measures.

The addition of A/O task measures increased R values in comparison with results from data that did not incorporate the A/O task. However it must be remembered that the A/O task measures were collected using four subjects, thus somewhat inflating the R value relative to the results seen when analyzing data from twelve subjects.

After completing some initial multiple regression analyses it was found that seat movement measures did not contribute to the prediction of drowsiness. The seat movement measures, NMRMOVS and THRESMVS, were then eliminated from further analyses.

Multiple regression analyses demonstrated that it was possible to track any portion of data. As can be seen in Figure 14, predicted data tracks the observed PERCLOS data quite accurately. The graphed tracking example resulting from multiple regression analysis has an R value of 0.872 as seen in the third row of Table 9. The regressors used for the analyses in row three include steering measures, accelerometer measures, LANDEVNAR, LNMNSQ, LANEX, and LNERRSQ.

Figure 15 shows a classification matrix that was generated from a ***thresholded*** multiple regression analysis of the dependent measure PERCLOS. The data that have been classified are the same are those graphed in Figure 14. The thresholds that were used for the purpose of classification in this case were the same as the thresholds used for the discriminant analysis procedure (see Figure 13). Figure 15 shows classifications and misclassifications of three categories of wakefulness. These categories include "Awake", "Questionable", and "Drowsy". The categories of wakefulness are presented along the left side of the table (observed) and across the top of the table (predicted). As an example of how to interpret this table find the "18" in the cell located under the predicted category of "Questionable" in the classification matrix. This cell contains 18 misclassifications due to the fact that those 18 data points were



R = 0.87159526 R^2 = 0.75967830 Adjusted R^2 = 0.75475703
F(6,293) = 154.37 p < 0.0000 Std. Error of estimate: 0.04849

	BETA	St. Err. of BETA	B	St. Err. of B	t(293)	p-level
Intercept			-0.003	0.004053	-0.694	0.488
INTACDEV	-0.109	0.030	-0.069	0.019114	-3.603	0.000
LANDEV	0.873	0.063	0.066	0.004763	13.798	0.000
LNERRSQ	-0.258	0.054	-0.002	0.000410	-4.820	0.000
STEXED	0.090	0.033	45.740	16.818827	2.720	0.007
NMRHOLD	-0.204	0.045	-0.004	0.000785	-4.494	0.000
THRSHLD	0.250	0.041	0.231	0.037904	6.098	0.000

Figure 14: Scatterplot of PERCLOS Data Shown With Regression Summary.

Predicted

	Group	Percent Correct	Awake	Questionable	Drowsy
	Awake	89.76	**184**	18	3
Observed	Questionable	47.27	7	**21**	16
	Asleep	62.75	3	16	**32**
	Total	79.00	194	55	51

PERCLOS (R = 0.872).

Apparent Accuracy Rate (large misclassifications): 0.98

Apparent Accuracy Rate (all misclassifications): 0.79

Figure 15: Classification Matrix Generated From Multiple Regression Analysis of PERCLOS Data. (Independent variables employed included Steering, Accelerometer, LANDEVNAR, LNMNSQ, LANEX, & LNERRSQ.)

classified as “Questionable” by the multiple regression equation but were actually in the “Awake” category. A “large error” is defined as any misclassification in which the predicted classification is two categories away from the observed (actual) classification. For example, data within a cell predicted as “drowsy” that was actually “awake” has been missclassified by two cells. The cells containing “184", “21”, and “32” are correct classifications, or hits.

As stated earlier, when performing the multiple regression analyses the B weights of various measures were first examined. By examining the B weights the experimenters were able to reduce linear dependency between variables. However, B, the nonstandardized numbers attained from multiple regression analyses, are the values that could be used for further application. The B values are coeffkients that can be used to create a drowsy driver detection algorithm.

Discriminant Analyses

The results of the discriminant analyses that were run corresponded, in general, with the results attained from the multiple regression analyses. In other words, a high R value resulting from multiple regression usually resulted in an accurate classification matrix. However, it was found that in some instances several of the variables that significantly contributed to drowsiness prediction with multiple regression were not significant with discriminant analysis. The dropping out of previously significant prediction measures was most profound when the set of variables being examined included lane measures or high pass heading measures.

Tables 10, 11, and 12 are summary tables of results obtained from the ‘discriminant analyses. Table 10 shows APARs (apparent accuracy rates) for large classification errors. Large errors are defined as misclassifications in which a prediction of “awake” is made when the subject is actually “drowsy” or vice versa. More complete results of the discriminant analyses can be seen in Wreggit, Kim, and Wierwille (1993).

Table 10: Summary Table of Two Category Discriminant Analyses Results Showing APAR..

	Independent Measures	Dependent Measures				
		AVEOBS	EYEMEAS	NEWDEF	PERCLOS	MASTER
D	Steering and Accelerometer	84.0	83.7	81.3	85.0	82.7
E	Steering, Accelerometer, & HPHDGDEV/VAR	84.7	84.3	81.7	89.7	85.7
F	Steering, Accelerometer, LANDEV/VAR, LNMNSQ, LANEX, & LNERRSQ	88.7	85.3	84.3	90.33	89.67
G	Steering, Accelerometer, & all lane measures (includes LNRTDEV/VAR)	✳ 88.0		✳ 83.0	▲	
H	Steering, Accelerometer, all lane measures, & DSYAWDEV/VAR	89.0	85.0	✳ 83.7		
I	A/O Task Measures Only	87.0	84.0	83.0	85.0	86.0
J	A/O Task, Steering, & Accelerometer	Accel. 92.0	86.0	Accel. 88.0	88.0	92.0
K	A/O Task, Steering, Accelerometer, & HPHDGDEV/VAR.	Accel. 94.0	▲		Accel. 91.0	Δ
L	A/O Task, LANDEV/VAR, LNMNSQ, LANEX, & LNERRSQ		Δ	87.0	89.0	Δ
M	A/O Task, Steering, Accelerometer, LANDEV/VAR, LNMNSQ, LANEX, & LNERRSQ	Δ			96.0	92.0
N	Heart, Steering, & Accelerometer	85.7	83.7	82.67		
O	Heart, Steering, Accelerometer, & HPHDGDEV/VAR	87.0		83.0	❐	❐
P	Heart, Steering, Accelerometer, LANDEV/VAR, LNMNSQ, LANEX, & LNERRSQ	89.3		85.7	❐	
Q	Heart, Steering, Accelerometer, & all lane measures (includes LNRTDEV/VAR)	✳ 87.7		▲		
R	Heart, Steering, Accelerometer, all lane measures, & DSYAWDEV/VAR	❐		steer/LnRT 84.0		
S	A/O Task and Heart			83.0		
T	A/O Task, Heart, Steering, & Accelerometer	Accel. 92.0		Accel. 85.0		
U	A/O Task, Heart, Steering, Accelerometer, & HPHDGDEV/VAR	❐			❐	❐
V	A/O Task, Heart, LANDEV/VAR, LNMNSQ, LANEX, & LNERRSQ			❐		❐
	A/O, Heart, Steering, Accelerometer, LANDEV/VAR, LNMNSQ, LANEX, & LNERRSQ					

KEY:

▲ In these discriminant analyses introduction of variable did not improve prediction value. See entry directly above for model with same prediction value and fewer terms.

❐ Heart Measures did not improve prediction value as compared with non-heart equivalent. See corresponding non-heart entry for model with same prediction value and fewer terms.

Δ A/O task measures did not improve prediction value as compared with non-A/O task measure equivalent. See corresponding non-A/O task measure entry for model with same prediction value and fewer terms.

✳ A cell designated with an asterisk could have been given the ▲ symbol. However, the asterisk denotes a substantially changed algorithm in term of measures used.

NOTES:
- Blank cells indicate that analysis was not computed because corresponding regression did not show improvement in R value.
- Any measure specified in a cell was deleted because of nonsignificance.
- Letter in left-hand column corresponds to appendix (Wreggit, Kim, and Wierwille, 1993) in which analysis is presented.

Table 11: Summary Table of Three Category Discriminant Analyses Results Showing APAR For All Classification Errors.

Independent Measures		Dependent Measures				
		AVEOBS	EYEMEAS	NEWDEF	PERCLOS	MASTER
D	Steering and Accelerometer	72.7	80.7	74.0	78.3	77.7
E	Steering, Accelerometer, & HPHDGDEV/VAR	73.0	81.0	73.3	81.3	80.3
F	Steering, Accelerometer, LANDEV/VAR, LNMNSQ, LANEX, & LNERRSQ	77.0	82.7	71.7	85.0	82.3
G	Steering, Accelerometer, & all lane measures (includes LNRTDEV/VAR)	✳ 74.7		75.3	▲	
H	Steering, Accelerometer, all lane measures, & DSYAWDEV/VAR	78.3	83.0	73.0		
I	A/O Task Measures Only	81.0	82.0	72.0	79.0	80.0
J	A/O Task, Steering, & Accelerometer	87.0	85.0	Accel. 89.0	77.0	85.0
K	A/O Task, Steering, Accelerometer, & HPHDGDEV/VAR.	Accel. 90.0	▲		Δ	Accel. 87.0
L	A/O Task, LANDEV/VAR, LNMNSQ, LANEX, & LNERRSQ		86.0	79.0	82.0	Δ
M	A/O Task, Steering, Accelerometer, LANDEV/VAR, LNMNSQ, LANEX, & LNERRSQ	Δ			89.0	90.0
N	Heart, Steering, & Accelerometer	72.7	79.0	75.0		
O	Heart, Steering, Accelerometer, & HPHDGDEV/VAR	73.0		76.3	❐	❐
P	Heart, Steering, Accelerometer, LANDEV/VAR, LNMNSQ, LANEX, & LNERRSQ	78.3		76.7	83.7	
Q	Heart, Steering, Accelerometer, & all lane measures (includes LNRTDEV/VAR)	✳ 77.7		▲		
R	Heart, Steering, Accelerometer, all lane measures, & DSYAWDEV/VAR	✳ 77.7		LnRT 77.3		
S	A/O Task and Heart			77.0		
T	A/O Task, Heart, Steering, & Accelerometer	Accel. 85.0		Accel. 89.0		
U	A/O Task, Heart, Steering, Accelerometer, & HPHDGDEV/VAR	❐			❐	❐
V	A/O Task, Heart, LANDEV/VAR, LNMNSQ, LANEX, & LNERRSQ			82.0		❐ Δ
	A/O, Heart, Steering, Accelerometer, LANDEV/VAR, LNMNSQ, LANEX, & LNERRSQ					

KEY:

▲ In these discriminant analyses introduction of variable did not improve prediction value. See entry directly above for model with same prediction value and fewer terms.

❐ Heart Measures did not improve prediction value as compared with non-heart equivalent. See corresponding non-heart entry for model with same prediction value and fewer terms.

Δ A/O task measures did not improve prediction value as compared with non-A/O task measure equivalent. See corresponding non-A/O task measure entry for model with same prediction value and fewer terms.

✳ A cell designated with an asterisk could have been given the ▲ symbol. However, the asterisk denotes a substantially changed algorithm in term of measures used.

NOTES:
- Blank cells indicate that analysis was not computed because corresponding regression did not show improvement in R value.
- Any measure specified in a cell was deleted because of nonsignificance.
- Letter in left-hand column corresponds to appendix (Wreggit, Kim, and Wierwille, 1993) in which analysis is presented.

Table 12: Summary Table of Three Category Discriminant Analyses Results Showing APAR for Large Classification Errors.

	Independent Measures	Dependent Measures				
		AVEOBS	EYEMEAS	NEWDEF	PERCLOS	MASTER
D	Steering and Accelerometer	93.3	87.0	91.7	94.0	95.0
E	Steering, Accelerometer, & HPHDGDEV/VAR	94.33	88.3	91.3	96.3	96.7
F	Steering, Accelerometer, LANDEV/VAR, LNMNSQ, LANEX, & LNERRSQ	96.7	90.0	92.0	97.3	97.0
G	Steering, Accelerometer, & all lane measures (includes LNRTDEV/VAR)	✳ 95.7		93.0	▲	
H	Steering, Accelerometer, all lane measures, & DSYAWDEV/VAR	✳ 95.7	90.3	✳ 92.7		
I	A/O Task Measures Only	89.0	87.0	91.0	95.0	94.0
J	A/O Task, Steering, & Accelerometer	93.0	90.0	Accel. 93.0	Accel. 96.0	99.0
K	A/O Task, Steering, Accelerometer, & HPHDGDEV/VAR.	Accel. 98.0	▲		Δ	Accel. 99.0
L	A/O Task, LANDEV/VAR, LNMNSQ, LANEX, & LNERRSQ		93.0	100.0	94.0	Δ
M	A/O Task, Steering, Accelerometer, LANDEV/VAR, LNMNSQ, LANEX, & LNERRSQ	Δ			99.0	99.0
N	Heart, Steering, & Accelerometer	94.3	86.3	93.00		
O	Heart, Steering, Accelerometer, & HPHDGDEV/VAR	95.3		93.7	❐	❐
P	Heart, Steering, Accelerometer, LANDEV/VAR, LNMNSQ, LANEX, & LNERRSQ	96.0		94.7	96.7	
Q	Heart, Steering, Accelerometer, & all lane measures (includes LNRTDEV/VAR)	96.33		▲		
R	Heart, Steering, Accelerometer, all lane measures, & DSYAWDEV/VAR	✳ 96.0		LnRT 95.0		
S	A/O Task and Heart			92.0		
T	A/O Task, Heart, Steering, & Accelerometer	Accel. 94.0		Accel. 96.0		
U	A/O Task, Heart, Steering, Accelerometer, & HPHDGDEV/VAR	❐			❐	❐
V	A/O Task, Heart, LANDEV/VAR, LNMNSQ, LANEX, & LNERRSQ			96.0		❐ Δ
	A/O, Heart, Steering, Accelerometer, LANDEV/VAR, LNMNSQ, LANEX, & LNERRSQ					

KEY:

▲ In these discriminant analyses introduction of variable did not improve prediction value. See entry directly above for model with same prediction value and fewer terms.

❐ Heart Measures did not improve prediction value as compared with non-heart equivalent. See corresponding non-heart entry for model with same prediction value and fewer terms.

Δ A/O task measures did not improve prediction value as compared with non-A/O task measure equivalent. See corresponding non-A/O task measure entry for model with same prediction value and fewer terms.

✳ A cell designated with an asterisk could have been given the ▲ symbol. However, the asterisk denotes a substantially changed algorithm in term of measures used.

NOTES:
- Blank cells indicate that analysis was not computed because corresponding regression did not show improvement in R value.
- Any measure specified in a cell was deleted because of nonsignificance.
- Letter in left-hand column corresponds to appendix (Wreggit, Kim, and Wierwille, 1993) in which analysis is presented.

Table 13 contains two columns of numbers. One columns consists of three-category thresholded regression results and the other consists of three-category discriminant analysis results. This table allows comparison of the results of the thresholded regression models with corresponding three-category discriminant analysis results. When comparing these results it can be seen that the gain in prediction accuracy from discriminant analyses when compared with that of multiple regression is negligible. Two-category thresholded multiple regression analyses were not carried out for comparison with the two-category discriminant analyses because the results would have corresponded closely with the three-category results.

Table 13: Comparison of Apparent Accuracy Rates for Thresholded Regression Models and Corresponding Discriminant Analysis Models. (Comparisons are for the Steering, Accelerometer, LANDEVNAR, LNMNSQ, LANEX, and LNERRSQ independent measure cases).

Definitional (Dependent) Measures	Type of Accuracy	Regression Results	Three-Category Discriminant Analyses
AVEOBS	APAR (Large Errors)	0.957	0.967
	APAR (All Errors)	0.753	0.770
EYEMEAS	APAR (Large Errors)	0.937	0.900
	APAR (Large Errors)	0.780	0.827
NEWDEF	APAR (All Errors)	0.957	0.920
	APAR (All Errors)	0.710	0.717
PERCLOS	APAR (Large Errors)	0.980	0.973
	APAR (All Errors)	0.790	0.850
MASTER	APAR (Large Errors)	0.980	0.970
	APAR (All Errors)	0.830	0.823
AVERAGES OF ABOVE	APAR (Large Errors)	0.962	0.946
	APAR (All Errors)	0.772	0.797

DISCUSSION AND CONCLUSIONS

In general, the five definitional (dependent) measures of drowsiness that were employed in the algorithm development phase were reasonably predictable. Discriminant analyses showed, in particular, that the number of large errors is relatively small. In fact, in a few cases the number of large errors is as low as one or two per 100 cases. In addition, the difference between discriminant analysis results and multiple regression results was quite small in many cases and nonexistent in others.

Models Including Heart Measures Versus Models Not Including Heart Measures

The potential gains are quite modest if it is assumed that we could secure a plethysmograph to an automobile driver. When comparing heart and non-heart models in Tables 10-12 (lines 1 through 5 versus 11 through 15) the number of open cells (cells containing no R value or APAR number), cells containing squares, and cells containing solid triangles in lines 11 through 15 demonstrates that in many cases there was no improvement when heart rate measures were introduced. The R value of NEWDEF improved by approximately 0.05, when heart measures are added. However, in most cases the addition of the heart rate variables contributes little or nothing to the prediction accuracy of drowsiness. On the whole, it is not worth encumbering the driver with a plethysmograph to obtain heart rate measures for the slight improvement in the prediction of drowsiness.

Models Including A/O-Task Measures Versus Models Not Including A/O-Task Measures

Models in which A/O-task measures have been introduced produce relatively high predictive values compared with non-A/O task measure results as seen in Tables 10 through 12. The number of large errors is in the range of 1% or less. Results of this nature suggest that the A/O task does contribute to prediction accuracy. However, it must be recognized that the A/O-task models are based on data from four subjects and therefore the models may have higher R values and APAR values because it is easier to fit a model to four subjects than to twelve subjects.

Overview

We conclude on the basis of Table 13 that regression models, after thresholding, are capable of producing comparable accuracy to three-category (two-threshold) discriminant analysis models. In fact, large errors are slightly fewer in multiple regression than in discriminant analysis. However, the total number of errors is slightly greater for multiple regression than for discriminant analysis.

While the models developed in this study are relatively accurate, the fact remains that they will produce some false alarms (a false alarm is defined as an outcome in which an alert driver is diagnosed as drowsy). The best estimate of the false alarm rate for drivers who have been sleep deprived is that given by the large error APARs that appear in Table 13. The results suggest that error rates of 2% to 3% are likely to occur. Error rates are of course dependent on the proportion of time that drivers are alert; questionable, and drowsy. (Error rates for alert drivers are likely to be lower because they would be less likely to produce model outputs near threshold.)

The fact that a finite false alarm rate remains suggests that a two-stage detection algorithm procedure should be used. In the first stage, the A/O task would not be performed and an algorithm appearing in rows one through five of Table 9 would be used. Once threshold is exceeded, indicating potential drowsiness, the driver would be asked to perform the A/O task. If the A/O-task algorithm then produced a value above threshold the driver would be assumed to be drowsy. A two-step algorithm of this type might produce sufficiently low false alarm rates so as to be acceptable for applications.

The first stage of detection involves driver-vehicle performance measures only. It is suggested that the first and third rows of Table 9 represent the most viable algorithms. The third row assumes the availability of a lane track and provides better accuracy than row one. However, if a lane track is not available the first row could be used. Algorithms in the first row of Table 9 include steering and lateral accelerometer measures. These two measures are assumed to be nearly 100% reliable and should be used exclusively if a valid lane track is not available.

With regard to the predictability of the definitional measures of drowsiness, results demonstrate that PERCLOS and MASTER are most predictable, followed by EYEMEAS, AVEOBS, and NEWDEF. These results are best seen in Table 13 but also show up in Tables 9, 10, 11, and 12. In general, the results suggest that the algorithms developed by regression and using a threshold with a two stage process should provide a viable, accurate, and low false alarm system of detection for drowsy drivers.

On the basis of Table 13 it would be concluded that a two-category regression model would have comparable accuracy to a two-category discriminant analysis model. Because of the comparable accuracy obtainable for regression models it is recommended that only regression models with thresholds be implemented in future validation and full scale studies. The advantage of using thresholded regression models is that the threshold(s) can be adjusted for sensitivity in operational settings. Discriminant analysis models, on the other hand, must be recomputed for each new setting of threshold. This would involve an on-line optimization process.

Chapter Five

Validation of Previously-Developed Drowsy-Driver Detection Algorithms

(This chapter represents an extended summary of work reported in the Fifth Semiannual Research Report, dated April 15, 1994, and referred to as Wreggit, Kirn, and Wierwille, 1994)

INTRODUCTION

The validation experiment was a multipurpose experiment, designed not only for validation, but also for the examination of several additional research issues. Chapter Six describes these additional issues, including corresponding analyses and results. The study described in the present chapter was directed at determining how well previously developed drowsy-driver detection algorithms (Wreggit, Kirn, and Wierwille, 1993) perform when they were applied to new data. Numerous algorithms were developed and have been previously reported (Wreggit, Kirn, and Wierwille, 1993). While estimates of algorithm accuracy were obtained along with the development of the algorithms, it was not certain that such estimates could be relied upon for new groups of drivers operating under similar conditions. Therefore, this study focuses on the application of typical algorithms to a new data set, to obtain a "validated" estimate of accuracy. Accordingly, this experiment was conducted having the primary purpose of algorithm validation, that is, determining algorithm classification accuracy for data from a new set of driver-subjects. Accuracy of typical algorithms applied to the new data set was determined through the use of multiple regression and classification matrices. The accuracy of the classification matrices constructed during the algorithm development study was compared with the accuracy of the classification matrices constructed in the validation study. A comparison of resulting R values from the two studies was also undertaken

Sleep deprived subjects drove an automobile simulator for approximately 2 l/2 hours during the night. The driving time and duration were approximately the same as was experienced by subjects in the algorithm development study.

In the algorithm development study four subjects did nothing but drive, another four performed occasional tasks as they drove, and another four performed a subsidiary task, called an A/O task, as they drove. The A/O task consisted of auditory presentation of words. The subjects responded by means of push-buttons labeled YES and NO, depending on whether or not the presented word contained the target letters A or O. In the validation experiment, all driver-subjects drove with cruise control for two quartiles of the run and without cruise control for the other two quartiles. They also performed the A/O task for two quartiles and did not perform the task during the other two quartiles. During the validation experiment the subjects did not interact with the dash board controls. Each driver/subject thus experienced four quartiles in which all combinations of cruise/no cruise and A/O task/no A/O task were presented. Order of presentation was counterbalanced across subjects. The reasons for using a slightly modified design in the validation experiment were:

1. It was desired to determine algorithm accuracy under similar, but not identical conditions, thereby "simulating" the likely conditions of an application, and
2. The data could be used for additional purposes, such as determining the effects of cruise control on algorithm detection accuracy.

METHOD

Subjects

In this validation study, data were collected from twelve subjects, as was the case for data collection for the algorithm development study (Wreggit, Kirn, and Wierwille, 1993). The subject population was located in the Blacksburg, Virginia area and the same screening procedures were used as the previous phase of the study. However, eight males and four females were used during the validation study instead of six males and six females. The use of twice as many males as females was determined to be a more accurate representation of the high risk driver population (Knipling and Wierwille, 1993). The subjects ranged in age from 18 to 47. The subjects were paid according to the hourly rate of the previous phase of the study and were involved with the experiment for approximately the same amount of time.

During data collection one subject stayed completely awake and had a heart rate of 90 beats per minute for the entire run. It was suspected by the experimenters that this subject may have taken caffeine pills or some other form of stimulant during a trip to the rest room prior to the driving session. This subject's data were not used. While running another subject the EEG electrodes loosened late in the run. Examination of the data led the experimenters to suspect that the EEG data had been corrupted and therefore, the subject's data were not used. The two problem subjects were replaced with two additional subjects, resulting in a total of twelve complete data sets.

Apparatus

The apparatus employed was identical to that of the previous phase of the study with one exception.

Simulator. The simulator was equipped with a cruise control system that allowed the experimenters to place the simulator in a cruise-control state which locked the velocity of the simulated automobile at 60 miles per hour. The cruise control could also be switched off by the experimenters, at which time it was necessary for the driver to maintain the speed of the automobile using the accelerator pedal.

Experimental Design

The experimental design involved a regression approach to data analysis. All drivers were subjected to the following conditions during driving: with cruise control/with task, with cruise control/without task, without cruise control/with task, and without cruise control/without task. The "task" in this case refers to the A/O subsidiary task described previously in the algorithm development chapter of this report. Each subject performed the A/O task for one-half of the entire run. Therefore, it was necessary for each subject to perform the A/O task for 72 minutes. While the A/O task performance measures were being collected, all other measures were being collected simultaneously. Subjects received counter balanced combinations of the conditions. Each condition lasted 36 minutes. The subjects did not interact with the instrument panel as was done by some subjects in the algorithm development phase.

The performance and physiological measures that were gathered during the study were the same as the performance and physiological measures included in the previously developed drowsiness detection algorithms.

Procedure

All subjects underwent the same pre-driving procedures as the subjects in the development phase and stayed at the Vehicle Analysis and Simulation Lab for approximately the same amount of time.

Experimental task. All subjects drove the simulated automobile as if it were an actual car. All subjects performed the same secondary (A/O) task that was employed during the algorithm development phase. In addition, a cruise control condition was incorporated into the driving task. When the cruise control was engaged the simulated automobile maintained 60 miles per hour. When the cruise control was not engaged the subject was asked to maintain approximately 60 miles per hour. Subjects drove for a total of 156 minutes.

Data Analysis Overview

All measures were first computed over one-minute intervals. Data manipulation procedures were then undertaken to prepare data for statistical analyses. Initially, the first two minutes from all measures were deleted. This was done so that the data to be analyzed did not include the time when subjects were “settling in” to the driving task. This procedure was consistent with the algorithm development phase.

All independent measures were baselined using the average of the first ten minutes of data (after the actual first two minutes had been deleted). The average of the first ten minutes of data was then subtracted from every subsequent data point within that measure. Each data point consisted of a one-minute average of data. After completion of the baselining procedure the ten minutes of data used for the baselining average were discarded. Following the baselining procedure the data were averaged across six-minute intervals. See Figure 16 for a pictorial overview of the data manipulation procedure.

As seen in Figure 17 the dependent (definitional) variable PERCLOS was graphed for each subject with threshold lines drawn. The threshold lines were developed during the algorithm development study and were placed over the new data set. In this graph, the first 24 points on the abscissa correspond to subject 1, the next 24 points correspond to subject 2, and so on. The corresponding graphs for AVEOBS, EYEMEAS, NEWDEF, and MASTER can be seen in Wreggit, Kim, and Wierwille (1994).

After data manipulation, previously developed drowsiness detection algorithms were applied to the new data set. Once algorithm outputs (predicted values) were calculated, a regression analysis was run between those values and the applicable definitional measure (observed) values. After completion of this procedure a comparison between the R values attained from the original data and new data-was carried out. The algorithms that were tested can be seen highlighted in Table 14. (Those with gray background were not tested.)

2 minutes: The first 2 minutes are deleted from all data

First 10 minutes after deletion of 2 previous minutes
10 minute average calculated

10 minutes used for baselining*
10 minutes discarded after baseline

6 minutes: average of 6 one-minute measure values

6 minutes: average of 6 one-minute measure values

6 minutes: average of 6 one-minute measure values

6 minutes: average of 6 one-minute measure values'

6 minutes: average of 6 one-minute measure values

6-minute averages continued until the entire set of one-minute segments (144 minutes) was manipulated. 156 total minutes were used due to the deletion of the first 12 minutes of data. Therefore, 24 data points of 6-minute averages were created.

* Baselining is a procedure in which the initial ten minutes of data are averaged and then subtracted from all subsequent one-minute segments. Baselining was carried out so that data relative to the subject's initial data values could be obtained.

Figure 16: Pre-Analysis Data Manipulation Procedures.

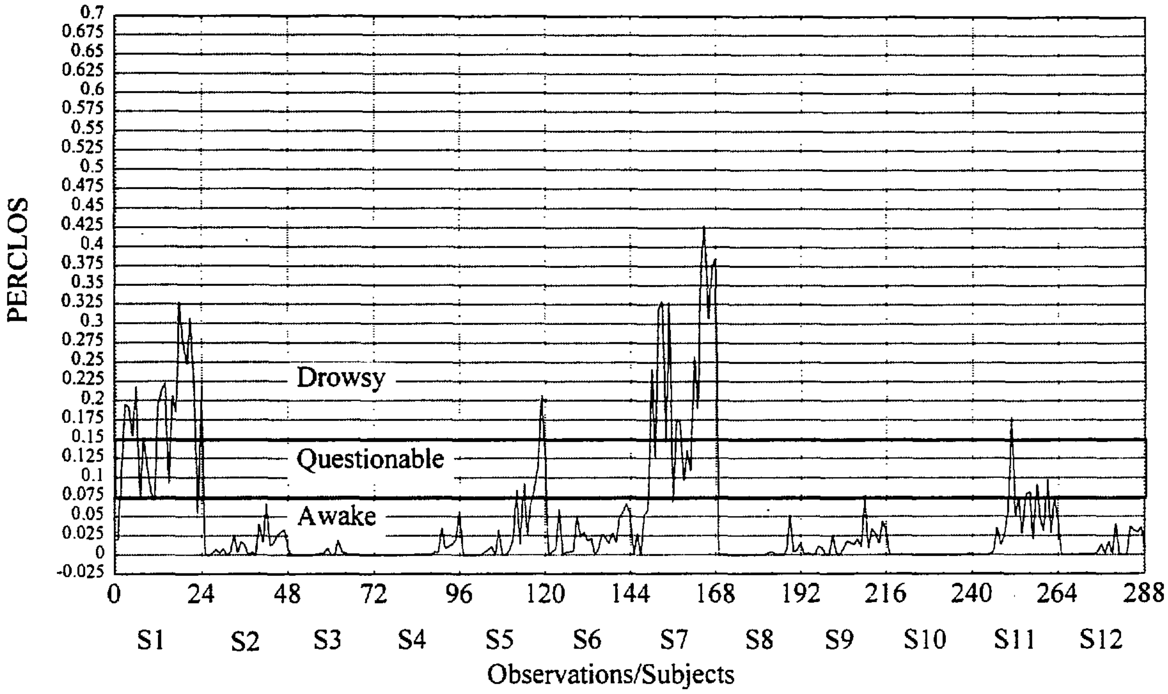


Figure 17: PERCLOS Data With Upper and Lower Criterion Lines (New Data).

Table 14: Summary Table of Multiple Regression Results (Calculated in the Development Phase) Showing Algorithms Used for Validation.

Independent Measures		Dependent Measures				
		AVEOBS	EYEMEAS	NEWDEF	PERCLOS	MASTER
D	Steering and Accelerometer	0.747	0.764	0.677	0.789	0.801
E	Steering, Accelerometer, & HPHDGDEV/VAR	0.793	0.809	0.700	0.847	0.852
F	Steering, Accelerometer, LANDEV/VAR, LNMNSQ, LANEX, & LNERRSQ	0.826	0.837	0.731	0.872	0.886
G	Steering, Accelerometer, & all lane measures (includes LNRTDEV/VAR)	0.824	▲	0.757	0.872	▲
H	Steering, Accelerometer, all lane measures, & DSYAWDEV/VAR	✳ 0.826	0.836	✳ 0.751	▲	▲
I	A/O Task Measures Only	0.761	0.768	0.660	0.810	0.822
J	A/O Task, Steering, & Accelerometer	Accel. 0.824	0.824	Accel. 0.740	0.836	0.876
K	A/O Task, Steering, Accelerometer, & HPHDGDEV/VAR.	0.917	Steering 0.855	▲	Accel. 0.868	0.903
L	A/O Task, LANDEV/VAR, LNMNSQ, LANEX, & LNERRSQ	Δ	0.874	0.768	0.875	0.903
M	A/O Task, Steering, Accelerometer, LANDEV/VAR, LNMNSQ, LANEX, & LNERRSQ	0.922	Δ	Δ	0.902	0.936
N	Heart, Steering, & Accelerometer	0.785	0.772	0.711	❐	❐
O	Heart, Steering, Accelerometer, & HPHDGDEV/VAR	0.813	❐	0.761	0.851	0.854
P	Heart, Steering, Accelerometer, LANDEV/VAR, LNMNSQ, LANEX, & LNERRSQ	0.838	❐	0.774	0.824	❐
Q	Heart, Steering, Accelerometer, & all lane measures (includes LNRTDEV/VAR)	✳ 0.826	❐	0.802	❐	❐
R	Heart, Steering, Accelerometer, all lane measures, & DSYAWDEV/VAR	Lane Rate 0.837	❐	0.797	▲	❐
S	A/O Task and Heart	❐	❐	0.774	❐	❐
T	A/O Task, Heart, Steering, & Accelerometer	Accel. 0.837	❐	0.810	❐	❐
U	A/O Task, Heart, Steering, Accelerometer, & HPHDGDEV/VAR	0.918	❐	▲	Accel. 0.880	0.909
V	A/O Task, Heart, LANDEV/VAR, LNMNSQ, LANEX, & LNERRSQ	❐	❐	0.823	❐	0.910
	A/O, Heart, Steering, Accelerometer, LANDEV/VAR, LNMNSQ, LANEX, & LNERRSQ	❐	Δ	ΔΔ	❐	❐

KEY:

▲ In regression analyses introduction of variable did not improve R value. See entry directly above for model with same R and fewer terms.

❐ Heart Measures did not improve regression as compared with non-heart equivalent. See corresponding non-heart entry for model with same R value and fewer terms.

Δ A/O task measures did not improve regression as compared with non-A/O task measure equivalent. See corresponding non-A/O task measure entry for model with same R value and fewer terms.

✳ A cell designated with an asterisk could have been given the ▲ symbol. However, the asterisk denotes a substantially changed algorithm in term of measures used.

NOTES: - Letter in left-hand column corresponds to appendix (Wreggit, Kim, and Wierwille, 1993) in which analysis is presented.

- Cells that are not grayed-out indicate algorithms that were validated by applying new data.

The comparison between the R values attained from the original data and the new data was accomplished using t-test and analysis of variance procedures. Multiple R values were used as data for these comparisons.

The algorithms that were chosen for validation were selected for several reasons. It was desirable for the R values to be relatively high and for the measures within the algorithms to be attainable in anbn-the-road vehicle. Also, it was necessary to choose algorithms that could be employed in a step-up, step-down procedure. For example, if all incoming signals to be used in an algorithm are valid, the best available algorithm for drowsiness detection would be used. However, if one of the sensors necessary for the best algorithm is not providing a valid signal, the next best algorithm that does not require the invalid signal would be used.

VALIDATION RESULTS:
DRIVER-VEHICLE PERFORMANCE MEASURES ONLY

This section describes the validation process for algorithms using driver-vehicle performance measures only. During the experimental runs there were intervals during which the A/O task was performed and there were intervals during which the A/O task was not performed. Similarly, there were intervals during which the cruise control was engaged and during which it was not engaged. Throughout these various intervals, driver-vehicle performance measures were computed. This section reports on the validation results using the driver-vehicle performance measures only. That is, it does not include measures taken from the A/O task itself and it does not include any attempt to include forward speed in algorithm validation. The term "all-data" indicates that performance data are included from all 156 minutes of each driver's data run, regardless of whether or not the A/O task was being performed and regardless of whether or not the simulated vehicle was in cruise. When specific sections of the data runs are referred to they are so designated. For example, the section of the run in which the A/O task was being performed and cruise was not engaged is referred to as "With Task, W/O Cruise."

Application of Algorithms to New Data

Table 15 is a summary of 1) results that were attained from multiple regression analyses of the original (algorithm development) data and 2) the correlation between new observed data and the algorithm output when the algorithm was applied to new data. The R values attained from the original data set are included in this table so that easy comparison between R values can be made. There was no general decrease in predictive power of the algorithms when applied to the new data $t(9) = 0.24$, $p > 0.05$. The average R values of the original and new data can be seen graphically in Figure 18.

The new data were divided into four categories, including combinations of cruise control and A/O secondary task so that the effects of cruise control and A/O task could be

Table 15: R Values From Multiple Regression Analyses of Original Data and R Values Achieved After Application of Algorithms to New Data.

	Independent Measures	Dependent Measures				
		AVEOBS	EYEMEAS	NEWDEF	PERCLOS	MASTER
D	Steering and Accelerometer	**(original)** **0.747** **Algorithm** **Dla** **(new)** **0.727**	(original) **0.764** Algorithm D2a (new) **0.777**	(original) **0.677** Algorithm D3a (new) **0.746**	**(original)** **0.789** **Algorithm** **D4a** **(new)** **0.800**	**(original)** **0.801** **Algorithm** **D5a** **(new)** **0.837**
F &	Steering, Accelerometer, LANDEVNAR, LNMNSQ LANEX, LNERRSQ	(original) **0.826** Algorithm F1a (new) **0.570**	**(original)** **0.837** **Algorithm** **F2a** **(new)** **0.838**	**(original)** **0.731** **Algorithm** **F3a** **(new)** **0.819**	**(original)** **0.872** **Algorithm** **F4a** **(new)** **0.862**	**(original)** **0.886** **Algorithm** **F5a** **(new)** **0.885**

NOTES: Letters in left column indicate appendices containing detailed analyses on original data set (Wreggit, Kim, and Wierwille, 1994).

Algorithm numbers located in each cell correspond to the multiple regression table within a given appendix (Wreggit, Kim, and Wierwille, 1994).

Classification matrices were created for the highlighted (bolded) R values (Wierwille, Kim, and Wreggit, 1994). (Also see Appendix A.)

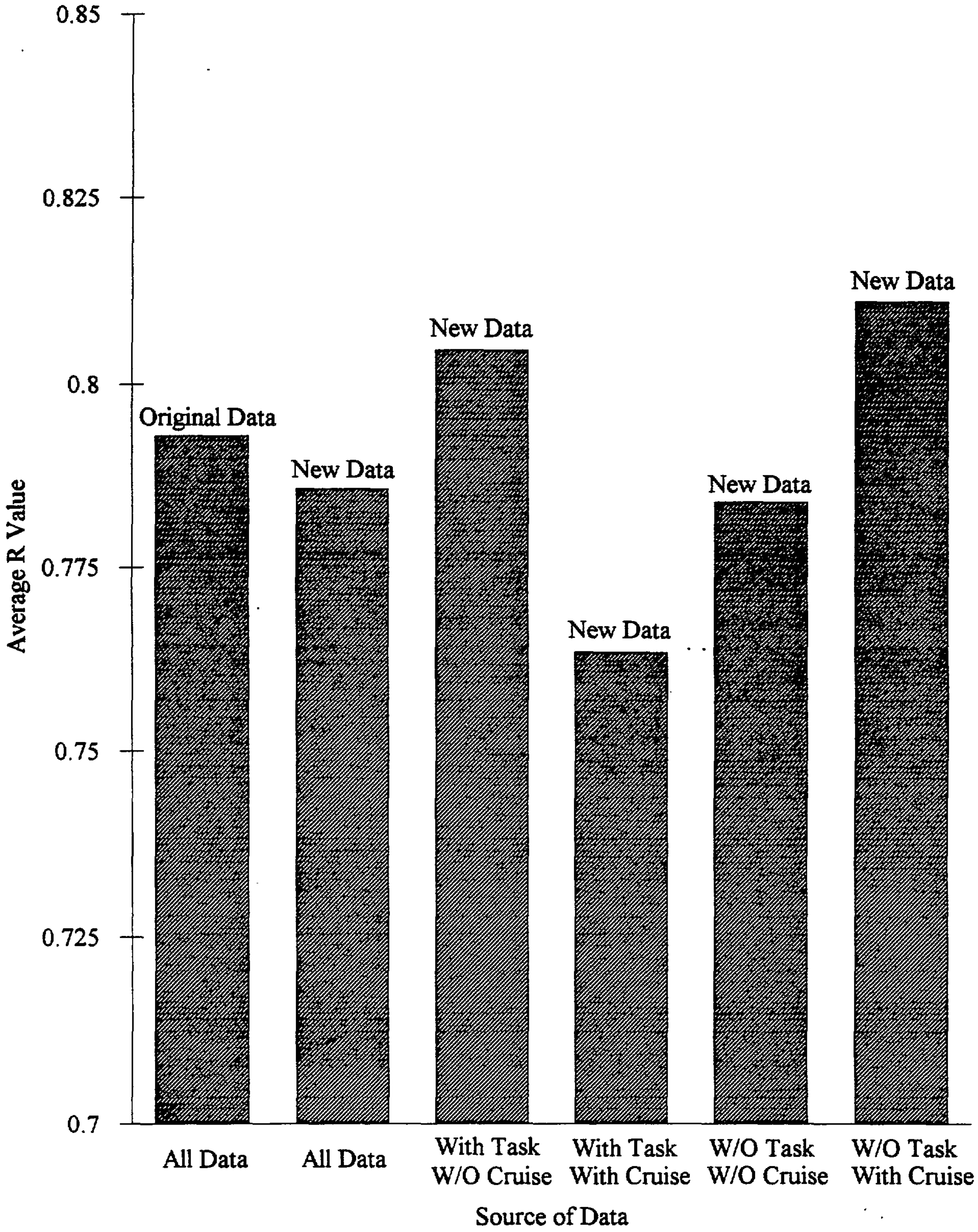


Figure 18: Average R Values for Ten Algorithms (See Table 15) Applied to Original Data, New Data, and Four Conditions From the New Data

examined (Figure 18). A 2 x 2 within subjects analysis of variance was performed to test for the effects of cruise control and A/O task performance on R values. No significant main effects were seen (Cruise Control: E(1, 9) = 0.177, $p > 0.05$, A/O Task: F(1, 9) = 0.129, $p > 0.05$). However, a significant cruise control by A/O task interaction was indicated by the results of the analysis of variance F(1,9) = 10.67, $p < 0.01$. To determine how the groups differed, a Tukey. HSD test was used. The results of the post hoc test showed that only the Without Task/with Cruise condition and the With Task/With Cruise condition were significantly different from one another at the a = 0.05 level. The differences between the other pairs of conditions were not significant at the a = 0.05 level.

Graphing the new observed data (definitional measure values) and the new predicted data (from application of algorithms) demonstrated that it was possible to track any portion of the new data with the previously developed detection algorithms. In Figure 19, predicted data tracks the observed PERCLOS data quite accurately. Corresponding graphs for AVEOBS, EYEMEAS, NEWDEF, and MASTER can be seen in Wreggit, Kim, and Wierwille (1993).

Figure 20 shows typical classification matrices that were generated from a thresholded multiple regression analysis of the definitional measure PERCLOS. The upper matrix shows classified original data (algorithm output) and the lower matrix shows classified new data (algorithm output). The data that were classified are the same as those graphed in Figure 19. The thresholds that were used for the purpose of classification were the same as those produced during the algorithm development phase. The thresholds for PERCLOS are illustrated in Figure 17. Table 16 is a summary of the apparent accuracy rates generated from the various classification matrices. This table shows that the difference between the apparent accuracy rates of original and new data were negligible. See Wreggit, Kirn, and Wierwille (1994) for the corresponding classification matrices.

Regression lines were drawn using the bolded original R values seen in Table 15 and the corresponding APAR values of the new data in Table 16. The two plots give a general

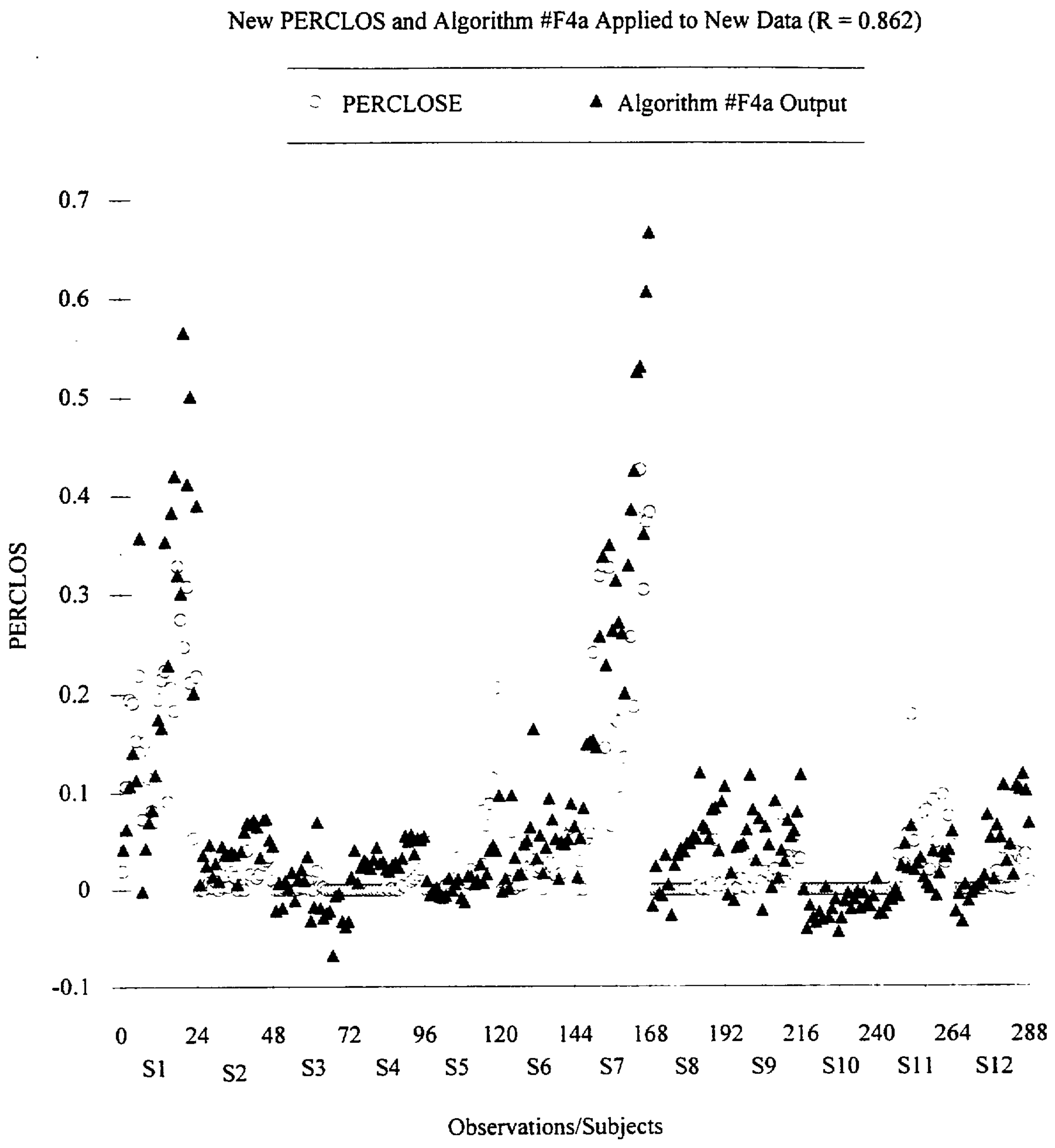
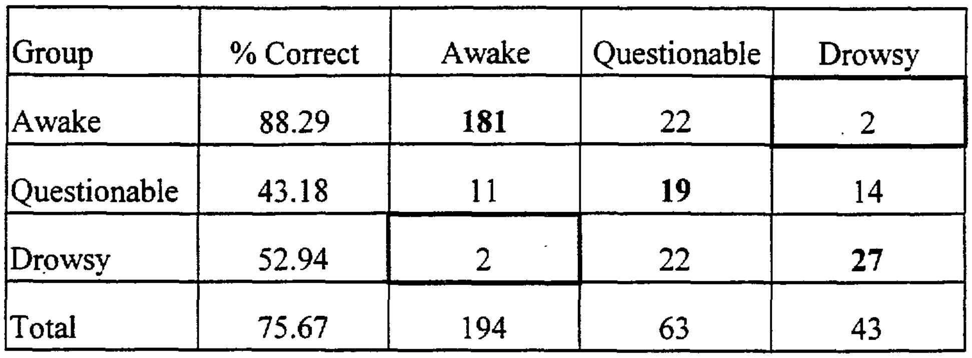


Figure 19: Scatter Plot of PERCLOS Data -- Predicted vs. Observed.

Predicted



	Group	% Correct	Awake	Questionable	Drowsy
Original	Awake	88.29	**181**	22	2
Observed	Questionable	43.18	11	**19**	14
	Drowsy	52.94	2	22	**27**
	Total	75.67	194	63	43

PERCLOS (R Value = 0.789)

Apparent Accuracy Rate (large misclassifications): 0.987
Apparent Accuracy Rate (all misclassifications): 0.757

Classification Matrix Generated From Multiple Regression Analysis of Original PERCLOS Data Resulting in Algorithm D4a. (Independent variables employed included Steering and Accelerometer.)

Predicted

	Group	% Correct	Awake	Questionable	Drowsy
New	Awake	79.32	188	40	9
Observed	Questionable	30.00	8	6	6
	Drowsy	90.32	1	2	28
	Total	77.08	197	48	43

PERCLOS (R Value = 0.800)

Apparent Accuracy Rate (large misclassifications): 0.965
Apparent Accuracy Rate (all misclassifications): 0.771

Algorithm D4a Applied to New Data and Compared with New Observed **PERCLOS** Data

Figure 20: Classification Matrices Showing Accuracy of Algorithm D4a When Applied to Original Data (Upper) and New Data (Lower)

Table 16: APAR Summary Table Generated from Classification Matrices of Original and New Data -- Including All Misclassifications and Large Misclassifications (APAR values correspond to R values in Table 15).

Independent Measures		Dependent Measures				
		AVEOBS	EYEMEAS	NEWDEF	PERCLOS	MASTER
D	Steering and Accelerometer	(original) Large 0.980 All 0.773 Algorithm D5a (new)			(original) Large 0.980 All 0.773 Algorithm D5a (new)	(original) Large 0.980 All 0.773 Algorithm D5a (new)
F	Steering, Accelerometer, LANDEV/VAR, LNMNSQ, LANEX, & LNERRSQ		(original) Large 0.980 All 0.773 Algorithm D5a (new)	(original) Large 0.980 All 0.773 Algorithm D5a (new)	(original) Large 0.980 All 0.773 Algorithm D5a (new)	(original) Large 0.980 All 0.773 Algorithm D5a (new)

NOTES: Letters in left column indicate appendices containing detailed analyses on original data set (Wreggit, Kirn, and Wierwille, 1994).

Algorithm numbers located in each cell correspond to the multiple regression table within a given appendix (Wreggit, Kirn, and Wierwille, 1994). (Also see Appendix A.)

idea of what APAR value can be achieved given certain drowsiness prediction R values for twelve subjects. In other words, if an algorithm based on data from twelve subjects is developed, it can be expected to produce APAR values in a validation study as provided by the regression lines in Figure 21. The reader is cautioned that the correlation coefficients associated with the data are not significant ($p > 0.05$) (probably as a result of small sample size), and therefore the prediction capabilities provided by Figure 2 1 are indicative and not conclusive.

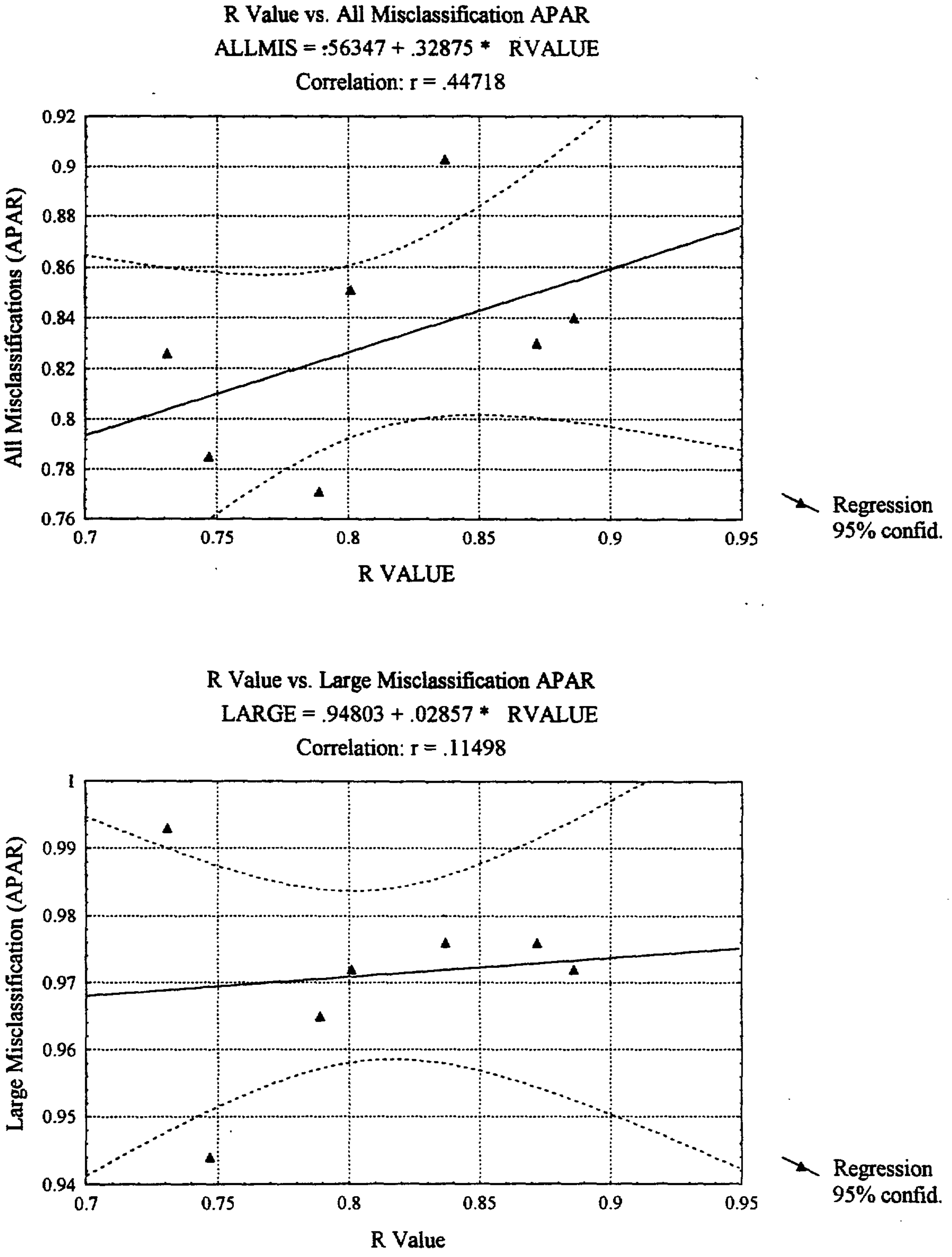


Figure 21: R Values vs. New Data APAR Values -- APAR Includes All Misclassifications (Upper Graph) and Large Misclassifications Only (Lower Graph).

VALIDATION RESULTS:

INCLUSION OF A/O TASK PERFORMANCE MEASURES

Models Containing A/O Task Performance Measures

This section contains the results of validation tests that include A/O task performance measures. In some cases these measures were used by themselves in developed algorithms and in other cases they were used in combination with driver-vehicle performance measures. Since the A/O task was performed during only half of each subject’ s data run, the data base used in this section is half the size of that used in the previous section.

Application of Algorithms (Employing A/O Task Measures) to New Data

As seen in Figure 22 the dependent (definitional) variable PERCLOS was re-graphed to include only the segments of time in which the subjects performed the A/O task. The threshold lines in the figure were developed during the algorithm development phase of the study and are the same as those seen in Figures 13 and 17. In the graphs, the first 12 points on the abscissa correspond to subject 1, the next 12 points correspond to subject 2, and so on. The corresponding graphs for AVEOBS, EYEMEAS, NEWDEF, and MASTER can be seen in Wreggit, Kirn, and Wierwille (1994).

Table 17 is similar to Table 16 in that is contains a description of the algorithms that were tested on the new data. However, Table 17 contains the original R values (those associated with application of the algorithms to the original data set) and new R values (those associated with application of the algorithms to the new data set) for the algorithms containing A/O data. The table also shows the appendices in which the algorithms were presented (Wreggit, Kim, and Wierwille, 1993).

Table 17 shows that there was a general decrease in predictive power of the algorithms when applied to the new A/O data (average R value = 0.606) as compared with the original data (average R value = 0.809) $t(7) = 6.21$, $p < 0.01$. This result is graphed in Figure 23. The figure also shows the effects of cruise control on the new R values. When cruise control was

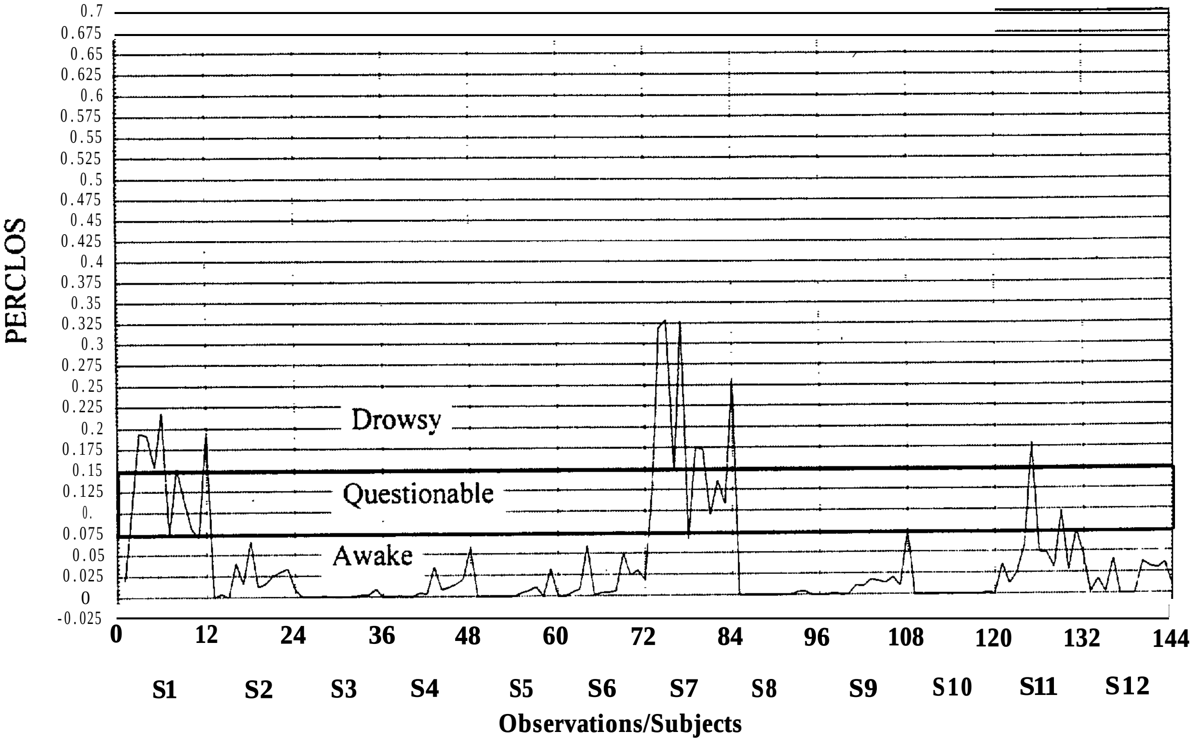


Figure 22: PERCLOS Data With Upper and Lower Criterion Lines (New Data During A/O Task).

Table 17: R Values From Multiple Regression Analyses of Original A/O Data and R Values Achieved After Algorithms Were Applied to New A/O Data.

Independent Measures		Dependent Measures				
		AVEOBS	EYEMEAS	NEWDEF	PERCLOS	MASTER
I	A/O Task Measures Only	(original) 0.761 Algorithm I1a (new) 0.595	(original) 0.768 Algorithm I2a (new) 0.570	(original) 0.660 Algorithm I3a (new) 0.422	(original) 0.810 Algorithm I4a (new) 0.447	(original) 0.822 Algorithm I5a (new) 0.570
J	A/O Task, Steering, & Accelerometer	-------------	-------------	-------------	(original) 0.836 Algorithm J4a (new) 0.599	-------------
L	A/O Task, LANDEV/VAR, LNMNSQ, LANEX, & LNERRSQ	-------------	-------------	-------------	(original) 0.875 Algorithm L3a (new) 0.796	-------------
M	A/O Task, Steering, Accelerometer, LANDEV/VAR, LNMNSQ, LANEX, & LNERRSQ	-------------	-------------	-------------	-------------	(original) 0.936 Algorithm M3a (new) 0.845

NOTES: Letters in left column indicate appendices containing detailed analyses on original data set Wreggit, Kirn, and Wierwille, 1994).

Algorithm numbers located in each cell correspond to the multiple regression table within a given appendix (Wreggit, Kirn, and Wierwille, 1994).

Classification matrices were created for the highlighted (bolded) R values(Wreggit, Kirn, and Wierwille, 1994). (Also see Appendix A for classification matrices and regression summaries corresponding to algorithms 14a, J4a, L3a, and M3a.)

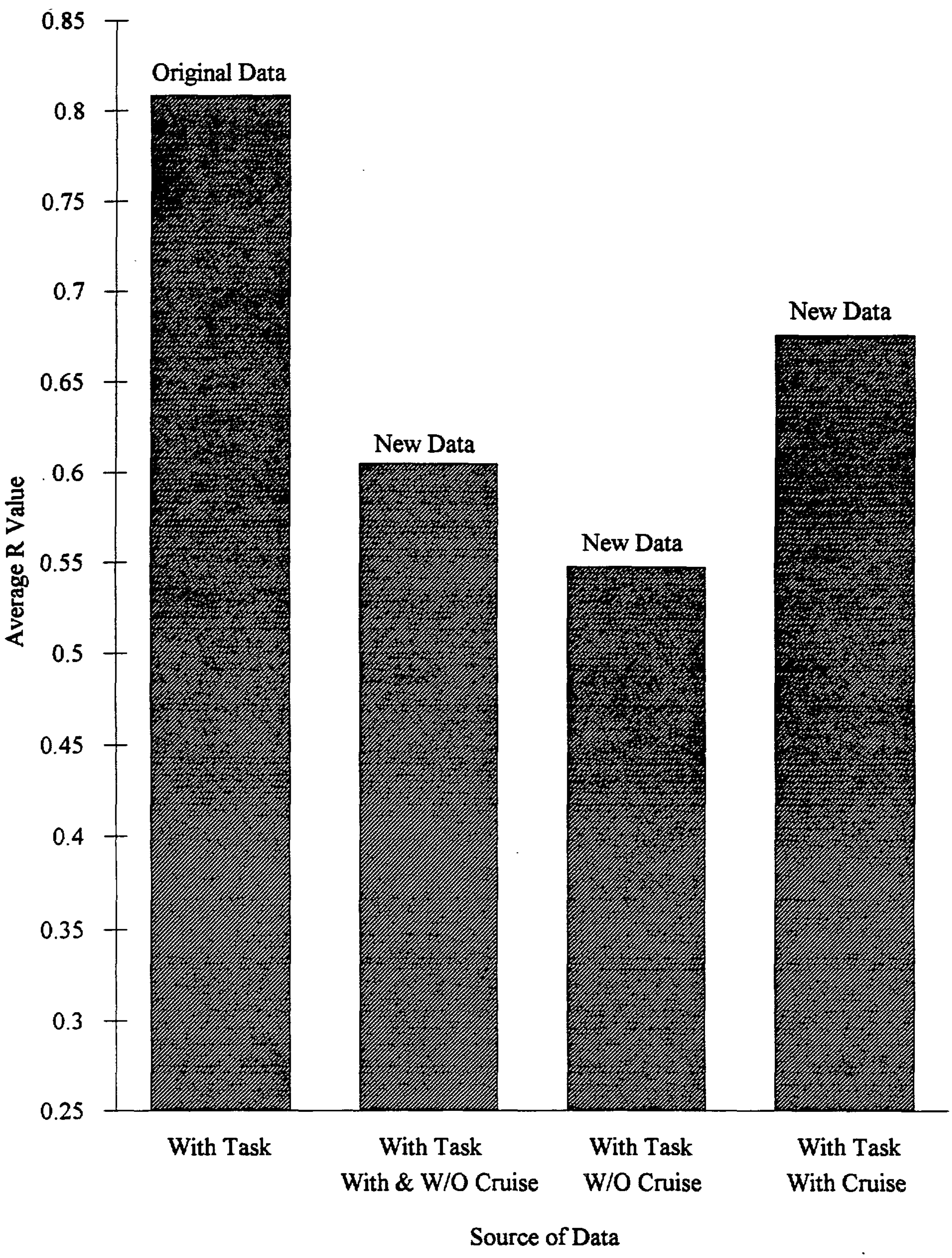


Figure 23: Average R Values for Eight Algorithms (Table 17) Applied to Original A/O Data, New A/O Data, New A/O Data With Cruise Control Engaged, and New A/O Data Without Cruise Control Engaged.

engaged, the new R values increased significantly (from an average of 0.549 (when not engaged) to an average of 0.677 (when engaged); t(7) = 2.50, $p < 0.05$.

To obtain a better understanding of drowsiness prediction using the new data. graphical comparisons were made between the definitional measures and the algorithm outputs applied to new data. Figure 24 shows a scatter plot of PERCLOS and algorithm J4a. As can be seen, the algorithm seems to do a reasonable job of tracking the variations in observed (definitional) measures, even though there are some obvious discrepancies. The corresponding graphs for AVEOBS, EYEMEAS, NEWDEF, and MASTER can be seen in Wreggit, Kim, and Wierwille (l994).

Figure 25 shows typical classification matrices that were generated from a thresholded multiple regression analysis of the definitional measure PERCLOS. The upper matrix shows classified original data (algorithm output) and the lower matrix shows classified new data (algorithm output). The data that were classified are the same as those graphed in Figure 24. The thresholds for PERCLOS are illustrated in Figure 22.

Table 18 is a summary of the apparent accuracy rates generated from the various classification matrices (a complete set of the developed matrices can be seen in Wreggit, Kim, and Wierwille, 1994). In general, the number of misclassifications appears to be smaller than the R values would seem to indicate.

Finally, regression lines were drawn using the original R values shown in Table 17 and the corresponding APAR values of the new data in Table 18. The regression lines are shown in Figure 26. The two plots give a general idea of the APAR value that can be achieved given certain drowsiness prediction R values. The reader is cautioned that the correlation coefficients associated with the data are not significant ($p > 0.05$), and therefore the prediction capabilities provided by Figure 26 are indicative and not conclusive. (Again, the probable reason for nonsignificance is sample size.)

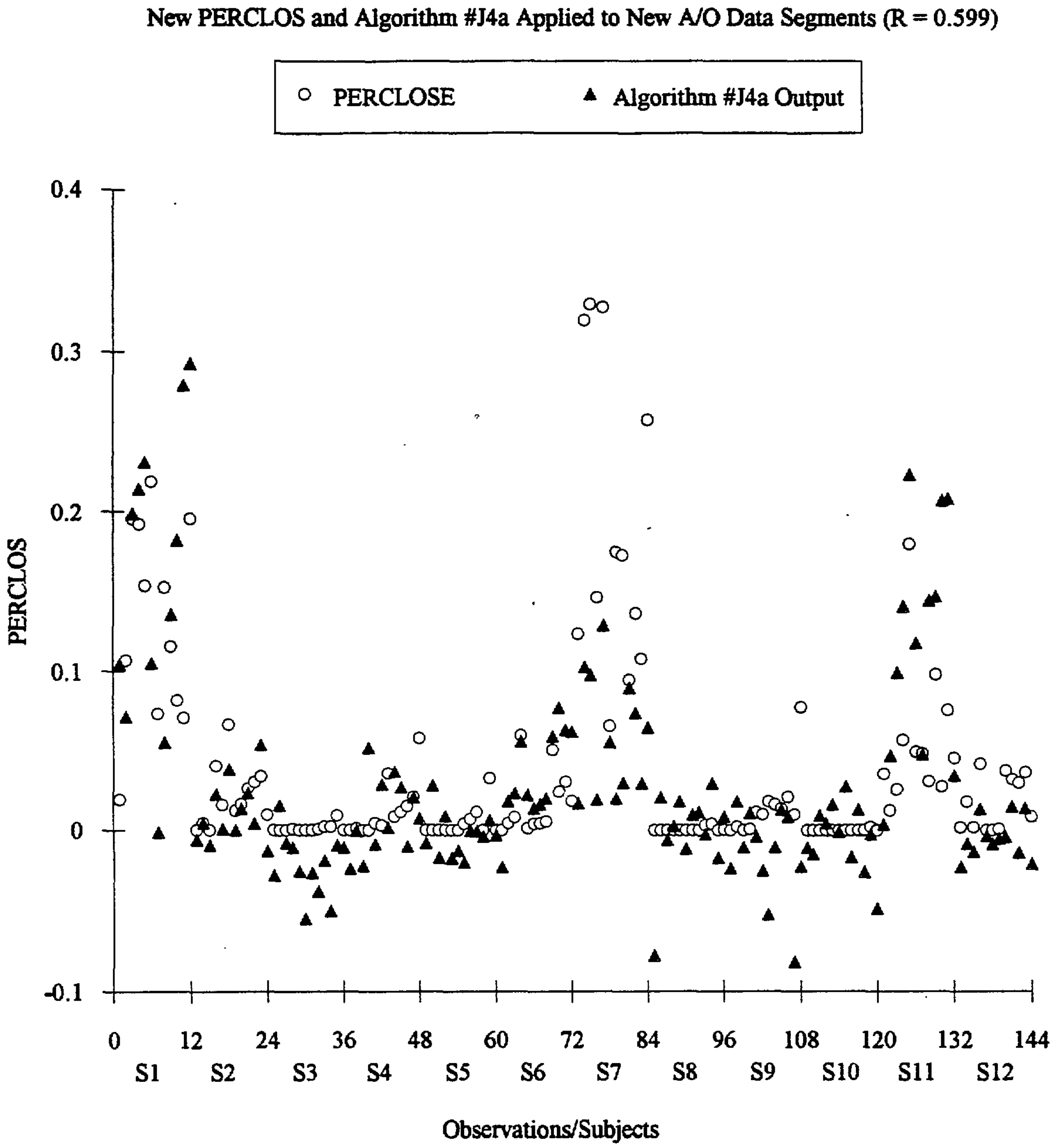


Figure 24: Scatter Plot of PERCLOS Data -- Predicted vs. Observed

Predicted

	Group	% Correct	Awake	Questionable	Drowsy
Original	Awake	94.29	**66**	3	1
Observed	Questionable	43.75	2	**7**	7
	Drowsy	42.86	2	6	**6**
	Total	79.00	70	16	14

PERCLOS (R Value = 0.836)

Apparent Accuracy Rate (large misclassifications):	0.970
Apparent Accuracy Rate (all misclassifications):	0.790

Classification Matrix Generated From Multiple Regression Analysis of Original PERCLOS Data Resulting in Algorithm J4a. (Independent variables employed included A/O Task, Steering, and Accelerometer.)

Predicted

	Group	% Correct	Awake	Questionable	Drowsy
New	Awake	93.33	**112**	6	2
Observed	Questionable	27.27	6	**3**	2
	Drowsy	38.46	4	4	**5**
	Total	83.33	122	13	9

PERCLOS (R Value = 0.599)

Apparent Accuracy Rate (large misclassifications):	0.958
Apparent Accuracy Rate (all misclassifications):	0.833

Algorithm J4a Applied to New Data and Compared with New Observed PERCLOS Data

Figure 25: Classification Matrices Showing, Accuracy of Algorithm J4a When Applied to Original Data (Upper) and New Data (Lower)

Table 18: APAR Summary Table Generated from Classification Matrices of Original and New A/O Data -- Including All Misclassifications and Large Misclassifications (APAR values correspond to R values in Table 17).

Independent Measures		Dependent Measures				
		AVEOBS	EYEMEAS	NEWDEF	PERCLOS	MASTER
I	A/O Task Measures Only	Algorithm I1a (new) Large 0.931 All 0.729	Algorithm I1a (new) Large 0.910 All 0.840	Algorithm I1a (new) Large 0.917 All 0.778	Algorithm I1a (new) Large 0.931 All 0.847	Algorithm I1a (new) Large 0.896 All 0.799
J	A/O Task, Steering, & Accelerometer	------------	------------	------------	(original) Large 0.970 All 0.790 Algorithm J4a (new) Large 0.958 All 0.833	------------
L	A/O Task, LANDEV/VAR, LNMNSQ, LANEX, & LNERRSQ	------------	------------	------------	Algorithm L3a (new) Large 0.979 All 0.868	------------
M	A/O Task, Steering, Accelerometer, LANDEV/VAR, LNMNSQ, LANEX, & LNERRSQ	------------	------------	------------	------------	(original) Large 1.000 All 0.850 Algorithm M3a (new) Large 0.944 All 0.799

NOTES: Letters in left column indicate appendices containing detailed analyses on original data set Wreggit, Kirn, and Wierwille, 1994).

Algorithm numbers located in each cell correspond to the multiple regression table within a given appendix (Wreggit, Kirn, and Wierwille, 1994).

See Appendix A for classification matrices and regression summaries corresponding to algorithms 14a, J4a, L3a, and M3a.

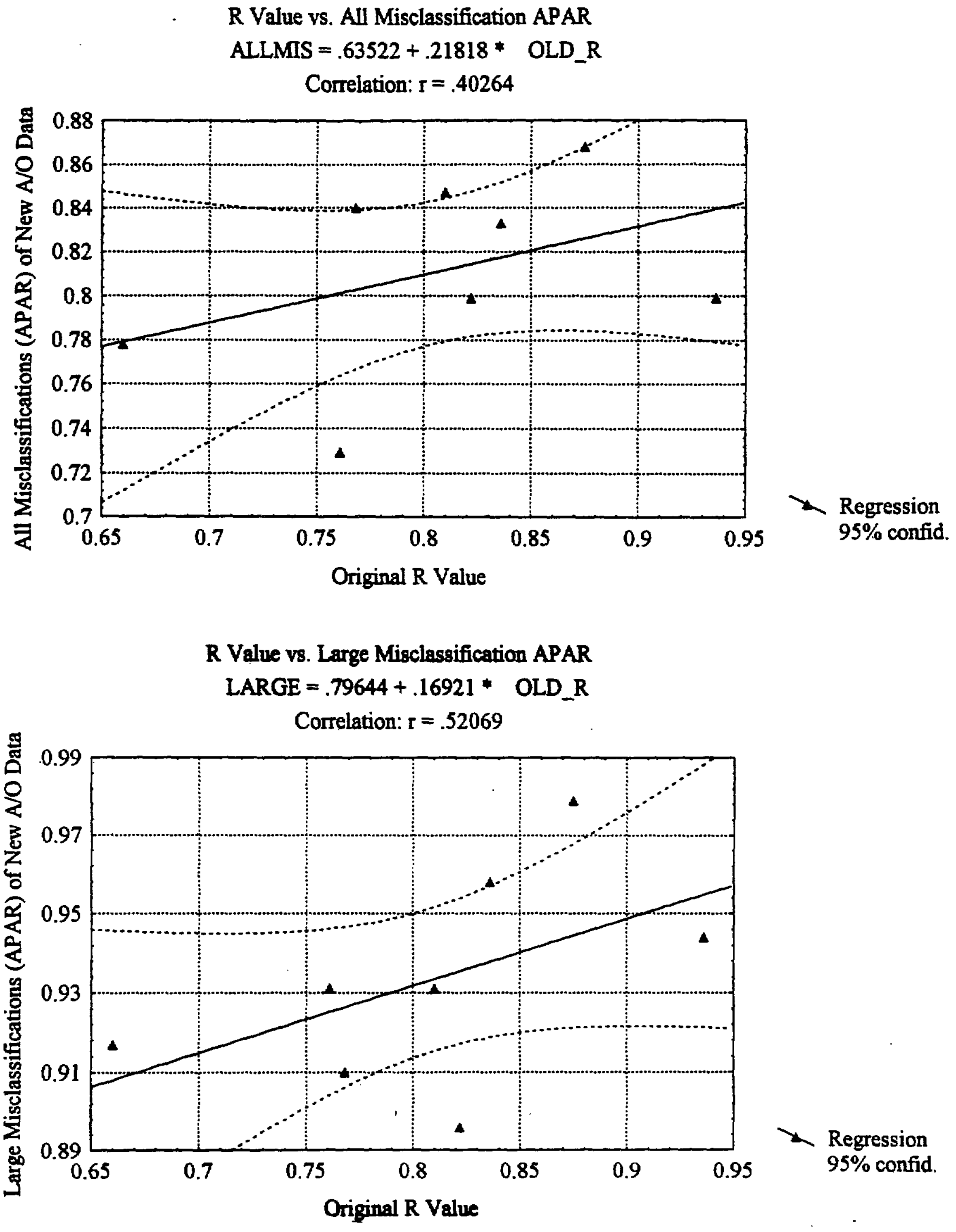


Figure 26: R Values vs. New APAR Values -- APAR Includes All Misclassifications (Upper Graph) and Large Misclassifications (Lower Graph).

DISCUSSION AND CONCLUSIONS OF VALIDATION PHASE

Validation of Algorithms Using Driver-Vehicle Performance Measures

When choosing the algorithms to validate, it was important that the component measures were very reliable and attainable in an on-the-road situation. The algorithms located in Appendices D and F (Wreggit, Kirn, and Wierwille, 1993) were chosen for the purpose of validation since they contained both reliable and probably attainable measures.

Another important aspect of the validated models is that their use allows for a step-up, step-down detection procedure. Some detection algorithms employed steering and lateral accelerometer measures and another set of detection algorithms employed steering, lateral accelerometer, and lane-related measures. Therefore, loss of a lane-related measure does not cause failure of the detection system. Rather, the system simply "steps-down" to a model that does not contain lane-related measures. This would be the case if one of the sensors necessary for the best algorithm did not provide a valid signal (i.e. lane sensors). A "step-up" procedure involves the use of newly validated signals (i.e. lane sensors pick up valid signal from the road).

The average R values achieved after application of drowsiness detection algorithms to new data were found to have no significant loss in drowsiness prediction compared with the original data upon which the algorithms were developed. Drowsiness classifications were accomplished with only small percentages of error. The algorithms that were validated using the driver-vehicle performance measures were found to be robust, in that no significant loss in drowsiness prediction was observed when the detection algorithms were applied to new data. This is an extremely important finding.

The Effect of Cruise Control and A/O Task on Detection Rate of Driver Performance Measure Algorithms. The detection algorithms from Appendices D and F (Wreggit, Kim, and Wierwille, 1994) were applied to the segments in time in which driver/subjects were under a cruise control condition combined with the A/O secondary task. It was found that when cruise control was engaged and the task was not being performed, the average drowsiness prediction R value was higher than when cruise control was engaged and the secondary task was being performed. The higher average R value for the Without Task/With Cruise condition is attributed to the fact that this condition is the most boring. The drivers did not have to monitor speed or interact with the push buttons mounted on the steering wheel while answering “yes” or “no” with the secondary task. It is hypothesized that the boredom experienced by the subjects tended to increase drowsiness. Subjects thus experienced a range of alertness that was greater than under the other conditions. In other words, subjects may have gone from alert to very drowsy within this condition. Therefore, the observed data were spread out, allowing the predicted data to track (fit the data) with higher relative success.

Validation of Algorithms Containing A/O Task Performance Measures

The detection algorithms which contained A/O task measures that were examined are shown in Table 17. The average R value for drowsiness detection, using data other than that which was used for the development of the algorithms, decreased significantly. This is likely a result of using only four subjects in the development of the prediction algorithms (based on A/O task performance data). The use of four subjects in the development stage may limit the predictive capabilities of the algorithms.

Another factor that may have contributed to the reduction of drowsiness prediction R values with new data involved an unrepresentatively small amount of drowsiness observed during the portions of runs in the new data in which the NO task was being performed. It was observed that the “awake” classifications were the great majority of the data. Since there was a relatively small domain of drowsiness (mostly “awake” and few “questionable” or “drowsy” observations) an unrepresentatively low R value may have occurred. Unfortunately, with the data that were collected during the validation phase, the algorithms were not exercised to an extent that would result in R values similar to the original R values. In other words, the new data were more tightly grouped.

Classification matrices were constructed using observed algorithm output and observed definitional measures of drowsiness for the new A/O task algorithms. The APAR results can be seen in Table 18. The classification matrices resulted in surprisingly high correct classification rates (APARs) given the relatively low R values of the prediction algorithms based on A/O task measures. The good results of the classification matrices suggest that an unrepresentative sample of drowsiness data was largely the cause for deflated R values instead of a limited predictive capability of the algorithms.

It was found that the average drowsiness-detection rate was greater for algorithms applied to new A/O data when cruise control was engaged as compared with new A/O data when cruise control was not engaged. One explanation for this finding is that drivers/subjects did not have to monitor their speed when cruise control was engaged. Therefore more resources could be allocated to the driving task and the A/O task. Since this may have been the case, alert drivers who were frequently monitoring the speed of the vehicle would have glanced at the speedometer often. With the greater amount of time available for subjects to glance at the speedometer the greater was the chance for the driving task and A/O task to degrade. In other words, when cruise control was engaged, the degradation in driving performance may have been purely due to the inattention or drowsiness of the driver.

Overview

With regard to the predictability of the definitional measures of drowsiness using the new data set, results demonstrate that MASTER and PERCLOS are the most predictable, followed by EYEMEAS, NEWDEF, and AVEOBS. This order of predictability is the same as with the original data except that AVEOBS and NEWDEF are reversed. However, a reason for this reversal may be that different drowsiness raters were used in the validation experiment. (AVEOBS is the average subjective rating of three raters).

The findings of this study are very encouraging, and the detection models look quite promising. It was estimated before the validation process that an average R value would be reduced by approximately 0.05 when the detection algorithms'were applied to the new data. Fortunately, the average R value loss was only 0.0069 across the validated algorithms.

We conclude on the basis of the validation procedures carried out that the detection algorithms based on steering and accelerometer measures, as well as on steering, accelerometer, and lane measures are quite robust and should be used in a future on-the-road study. Even though the algorithms were developed with a certain amount of "noise", such as interacting with instrument panel controls while driving, they do an excellent job of drowsiness prediction when applied to new data.

Chapter Six: Additional Analyses of the Algorithm Validation Data -- Simulator Study of the Effects of Cruise Control, Secondary Task, and Velocity-Related Measures on Driver Drowsiness and Drowsiness Detection

{This chapter represents an extended summary of work reported in the Sixth Semiannual Research Report, data October 15, 1994 and referred to as Kirn, Wreggit, and Wierwille, 1994)

INTRODUCTION

The validation experiment described in Chapter Five (Wreggit, Kim, and Wierwille, 1994) addresses the accuracy of previously developed drowsy-driver detection algorithms when applied to new driver-subjects. However, the validation experiment was a multipurpose experiment, designed not only for validation of algorithms, but also for the examination of several additional research issues. The present chapter describes these additional issues, the corresponding analyses that were performed, and the corresponding results that were obtained. It should be noted here that the data collected for use in the previously discussed algorithm-validation study (Chapter Five) are the same data used in this present study.

During the algorithm development portion of the research project (Chapter Four; Wreggit, Kim, and Wierwille, 1993), it was observed that drivers tended to vary their speed when they became drowsy. Velocity-related measures had not been gathered during that phase of the project, and therefore, such measures could not be included in the main algorithm development portion of the research. The question that arose, then, was whether or not velocity-related measures could contribute significantly to the accuracy of drowsy-driver detection algorithms. To answer the question, velocity-related measures were implemented in the validation experiment.

Three velocity-related measures were obtained during the validation phase, including: forward velocity standard deviation (FVELSD), forward acceleration standard deviation (FACCSD), and accelerator position standard deviation (PEDDEV). New algorithms were developed in this supplemental study using the validation experiment data in two ways: without velocity-related measures and with velocity-related measures. Thus, a direct comparison could be made that would allow assessment of any gains in accuracy obtainable using velocity-related measures.

Another question that remained regarding the A/O task was whether or not the task had an alerting effect on the driver. If so, the task would serve the dual purposes of a drowsiness detection aid and an alertness aid. Originally, the purpose of the A/O task was to provide an independent assessment of the level of drowsiness by having the driver respond to a task of low cognitive content. However, further investigation was carried out to determine if the A/O task does alert drowsy drivers.

Finally, the design of the validation experiment included segments in which cruise control was engaged. In this condition, the driver did not have to control speed, just as in an actual vehicle with cruise control engaged. Since each driver experienced both cruise control and non-cruise control conditions, direct comparisons of alertness could be made. Thus, the question of whether or not cruise control usage contributed to level of drowsiness could be answered.

In summary, there were three main questions to be answered by the additional analyses performed on the data from the validation experiment:

1. Do forward-velocity measures covary with level of drowsiness, and do they improve drowsiness-detection algorithm accuracy? If so, by how much do they improve accuracy?
2. Does the A/O task, which can be used as a drowsiness detection discriminator, have an alerting effect on the driver? and.
3. Does the use of cruise control increase the level of drowsiness in sleep-deprived drivers?

METHOD

Subjects

Subjects were the same as those described in Chapter Five (Wreggit, Kim, and Wierwille, 1994)

Apparatus.

The apparatus employed was the same as described in Chapter Five (Wreggit, Kim, and Wierwille, 1994)

Experimental Design

The additional analyses of the data collected during the algorithm validation phase employed a 2 X 2 X 6 complete factorial within-subject design. The first two factors and levels were as follows:

1. Speed control
   a. Speed controlled by driver
   b. Cruise control engaged; Speed automatically set at 60 m.p.h.
2. Subsidiary Task
   a. No subsidiary task
   b. Auditory subsidiary task requiring response

The third factor that was considered was time interval which had six levels. The experimental session was divided into four sections of 36 minutes each. Within each section, six six-minute averages of the various dependent measures were calculated to examine the effect of time on driving performance. In each section the subject underwent one of the four possible conditions:

1. No Cruise Control, No Secondary Task
2. No Cruise Control, Secondary Task
3. Cruise Control Engaged, No Secondary Task
4. Cruise Control Engaged, Secondary Task

Presentation order of the conditions for each subject was determined by selecting one line from three different 4 X 4 Latin squares. Four of the male subjects completed one Latin square, the other four males completed the second, and the four female subjects completed the third.

Several categories of measures were gathered for analysis in this experiment. The collected measures are described in the algorithm development portion (Chapter Four; Wreggit, Kirn, and Wierwille, 1993) of this paper and are the same those employed in the algorithm-validation study. The development and validation studies, however, did not employ velocity-related measures though these measures were collected during the algorithm-validation phase. The collected velocity-related measures are described below.

- FVELSD: The standard deviation of the forward velocity of the vehicle.
- FACCSD: The standard deviation of the forward acceleration of the vehicle
- PEDDEV: The standard deviation of the position of the accelerator pedal relative to the released position.

Procedure

Subject procedure. All driver-subjects underwent the same pre-driving procedures as the driver-subjects in the algorithm development phase and stayed at the Vehicle Analysis and Simulation Laboratory for approximately the same amount of time.

Experimental task. All subjects drove the simulated automobile as if it were an actual car. The driver-subjects were instructed to drive within the right lane at all times during the run. All subjects performed the same secondary (A/O) task that was employed during the algorithm development phase. In addition, a cruise control condition was incorporated into the driving task. When the cruise control was engaged the simulated automobile maintained 60 miles per hour. When the cruise control was not engaged the subject was asked to maintain approximately 60 miles per hour. Subjects drove for a total of 156 minutes.

As previously mentioned, the experimental task conditions changed every 36 minutes (after the first condition). Depending on the condition presented, the subject was asked either to:

1. Respond to the secondary task while monitoring and maintaining a speed of 60 m.p.h.
2. Respond to the secondary task while cruise control was engaged.
3. Simply monitor and maintain 60 m.p.h.
4. Simply stay in the right lane while cruise control was engaged.

Data Analysis Overview

The pre-analysis data reduction procedures were the same as those described in Chapter Five (see Figure 16).

After data reduction, several different analyses were run to answer the various research questions of the study. Unequal n's analyses were used to determine the relationship between speed variability and drowsiness. Each analysis involved a three part procedure. Each data point (six minute average) for each subject was classified as awake (A), questionable (Q), or drowsy (D) for each of the five drowsiness measures. This was accomplished using the same threshold criteria set by Wreggit, Kim, and Wierwille (1993).

Once the drowsiness measures had been classified, it was possible to classify the corresponding data point for each velocity-related measure. Thus within each drowsiness measure, there were three groups (A, Q. and D) of unequal number that could be compared in terms of variation of velocity-related measures. The tests that were used to make these comparisons were one-way parametric ANOVAs and one-way Kruskal-Wallis nonparametric ANOVAs. The nonparametric tests were used when the assumptions for the parametric tests were not met (usually lack of homogeneity of variance).

Next, Pearson product-moment correlation coefficients (r) were calculated to determine whether or not velocity-related measures were reliable indicators of drowsiness. First the non-cruise data were divided into task, no task, and all non-cruise data. Pearson r values were then found between the variation of each velocity-related measure and the magnitude of each drowsiness measure for the three different groups. The r values were then compared and tested for significance between groups.

Thirdly, 2 X 2 X 6 analyses of variance were conducted to examine the effects and interactions of cruise control, secondary task, and time interval on drowsiness and lane keeping. To examine these effects on drowsiness, the five definitional measures AVEOBS, EYEMEAS, NEWDEF, PERCLOS, and, MASTER were used as dependent measures. Similarly, to examine the effects of cruise and A/O task on lane keeping, the previously described lane related measures LANDEV, LNMNSQ, LANEX, and LNERRSQ were used as dependent measures.

The final set of analyses was used to examine whether or not velocity-related measures would improve drowsiness detection algorithms. Two cases of multiple regression analyses were used for this purpose. In case 1, multiple regression was used to develop algorithms without the inclusion of velocity-related measures. In case 2, velocity-related measures were included with the other measures to develop detection algorithms. Multiple correlation coefficients and apparent accuracy rates of the different cases were compared for each velocity measure and each drowsiness measure.

RESULTS

Unequal ’ Analyses

Both parametric and Kruskal-Wallis nonparametric ANOVAs were run. Before conducting either test, the data for the groups “Awake”, “Questionable”, and “Drowsy” were examined for normality and homogeneity of variance using both Levene’s test and the Hartley F-Max test. If results from either of these tests exhibited heterogeneity of variance, a plot of means and standard deviations was examined for high correlation. Based on the findings of these analyses, a parametric test, nonparametric test, or both (if the tests of assumptions were inconclusive) were run to differentiate between the groups. The summarized results can be seen in Kim, Wreggit, and Wierwille (1994).

Regardless of the type of test used (parametric or nonparametric), the results followed a distinct pattern. For the data set in which subjects completed the secondary A/O task, there were only two significant differences (a = 0.05) between groups out of a possible 15 (five drowsiness measures across the three velocity-related measures). More specifically, there were two times that forward acceleration (FACCSD) differed between the “Awake”, “Questionable”, and “Drowsy” groups. No significant differences between the groups were found in either forward velocity (FVELSD) or pedal deviations (PEDDEV) under the task condition.

In contrast, under the no task condition (which refers to the absence of the A/O task), significance (a = 0.05) between groups was seen in 14 out of the 15 possible cases. For PEDDEV and FVELSD, there was a significant difference between at least two groups in all cases. As for FACCSD, there was a significant difference in forward acceleration between at least two groups for four out of five cases.

When all non-cruise data were examined as a whole data set, there was also a large domain of significance found between groups. Once again, at least two groups differed significantly in every case with regard to FVELSD and PEDDEV. With FACCSD, parametric and nonparametric tests tended to give different results. Nonparametric tests exhibited significance in four cases. However, parametric tests failed to show significance between any groups at $a = 0.05$.

Correlation Analyses

Pearson product-moment correlation (r) values were found between each velocity-related measure and each drowsiness measure. Similar to the previous analyses, these correlations were computed for task condition data, no task condition data, and combined data separately and compared. Table 19 is a summary of the results.

Using Fisher's Z transformations, it was possible to test for significant differences between the groups. As one can also see from Table 19, the correlations in the no-task condition were significantly higher than the task grouping, combined grouping, or both in 13 out of 15 cases. In all cases within the no task group, correlations between drowsiness measure and velocity-related measure were moderately high, whereas in the task group correlations between drowsiness measure and velocity-related measure were weak in all cases but one. The correlations found with the combined data were mostly weak.

To further compare the groups, regression lines plotted for each group's d&a can be seen in Figure 27. The figure shows results obtained for PERCLOS versus FVELSD and is typical of the other plots.

Analyses of Variance

Table 20 contains a summary of the results of the 2 X 2 X 6 ANOVAs that were run to test the main effects and interactions of cruise control, secondary task, and time interval on drowsiness and lane keeping. As mentioned previously, there were two levels of cruise control (engaged and disengaged), there were two levels of secondary task (present and absent) and there were six levels of time interval (six six-minute intervals within each condition).

Five separate dependent measures were used as measures of drowsiness: AVEOBS, EYEMEAS, NEWDEF, PERCLOS, and MASTER. There were no main effects for either cruise or task for any drowsiness measure. In addition, no two- or three-way interactions

Table 19: Summary Table of Correlation Analyses

Pearson product-moment correlation coefficients (r) for drowsiness measure and longitudinal speed variation (FVELSD) with associated test of significance as a function of A/O task, no task, or combined data conditions.

Drowsiness Measure	A/O Task	No Task	Combined Data
AVEOBS	0.345§	0.6174	0.441
EYEMEAS	0.198§	0.606”	0.347§
NEWDEF	0.129§	0.617*	0.310§
PERCLOS	0.147§	0.638*	0.327§
MASTER	0.223 §	0.666*	0.387§

Pearson product-moment correlation coefficients (r) for drowsiness measure and longitudinal acceleration standard deviation (FACCSD) with associated test of significance as a function of A/O task, no task, or combined data conditions.

Drowsiness Measure	A/O Task	No Task	Combined Data
AVEOBS	0.145	0.445	0.203 .
EYEMEAS	0.040§	0.485*	0.142§
NEWDEF	0.053§	0.529*	0.161§
PERCLOS	0.086§	0.506*	0.177§
MASTER	0.088§	0.527*	0.185§

Pearson product-moment correlation coefficients (r) for drowsiness measure and accelerator pedal deviation (PEDDEV) with associated test of significance as a function of A/O task, no task, or combined data conditions.

Drowsiness Measure	A/O Task	No Task	Combined Data
AVEOBS	0.406	0.513	0.443
EYEMEAS	0.218§	0.578√	0.380
NEWDEF	0.131§	0.562√	0.331
PERCLOS	0.143§	0.589√	0.356
MASTER	0.241§	0.601√	0.410

* r value differs significantly from all other r values for a given drowsiness measure (row)

§ r value differs significantly from r value under “No Task” condition for a given drowsiness measure (row)

√ r value differs significantly from r value under “A/O Task” condition for a given drowsiness measure (row)

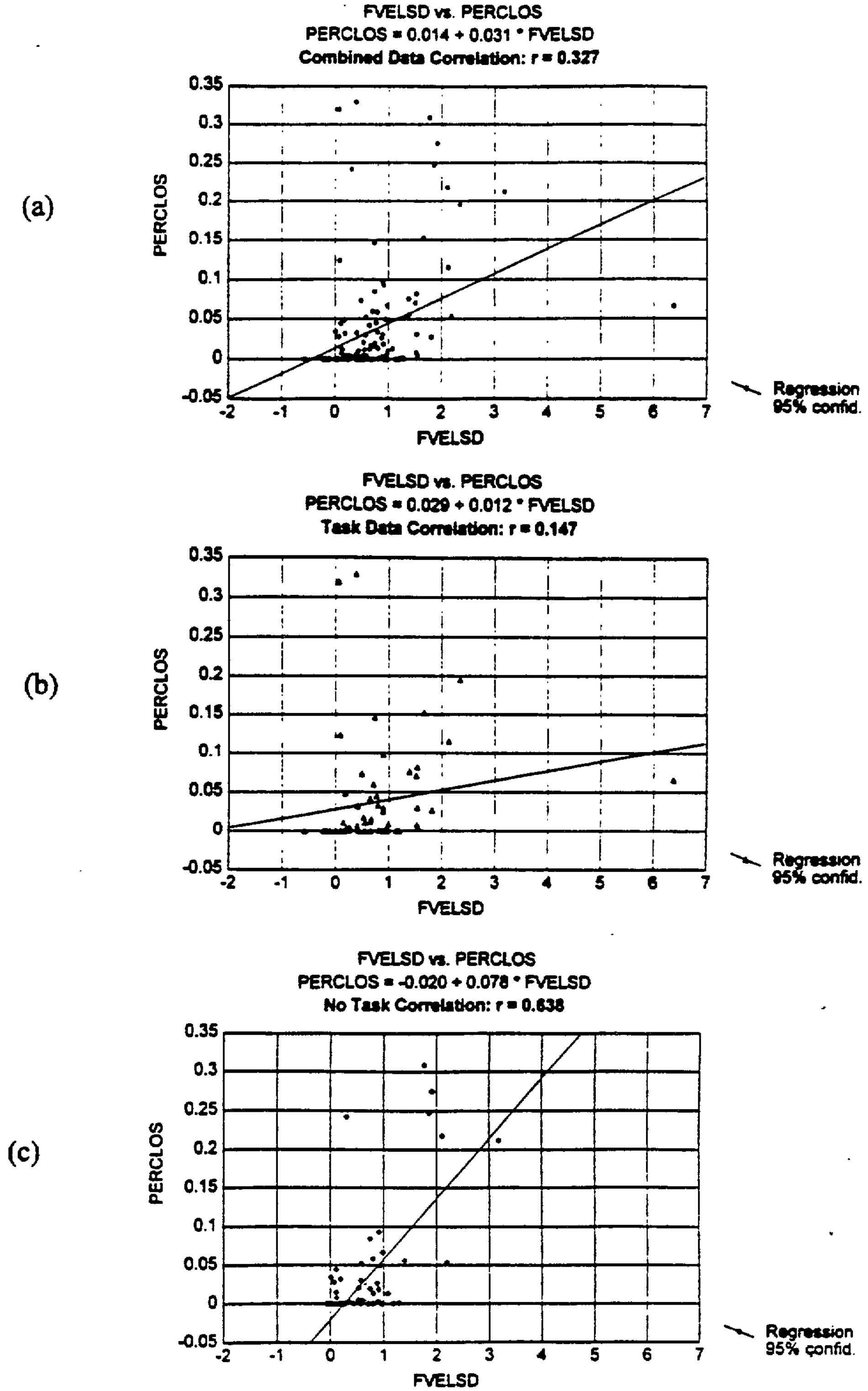


Figure 27: Comparison of Regression Lines for (a) Combined, (b) A/O Task, and (c) No Task Data

Table 20: Summary Table of ANOVA Results

summary table of p-values of main effects for
2 (A/O Task) x 2 (Cruise Control) x 6 (Interval) ANOVAs

Drowsiness Measure	**A/O Task**	**Cruise Control**	**Interval**
AVEOBS	0.693	0.520	1.14 E-13*
EYEMEAS	0.847	0.098	6.99 E-7*
NEWDEF	0.874	0.114	4.79 E-6*
PERCLOS	0.182	0.084	0.049*
MASTER	0.325	0.103	0.026*

Performance Measure	**A/O Task**	**Cruise Control**	Interval
LANDEV	0.270	0.327	0.005*
LANEX	0.507	0.220	3.82 E-5*
LNERRSQ	0.192	0.376	0.204
LNMNSQ	0.197	0.385	0.310

Note: No interaction was significant (a = 0.05)

* significant (a = 0.05)

were found to be significant. However, a significant main effect (*a* = 0.05) of time interval was found in every case. A significant increase in drowsiness-measure values between levels of time interval was seen.

Dependent measures for lane keeping included LANDEV, LNMNSQ, LANEX, and LNERRSQ. Results were similar to measures of drowsiness in that there were no main effects found for either cruise control or secondary task conditions. Likewise, no significant two- or three-way interactions were found. However, there was a significant main effect increase (a = 0.05) of both LANDEV and LANEX lateral performance measures over time interval.

Multiple Regression

Multiple regression analyses were run to examine whether or not velocity-related measures could contribute to drowsiness detection algorithms. Tables 21, 22, and 23 are summaries of the results for the individual velocity-related measures (FVELSD, FACCSD, and PEDDEV). Table 24 is a summary of the results when all velocity-related measures were included together in the algorithm development. As mentioned previously, algorithms were developed in two different cases in order to evaluate the predictive strength of the velocity-related measures. The following is a brief description of the procedures that were used for generating the results found in Tables 21, 22, 23, and 24.

On the left half of each table, algorithms for both Case 1 and Case 2 were developed using the accelerometer and steering measures that were defined previously. In Case 1, these measures alone were used for algorithm development. Using each drowsiness measure as a separate dependent variable, backwards stepwise regression and a re-substitution process were conducted. The algorithm was developed when all remaining independent measures were significant.

In Case 2, velocity-related measures were added to the accelerometer and steering measures for algorithm development. Once again backwards stepwise regression and re-

Table 21: Summary Table of Multiple Regression Analyses Results Showing R Values for Forward Velocity Standard Deviation (FVELSD)

	Steering and Accel. Measures		Steering, Accel., and Lane Measures	
	Case 1	Case 2	Case 1	Case 2
AVEOBS	0.799	0.799*	0.845	0.845*
EYEMEAS	0.862	0.862*	0.892	0.892*
NEWDEF	0.834	0.834*	0.851	0.866. FVELSD
PERCLOS	0.871	0.876 FVELSD	0.911	0.923 **FVELSD**
MASTER	0.897	0.897*	0.931	0.931*

* Algorithm same as Case 1 (FVELSD provided no improvement and was deleted)

FVELSD -- boldface indicates that FVELSD contributed to significant increase in algorithm R value

Table 22: Summary Table of Multiple Regression Analyses Results Showing R Values for Forward Acceleration Standard Deviation (FACCSD)

	Steering and Accel. Measures		Steering, Accel., and Lane Measures	
	Case 1	Case 2	Case 1	Case 2
AVEOBS	0.799	0.799*	0.845	0.845*
EYEMEAS	0.862	0.862*	0.892	0.892*
NEWDEF	0.834	0.834*	0.851	0.868 FACCSD
PERCLOS	0.871	0.871*	0.911	0.927 FACCSD
MASTER	0.897	0.897*	0.93 1	0.933 FACCSD

* Algorithm same as Case 1(FACCSD provided no improvement and was deleted)

FACCSD -- boldface indicates that FACCSD contributed to significant increase in algorithm R value

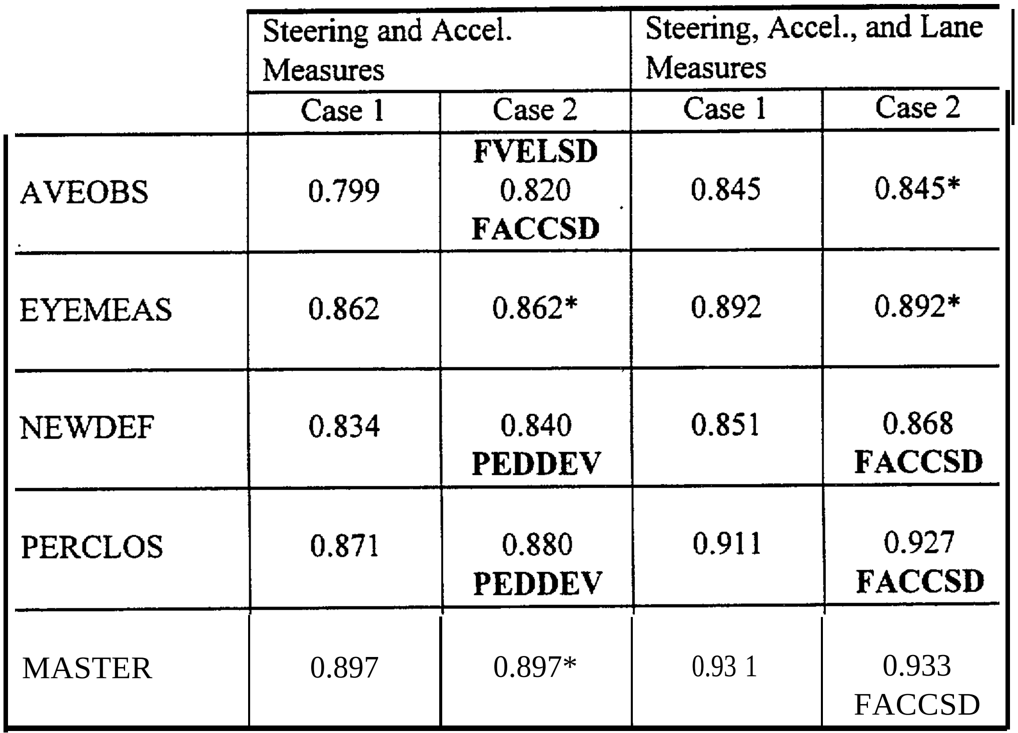
Table 23: Summary Table of Multiple Regression Analyses Results Showing R Values for Accelerator Pedal Movement Standard Deviation (PEDDEV)

	Steering and Accel. Measures		Steering, Accel., and Lane Measures	
	Case 1	Case 2	Case 1	Case 2
AVEOBS	0.799	0.799’	0.845	0.844*
EYEMEAS	0.862	0.862*	0.892	0.892*
NEWDEF	0.834	0.840 PEDDEV	0.851	0.859 PEDDEV
PERCLOS	0.871	0.880 PEDDEV	0.911	0.915 PEDDEV
MASTER	0.897	0.897*	0.931	0.931*

* Algorithm same as Case 1 (PEDDEV provided no improvement and was deleted)

PEDDEV -- boldface indicates that PEDDEV contributed to significant increase in algorithm R value

Table 24: Summary Table of Multiple Regression Analyses Results Showing R Values for All Velocity-Related Measures (FVELSD, FACCSD, and PEDDEV)



	Steering and Accel. Measures		Steering, Accel., and Lane Measures	
	Case 1	Case 2	Case 1	Case 2
AVEOBS	0.799	**FVELSD** 0.820 **FACCSD**	0.845	0.845*
EYEMEAS	0.862	0.862*	0.892	0.892*
NEWDEF	0.834	0.840 **PEDDEV**	0.851	0.868 **FACCSD**
PERCLOS	0.871	0.880 **PEDDEV**	0.911	0.927 **FACCSD**
MASTER	0.897	0.897*	0.93 1	0.933 FACCSD

* Algorithm same as Case 1 (None of the longitudinal measures improved algorithm accuracy- All were deleted)

FVELSD -- boldface indicates that FVELSD contributed to significant increase in algorithm R value
FACCSD -- boldface indicates that FACCSD contributed to significant increase in algorithm R value
PEDDEV -- boldface indicates that PEDDEV contributed to significant increase in algorithm R value

Multiple longitudinal measures listed in the table indicate that the combination of measures contributed significantly to algorithm R value substitution were used. Those cases in which the velocity-related measures remained significant and contributed to the algorithm are listed in bold print in the tables.

For the right half of each table, lane-related measures were added to the independent measures for both cases, The definitions for these measures can also be found in the Experimental Design section. The procedures for developing the algorithms for Case 1 and Case 2 are identical to those used for the left half of the chart (velocity-related measures added in Case 2 only).

Velocity-related measures, added individually, contributed to 10 out of 30 drowsiness detection algorithms. The amounts contributed ranged from 0.002 to 0.017. When all three velocity-related measures were used in the regression, one out of ten algorithms was improved by 0.02 1.

To better understand the additional predictive strength of the velocity-related measures, classification matrices of the Case 1 and Case 2 algorithms were constructed. These can be seen with the associated algorithms for PERCLOS in Figures 28, 29, and 30. See Kirn, Wreggit, and Wierwille for a complete set of APARs and algorithm results. These matrices represent data that have been classified as "Awake, "Questionable", or "Drowsy" as was done previously for the unequal n's analyses. The bolded numbers in the classification matrices have been classified correctly. The cells with bolded borders contain the number of large misclassifications.

Steering, Accelerometer, and Selected Lane Measures - Case 1 (Algorithm developed <u>without</u> FVELSD)

Regression Summary **foi** Dependent Variable: **PERCLOS**

R = 0.91122539 R2 = 0.83033171 Adjusted R2 = 0.82418431 F(5. 138) = 135.07 p < 0.0000 Std. error of estimate: 0.02996

	Beta	St. Err. of Beta	B	St. Err of B	t(138)	p-level
Intercept			-0.00455	0.00330	-1.380	0.1698
INTACDEV	-0.2149	0.0373	-0.09539	0.01655	-5.763	0.0000
LANDEV	0.5768	0.0812	0.03680	0.00518	7.104	0.0000
LANEX	0.2062	0.0903	0.13398	0.05869	2.283	0.0240
LGREV	0.3921	0.0942	0.01547	0.00372	4.161	0.0001
STEXED	-0.3238	0.0620	-29.59937	5.66390	-5.226	0.0000

Predicted

	Group	% Correct	Awake	Questionable	Drowsy
	Awake	95.20	**119**	5	1
Observed	Questionable	37.50	4	**3**	1
	Drowsy	81.82	1	1	**9**
	Total	90.97	124	9	11

PERCLOS (R Value = 0.911)

Apparent Accuracy Rate (large misclassifications): 0.986

ApparentAccuracyRate(allmisclassifications): 0.910

Steering, Accelerometer, and Selected Lane Measures - Case 2 (Algorithm developed with FVELSD)

Regression Summary for Dependent Variable: **PERCLOS**

R = 0.92265664 R2 = 0.85129528 Adjusted R2 = 0.84364136 F(7.136) = 111.22 p < 0.0000 Std. error of estimate: 0.02826

	Beta	St. Err. of Beta	B	St. Err of B	t(138)	p-level
Intercept			0.01113	0.00351	3.174	0.0019
INTACDEV	-0.2084	0.0356	-0.09250	0.01580	-5.856	0.0000
FVELSD	-0.2363	0.0411	-0.02260	0.00393	-5.757	0.0000
LANVAR	1.1831	0.2118	0.00878	0.00157	5.586	0.0000
LANEX	0.2922	0.0813	0.18989	0.05281	3.596	0.0005
LNERRSQ	-0 7567	0.2233	-0.00602	0.00178	-3.389	0.0009
LGREV	0.4658	0.0928	0.01838	0.00366	5.017	0.0000
STEXED	-0.1961	0.0911	-17.92416	8.33013	-2.152	0.0332

Predicted

	Group	% Correct	Awake	Questionable	Drowsy
	Awake	95.20	**119**	6	0
Observed	Questionable	37.50	4	**3**	1
	Drowsy	81.82	1	1	**9**
	Total	90.97	124	10	10

PERCLOS (R Value = 0.923)

Apparent Accuracy Rate (large misclassifications): 0.993

ApparentAccuracyRate(allmisclassifications): 0.910 .

Figure 28: Case 1 and Case 2 Algorithm Comparison -- Multiple Regression Results (Independent Variables Included Steering, Accelerometer, and Lane Measures. FVELSD Included in Case 2 Only.)

Steering, Accelerometer, and Selected Lane Measures - **Case 1** (Algorithm developed without FACCSD)

Regression Summary for Dependent Variable: PERCLOS

R = 0.91122539 R^2 = 0.83033171 Adjusted R^2 = 0.82418431 F(5, 138) = 135.07 p < 0.0000 Std. error of estimate: 0.02996

	Beta	St. Err. of Beta	B	St. Err of B	t(138)	p-level
Intercept			-0.00455	0.00330	-1.380	0.1698
INTACDEV	-0.2149	0.0373	-0.09539	0.01655	-5.763	0.0000
LANDEV	0.5768	0.0812	0.03680	0.00518	7.104	0.0000
LANEX	0.2062	0.0903	0.13398	0.05869	2.283	0.0240
LGREV	0.3921	0.0942	0.01547	0.00372	4.161	0.0001
STEXED	-0.3238	0.0620	-29.59937	5.66390	-5.226	0.0000

Predicted

	Group	% Correct	Awake	Questionable	Drowsy
Observed	Awake	95.20	**119**	5	1
	Questionable	37.50	4	**3**	1
	Drowsy	81.82	1	1	**9**
	Total	90.97	124	9	11

PERCLOS (R Value = 0.911)

Apparent Accuracy Rate (large misclassifications): 0.986

Apparent Accuracy Rate (all misclassifications): 0.910

Steering, Accelerometer, and Selected Lane Measures - Case 2 (Algorithm developed **with FACCSD)**

Regression Summary for Dependent Variable: PERCLOS

R = 0.92738791 R^2 = 0.86004834 Adjusted R^2 = 0.85284495 F(7. 136) = 119.39 p < 0.0000 Std. error of estimate: 0.02741

	Beta	St. Err. of Beta	B	St. Err of B	t(138)	p-level
Intercept			0.00680	0.00307	2.215	0.0284
INTACDEV	-0.1932	0.0347	-0.08575	0.01540	-5.569	0.0000
FACCSD	-0.2540	0.0384	-0.14261	0.02157	-6.612	0.0000
LANVAR	1.2794	0.2078	0.00950	0.00154	6.157	0.0000
LANEX	0.2860	0.0789	0.18588	0.05125	3.627	0.0004
LNERRSQ	-0.7088	0.2152	-0.00564	0.00171	-3.293	0.0013
LGREV	0.3213	0.0904	0.01268	0.00357	3.555	0.0005
STEXED	-0.2219	0.0881	-20.28603	8.04976	-2.520	0.0129

Predicted

	Group	% Correct	Awake	Questionable	Drowsy
Observed	Awake	96.00	**120**	5	0
	Questionable	37.50	4	**3**	1
	Drowsy	81.82	1	1	**9**
	Total	91.67	125	9	10

PERCLOS (R Value = 0.927)

Apparent Accuracy Rate (large misclassifications): 0.993

ApparentAccuracyRate(allmisclassifications): 0.917

Figure 29: Case 1 and Case 2 Algorithm Comparison -- Multiple Regression Results (Independent Variables Included Steering, Accelerometer, and Lane Measures. FACCSD Included in Case 2 Only.)

Steering, Accelerometer, and Selected Lane Measures - Case 1 (Algorithm developed without PEDDEV)

Regression Summary for Dependent Variable: **PERCLOS**

R= 0.91122539 R2 = 0.83033171 Adjusted R2 = 0.82418431 F(5. 138) = 135.07 p < 0.0000 Std. error ofestimate: 0.02996

	Beta	St. Err. of Beta	B	St. Err of B	t(138)	p-level
Intercept			-0.00455	0.00330	-1.380	0.1698
INTACDEV	-0.2149	0.0373	-0.09539	0.01655	-5.763	0.0000
LANDEV	0.5768	0.0812	0.03680	0.00518	7.104	0.0000
LANEX	0.2062	0.0903	0.13398	0.05869	2.283	0.0240
LGREV	0.3921	0.0942	0.01547	0.00372	4.161	0.0001
STEXED	-0.3238	0.0620	-29.59937	5.66390	-5.226	0.0000

Predicted

	Group	% Correct	Awake	Questionable	Drowsy
	Awake	95.20	**119**	5	1
Observed	Questionable	37.50	4	**3**	1
	Drowsy	81.82	1	1	**9**
	Total	90.97	124	9	11

PERCLOS (R Value = 0.911)

Apparent Accuracy Rate (large misclassifications): 0.986

ApparentAccuracyRate(allmisclassifications): 0.910

Steering, Accelerometer, and Selected Lane Measures - **Case** 2 (Algorithm developed with PEDDEV)

Regression Summary for Dependent Variable: **PERCLOS**

R = 0.92738791 R2 = 0.86004834 Adjusted R2 = 0.85284495 F(7. 136) = 119.39 p < 0.0000 Std. error of estimate: 0.02741

	Beta	St. Err. of Beta	B	St. Err of B	t(138)	p-level
lnterceot			0.00309	0.00442	0.699	0.4854
INTACDEV	-0.2195	0.0366	-0.09742	0.01625	-5.994	0.0000
LANDEV	0.5296	0.0818	0.03379	0.00522	6.477	0.0000
LANEX	0.2331	0.0892	0.15151	0.05798	2.613	0.0100
PEDDEV	-0.1215	0.0478	.-0.00527	0.00207	-2.542	0.0121
LGREV	0.4474	0.0950	0.01765	0.00375	4.711	0.0000
STEXED	-0.2744	0.0638	-25.08181	5.83238	-4.300	0.0000

Predicted

	Group	% Correct	Awake	Questionable	Drowsy
	Awake	95.20	**119**	5	1
Observed	Questionable	37.50	4	**3**	1
	Drowsy	81.82	1	1	**9**
	Total	90.97	124	9	11

PERCLOS (R Value = 0.915)

Apparent Accuracy Rate (large misclassifications): 0.986

Apparent Accuracy Rate (all misclassifications): 0.910

Figure 30: Case 1 and Case 2 Algorithm Comparison -- Multiple Regression Results (Independent Variables Included Steering, Accelerometer, and Lane Measures. PEDDEV Included in Case 2 Only.)

DISCUSSION

In general, velocity-related measures were found to be moderately good indicators and predictors of drowsiness under certain conditions. Under the “no task” condition, unequal n’s analyses and correlation analyses showed promising results. In addition, all three velocity-related measures contributed slightly to the predictive power of multiple regression algorithms. Analysis of variance results showed no main effect for either cruise control or secondary task on drowsiness or lane-keeping. However, the ANOVAs did show a main . effect for time interval.

Speed Variability vs. Drowsiness

Unequal n’s analyses suggested that speed variability increased with drowsiness when no secondary task was present. In every case, speed variability was greater among data classified as “drowsy” than data classified as “awake” for the no task condition. The results of further analyses showed that the same holds true for forward acceleration variability and accelerator pedal movement variability.

Similarly, when the task condition and the no task condition are combined to form all non-cruise data, speed variability, forward acceleration, and accelerator pedal movement are significantly different between the “Drowsy” and “Awake” data sets in most cases. However, when only the task condition data is examined, there are no significant differences between “Drowsy” and “Awake” data for either speed variability or pedal movement.

In summary, when driver-subjects are not given a secondary task, the variability for all velocity-related measures increase with drowsiness. When subjects are given a task, very little variability is seen in speed, acceleration, or pedal movement. The results suggest that the secondary task may have kept subjects more stimulated and thus helped them monitor and control their speed. In the conditions in which the secondary task is absent, it was possible that subjects become under loaded and lost their ability to concentrate on speed maintenance.

Velocity-Related Measures as Indicators of Drowsiness

A strong positive correlation between drowsiness and velocity-related measures was not established for the entire sample. However, the data from the no-task condition showed moderately high positive correlations with means of 0.629 for speed variation, 0.498 for acceleration, and 0.569 for pedal movement, Correlations for the task condition were weak.

As with the previous analyses, velocity-related measures proved to be more promising in the no-task condition. An increase in drowsiness was moderately associated with an increase in speed variation only when a secondary task was not presented.

From these data one can conclude that, in general, velocity-related measures are fairly weak indicators of drowsiness. However, when subjects are not responding to a task, velocity-related measures are moderately strong indicators of drowsiness,

Cruise Control, Secondary Task, and Time Interval vs. Drowsiness and Lane-Keeping

Five 2 X 2 X 6 ANOVAs were run to examine the effects and interactions of cruise control, secondary A/O task, and time interval on drowsiness. One ANOVA was run for each drowsiness measure. Time interval was found to have the only significant main effect on drowsiness. No two- or three-way interactions were found to be significant.

These findings were somewhat surprising since the secondary A/O task seemed to have a large effect on speed maintenance ability. However, the previous analyses utilized only half of the gathered data (non-cruise control). Analysis of all the data showed that for most subjects, drowsiness level varied greatly. Although the mean drowsiness level was higher in the no-task condition than in the task condition, there was a great deal of variance. Therefore, drowsiness level in the no-task condition was not significantly higher than the drowsiness level in the task condition. This same trend occurred with the cruise control factor. Although the mean value for most drowsiness measures was higher in the cruise control conditions, no significant effect (a = 0.05) was found due to large variance of drowsiness in the cruise control condition.

In terms of lane-keeping, the results offer similar suggestions. From the ANOVAs, it appears that time interval was the only factor significantly affecting drivers' drowsiness. Drivers' abilities to stay within lane boundaries did not seem to be significantly affected by the presence or absence of secondary task or cruise control.

Detection Models Including Velocity-Related Measures vs. Detection Models Not Including Velocity-Related Measures

The potential gains from velocity-related measures are quite modest. For the algorithms that were improved by one or more velocity measures, the average gain in correlation was only 0.010. Examination of classification matrices revealed very small improvement in accuracy for the algorithms to which they contributed.

In most cases, velocity-related measures did not contribute to detection accuracy. However, the installation of these measures is both unobtrusive and not overly complex. Ultimately, it is a tradeoff of costs and benefits as to whether very small improvements in detection accuracy justify the added cost.

CONCLUSIONS

This experiment helped to shed light upon questions related to velocity-related performance measures, cruise control, and auditory secondary tasks as they apply to drowsiness and driving. With regard to velocity-related measures, drivers seemed to vary their speed when they became quite drowsy. This can be seen best when drivers are not performing a secondary task and are possibly under loaded. However, velocity-related measures were not found to be good indicators of drowsiness. It is likely that speed variance while drivers are alert was too similar to variance of speed while drivers were drowsy to be a reliable indicator of drowsiness. However, velocity-related measures became much better indicators when drivers were exposed to the no-task condition in comparison to task conditions.

This research suggests that improvements in detection algorithms from the addition of velocity-related measures will be modest at best . This is not to say that velocity-related measures cannot be very strong predictors at times, but overall these performance measures add only small amounts of predictive information.

With regard to the secondary task. the findings from this study suggest that future research is required. Although driver drowsiness does not appear to be affected by the presence of a secondary task, drivers' ability to maintain speed is improved by the presence of a secondary task. The lack of a secondary task or other stimulation does not induce drowsiness, but it may help induce inattention. Perhaps the presence of a nonstressful secondary task would help keep driver attention from waning.

Chapter Seven

Part One: Further Algorithm Refinement and Investigation – Effects of Using Higher Order Algorithms on Drowsy Driver Detection Accuracy

(Work reported in this part of Chapter Seven has not appeared in previous semiannual research reports. This work was carried out by Rollin J. Fairbanks and Walter W. Wierwille. It is referred to as Fairbanks and Wierwille, 1994)

INTRODUCTION

This study focused on the effects of using higher-order (non-linear) algorithms on drowsy-driver detection accuracy. Measures from the development phase (Wreggit, Kirn, and Wierwille, 1993) and validation phase (Wreggit, Kirn, and Wierwille, 1994) were squared or multiplied with each other to obtain cross products. The second order terms, combined with first order terms, were used to calculate predictive algorithms using data from the development phase. The developed algorithms were then applied to the validation data.

The main purpose of this follow-up study was to examine the potential for improvement in algorithm accuracy with the addition of second order terms to the drowsiness-detection algorithms. Three groups of independent measures were selected from the development phase (Chapter 4) data and used to estimate the dependent measure "PERCLOS." Multiple regression analyses were performed using linear (first order) terms only, linear and cross product (first order and partial second order) terms, and all first and second order (full second order) terms from each of the three groups of independent measures. The nine algorithms developed from this process were applied to the data collected during validation phase (Chapter 5) of the main study. The accuracy of each type of algorithm (linear, cross product, or full second order) was determined by examining multiple regression Pearson-product-moment correlation (R) values as well as by classification matrices.

Although not conclusively proven by the present study, the results do support the hypothesis that higher-order algorithms produce more and larger outliers when applied to new data than do linear algorithms. The experimenters involved with this study were interested in quantifying the effects of the prediction outliers on classification accuracy, therefore prediction outliers were limited to the maximum and minimum scores of the observed data. Subsequently, a comparison of classification accuracy was conducted between data with outliers present and data with no outliers.

METHOD

Data collected during the main study were selected to be re-analyzed. The selected data included independent and dependent measures from all subjects in both phases of the main study.

Three groups of independent measures were selected from those used in phase II of the driver drowsiness main study. There were several categories of measures used in phase II which included seat movement, steering, accelerometer, lane, heading, and subsidiary (A/O) task-related measures, as well as brain wave activity and heart rate measures. These measures are described in detail in the algorithm development report (Wreggit, Kirn, and Wierwille, 1993). Three groups of independent measures employed in this study included variables most often appearing in the previously developed algorithms. The names of the selected variables used for the present study and their descriptions are as follow (Wreggit, Kim, and Wierwille, 1993):

Steering-related measures:

- NMRHOLD: The number of times the hold circuit output on the steering wheel exceeded a threshold value (corresponding to holding the steering wheel still for 0.4 second or longer).
- THRSHLD: The proportion of total time that the hold circuit on the steering wheel exceeded a threshold value.
- STVELV: The variance of steering velocity.
- LGREV: The number of times that steering excursions exceeded 15 degrees after steering velocity passed through zero.
- MDREV: The number of times that steering excursions exceeded 5 degrees (but less than 15 degrees) after steering velocity passed through zero.
- SMREV: The number of times that steering excursions exceeded 1 degree (but less than 5 degrees) after steering velocity passed through zero.

- STEXED: The proportion of time that steering velocity exceeded 150 degrees per second.

Accelerometer-related measure:

- INTACDEV: The standard deviation of the lateral velocity of the vehicle. (This signal was obtained by passing the smoothed accelerometer signal through an additional low pass filter (leaking integrator) with a comer frequency of 0.004 Hz.)

Lane-related measures: ,

- LANDEV: The standard deviation of lateral position relative to the lane.
- LNRTDEV: The standard deviation of the time derivative of lane position.
- LANEX: The proportion of time that any part of the vehicle exceeded a lane boundary.
- LNERRSQ: The mean square of the difference between the outside edge of the vehicle and the lane edge when the vehicle exceeded the lane. When the vehicle did not exceed the lane, the contribution to the measure was zero.

The definitional measure of drowsiness, PERCLOS, was selected to be used as the dependent measure. PERCLOS is defined as the proportion of time that the eyes of a driver/subject are closed 80% or more. This measure was collected during both phases of the main study. Although five definitional measures were used as dependent measures in algorithm development and validation phases, PERCLOS was chosen as the dependent. measure for this study for the following reasons:

- It is desirable to use only one definitional measure to ensure control across conditions, thus allowing reliable comparisons between the various non linear and linear algorithms.

- PERCLOS is typical of the definitional measures and is one of the most likely to be used in implementation.
- It was found in phase III of the main study that PERCLOS was one of the most reliable definitional measures between subjects (Wreggit, Kim, and Wierwille, 1993).

The data representing these independent variables from the twelve subjects of phase II of the main study were divided into three variable groups as illustrated in Table 25. The data from each group were then used to develop predictive algorithms using PERCLOS as the dependent variable.

Each group of variables was expanded to include linear terms (X, Y, Z, etc.), cross product terms (XY, XZ, YZ, etc.), and squared terms (X2, Y2, Z2, etc.). These variables were divided into three subgroups with designations LIN, LINCROSS, and FULL, described as follow:

1. Subgroup LIN includes linear terms only,
2. Subgroup LINCROSS includes linear terms and cross product terms, and
3. Subgroup FULL includes linear terms, cross product terms, and squared terms.

Data Analysis

Backwards stepwise multiple regression and re-substitution were performed on each of the nine subgroups of collected data to find optimized combinations of variables that would best predict the values of PERCLOS. Pearson-product-moment correlation (R) analysis and classification matrices were used to analyze the results of these algorithms. Although the use of discriminant analysis was considered, it had been shown that this technique results in negligible gain over the results of multiple regression (Wreggit, Kim and Wierwille, 1993). Therefore, discriminant analysis was not used.

Multiple regression. In each multiple regression analysis the B weights of the various measures were first examined. Pairs of measures that were linearly related would exhibit large offsetting B coefficients. One member of the pair was then removed. Thereafter, the

Table 25: Summary of Variable Groups Used in Algorithm Development.

VARIABLE GROUP 1

(Used for regressions to develop algorithms: 1-LIN, 1-LINCROSS, and 1 -FULL)

Dependent Variable: PERCLOS

Independent Variables:	A-	STVELV	(steering-related measure)
	B-	LGREV	(steering-related measure)
	C -	MDREV	(steering-related measure)
	D-	SMREV	(steering-related measure)
	E-	STEXED	(steering-related measure)
	F-	NMRHOLD	(steering-related measure)
	G-	THRSHLD	(steering-related measure)

VARIABLE GROUP 2

(Used for regressions to develop algorithms: 2-LIN, 2-LINCROSS, and 2-FULL)

Dependent Variable:. PERCLOS

Independent Variables:	A-	LGREV	(steering-related measure)
	B-	STEXED	(steering-related measure)
	C-	NMRHOLD	(steering-related measure)
	D-	THRSHLD	(steering-related measure)
	E-	INTACDEV	(accelerometer-related measure)
	F-	LANDEV	(lane-related measure)
	G-	LINERRSQ	(lane-related measure)

VARIABLE GROUP 3

(Used for regressions to develop algorithms: 3-LIN, 3-LINCROSS, and 3-FULL)

Dependent Variable: PERCLOS

Independent Variables:	A	LGREV	(steering-related measure)
	B	STEXED	(steering-related measure)
	C-	NMRHOLD	(steering-related measure)
	D-	THRSHLD	(steering-related measure)
	E-	MDREV	(steering-related measure)
	F-	INTACDEV	(accelerometer-related measure)
	.G-	LANDEV	(lane-related measure)
	H-	LINERRSQ	(lane-related measure)
	I-	LINRTDEV	(lane-related measure)
	J-	LANEX	(lane-related measure)

elimination of nonsignificant measures ($p > 0.05$) began, starting with the measure having the smallest F-ratio. During each step of this process the Bweights continued to be examined and correction for linearly related measures was made. Once all remaining independent measures were found to be significant ($p < 0.05$), various measures were substituted back into the set. This backward stepwise/re-substitution approach to multiple regression produced the final set of results for each subgroup of variables. Once these results were attained, B weights were used as coefficients for the corresponding independent variable to form predictive algorithms with respect to PERCLOS.

The stability of these predictive algorithms was examined using data collected during phase III (the validation phase) of the main study. The algorithm outputs (predicted) PERCLOS were produced and compared to the actual (observed) PERCLOS. Algorithm accuracy was measured using multiple correlation Pearson-product-moment correlation (R) values and classification matrices.

R values. The algorithms were re-applied to the data which were used in their development, and the resulting predicted PERCLOS data were compared to the actual (observed) PERCLOS data. Algorithm accuracy was measured using Pearson-product-moment correlation (R) values and correlation matrices.

"Clipped" R values. To minimize the confounding effect of outliers within the algorithm output data sets output values greater than 0.45 were set equal to 0.45, and those with values less than zero (0.0) were set equal to zero (0.0). These values are based on the maximum and minimum values of the actual (observed) PERCLOS data. As was done previously, algorithm accuracy was measured using multiple correlation Pearson-product-moment correlation (R) values.

Classification matrices. As indicated, algorithm accuracy was also examined using classification matrices. The threshold levels employed were the same as those used in the algorithm development phase of the main study (Wreggit, Kim, and Wierwille, 1993). The

PERCLOS data were classified into three categories of drowsiness (“awake,” “questionable,” and “drowsy”) according to the following criteria:

Classification	PERCLOS Value
Awake	PERCLOS < 0.075
Questionable	0.075 < PERCLOS < 0.15
Drowsy	PERCLOS > 0.15

Observed (actual) PERCLOS classification data were compared with predicted (algorithm output) data, and the results were summarized in classification matrices. Misclassifications, or data sets in which the predicted category did not match the observed category, were examined and further divided into “large error” misclassifications and “all error” misclassifications. “Large error” misclassifications were defined as any misclassification in which the predicted classifications are two categories away from the observed (actual) classification. To summarize the results of these analyses, Apparent Accuracy Rates (APARs) were calculated for both “large error” misclassifications and “all” misclassifications. The APAR for large misclassifications is the proportion of predicted PERCLOS classifications which are not large errors, and the APAR for all misclassifications is the proportion of predicted PERCLOS classifications which are correct. Classification matrix analysis was not conducted using “clipped” algorithm output (predicted) PERCLOS data since there would have been no differences when compared with unclipped data.

RESULTS

In total, nine algorithms were developed in this study. These algorithms were derived from three separate groups of variables which were each expanded to subgroups of linear (LIN) terms only, linear and cross product (LINCROSS) terms, and all first and second order (FULL) terms from each of the three groups of independent measures. These algorithms were applied to the original (development) data set and to the new (validation) data set from the main study. The results from these data sets were analyzed using Pearson-product-moment correlation (R) values and classification matrices as described earlier.

Multiple Regression

Table 26 contains an example of typical results obtained from multiple regression analyses. First order terms are labeled with the appropriate variable name, while second order terms are signified with letters which correspond to the letter designations noted in Table 25. R values are shown at the top, B weights (non-standardized) are listed in the fourth column and were used as coefficients for the corresponding variable for the purpose of algorithm development. For a complete set of multiple regression tables see Appendix A in Fairbanks and Wierwille (1994).

Pearson-Product-Moment Correlation (R) Values

Correlation (R) values resulting from analysis of the output PERCLOS data (and "clipped" output PERCLOS data) versus actual PERCLOS values for each of the nine algorithms are summarized in Table 27. These results are presented for both the original data set and validation data to allow comparison. Average values for each type algorithm (linear, linear/cross product, and full second order) are presented in the lower section of the table.

Table 26: Multiple Regression Summary for Algorithm 1 -FULL and Dependent Variable PERCLOS

R= 0.86052156 R2= 0.74049735 Adjusted R2= 0.72679123
F(15,284)=54.027 p<0.0000 Std. Error of estimate: 0.05 118

	BETA	St. Err. of BETA	B	St. Err. of B	t(284)	p-level
Intercept			0.00091	0.006	0.158	0.875
STVELV	1.141	0.152	0.00373	0.000	7.512	0.000
MDREV	-0.240	0.068	-0.00548	0.002	-3.513	0.001
SMREV	-0.435	0.090	-0.00322	0.001	-4.812	0.000
STEXED	-0.386	0.117	-196.48261	59.592	-3.297	0.001
NMRHOLD	-0.299	0.107	-0.00516	0.002	-2.796	0.006
A x D	0.277	0.088	0.00006	0.000	3.166	0.002
B x E	0.493	0.120	92.44439	22.431	4.121	0.000
C x F	0.236	0.090	0.00061	0.000	2.607	0.010
C x G	0.162	0.078	0.02384	0.012	2.070	0.039
D x G	-1.068	0.171	-0.04143	0.007	-6.240	0.000
E x G	0.106	0.050	444.52088	211.382	2.103	0.036
F x G	-0.374	0.088	-0.05290	0.012	-4.274	0.000
A-SQUARED	-0.569	0.117	-0.00002	0.000	-4.851	0.000
D-SQUARED	-0.970	0.222	-0.00025	0.000	-4.380	0.000
F-SQUARED	0.348	0.055	0.00094	0.000	6.294	0.000

* Measures designated by a capital letter are defined in Table 19, Variable Group 1.

Table 27: Summary of Results

VARIABLE GROUP	TERMS USED IN REGRESSION	ACCURACY MEASURE TYPE	ORIGINAL DATA SET	VALIDATION DATA SET
VARIABLE GROUP 1: Steering Measures Only	Linear Only (1-LIN)	**R Value**	**0.779**	**0.791**
		Clipped R Value		0.829
		APAR (large errors)	**0.983**	**0.965**
		APAR (all errors)	0.760	0.747
	Cross Products and Linear (1-LINCROSS)	**R Value**	**0.835**	**0.141**
		Clipped R Value		0.739
		APAR (large errors)	**0.980**	**0.955**
		APAR (all errors)	0.797	0.826
	Full 2nd Order (1-FULL)	**R Value**	**0.860**	**0.750**
		Clipped R Value		0.830
		APAR (large errors)	**0.983**	**0.965**
		APAR (all errors)	0.790	0.823
VARIABLE GROUP 2: Steering, Accel, and Lane (small set)	Linear Only (2-LIN)	**R Value**	**0.872**	**0.863**
		Clipped R Value		0.862
		APAR (large errors)	**0.980**	**0.979**
		APAR (all errors)	0.790	0.830
	Cross Products and Linear (2-LINCROSS)	**R Value**	**0.904**	**0.402**
		Clipped R Value		0.678
		APAR (large errors)	**0.980**	**0.962**
		APAR (all errors)	0.807	0.858
	Full 2nd Order (2-FULL)	**R Value**	**0.912**	**0.415**
		Clipped R Value		0.701
		APAR (large errors)	**0.980**	**0.965**
		APAR (all errors)	0.810	0.861
VARIABLE GROUP 3: Steering, Accel. and Lane (large set)	Linear Only (3-LIN)	**R Value**	**0.872**	**0.863**
		Clipped R Value		0.862
		APAR (large errors)	**0.980**	**0.979**
		APAR (all errors)	0.790	0.830
	Cross Products and Linear (3-LINCROSS)	**R Value**	**0.925**	**0.694**
		Clipped R Value		0.845
		APAR (large errors)	**0.993**	**0.976**
		APAR (all errors)	0.820	0.872
	Full 2nd Order (3-FULL)	**R Value**	**0.929**	**0.590**
		Clipped R Value		0.853
		APAR (large errors)	**0.993**	**0.976**
		APAR (all errors)	0.840	0.833
AVERAGE VALUES	Linear Only (LIN)	**R Value**	**0.841**	**0.839**
		Clipped R Value		0.851
		APAR (large errors)	**0.981**	**0.974**
		APAR (all errors)	0.780	0.802
	Cross Products and Linear (LINCROSS)	**R Value**	**0.888**	**0.412**
		Clipped R Value		0.754
		APAR (large errors)	**0.984**	**0.964**
		APAR (all errors)	0.808	0.852
	Full 2nd Order (FULL)	**R Value**	**0.900**	**0.585**
		Clipped R Value		0.795
		APAR (large errors)	**0.985**	**0.969**
		APAR (all errors)	0.813	0.839

"Clipped" Algorithm Output PERCLOS Values

The R values resulting from correlation analyses using "clipped" output data were notably higher than R values resulting from the use of non-clipped data. The average clipped R value = 0.800 and the average non-clipped R value = 0.6 12 for nine validated algorithms. The R values used for the averages are in Table 27.

Classification Matrices

All APAR values are summarized in Table 27.

DISCUSSION and CONCLUSIONS

Interpretation of Results

Stability of algorithms when applied to new subject data R values and APAR values in comparison of the results from the original and validation data sets (Table 27) suggest that, in general, all nine algorithms maintained predictive propensity when applied to new data. Additionally, R values using “clipped” data exhibited an even greater stability between subjects. It should be noted that APAR values were calculated using “pure” algorithm output data (“clipped” data were only used in R value calculations). In general, all algorithms were able to predict appropriate classification of data in 78% to 85% of all cases.

Effect of higher order algorithms on accuracy of prediction In all cases (variable groups 1, 2, and 3) the inclusion of higher order terms in the multiple regression process did not increase the predictive abilities of the resulting algorithms. In fact, when applied to new data for validation, the algorithms which used linear only terms produced higher R values in every case. Although small improvements in average APAR values for all errors occur between first and second order algorithms applied to original data. there is no similar improvement when applied to validation data.

Outliers. It was found that higher order algorithms may have a greater propensity to produce outliers than linear algorithms, when the algorithms were applied to new (validation) data. This propensity was probably a result of multiplying measures together that have moderate statistical instability. Such measures would occasionally exhibit larger derivations, causing extreme values in algorithm output.

Table 27 shows that clipping (limiting) algorithm output to a feasible range can produce large increases in R values when higher order algorithms are applied to new data. For example, in the case of 1 -LINCROSS, the value of R increased from 0.141 to 0.739 when clipping was applied. The fact that clipping can produce substantial increases in R values further supports the hypothesis that higher order models have a greater number and larger outliers when applied to new data.

Outliers have less of an effect on classification matrices. The reason for this is that once the output of an algorithm exceeds a threshold, it does not matter whether it exceeds the threshold by a small amount or a large amount. The classification selected is the same. Nevertheless, outliers are a symptom of underlying instability. Thus, it could be hypothesized that the measures devised for drowsy driver detection do not have sufficient statistical reliability or stability to benefit from algorithms using higher-order terms

Conclusions

The results of this study suggest that the use of second order terms in driver drowsiness detection algorithms does not result in detection accuracy improvement when the algorithms are applied. This is a surprising result and it underscores the importance of testing newly derived algorithms on a second set of data (that is, a validation set). Had this not been done, it would have been concluded that higher order terms were capable of providing detection accuracy improvements.

The results of this study also do not bode well for even more sophisticated detection algorithms, such as pattern recognition or neural networks. Since these latter approaches are actually sophisticated nonlinear optimization procedures, there is a possibility that they would not provide improvement in detection accuracy (over less sophisticated techniques) when they are applied to new (validation) data. At the very least, it can be stated, based on the results of the present study, that all such sophisticated algorithms must be applied to a second set of data for classification accuracy evaluation. Otherwise, when such algorithms are applied in a field experiment, their detection capabilities may be found wanting and the reasons may not be fully understood.

Finally, although not conclusively proven by the present study, the results do support the hypothesis that higher-order algorithms produce more and larger outliers when applied to new data than do linear algorithms.

Chapter Seven

Part Two: Further Algorithm Refinement and Investigation – A Comparison of R Values Obtained from the Application of Algorithms to Original A/O Data, New A/O Data, and New Clipped A/O Data

(Work reported in this part of Chapter Seven has not appeared in previous semiannual research reports. This work was carried out by Steven S. Wreggit and Walter W. Wierwille. It is referred to as Wreggit and Wierwille, 1994a.)

INTRODUCTION

The purpose of this follow-up study was to examine the possibility of potential improvement in A/O based algorithm prediction accuracy by confining algorithm output values to the minimum and maximum values of the observed data. However, it is important to note that there were few outliers in the A/O algorithm output data.

Various A/O task algorithms were developed and validated in previous phases of this study (Wreggit, Kirn, and Wierwille, 1993 and Wreggit, Kirn, and Wierwille, 1994) and it was found that the R values in the validation phase (using new A/O data) were significantly lower than the R values obtained in the development phase (using original A/O data). When the results of the Fairbanks and Wierwille (1994) study became available, it was felt that the significant decrease in R values from the development phase (using A/O data) to the validation phase (using A/O data) could be due to the effects of prediction outliers. Therefore, this supplemental study was undertaken.

The algorithm output values were limited to the minimum and maximum values of the corresponding observed data. In other words, the outliers were "clipped" from the data and set to a value equal to the largest and smallest observed data. Therefore, no outliers were present when the subsequent correlation analyses were run. A comparison of R values obtained from analyses of original data, new data, and "clipped" data were examined.

METHOD

Data Analysis

Certain algorithm output data and the observed definitional measures that were collected and calculated during the validation phase were used in this follow-up study. The data employed in this study consisted of the algorithm output data from eight previously developed A/O task algorithms and the four definitional measures of drowsiness.

Algorithm output data were limited to the minimum and maximum values of the observed data so that no prediction outliers were present in the data set.

Correlations between the algorithm output (prediction data) and observed data were run. The resulting R values were compared with the R values attained during analyses of the new A/O data.

RESULTS and CONCLUSIONS

Table 28 shows the R values that were obtained during the analyses of the new data and new clipped data as well as the original data. The original data were not “clipped” as were the new (validation) data. It can be seen in Table 28 that the R values resulting from the new data and new clipped data are practically the same. The average R values for each data type is as follows:

Original: Average R = 0.809
New: Average R = 0.606
New Clipped: Average R = 0.608

It can be concluded from the results of this follow-up study that the significant decrease in A/O task performance based algorithm prediction strength was not due to prediction outliers. Inspection of the data revealed that the number and magnitude of prediction outliers was minimal. However, in the previous phase of this study in which higher-order algorithm outputs were “clipped”, an increase in drowsiness prediction occurred. It was found by Fairbanks and Wierwille (1994) that higher order algorithms may have a greater propensity to produce outliers than linear algorithms. This propensity was probably a result of multiplying measures together. Therefore, if linear algorithms are employed there is no need to limit the upper and lower values of the prediction data.

Table 28: R Values From Multiple Regression Analyses When Algorithms were Applied to Original A/O Data, New A/O Data, and New Clipped* A/O Data.

Independent Measures		Dependent Measures				
		AVEOBS	EYEMEAS	NEWDEF	PERCLOS	MASTER
I	A/O Task Measures Only	Algorithm I1a (original) 0.761 (new) 0.595 (new clipped) 0.607	Algorithm I2a (original) 0.768 (new) 0.570 (new clipped) 0.572	Algorithm I3a (original) 0.660 (new) 0.422 (new clipped) 0.426	Algorithm I4a (original) 0.810 (new) 0.447 (new clipped) 0.437	Algorithm I5a (original) 0.822 (new) 0.570 (new clipped) 0.574
J	A/O Task, Steering, & Accelerometer	-------------	-------------	-------------	Algorithm J4a (original) 0.836 (new) 0.599 (new clipped) 0.595	-------------
L	A/O Task, LANDEV/VAR, LNMNSQ, LANEX, & LNERRSQ	-------------	-------------	-------------	Algorithm L3a (original) 0.875 (new) 0.796 (new clipped) 0.595	-------------
M	A/O Task, Steering, Accelerometer, LANDEV/VAR LNMNSQ, LANEX, & LNERRSQ	-------------	-------------	-------------	-------------	Algorithm M3a (original) 0.936 (new) 0.845 (new clipped) 0.849

* Clipped refers to data sets that contained no data greater or less than the maximum values of the observed data. Thus, any outliers that were present were *clipped* out of the data set.

Chapter Seven

Part Three: Further Algorithm Refinement and Investigation – An Investigation of False Alarm Rates When Applying Detection Algorithms to Alert-Driver Segments

(Work reported in this part of Chapter Seven has not appeared in previous semiannual reports. This work was carried out by Steven S. Wreggit and Walter W. Wierwille. It is referred to as Wreggit and Wierwille, 1994b.)

INTRODUCTION

The false alarm rate of any warning system must be reasonably low to be practical and marketable. Users of a warning system with a high false alarm rate would easily become annoyed and may become habituated to frequent false signals, thus ignoring or disbelieving a true warning of impending driver impairment.

This follow-up study addresses the concern of how well several typical algorithms perform when drivers are alert. The algorithms that were employed in this follow-up study were based on the definitional measures PERCLOS and AVEOBS. The goal of this study was to determine the false alarm rate produced by several algorithms if observed alert data were used exclusively.

It is important to note here that in the algorithm development and validation studies, the false alarm rates may have been artificially high (compared with an actual on-the-road situation) since the subject-drivers had been partially sleep deprived. Since very drowsy subject-drivers were employed, the observed level of alertness would have, in many cases, been very close to the "drowsy threshold" (the level of alertness determined previously that indicates impairment).

METHOD

Data Analysis

Certain algorithm output data and the observed definitional measures that were collected and calculated during the validation phase were used in this follow-up study. The data employed in this study consisted of the observed data from the definitional measures PERCLOS and AVEOBS. The previously collected defmtional measures contained alert data segments and drowsy data segments.

Since the purpose of this study was to examine the accuracy of previously developed algorithms when applied to alert data only, the non-alert segments were deleted. Any observed PERCLOS data greater than or equal to 0.030 were deleted. Any observed AVEOBS data greater than or equal to 35.0 were deleted. It should be noted here that the cut-off points employed in this study were different than the thresholds used in the algorithm development and validation phases. The reason for this difference is that only “very alert” segments were used in this false alarm rate examination.

RESULTS and CONCLUSIONS

Table 29 shows three classification matrices and corresponding R values based on alert data. It can be seen that the R values are quite low while the APARs are very high (in some cases 1.0). The R values were expected to be low since the data were so tightly grouped, thus the use of alert segments only did not allow for the detection algorithms to be exercised to their fullest. In other words, the factor that may have contributed to the reduction of drowsiness prediction R values was that a very small range of drowsiness/alertness was observed,

It can be concluded from these results from the typical algorithms D1a, D4a, and F4a that low false alarm rates can be achieved when drowsy-driver detection algorithms are applied during alert segments exclusively. This finding is important since drivers are alert a majority of the time. The very low false alarm rates achieved in this follow-up study are a significant finding because they represent false alarm rates that would be typical in an alert driving situation. The false alarm rates during the validation study were slightly higher than would be typical since the drivers/subjects were partially sleep deprived. Of course, false alarm rates and classification accuracies in an actual application can be expected to differ from those presented in this report, because the relative numbers of alert, questionable, and drowsy epochs for actual driving are unknown.

Table 29: Classification Matrices with Awake Data Only -- Demonstration of False **Alarm** Rate .

Algorithm #D 1 a vs. AVEOBS

				Predicted	
	Group	% Correct	Awake	Questionable	Drowsy
Observed	Awake	98.97	96	- 1	0
	Questionable	N/A	0	0	0
	Drowsy	N/A	0	0	0
	Total	98.97	96	1	0

R Value = 0.157
Apparent Accuracy Rate (large misclassifications): 1 .0000
Apparent Accuracy Rate (all misclassifications): 0.9897

Algorithm #D4a vs. PERCLOS'

				Predicted	
	Group	% Correct	Awake	Questionable	Drowsy
Observed	Awake	86.57	174	23	**4**
	Questionable	N/A	0	0	0
	Drowsy	N/A	0	0	0
	Total	86.57	174	23	4

R Value = 0.370
Apparent Accuracy Rate (large misclassifications): 0.980
Apparent Accuracy Rate (all misclassifications): 0.866

Algorithm #F4a vs. PERCLOS

				Predicted	
	Group	% Correct	Awake	Questionable	drowsy
Observed	Awake	94.53	190	11	0
	Questionable	N/A	0	0	0
	drowsy.	N/A	0	0	0
	Total	94.53	190	11	0

R Value = 0.499
Apparent Accuracy Rate (large misclassifications): 1.000
Apparent Accuracy Rate (all misclassifications): 0.945

Chapter Eight

Summary of Findings and Recommendations

INTRODUCTION

Because numerous investigations were carried out during this three-year study and a great number of important results were obtained, it was felt beneficial to summarize the most important findings in one place. Therefore, this chapter provides in brief form the major findings of the research. Any research team undertaking further research on drowsy-driver detection should examine this summary carefully since not doing so may result in substantial lost effort.

It should be remembered that all of the results obtained are for research conducted in a validated simulator using ordinary young drivers in a state of partial sleep deprivation. These results are believed to be indicative of actual driving under similar on-the-road night time conditions.

It must be pointed out that the automobile simulator located in the Vehicle Analysis and Simulation Lab at Virginia Polytechnic Institute and State University does accurately represent actual driving. In other words, this simulator handles and feels like an actual automobile. Furthermore, this simulator has been validated so that quantitative values similar or equal to corresponding full-scale (field-test) results can be obtained. The results of this research are believed to be accurate due to the realism and validated performance of the automobile simulator used.

MAINFINDINGS

1. Appropriate operational definitions of drowsiness are important components when developing drowsiness detection algorithms. Four definitions have been evolved and have been used in this research. (A fifth measure, consisting of the average of the standardized values of the other four was also used) They need not be obtainable in an on-the-road setting.
2. No single measure currently obtainable in an actual automobile is capable of producing sufficient accuracy to serve as a measure for drowsiness detection. However, combinations of operational measures are capable of providing reasonably accurate drowsiness detection. The measures showing the greatest promise as components in a detection system are lane- and heading-related measures, steering-related measures, and lateral accelerometer-related measures.
3. The algorithms capable of the greatest detection accuracy have the following general characteristics:
   a. They are composed of four to seven component measures.
   b. They were derived from the above cited measure sets.
   c. Measures were initially obtained over one-minute intervals.
   d. Averaging six consecutive one-minute-interval measures produces the highest drowsiness prediction accuracy.
   e. Algorithms were derived by means of multiple linear regression with thresholds applied subsequently.
4. Discriminant analysis procedures did not produce better results than multiple regression followed by thresholding.
5. Prediction models that included heart-rate measures, in most cases, were not more accurate at predicting drowsiness than models that did not include heart-rate measures.

On the whole, it is not worth encumbering the driver with a plethysmograph to obtain heart-rate measures for the slight improvement in the prediction of drowsiness.

6. A secondary task seems to provide an alternative method for the detection of driver drowsiness, however, a small loss of accuracy was observed when secondary task based algorithms were applied to the new set of data in the validation phase.
7. The relative predictive strengths of the five definitional measures of drowsiness used in this study varied somewhat. In general, the most accurate dependent measures were MASTER and PERCLOS, followed in decreasing order by AVEOBS, EYEMEAS, and NEWDEF.
8. Experienced raters were able to produce a good operational measure of drowsiness. This was accomplished by viewing videotaped images of driver-subjects and rating each one-minute segment for level of drowsiness. Three raters were employed and rated independently of one another. The scores from all raters were averaged for each one-minute segment to create the definitional measure of drowsiness called AVEOBS.
9. The accuracy of the algorithms that classified levels of drowsiness most accurately are characterized by the following:
    a. The average apparent accuracy rate (APAR) for all errors when developed algorithms were applied to a new set of data and when dual thresholds were used was approximately 0.829. (This average was calculated using seven values in Table 16).
    b. The average APAR for large errors only when developed algorithms were applied to a new data set and when dual thresholds were used was approximately 0.971 (This average was calculated using seven values in Table 16).
10. The first two minutes of each data set should not be used for data analysis and should be deleted. This procedure was found to be necessary since many drivers had a “settling in” time of approximately two minutes. In other words, once the driver-subjects were placed in the simulator, even after they had experienced a lengthy

practice run (approximately 10- 15 minutes), it was observed that they still needed around two minutes to begin driving normally. In an on-the-road setting drivers would usually use this time for reaching speed and settling into the driving task.

11. Baselining is desirable when developing the data sets for analysis. The process of baselining was used to account for individual differences in physiological characteristics, driving ability, and capability to perform certain tasks. The data used for the baselining procedure were the initial ten minutes of data (after the first two minutes were deleted). These data are averaged and then subtracted from all subsequent one-minute segments for each driver's data set. Therefore, baselining was carried out so that data relative to the subject's initial data values could be obtained.

12. When typical algorithms based on driving performance were applied to a new data set using different subjects driving under similar but not identical conditions, no loss in. detection accuracy resulted. Both R values and classification matrix accuracies maintained their values.

13. The use of twelve representative subjects was sufficient to characterize algorithms for general use. Care must be taken to ensure that bouts of drowsiness do in fact occur in several of the drivers. Otherwise, algorithms obtained will not be properly "trained" for drowsiness detection.

14. There was a degradation in R values for previously developed algorithms that included A/O (secondary) task measures, when the algorithms were applied to new data. The drop in value averaged 0.2. However, classification matrices did not exhibit a correspondingly large decrease in accuracy. Instead, their reduction in accuracy was small.

15. The reduction in R values when NO task algorithms were applied to a new set of data was probably a result of using only four subjects to develop the algorithms, or possibly a result of the limited number of bouts of drowsiness in the new (validation) data.

16. R values for validation results may underestimate the capabilities of algorithms to classify correctly, especially when a small subject sample is used to develop the algorithms.

17. The dullest of driving conditions showed signs of inducing drowsiness. The combination of barren highway, engaged cruise control, no traffic, and the absence of a secondary task was found to induce drowsiness most readily and with greatest frequency.

18. If drivers who are being monitored by a drowsiness-detection device are in fact alert, the false alarm rates as presented in Chapters 4 and 5 are generally larger than will actually be encountered. The false alarm rates generated in these chapters are for drivers who are partially sleep deprived. However, a follow-up study was conducted that demonstrated that if the developed detection algorithms were applied to only alert driving segments, a lower false alarm rate would result (see Chapter Seven: Part Three). It should be noted here, however, that "real-world" accuracy rates are likely to differ from the accuracy rates of the main analyses and follow-up study because the ratio of alert, questionable, and drowsy epochs of normal drivers is unknown.

19. Higher order detection algorithms (second order in particular) do not provide improved accuracy when applied to a new data set.

20. Longitudinal measures do not provide any appreciable improvement in detection accuracy. Driver speed variation does not provide independent information not already available in other measures.

21. Classification matrices based on certain algorithms resulted in high correct classification rates (APARs) even though relatively low R values were obtained from multiple regression analyses. This occurred when A/O task data were used for algorithm development. During these segments the driver-subjects were more alert. The good results of the classification matrices suggest that if alert drivers make up the subject pool a deflation in R values will be seen. Thus, R values should not be relied

upon solely as an indication of the accuracy of detection algorithm. Further steps must be taken in the investigation of algorithm accuracy by way of classification matrix analyses.

22. When developing the algorithms by linear regression methods, extreme care must be taken with problems of colinearity. The best procedure found to deal with this problem was to examine statistical output for nearly equal but opposite B coefficients and eliminate the measure that contributed least to the prediction accuracy of the algorithm being developed. Failure to heed this warning will result in prediction algorithms that appear to possess high predictive accuracy but subsequently show a large drop in accuracy when applied to a new set of data.

RECOMMENDATIONS

Efforts are currently being undertaken in a project extension to develop and implement a drowsy-driver detection system. The first phase focuses on the effectiveness of various types of on-board warning systems. The purpose of the warning system should be to alert the driver that he or she is becoming drowsy. The system would also be used as a means of arousing the driver.

An effective warning must get the attention of a driver even if he or she is drowsy. However, warnings for a drowsy-driver detection device must not be so intrusive and jarring that they startle the driver. Another consideration pertaining to the intrusiveness of the warning is the degree of driver annoyance. However, the warning must not be so conservative that it fails to result in the desired effect of alerting or arousing a driver.

The research to be conducted will involve the use of performance algorithms to detect an increase in subject-driver drowsiness. The algorithms that will be used are those developed by Wreggit, Kim, and Wierwille (1993). Once the detection algorithms have classified a subject-driver as drowsy, a warning will inform the driver that he or she is exhibiting signs of impaired driver performance. A “full alarm” will activate if the driver does not manually reset the system.

One of the objectives of the research will be to determine the optimal tone and/or voice warning to be used for the initial warning. The option to reset the system will give the driver the opportunity to avoid exposure to the “full alarm”. The action of resetting the warning may also interrupt a driver’s increasing drowsiness for at least a short period of time. When the driver depresses the reset button the initial warning system will be disengaged for approximately five minutes.

Another objective of the research is to determine the optimal full-alarm signal to be used. Auditory displays such as various modulated tones and rumble strip-like sounds may be investigated along with steering-column vibration and driver’s seat-vibration. The alarm will continue to be displayed with increasing intensity until it is manually deactivated by the driver. Once the driver has deactivated the alarm the detection algorithms will be disengaged for approximately five minutes. The driver will then be given the option of selecting drowsiness countermeasures. Once the multi-stage warning and alerting system described above has been developed in the automobile simulator there is still a need for full-scale implementation and testing of the in-car driver-drowsiness detection and alerting system.

REFERENCES

Brown, I. D. (1965). A comparison of two subsidiary tasks used to measure fatigue in car drivers. Ergonomics, 8, 467-473.

Brown, I. D. (1966). Effects of prolonged driving upon driving skill and performance of a subsidiary task. Industrial Medicine and Surgery 35, 760-765.

Brown, I. D., Simmonds, D. C. V., and Tickner, A. H. (1967). Measurement of control skills, vigilance and performance of a subsidiary task during 12 hours of car driving. Ergonomics, 10, 665-673.

Carroll, J. S., Blisewise, D. L., and Dement, W. C. (1989). A method for checking interobserver reliability in observational sleep studies. Sleep, 12(4), 363-367.

Carskadon, M. A. (1980). A manual for polysomnography (PSG) technicians. Stanford, California: Stanford University School of Medicine, Department of Psychiatry and Behavioral Sciences.

Dement, W. C. (1975) Proposals for future Research. In G. Lairy and P. Salzarulo (Eds.), The Experimental Study of Human Sleep: Methodological Problems (pp. 435-443). New York: Elsevier Scientific Publishing Company.

Dingus, T. A., Hardee, L. H. and Wierwille, W. W. (1985). Detection of drowsv and intoxicated drivers based on highway driving performance measures. (IEOR Department Report #8402). Vehicle Simulation Laboratory, Human Factors Group. Virginia Polytechnic Institute and State University, Blacksburg, Virginia.

Dureman, E. I. and Boden, C. (1972). Fatigue in simulated car driving. In Y. Seko (Trans.), Present technological status of detecting drowsy driving patterns. Jidosha Gijutsu, 30(5), 547-554. Central Research Institute, Nissan Motor Company.

Ellsworth. L. A., Wreggit, S. S., and Wierwille, W. W. (1993). Research on vehicle-based driver status/performance monitoring. (DTNH 22-91 -Y-07266) Third Semiannual Research Report. Virginia Polytechnic Institute and State University, Blacksburg, VA: Department of Industrial and Systems Engineering, Report No. 93-02, April.

Ellsworth, L. A.-(1993). Experimental evaluation of subjective ratings of drowsiness and development of drowsiness definitions. Unpublished masters thesis, Virginia Polytechnic Institute and State University, Blacksburg, VA.

Endo, Inomata, and Sugiyama (1978). Distinctive EOG characteristics observed during automobile driving. In Y. Seko (Trans.), Present technological status of detecting drowsy driving patterns. Jidosha Giiutsu (Vol. 30, No. 5, pp. 547-554). Central Research Institute, Nissan Motor Company.

Erwin, C. W. (1976, February). Studies of drowsiness (Final report). Durham, North Carolina: The National Driving Center.

Erwin, C. W., Hartwell, J. W., Volow, and Alberti, G. S. (1976). Electrodermal change as a predictor of sleep. In Erwin, C. W. (1976, February). Studies of drowsiness (Final Report). Durham, North Carolina: The National Driving Center.

Erwin, C. W. (1976, February). Studies of drowsiness (Final report). Durham, North Carolina: The National Driving Center.

Erwin, C. W., Hartwell, J. W., Volow, M. R., and Alberti, G. S. (1976). Electrodermal change as a predictor of sleep. In Erwin, C. W. (1976). Studies of drowsiness (Final report). Durham, North Carolina: The National Driving Center, (February).

Fairbanks, R. J. and Wierwille, W. W. (1994). Effects of using higher-order algorithms on drowsy driver detection accuracy. Virginia Polytechnic Institute and State University, Blacksburg, VA: Department of Industrial and Systems Engineering, Supplemental Research Report, October.

Fontana, F. (1765) Dei moti dell' iride. In C. Guilleminault (Ed.), Sleep and Waking Disorders: Indications and Techniques. Menlo Park, CA: Addison-Wesley.

Hardee, L. H., Dingus, T. A. and Wierwille, W. W. (1985). A comparison of three subsidiary tasks used as driver drowsiness countermeasures. (IEOR Department Report #8505). Vehicle Simulation Laboratory, Human Factors Group. Virginia Polytechnic Institute and State University, Blacksburg, Virginia.

Hauri, P. H. (1982). The sleep disorders. (2nd ed.). Kalamazoo, Michigan: Upjohn.

Haworth, N. L., Vulcan, P., Triggs, T. J. and Fildes, B. N. (1989). Driver fatigue research Development of methodology. Accident Research Center, Monash University, Australia.

Haworth, N. L. and Vulcan, P. (1990). Testing of commercially available fatigue monitors. (Draft Report). Accident Research Center Monash University, Australia.

Hiroshige, Y., and Niyata, Y. (1990). Slow eye movements and transition period EEG sleep stages during daytime sleep. In Planque, S., Chaput, D, Petit, C., and Tarriere, C. (1991, November 4-7). Analysis of EOG and EEG signals to detect lapses of alertness in car simulation driving. Paper presented at the 13th ESV Conference, Paris, France.

Hulbert, S. F. Blood sugar level and fatigue effects on a simulated driving task. Engineering report 63-43, UCLA, October, 1963. Cited in Hulbert, S. Effects of driver fatigue. Human Factors in Highway Traffic Research, 228-302, New York: Wiley, 1972.

Hulbert, S. F. (1972). Effects of driver fatigue. In T. W. Forbes (Ed.) Human factors in highway traffic safety research. New York: Wiley and Sons, 1972.

Huntley, M. S., and Centybear, T. M. (1974). Alcohol, sleep deprivation, and driving speed effects upon control use during driving. Human Factors, 16, 19-28.

Kim, C. L., Wreggit, S. S., and Wierwille, W. W. (1994). Research on vehicle-based driver status/performance monitoring (DTNH 22-91 -Y-07266) Sixth Semiannual Research Report. Virginia Polytechnic Institute and State University, Blacksburg, VA: Department of Industrial and Systems Engineering, Report No. 94-02, October.

Knipling, R. R., and Wang, J. S. Crashes and fatalities related to driver drowsiness/fatigue, NHTSA Research Note, November. 1994.

Knipling, R. R.. and Wierwille, W. W. (1994, April). Vehicle-based drowsy driver detection. Current status and future prospects. Paper presented at the IVHS America Fourth Annual Meeting, Atlanta, Georgia.

Knipling, R. R. and Wierwille, W. W. (1993, December 9-10). U. S. IVHS research: Vehicle-based drowsy driver detection. Paper presented at the Vigilance and Transport Conference, Lyon, France; sponsored by the French National Institute for Transport and Safety Research (INRETS).

Kurokawa, K. and Wierwille, W. W. (1990). Validation of a driving simulation facility for instrument panel task performance. In Proceedings of the 34th Annual Meeting of the Human Factors Society, (pp. 1299-1303). Santa Monica, CA: Human Factors Society.

Kuroki, Kitakawa, and Oe. (1974). Mental and physical responses as evidenced by the EEG and cerebral discharge induction potential during driving. In Seko, Y. (1984). Present technological status of detecting drowsy driving patterns. Jidosha Gijutsu, 30(5), 547-554. Central Research Institute, Nissan Motor Company.

Laurel, H., and Lisper, H. 0. (1978). A validation of subsidiary reaction time against detection of roadside obstacles during prolonged driving. Ergonomics, 21, 8 l-88.

Leonard, J. and Wierwille, W. W. (1975). Human performance validation of simulators: Theory and experimental verification. In Proceedings of the 19th Annual Meeting of the Human Factors Society, (pp. 446-456). Santa Monica, CA: Human Factors Society.

Lisper, H. O., Laurell, H., and Van Loon. J. (1986). Relation between time to falling asleep behind the wheel on a closed track and changes in subsidiary RT during prolonged driving on a motor way. Ergonomics, 29, 445-453.

Lowenstein, O., and Loewenfeld, I. (1963). Pupillary movements during acute and chronic fatigue: A new test for the objective evaluation of tiredness. In C. Guilleminault (Ed.), Sleep and Waking Disorders: Indications and Techniques. Menlo Park, CA: Addison-Wesley.

Lowenstein, O., and Loewenfeld, I. (1964). The sleep-wake cycle and pupillary activity. In C. Guilleminault (Ed.), Sleep and Waking Disorders: Indications and Techniques. Menlo Park, CA: Addison-Wesley.

Mast, T. M., Jones, H. V., and Heimstra, N. W. (1966). Effects of fatigue on performance in a driving device. Highway Research Record, 122, 93 (Abridgment). Cited in Haworth, N. L., Vulcan, P., Triggs, T. J. and Fildes, B. N. (1989). Driver fatigue research: Development of methodology. Accident Research Center, Monash University, Australia.

Muto, W. H. and Wierwille, W. W. (1982). The effects of repeated emergency response trials on performance during extended-duration simulated driving. Human Factors, 24, 693-698.

Office of Crash Avoidance Research (1991). Report No. 4: Drowsy/fatigued driver crashes. (IVHS/Crash Avoidance Countermeasure-Target Crash Problem Size Assessment and Statistical Description) Washington, D. C.: NHTSA, OCAR, September.

Ogilvie, R. D., Wilkinson, R. T., and Allison, S. (1989). The detection of sleep onset: behavioral, physiological, and subjective convergence. Sleep, 12(5), 458-474.

Planque, S., Chaput, D, Petit, C., and Tarriere, C. (1991. November 4-7). Analysis of EOG and EEG signals to detect lapses of alertness in car simulation driving. Paper presented at the 13th ESV Conference, Paris, France.

Plan, F. N. (1963). A new method of measuring the effects of continued driving performance. Highway Research Record, 25, 33-57. Cited in Safford, R., and Rockwell, T. H. (1967). Performance decrements in twenty-four hour driving. Highway Research Record, 163, 68-79.

Riemersma, J. B. J., Sanders, A. F., Wildervanck, C., and Gaillard, A. W. (1977). Performance decrement during prolonged night driving. In R. R. Mackie (Ed.), Vigilance: Theory, operational performance and physiological correlates to New York Plenum Press.

Ryder, J. M., Malin, S. A., and Kinsley, C. H. (1981). The effects of fatigue and alcohol on highway safety. National Highway Traffic Safety Administration Report No. DOT-HS-805-854. Cited in Dingus, T. A., Hardee, L. H. and Wierwille, W. W. (1985). Detection of drowsy and intoxicated drivers based on highway driving performance measures. (IEOR Department Report #8402). Vehicle Simulation Laboratory, Human Factors Group. Virginia Polytechnic Institute and State University, Blacksburg, Virginia.

Safford, R., and Rockwell, T. H. (1967). Performance decrements in twenty-four hour driving. Highway Research Record, 163, 68-79.

Santamaria, J., and Chiappa, K. H. (1987). The EEG of drowsiness in normal adult. In Planque, S., Chaput, D, Petit, C., and Tarriere, C. (1991, November 4-7). Analysis of EOG and EEG signals to detect lapses of alertness in car simulation driving. Paper presented at the 13th ESV Conference, Paris, France.

Seko, Y. (1984). Present technological status of detecting drowsy driving patterns. Jidosha Gijutsu. 30(5), 547-554. Central Research Institute, Nissan Motor Company.

Skipper, J. H., Wierwille, W. W., and Hardee, L. (1984). An investigation of low-level stimulus-induced measures of driver drowsiness. (IEOR Department Report #8402). Vehicle Simulation Laboratory, Human Factors Group. Virginia Polytechnic Institute and State University, Blacksburg, Virginia.

Sugarman, R. C., and Cozad, C. P. (1972). Road test of alertness variables. (NTIS No. PB-215 450/8). Washington, D. C.: National Highway Traffic Safety Administration, U. S. Department of Transportation.

Sussman, R. D., Sugarman, R. C., and Knight, J. R. (1971). Use of simulation in a study investigating alertness during long-distance, low-event driving. Highway Research Record, 364, 27-32

Thorpy, M. J. and Ledereich, P. S. (1990). Medical treatment of obstructive sleep apnea. In M. J. Thorpy (Ed.) Handbook of sleep disorders. Bronx, New York: Monte Fiore Medical Center and Albert Einstein College of Medicine, 1990.

Tilley, D. H., Erwin, C. W., and Gianturco, D. T. (1973). Drowsiness and driving: Preliminary report of a population survey. Society of Automotive Engineers, International Automotive Engineering Congress, Detroit Michigan, January 8- 12, Report No. 730121.

Torsvall, L., and Akerstedt, T. (1988). Extreme sleepiness: Quantification of EOG and EEG parameters. International Journal of Neuros. 38, 435-441.

Virginia traffic crash facts. (1990) Department of Motor Vehicles. Richmond, VA.

Volow, M. R., and Erwin C. W. (1973, January 8-12). The heart rate variability correlated of spontaneous drowsiness onset. Society of Automotive Engineers International Automotive Engineering Congress, Detroit, Michigan, January 8-12, Report No. 730124.

Wierwille, W. W. and Muto, W. H. (1981). Significant changes in driver-vehicle response measures for extended simulated driving tasks. Paper presented at the First European Annual Conference on Human Decision Making and Manual Control. Netherlands, May 25-27, 1981, 298-314.

Wierwille, W. W. and Ellsworth, L. E. (1992). Research on Vehicle-based Driver status/performance monitoring (DTNH 22-91 -Y-07266) Second Semiannual Research Report. Virginia Polytechnic Institute and State University, Blacksburg, VA: Department of Industrial and Systems Engineering, Report No. 92-05, October.

Wierwille, W. W., Wreggit. S. S.. and Mitchell, M. W. (1992). Research on vehicle-based driver status/performance monitoring (DTNH 22-91-Y-07266) First Semiannual Research Report. Virginia Polytechnic Institute and State University, Blacksburg, VA: Department of Industrial and Systems Engineering, Report No. 92-01, April.

Wreggit, S. S., Kim, C. L., and Wierwille, W. W. (1993). Research on vehicle-based driver status/performance monitoring (DTNH 22-91-Y-07266). Fourth Semiannual Research Report. Virginia Polytechnic Institute and State University, Blacksburg, VA: Department of Industrial and Systems Engineering, Report No. 93-06, October.

Wreggit, S. S., Kim, C. L., and Wierwille, W. W. (1994). Research on vehicle-based driver status/performance monitoring (DTNH 22-91-Y-07266) Fifth Semiannual Research Report. Virginia Polytechnic Institute and State University, Blacksburg, VA: Department of Industrial and Systems Engineering, Report No. 94-01, April.

Wreggit, S. S. and Wierwille, W. W. (1994a). A comparison of R values obtained from the application of algorithms to original A/O data, new A/O data, and new clipped A/O data. Virginia Polytechnic Institute and State University, Blacksburg, VA: Department of Industrial and Systems Engineering, Supplemental Research Report, August.

Wreggit, S. S. and Wierwille, W. W. (1994b). An investigation of false alarm rates when applying detection algorithms to alert-driver segments. Virginia Polytechnic Institute and State University, Blacksburg, VA: Department of Industrial and Systems Engineering, Supplemental Research Report, August.

Yabuta, K., Iizuka, H., Yanagishima, T., Kataoka. Y., and Seno, T. (1985). The development of drowsiness warning devices. The Tenth International Technical Conference on Experimental Safety Vehicles. Nissan Motor Company.

Appendix A

Regression Summaries and Classification Matrices for Selected Algorithms

(Numbering of algorithms is the same as in previous technical reports)

Regression Summary for Dependent Variable: AVEOBS
R = 0.74732 R2= 0.55849 Adjusted R2 = 0.54791
F(7,292)= 52.768 p < 0.0000 Std. Error of estimate: 16.765

	BETA	St. Err. of BETA	B	St. Err. of B	t(292)	p-level
Intercept			29.060127	1.597	18.200	0.000
ACCVAR	-0.155	0.068	-1.738049	0.763	-2.278	0.023
INTACDEV	-0.207	0.058	-33.279385	9.389	-3.544	0.000
ACEXEED	0.117	0.050	290.205383	124.826	2.325	0.021
STVELV	-0.238	0.089	-0.197639	0.074	-2.670	0.008
LGREV	0.561	0.078	14.484322	2.010	7.207	**0.000**
MDREV	0.537	0.065	3.120561	0 778	8.264	0.000
THRSHLD	0.213	0.044	50.180883	10.407	4.822	**0.000**

Predicted

	Group	% Correct	Awake	Questionable	Drowsy
Original	Awake	93.41	**156**	6	5
Observed	Questionable	18.87	36	**10**	7
	Drowsy	50.00	19	21	40
	Total	68.67	211	37	52

AVEOBS (R Value = 0.747)

Apparent Accuracy Rate (large misclassifications): 0.920
Apparent Accuracy Rate (all misclassifications): 0.687

Classification Matrix Generated From Multiple Regression Analysis of **Original AVEOBS** Data Resulting in **Algorithm Dla.** (Independent variables employed included Steering and Accelerometer.)

Predicted

	Group	% Correct	Awake	Questionable	Drowsy
New	Awake	96.43	189	5	2
Observed	Questionable	12.20	29	5	7
	Drowsy	62.75	14	5	32
	Total	78.47	232	15	41

AVEOBS (R Value = 0.727)

Apparent Accuracy Rate (large misclassifications): 0.944
Apparent Accuracy Rate (all misclassifications): 0.785

Algorithm Dla Applied to New Data and Compared with New Observed AVEOBS Data

Figure Al : Regression Summary and Classification Matrices Showing Accuracy of Algorithm D 1 a When Applied to Original Data and New Data.

Regression Summary for Dependent Variable: PERCLOS
R = 0.78910 R2= 0.62268 Adjusted R2 = 0.61626
F(5,294) = 97.038 p < 0.0000 Std. Error of estimate: 0.06065

	BETA	St. Err. of BETA	B	St. Err. of B	t(294)	p-level
Intercept			0.014132	0.005	2.974	0.003
INTACDEV	-0.134	0.038	-0.084597	0.024	-3.488	0.001
LGREV	0.395	0.050	0.040055	0.005	7.845	0.000
STEXED	0.146	0.042	74.427007	21.231	3.506	0.001
NMRHOLD	-0.427	0.054	-0.007378	0.001	-7.985	0.000
THRSHLD	0.450	0.047	0.416209	0.044	9.491	0.000

Predicted

	Group	% Correct	Awake	Questionable	Drowsy
Original	Awake	88.29	**181**	22	2
Observed	Questionable	43.18	11	**19**	14
	Drowsy	52.94	2	22	**27**
	Total	75.67	194	63	43

PERCLOS (R Value = 0.789)

Apparent Accuracy Rate (large misclassifications):	0.987
Apparent Accuracy Rate (all misclassifications):	0.757

Classification Matrix Generated From Multiple Regression Analysis of **Original PERCLOS** Data Resulting in **Algorithm D4a.** (Independent variables employed included Steering and Accelerometer.)

Predicted

	Group	% Correct	Awake	Questionable	Drowsy
New	Awake	79.32	188	40	9
Observed	Questionable	30.00	8	6	6
	Drowsy	90.32	1	‘2	28
	Total	77.08	197	48	43

PERCLOS (R Value = 0.800)

Apparent Accuracy Rate (large misclassifications):	0.965
Apparent Accuracy Rate (all misclassifications):	0.771

Algorithm D4a Applied to New Data and Compared with **New** Observed **PERCLOS** Data

Figure A2: Regression Summary and Classification Matrices Showing Accuracy of Algorithm D4a When Applied to Original Data and New Data.

Regression Summary for Dependent Variable: MASTER

R = 0.80116 R2'= 0.64185 Adjusted R2 = 0.63452

F(6,293) = 87.5 18 p < 0.0000 Std. Errorof estimate: 2.1481

	BETA	St. Err. of BETA	B	St. Err. of B	t(293)	p-level
Intercept			-2.374181	0.206	-11.503	0.000
INTACDEV	-0.188	0.038	-4.325066	0.875	-4.942	0.000
LGREV	0.448	0.054	1.651363	0.198	8.340	0.000
MDREV	0.149	0.051	0.123422	0.042	2.930	0.004
STEXED	0.090	0.041	1672.678369	751.930	2.225	0.027
NMRHOLD	-0.314	0.054	-0.196918	0.034	-5.821	**0.000**
THRSHLD	0.357	0.047	11.974089	1.579	7.582	**0.000**

Predicted

	Group	% Correct	Awake	Questionable	Drowsy
Original	Awake	93.53	188	10	3
Observed	Questionable	33.33	14	14	14
	Drowsy	52.63	3	24	30
	Total	77.33	205	48	47

MASTER (R Value = 0.801)

Apparent Accuracy Rate (large misclassifications):	0.980
Apparent Accuracy Rate (all misclassifications):	0.773

Classification Matrix Generated From Multiple Regression Analysis of **Original MASTER** Data Resulting in **Algorithm D5a.** (Independent variables employed included Steering and Accelerometer.)

Predicted

	Group	% Correct	Awake	Questionable	Drowsy
New	Awake	93.52	202	10	4
Observed	Questionable	25.93	16	7	4
	Drowsy	80.00	4	5	36
	Total	85.07	222	22	44

MASTER (R Value = 0.837)

Apparent Accuracy Rate (large misclassifications):	0.972
Apparent Accuracy Rate (all misclassifications):	0.851

Algorithm D5a Applied to New Data and Compared **with New** Observed **MASTER** Data

Figure A3 : Regression Summary and Classification Matrices Showing Accuracy of Algorithm D5a When Applied to Original Data and New Data.

Regression Summary for Dependknt Variable: PERCLOS
R = 0.84691010 R2 = 0.71725672 Adjusted R2 = 0.71244816
F(5,294) = 149.16 p < 0.0000 Std. Error of estimate: 0.05250

	BETA	St. Err. of BETA	B	St. Err. of B	t(294)	p-level
Intercept			-0.000348	0.004	-0.078	0.938
ACCVAR	-0.128	0.035	-0.005646	0.002	-3.615	0.000
HPHDGDE	0.616	0.045	0.182652	0.013	13.600	0.000
STEXED	0.112	0.035	56.959348	18.000	3.164	0.002
NMRHOLD	-0.296	0.048	-0.005112	0.001	-6.154	0.000
THRSHLD	0.320	0.043	0.295479	0.040	7.463	0.000

Note: classification matrices not developed for this algorithm.

Figure A4: Regression Summary for Algorithm E4a.

Regression Summary for Dependent Variable: AVEOBS

R = 0.82577937 R2 = 0.68191157 Adjusted R2 = 0.67428617

F(7,292) = 89.426 p < 0.0000 Std. Error of estimate: 14.230

	BETA	St. Err. of BETA	B	St. Err. of B	t(292)	p-level
Intercept			25.645817	1.390	18.457	0.000
ACCDEV	-0.350	0.045	-14.565780	1.874	-7.773	0.000
ACEXEED	0.099	0.039	246.164352	95.978	2.565	0.011
LANDEV	1.142	0.079	21.903765	1.516	14.450	0.000
LNERRSQ	-0.667	0.065	-1.300765	0.127	-10.229	0.000
STVELV	-0.146	0.064	-0.121066	0.053	-2.268	0.024
MDREV	0.503	0.059	2.919365	0.343	8.517	0.000
THRSHLD	0.128	0.038	30.226578	9.054	3.339	0.001

Note: classification matrices not developed for this algorithm.

Figure A5 : Regression Summary for Algorithm Fl a.

Regression Summary for Dependent Variable: EYEMEAS
R = 0.83700489’ R2 = 0.70057719 Adjusted R2 = 0.69339924
F(7,292) = 97.601 p < 0.0000 Std. Error of estimate: 768.40

	BETA	St. Err. of BETA	B	St. Err. of B	t(292)	p-level
Intercept			967.482741	73.780	13.113	0.000
ACCDEV	-0.336	0.042	-779.208055	97.170	-8.019	0.000
LNMNSQ	-0.372	0.084	-28.996034	6.559	-4.421	0.000
LANDEV	0.738	0.120	787.576776	127.703	6.167	0.000
LANEX	0.269	0.063	3048.878638	709.593	4.297	0.000
STVELV	0.157	0.062	7.247299	2.854	2.540	0.012
MDREV	0.230	0.057	74.233060	18.585	3.994	0.000
THRSHLD	0.140	0.038	1828.608058	493.389	3.706	0.000

			Predicted		
	Group	% Correct	Awake	Questionable	Drowsy
Original	Awake	90.10	182	12	8
Observed	Questionable	10.00	9	2	9
	Drowsy	64.10	11	17	50
	Total	78.00	202	31	67

EYEMEAS (R Value = 0.837)

Apparent Accuracy Rate (large misclassifications): 0.963
Apparent Accuracy Rate (all misclassifications): 0.780

Classification Matrix Generated From Multiple Regression Analysis of **Original EYEMEAS** Data Resulting in **Algorithm F2a.** (Independent variables employed included Steering, Accelerometer, LANDEVNAR, LNMNSQ. LANEX, & LNERRSQ.)

			Predicted		
	Group	% Correct	Awake	Questionable	Drowsy
New	Awake	95.51	134	7	4
Observed	Questionable	12.50	0	1	7
	Drowsy	71.43	3	7	25
	Total	90.28	237	15	36

EYEMEAS (R Value = 0.838)

Apparent Accuracy Rate (large misclassifications): 0.976
Apparent Accuracy Rate (all misclassifications): 0.903

Algorithm F2a Applied to New Data and Compared with New Observed **EYEMEAS** Data

Figure A6: Regression Summary and Classification Matrices Showing Accuracy of Algorithm F2a When Applied to Original Data and New Data.

Regression Summary for Dependent Variable: NEWDEF
R = 0.73127598‘ R2 = 0.53476456 Adjusted R2 = 0.52845628
F(4,295) = 84.772 p < 0.0000 Std. Error of estimate: 1.1789

	BETA	St. Err. of BETA	B	St. Err. of B	t(295)	p-level
Intercept			-0.518427	0.090	-5.740	0.000
INTACDEV	-0.153	0.042	-1.693902	0.463	-3.660	0.000
LANVAR	0.255	0.055	0.031894	0.007	4.660	0.000
LANEX	0.350	0.057	4.908173	0.803	6.109	0.000
STVELV	0.250	0.052	0.014324	0.003	4.802	0.000

Predicted

	Group	% Correct	Awake	Questionable	Drowsy
Original	Awake	**83.42**	**161**	**26**	**6**
Observed	Questionable	**35.29**	**25**	**18**	**8**
	Drowsy	**62.50**	**8**	**13**	**34**
	Total	71.33	194	**57**	**49**

NEWDEF (R Value = 0.73 1)

Apparent Accuracy Rate (large misclassifications): 0.953
Apparent Accuracy Rate (all misclassifications): 0.713

Classification Matrix Generated From Multiple Regression Analysis of **Original NEWDEF** Data Resulting in **Algorithm F3a.** (Independent variables employed included Steering, Accelerometer, LANDEVNAR, LNMNSQ, LANEX, & LNERRSQ.)

Predicted

	Group	% Correct	Awake	Questionable	Drowsy
New	Awake	**93.36**	**197**	13	1
Observed	Questionabl	e 26.19	**23**	**11**	**8**
	Drowsy	85.71	1	**4**	**30**
	Total	**82.64**	**221**	**28**	**39**

NEWDEF (R Value = 0.8 19)

Apparent Accuracy Rate (large misclassifications): **0.993**
Apparent Accuracy Rate (all misclassifications): **0.826**

Algorithm F3a Applied to New Data and Compared with New Observed **NEWDEF** Data

Figure A7: Regression Summary and Classification Matrices Showing Accuracy of Algorithm F3a When Applied to Original Data and New Data.

Regression Summary for Dependent Variable: PERCLOS
R = 0.87159526’ R2 = 0.75967830 Adjusted R2 = 0.75475703
F(6,293) = 154.37 p < 0.0000 Std. Error of estimate: 0.04849

	BETA	St. Err. of BETA	B	St. Err. of B	t(293)	p-level
Intercept			-0.003	0.004053	-0.694	0.488
INTACDEV	-0.109	0.030	-0.069	0.019114	-3.603	0.000
LANDEV	0.873	0.063	0.066	0.004763	13.798	0.000
LNERRSQ	-0.258	0.054	-0.002	0.000410	-4.820	0.000
STEXED	0.090	0.033	45.740	16.818827	2.720	0.007
NMRHOLD	-0.204	0.045	-0.004	0.000785	-4.494	0.000
THRSHLD	0.250	0.041	0.23 1	0.037904	6.098	0.000

Predicted

	Group	% Correct	Awake	Questionable	Drowsy
Original	Awake	89.76	184	18	3
Observed	Questionable	47.73	7	21	16
	Drowsy	62.75	3	16	32
	Total	79.00	194	55	51

PERCLOS (R Value = 0.872)

Apparent Accuracy Rate (large misclassifications): 0.980
Apparent Accuracy Rate (all misclassifications): 0.790

Classification Matrix Generated From Multiple Regression Analysis of **Original PERCLOS** Data Resulting in **Algorithm F4a.** (Independent variables employed included Steering, Accelerometer, LANDEVNAR, LNMNSQ, LANEX, & LNERRSQ.)

Predicted

	Group	% Correct	Awake	Questionable	Drowsy
New	Awake	89.03	**211**	22	4
Observed	Questionable	15.00	12	3	5
	Drowsy	80.65	3	3	25
	Total	82.99	226	28	34

PERCLOS (R Value = 0.862)

Apparent Accuracy Rate (large misclassifications): 0.976
Apparent Accuracy Rate (all misclassifications): 0.830

Algorithm F4a Applied to ‘New Data and Compared with **New** Observed **PERCLOS** Data

Figure A8: Regression Summary and Classification Matrices Showing Accuracy of Algorithm F4a When Applied to Original Data and New Data.

Regression Summary for Dkpendent Variable: MASTER
R = 0.88641410 R2 = 0.78572996 Adjusted R2 = 0.77908020
F(9,290) = 118.16 p < 0.0000 Std. Error of estimate: 1.6701

	BETA	St. Err. of BETA	B	St. Err. of B	t(290)	p-level
Intercept			-2.982588	0.170	-17.551	0.000
ACCVAR	-0.163	0.047	-0.259953	0.075	-3.466	0.001
INTACDEV	-0.091	0.042	-2.087452	0.972	-2.147	0.033
LANDEV	0.757	0.101	2.069666	0.275	7.515	0.000
LANEX	0.174	0.059	5.049306	1.719	2.938	6.004
LNERRSQ	-0.298	0.061	-0.082919	0.017	-4.886	0.000
STVELV	0.116	0.052	0.013713	0.006	2.205	0.028
MDREV	0.161	0.049	0.133662	0.040	3.307	0.001
NMRHOLD	-0.100	0.046	-0.062747	0.029	-2.189	0.029
THRSHLD	0.168	0.039	5.636318	1.315	4.287	0.000

			Predicted		
	Group	% Correct	Awake	Questionable	Drowsy
Original	Awake	94.53	**190**	9	2
Observed	Questionable	45.24	9	**19**	14
	Drowsy	71.93	3	13	41
	Total	83.33	202	41	57

MASTER (R Value = 0.886)
Apparent Accuracy Rate (large misclassifications): 0.983
Apparent Accuracy Rate (all misclassifications): 0.833

Classification Matrix Generated From Multiple Regression Analysis of **Original MASTER** Data Resulting in **Algorithm F5a.** (Independent variables employed included Steering, Accelerometer, LANDEVNAR, LNMNSQ, LANEX, & LNERRSQ.)

			Predicted		
	Group	% Correct	Awake	Questionable	Drowsy
New	Awake	96.76	209	5	2
Observed	Questionable	18.52	22	5	0
	Drowsy	62.22	6	11	28
	Total	84.03	237	21	30

MASTER (R Value = 0.885)
Apparent Accuracy Rate (large misclassifications): 0.972
Apparent Accuracy Rate (all misclassifications): 0.840

Algorithm F5a Applied to New Data and Compared with **New** Observed **MASTER** Data

Figure A9: Regression Summary and Classification Matrices Showing Accuracy of Algorithm F5a When Applied to Original Data and New Data.

Regression Summary for Dependent Variable: PERCLOS
R= 0.80983889 R2 = 0.65583902 Adjusted R2 = 0.64508399
F(3,96) = 60.980 p < 0.00000 Std.Error of estimate: 0.05334

	BETA	St. Err. of BETA	B	St. Err. of B	t(96)	p-level
Intercept			0.002801	0.008	0.362	0.718
AOTIME	1.588	0.186	0.113237	0.013	8.551	0.000
NMWRONG	-1.215	0.192	-0.297367	0.047	-6.338	0.000
NMNR	0.358	0.092	0.160588	0.041	3.899	0.000

Note: Classification matrix not developed for original data for this algorithm.

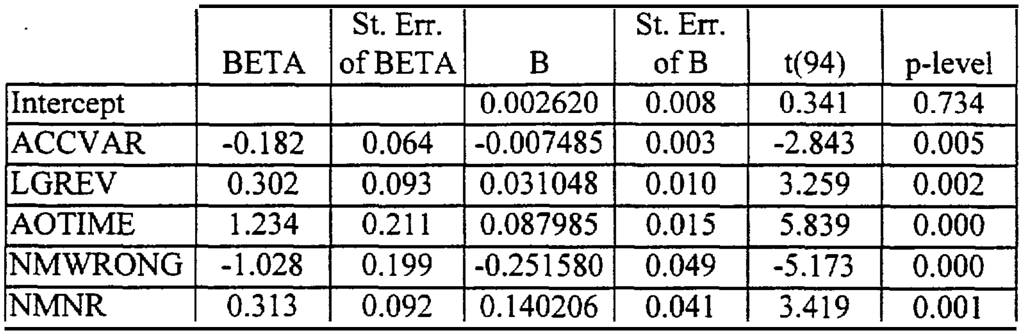
			Predicted		
	Group	% Correct	Awake	Questionable	Drowsy
New	Awake	95.83	**115**	3	2
Observed	Questionable	36.36	6	**4**	1
	Drowsy	23.08	8	2	**3**
	Total	84.72	129	9	6

PERCLOS (R Value = 0.447)
Apparent Accuracy Rate (large misclassifications): 0.93 1
Apparent Accuracy Rate (all misclassifications): 0.847

Algorithm 14a Applied to New Data and Compared with New Observed **PERCLOS** Data

Figure Al 0: Regression Summary and Classification Matrix Showing Accuracy of Algorithm 14a When Applied to New Data.

Regression Summary for Dependent Variable: PERCLOS
R = 0.83585799 ‘R2 = 0.69865857 Adjusted R2 = 0.68262977
F(5,94) = 43.588 p < 0.00000 Std.Error of estimate: 0.05044



	BETA	St. Err. of BETA	B	St. Err. of B	t(94)	p-level
Intercept			0.002620	0.008	0.341	0.734
ACCVAR	-0.182	0.064	-0.007485	0.003	-2.843	0.005
LGREV	0.302	0.093	0.031048	0.010	3.259	0.002
AOTIME	1.234	0.211	0.087985	0.015	5.839	0.000
NMWRONG	-1.028	0.199	-0.251580	0.049	-5.173	0.000
NMNR	0.313	0.092	0.140206	0.041	3.419	0.001

Predicted

	Group	% Correct	Awake	Questionable	Drowsy
Original	Awake	94.29	66	3	1
Observed	Questionable	43.75	2	7	7
	Drowsy	42.86	2	6	6
	Total	79.00	70	16	14

PERCLOS (R Value = 0.836)

Apparent Accuracy Rate (large misclassifications):	0.970
Apparent Accuracy Rate (all misclassifications):	0.790

Classification Matrix Generated From Multiple Regression Analysis of **Original PERCLOS** Data Resulting in **Algorithm J4a.** (Independent variables employed included A/O Task, Steering, and Accelerometer.)

Predicted

	Group	% Correct	Awake	Questionable	Drowsy
New	Awake	93.33	**112**	6	2
Observed	Questionable	27.27	6	3	2
	Drowsy	38.46	4	4	5
	Total	83.33	122	13	9

PERCLOS (R Value = 0.599)

Apparent Accuracy Rate (large misclassifications):	0.958
Apparent Accuracy Rate (all misclassifications):	0.833

Algorithm J4a Applied to New Data and Compared with **New** Observed **PERCLOS** Data

Figure A11: Regression Summary and Classification Matrices Showing Accuracy of Algorithm J4a When Applied to Original Data and New Data.

Regression Summary for Dependent Variable: PERCLOS
R= 0.874877 17 R2= 0.76541006 Adjusted R2 = 0.75027522
F(6,93) = 50.573 p< 0.00000 Std.Error of estimate: 0.04474

	BETA	St. Err. of BETA	B	St. Err. of B	t(93)	p-level
Intercept			-0.004308	0.007	-0.648	0.519
LANVAR	0.924	0.246	0.009506	0.003	3.749	0.000
LANEX	0.310	0.105	0.195697	0.066	2.964	0.004
LNERRSQ	-0.641	0.174	-0.007871	0.002	-3.683	0.000
AOTIME	0.548	0.236	0.039067	0.017	2.326	0.022
NMWRONG	-0.591	0.197	-0.144499	0.048	-3.002	0.003
NMNR	0.286	0.095	0.127975	0.042	3.014	0.003

Note: Classification matrix not developed for original data for this algorithm.

			Predicted		
	Group	% Correct	Awake	Questionable	Drowsy
New	Awake	94.17	**113**	**5**	**2**
Observed	Questionable	**54.55**	**3**	**6**	**2**
	Drowsy	46.15	1	**6**	**6**
	Total	86.81	117	17	10

PERCLOS (R Value = 0.796)

Apparent Accuracy Rate (large misclassifications):	0.979
Apparent Accuracy Rate (all misclassifications):	**0.868**

Algorithm L3a Applied to New Data and Compared with New Observed **PERCLOS**

Figure A12: Regression Summary and Classification Matrix Showing Accuracy of, Algorithm L3a When Applied to New Data.

Regression Summary for Dependent Variable: MASTER
R = 0.93610768 R2 = 0.87629758 Adjusted R2 = 0.86542265
F(8,91) = 80.580 p < 0.00000 Std.Error of estimate: 1.3 177

		St. Err.		St. Err.		
	BETA	of BETA	B	of B	t(91)	p-level
Intercept			-3.437887	0.232	-14.806	0.000
ACCDEV	-0.185	0.043	-1.137986	0.264	-4.3 13	0.000
LANDEV	0.847	0.139	2.583602	0.425	6.077	0.000
LANEX	0.368	0.096	9.328965	2.430	3.840	0.000
LNERRSQ	-0.455	0.081	-0.224128	0.040	-5.622	0.000
THRSHLD	0.133	0.044	5.008195	1.659	3.019	0.003
AOTIME	0.509	0.181	1.456495	0.519	2.805	0.006
NMWRONG	-0.572	0.149	-5.613958	1.466	-3.828	0.000
NMNR	0.192	0.074	3.458096	1.332	2.597	0.011

			Predicted		
	Group	% Correct	Awake	Questionable	Drowsy
Original	Awake	95.59	65	3	0
Observed	Questionable	50.00	3	7	4
	Drowsy	72.22	0	5	13
	Total	85.00	68	15	17

MASTER (R Value = 0.936)

Apparent Accuracy Rate (large misclassifications): 1 .000
Apparent Accuracy Rate (all misclassifications): 0.850

Classification Matrix Generated From Multiple Regression Analysis of **Original MASTER** Data Resulting in **Algorithm M3a.** (Independent variables employed included A/O Task, Steering, Accelerometer, LANDEVNAR, LNMNSQ, LANEX, & LNERRSQ.)

			Predicted		
	Group	% Correct	Awake	Questionable	Drowsy
New	Awake	100.00	**108**	**0**	**0**
Observed	Questionable	0.00	13	0	0
	Drowsy	30.43	8	8	7
	Total	79.86	129	8	7

MASTER (R Value = 0.845)

Apparent Accuracy Rate (large misclassifications): 0.944
Apparent Accuracy Rate (all misclassifications): 0.799

Algorithm M3a Applied to New Data and Compared with **New** Observed **MASTER** Data

Figure Al 3 : Regression Summary and Classification Matrices Showing Accuracy of Algorithm M3a When Applied to Original Data and New Data.