DALE E. OLSEN, NORMAN ANSLEY, IAN E. FELDBERG, JOHN C. HARRIS, and JOHN A. CRISTION

RECENT DEVELOPMENTS IN POLYGRAPH TECHNOLOGY

Are polygraphs able to detect lies? If so, how well? How do they work? This article provides background information on polygraph applications, accuracy, test scoring, and test formats. The Applied Physics Laboratory is using statistical techniques and personal computers in an effort to improve the scoring of polygraph data. Techniques for discriminating truth from deception are described. The results of an application of the first algorithm developed at APL are provided and compared with other techniques currently in use. The Laboratory is also developing a personal computer system for converting polygraph tracings on paper charts into digitized data suitable for analysis by the developing scoring methods.

INTRODUCTION

On any given day, more than a thousand suspects in criminal investigations will voluntarily take polygraph examinations in the hope of being cleared. The analyses performed are based on the assumption that, when deception is attempted, small changes in human physiology occur as a result of either cognitive processing or emotional stress. Tests are administered by more than two thousand trained and experienced examiners in the United States, Canada, Japan, India, Israel, Saudi Arabia, Turkey, and many other nations. Somewhere between 40% and 60% of those who take the tests will be cleared on the basis of an examiner's decision of "No deception indicated." For those who are not cleared, the criminal investigative process will continue. Polygraphs affect the lives of many people, from those who are the victims of criminals to those who are suspects.

A new project at APL is concerned with improving the methods used to analyze polygraph data. Under contract with a government defense agency, APL has a research program to develop statistical methods to analyze digitized physiological signals received and processed by a polygraph instrument. In addition, to bridge the gap between traditional methods and the implementation of new technology, APL is developing an optical scanner system for converting standard polygraph pen tracings into a digital format that will permit us to apply our methods to analyze the data recorded on paper charts. Standard recording devices currently do not produce digitized physiological signals.

The use of polygraphs gives rise to many questions and concerns, the most basic of which is, Do they work? This article provides background information on polygraph technology, including applications, test formats, and polygraph accuracy. Analytic methods for discriminating truth from deception are described, and the results of applying the first APL-developed method are reported. The digitization of paper records presents an interesting challenge; this article describes a method of solution.

BACKGROUND

Polygraph Applications

The primary use of the polygraph test is during the investigative stage of the criminal justice process, but polygraph results are sometimes presented in court as evidence, and polygraph tests do play a small role in parole and probation supervision. In addition to the significant role in criminal justice, polygraph examinations are also used for national security, intelligence, and counterintelligence activities of the United States and foreign nations. Thousands of federal screening examinations are used annually to grant or deny clearance and access to sensitive operations and material.

Test Formats

A polygraph test format is an ordered combination of relevant questions about an issue, control questions that provide physiological responses for comparison, and irrelevant (or neutral) questions that also provide responses or the lack of responses for comparison, or act as a buffer. All questions asked during a polygraph test are reviewed and discussed with the examinee and reworded when necessary to assure understanding, accommodate partial admissions, and present a dichotomy answerable with a definite "yes" or "no." During the test, the questions are delivered in a monotone voice to avoid emphasis on one question or another.

Polygraph examiners have a choice of several standard test formats. The examiner's decision will be based on test objectives, experience, and training. Three classes of test formats are used: control question tests, concealed knowledge tests, and relevant–irrelevant tests. Each format consists of a prescribed series of questions that together make up a chart. Two to five charts make up a test.

Control Question Tests. The majority of criminal investigation tests are conducted by using one of the possible formats of control question tests. These tests consist of a series of control, relevant, and irrelevant questions.

Each question series is repeated two to five times, and each series produces a separate chart. An example of a relevant question is, Did you embezzle any of the missing $12,000? For this test format, the corresponding control questions will be about stealing; the questions are threatening to the subject but are not about the theft at issue. An example is, Before you were employed at this bank, did you ever steal money or property from an employer? The control and relevant questions will be compared. Irrelevant questions will also be asked that will probably be answered truthfully, are not stressful, and act as buffers. Do you reside in Maryland? or Do they call you Jim? are examples of irrelevant questions.

Concealed Knowledge Tests. If the police have facts about a crime that have not become public or common knowledge (facts that would be known to the guilty subject but not to the innocent), they will use a concealed knowledge test.

In a murder case in Maryland, newspapers had revealed only that a woman was found murdered in a motel room. All suspects denied knowledge of the details of her death. The concealed knowledge test format was used in the polygraphs given. The focus was on the method of the murder. The first series comprised four questions: Was Sally stabbed? Was Sally strangled? Was Sally shot? Was Sally poisoned? After three repetitions, each in the same sequence but recorded on separate charts, the examinee was told that the victim was strangled. The second series tested the examinee for concealed knowledge of the means of strangulation. It included six questions: Was Sally strangled with a heavy rope? Was Sally strangled with a jewelry chain? Was Sally strangled with a man's belt? Was Sally strangled with a venetian blind cord? Was Sally strangled with a woman's stocking? Was Sally strangled with a lamp cord? This series was repeated three times on separate charts.

To the first two suspects, the questions had equal emotional value because they did not know before the test which answer was correct. The third examinee had good reason to conceal knowledge of the details, because to admit knowing details would implicate him in the crime. Consequently, his physiological responses were much greater to the question about strangling in the first series and to strangling with a venetian blind cord in the second. After the test, he confessed (personal communication, W. T. Travers, 1979).

In another version of the concealed knowledge test, the examiner does not know the critical item but believes the examinee does know. In one kidnapping case, the examinee was the prime suspect. The examiner used a map marked in grids and asked the examinee where the bodies were buried, specifying each grid location in series. When a section was identified on the basis of the subject's reactions, a map of that section, also divided into grids, was used to narrow the focus. In that manner, the location of the bodies was revealed.[1] This type of test is also used to locate stolen vehicles, money, and other goods, as well as persons in hiding.

Relevant–Irrelevant Tests. A relevant–irrelevant test differs from a control question test in several ways. It has few, if any, control questions on each chart, the sequence of questions usually varies from chart to chart, and the amplitude of reactions to relevant questions is not compared with the amplitude of reactions to the control questions. This type of test is widely used for multiple-issue testing, such as that used for commercial and counterintelligence screening.

Physiological Measures

Regardless of the test format used, three physiological measurements are normally recorded:

1. Volumetric measures taken from the upper arm: A standard blood pressure cuff (see Fig. 1) is placed on the arm over the brachial artery and inflated to about 60 mm Hg pressure for an indirect measure of blood pressure variables, together with the strength and rate of pulsation from the heart.[2]

2. Respiratory measures taken from expansion and contraction of the thoracic and abdominal areas using rubber tubes placed around the subject (see Fig. 1): The resulting data are closely related to the amount of gaseous exchange.[3]

3. Skin conductivity (or resistance) measures of electrodermal activity, largely influenced by eccrine (sweat) gland activity: Electrodes are attached to two fingers of the same hand (see Fig. 2), and a galvanometer records the measured skin conductance or resistance to an electrical current.[4,5]

Accuracy of Polygraph Decisions

Real Cases. When some charts are scored, the examiner cannot make a clear decision and must score the chart as inconclusive. From analyzing charts where decisions were made, a defense agency completed a report on polygraph validity based on all the studies of real cases conducted since 1980. Examiner decisions were compared with other results such as confessions, evidence, and judicial disposition. Ten studies, which considered the outcome of 2,042 cases, were reviewed. It must be

Figure 1. Measuring a subject's cardiovascular and respiratory responses.

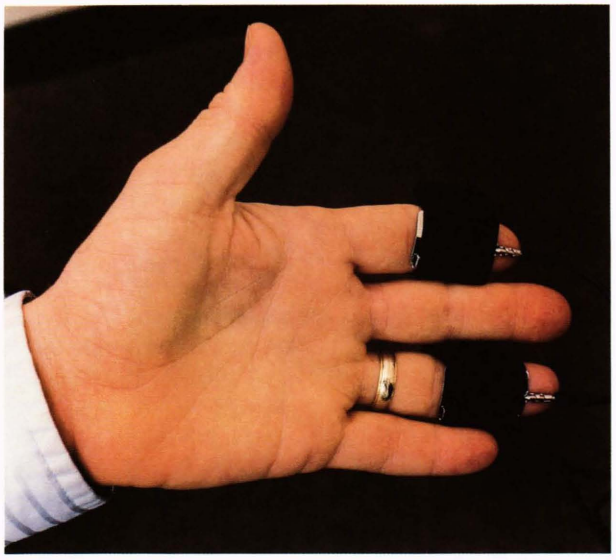


Figure 2. Measuring skin conductivity or resistance using a galvanometer.

pointed out, however, that studies of real polygraph tests are necessarily flawed by the fact that the guilt or innocence of the subject must be determined, and correct calls are more easily confirmed than incorrect calls. For example, if the test shows that the subject is guilty, the examiner will often obtain a confession. Thus, tests scored as guilty are more often confirmed if the subject is guilty. If the test shows that the subject is innocent, other people may be investigated. If another person is found to be guilty, the test becomes confirmed innocent. With this in mind, and assuming that every disagreement was a polygraph error, the results indicate an accuracy (or validity) of 98% for the 2,042 confirmed cases. For deceptive cases, the accuracy was also 98%, and for nondeceptive cases, 97%.

Mock-crime studies, such as the one described later in this article, generally have correct calls about 85% of the time. Because of the nature of mock-crime studies, it is believed that real-crime tests are scored as accurately or more accurately. The accuracy of polygraph decisions for real cases, then, is somewhere between the 85% demonstrated with mock-crime studies and the 98% demonstrated with confirmed charts.

Special Case: Psychopathic Liars. It is widely believed that psychopathic liars can beat a polygraph test because they are not bothered by lying. Psychophysiologists who have studied psychopaths believe they are especially reactive, at least in electrodermal responding.[6,7] All of the studies focused on this issue indicate that the detection rates do not differ significantly for psychopaths and nonpsychopaths.[8-12]

AUTOMATIC SCORING

In 1973, Kubis[13] presented the concept of quantifying polygraph patterns for computer analysis. Later, analysts[14] at the psychology laboratory at the University of Utah began to develop a computerized scoring algorithm, employing a few of the many variables available in physiological patterns. Their research suggested that the most useful measures were the amplitude and duration of the electrodermal response, the rise and fall of the cardiovascular pattern (related to blood pressure changes), and the length of the respiration tracing within a fixed time sequence.[15] These responses were incorporated into a special-purpose computer analytic system, marketed under the name CAPS (Computer Assisted Polygraph System).[14,16]

A novel aspect of the CAPS system is the introduction of decisions based on a probability figure. For example, deception might be indicated with a probability of 0.89. The probabilities were developed by using both laboratory and field polygraph data, the latter being tests conducted by the U.S. Secret Service that were confirmed by confession.[16] Two other features of the CAPS system are its ability to rank-order reactions and its analytic system, which gives the greatest weight to electrodermal responses, less to respiratory responses, and the least to cardiovascular responses. Before this work, scoring systems gave equal weight to responses from each of the three physiological recordings. A deficiency of the CAPS system is that the data are taken from a field polygraph instrument that is often nonlinear, and the analog-to-digital conversion (ADC) is performed after some processing. New instruments will reduce distortions in the data by performing the ADC before any processing, displaying, or printing.

In 1989, Axciton Systems, Inc. of Houston, Texas, developed a new commercial computerized polygraph (see Fig. 3). This system features a computer that processes the physiological signals directly, scrolls the physiological data across a screen in real time during testing, provides for a later printout, records the test on a hard drive or a floppy disk, and provides a system for ranking subject responses. The system has been field tested with real cases in a Texas police department and is user friendly. The charts, printed after the test, look like standard polygraph charts and can be hand-scored by traditional methods. The availability of the Axciton system and the probability of other new computerized polygraphs becoming available make the APL research timely, because instruments on the market will be able to incorporate the results of our work.

RESEARCH AT APL

The development of an automatic scoring algorithm for use with new computerized data collection systems is the focus of our research. The algorithm will be able to use more sophisticated techniques than human examiners, should be more accurate (higher validity), and will ensure consistency from case to case. It is expected that the accuracy will improve as the automatic scoring techniques evolve.

The first algorithm developed uses data collected with the CAPS system and is applicable for the one control question format used to generate the data. Future algorithms will address different question formats and will use digitized data that have been recovered from paper charts. Most polygraph data exist only on charts and will continue to be collected as tracings on paper for much of

the remainder of this century. As a second task, APL is developing a system to scan the paper charts and convert these time series into digitized data to be used for specialized algorithm development and evaluation. Since information is lost when the pens transcribe the signals on paper charts, the data from the scanner will be scored differently.

Algorithm Development for Automatic Scoring

We are using an empirical approach for the development of an algorithm. Data are collected from both deceptive and nondeceptive subjects. The study of the physiological measures requires using digitized data from laboratory studies in which we know whether the subject is telling the truth or being deceptive. Ultimately, the scoring algorithm must be enhanced and validated by using data from real polygraph tests in which the decision of truth or deception has been verified by independent means.

The output of each recording channel (physiological measurements) is represented by a time series of values. Features (also referred to here as parameters) of the time series, which characterize the responses to questions, are studied to determine which can be used to distinguish deceptive and nondeceptive subjects. An example will help to explain the method. Figures 4 and 5 are samples of charts from a mock-crime study. The figures can be used to see how characteristics of skin conductivity are used to detect deception. The subject may be telling the truth or lying about having seen a psychologist, falsifying an employment application, or stealing a watch. Notice that the skin conductivity increases when the subject is asked about the watch. In fact, the subject did steal the watch. The rise is associated with deception and can be characterized by a rapid increase in the amplitude of skin conductivity. (The skin conductivity tracing is plotted 5 to 7 seconds behind the other tracings to reduce pen collisions.) If the pattern of increases is repeated over many cases, it suggests that the change in skin conductivity amplitude, after a question is asked, may be a valuable parameter in discriminating between truth and deception.

Many different features can be used to characterize the time series of responses to the questions. Once a list of features is chosen, it is necessary to select the ones that best separate deceptive and nondeceptive subjects and to develop a discrimination rule based on the parameters. To begin the efforts, polygraph data are required.

Training Data from Mock Crimes. To acquire data for developing an algorithm, a pool of volunteers is selected. Half are randomly chosen and asked to commit a mock crime. All of the volunteers then take a polygraph test, and all deny committing the crime. The polygraph examiners have no knowledge of the guilt of the subjects. They score each test as inconclusive, deceptive, or nondeceptive. Data collected in this type of experimental setting are often used to evaluate different polygraph techniques. Although many mock-crime experiments have been performed, few have produced digitized data suitable for use in developing an automatic scoring system. Suitable data are available, however, from two experiments. One of these data sets has been used to develop an algorithm.

Data from mock-crime studies have the advantage of a well-established knowledge of which subjects are attempting deception. The examiners, however, typically provide more inconclusive and incorrect results for mock crimes than they do for real, verified criminal cases. One possible explanation is that the control questions regarding an individual's character may be more of a threat to the subject than the relevant question concerning a mock crime. As a result, a guilty person may be scored as inconclusive or nondeceptive. For these mock crimes, the threat to the person during a relevant question is substantially diminished, and the reactions being measured are more easily overcome by physiological reactions not related to the attempted deception (low signal-to-noise ratio). Even so, polygraph tests for mock-crime studies usually produce correct conclusions.

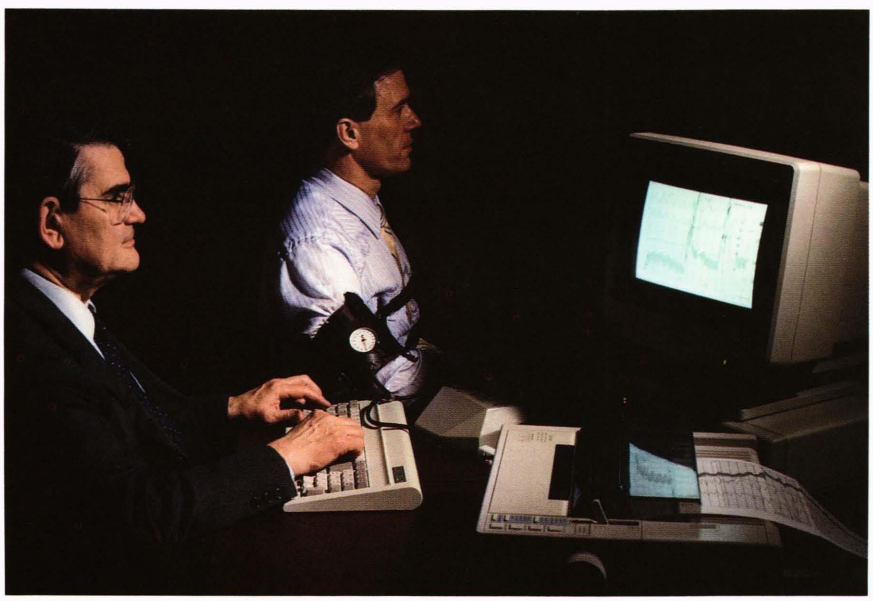


Figure 3. Demonstration of a polygraph examination using the Axciton computerized polygraph system.

Removing Artifacts. Even when data are collected in a laboratory setting, artifacts are present. Artifacts, or data distortions, are often the result of a data collection problem. Figure 6 provides examples of two types of problems: centering artifacts and a clipped series. A centering artifact occurs when the examiner adjusts the polygraph instrument, and a clipped series occurs when the data exceed the dynamic range of the ADC. Problems associated with these and other artifacts are being addressed.

Characterizing the Response. When this project started, the only digitized data available were provided by the Department of Defense Polygraph Institute at Fort McClellan, Alabama. The data consisted of a set of parameters produced by CAPS.[14] As a result, the initial work to develop a discrimination rule has been based only on the features selected by CAPS. This system characterizes each blood pressure response and each skin conductivity response by using twenty parameters, and it characterizes each of the two respiration responses by using only the tracing length for a 10-s period. The twenty parameters characterizing blood pressure and skin conductivity consist of various measures of the area under the response curve; times relating response onset to peak amplitude, or recovery; rise and recovery rates; and the tracing length for a 20-s period.

These parameters characterize the data sufficiently to detect deception as well as human examiners do. Other parameters should be investigated. The APL researchers plan to look at the time-series history just before the question to predict the physiological measurements that would be expected if the subject were not deceptive. This sort of prediction could be used to establish a better baseline for computing parameters such as the area under the curve. Individual characteristics will also be treated

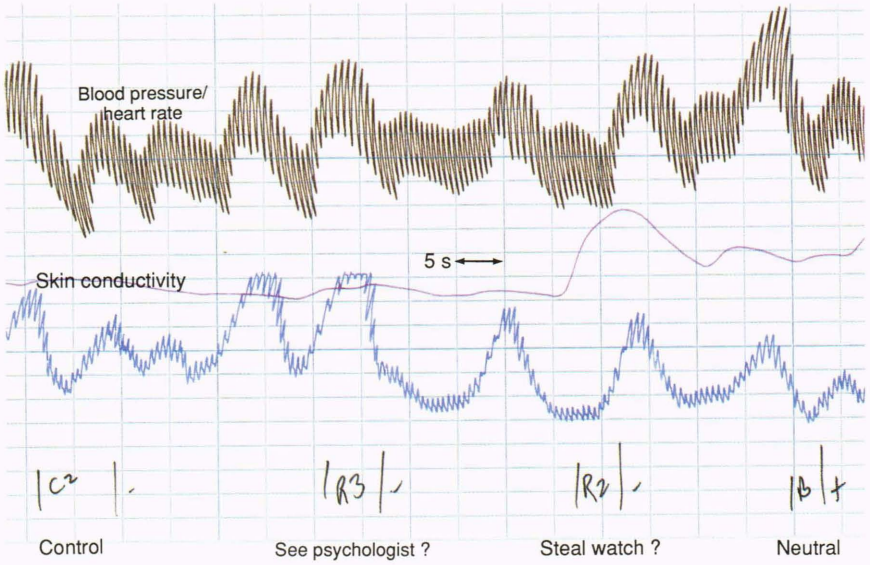


Figure 4. Sample polygraph chart showing responses to control and relevant questions from a mock-crime study.

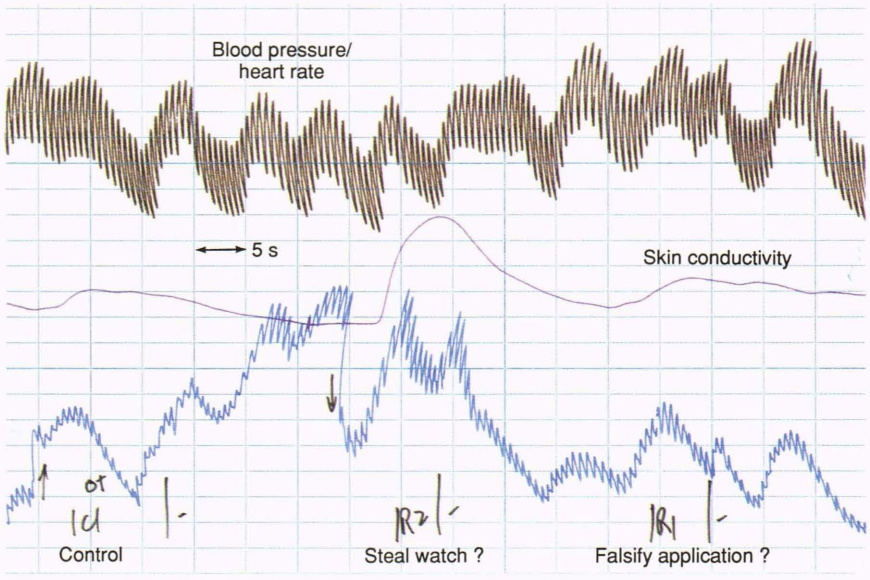


Figure 5. Another sample polygraph chart showing responses to control and relevant questions.

when calculating parameters. For example, some individuals react to questions immediately, whereas others require a few seconds. Reaction-time information is available from the data and should be considered when characterizing the responses. This type of information is used by human examiners but not by the CAPS system.

Other types of parameters are not easily evaluated by the human examiner. Frequency information is nearly impossible to evaluate from paper tracings but is easily obtained by using a computer. Detrending algorithms can be used before the parameters are calculated to remove nominal physiological changes with time. Since algorithms for discrimination studies (neural networks, discriminant analysis, and logit analysis) all provide reasonable discrimination rules, we expect that the biggest improvements in the scoring of the standard measurements will come from a careful development of characterizing parameters.

From the four basic channels of information, more than one hundred different characterizations of the physiological response to a question will eventually be evaluated to find the best subset for discriminating between truth and attempted deception. The vector of characteristics is called a feature vector and is computed for each question. We have not yet described how to select the most useful elements of the vector of characteristics and how to combine them to provide a probability of deception.

Methodologies for Discrimination. Three feature-vector–based methods for discriminating between deceptive and nondeceptive subjects have been considered. The first, the use of artificial neural networks, requires training a network by using feature vectors as input and zero-to-one target (output) vectors to separate innocent and deceptive subjects. Once the network has been trained, new feature vectors produce a number between zero and one. A value near zero indicates that the subject was nondeceptive, and a value near one indicates that the subject attempted deception. This approach produces a useful solution, but it does not provide a probability of deception or clear information about which feature-vector elements are statistically significant, and it does not offer the insights that other methods do. Nevertheless, this approach provides a rich modeling capability and is being investigated.

A second method, statistical discriminant analysis (used by the CAPS system[16]), usually assumes the features can be modeled using a normal probability distribution. During training, statistical models for feature vectors from both deceptive and nondeceptive subjects are built. The method uses these models to compute both the likelihood that a subject is deceptive and the likelihood that the same subject is nondeceptive, and then it makes a decision on the basis of which is more likely. Another application of statistical discriminant analysis, involving kernel estimation, has been applied by Dustin (personal communication, 1990) to polygraph analysis by using techniques described by Priebe and Marchette.[17] This application does not assume normality of feature-vector elements and has provided promising results.

We chose a third method, known as logit regression, for the first algorithm developed at APL. This method provides insights into how the algorithm works. For example, it provides information about how the different channels of data are weighted in making a decision. It allows the investigators to easily explore many ideas and examine factors such as age and sex. In a manner similar to that of linear regression, logit regression selects the best elements of the feature vector to incorporate into the algorithm. It does this while considering the elements already selected for the models, making sure that new elements do not contain essentially the same information. This method does not assume that the feature vectors have a normal distribution.

Logistic Regression. Before explaining logit regression, linear regression will be reviewed briefly. For selected values of $x_1, x_2, \ldots, x_p$, a value y is measured. It is assumed that, except for noise, a linear relationship exists between the x_i values and y; that is,

$$y = a_0 + a_1 x_1, + \ldots + a_p x_p + \text{noise}.$$

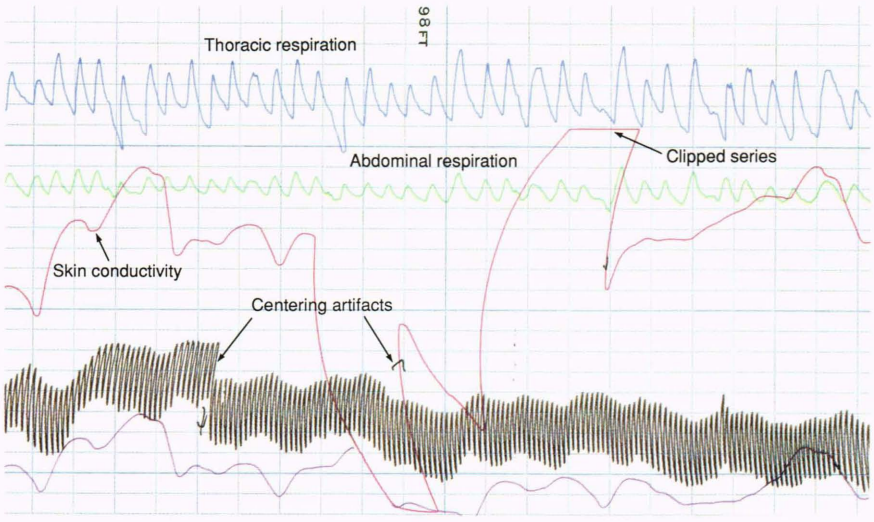


Figure 6. Sample artifacts present on a polygraph chart.

For a series of measured y values, corresponding to different values of x_i's, the coefficients, a_i's, are selected to minimize the squared difference between the observed y and that predicted by using the model; that is, the a_i's are selected to minimize

$$\sum [y - (a_0 + a_1x_1 + \ldots + a_px_p)]^2$$

for all the measured values of y.

The x_i's correspond to the feature-vector elements in this model. The y's would have a value of one for deceptive subjects and zero otherwise. This model allows the use of substitute or additional functions of the feature-vector elements so that models such as

$$\hat{y} = a_0 + a_1x_1 + a_2x_2 + a_3x_1^2 + a_4x_1x_2 \qquad (1)$$

can be developed by letting $x_3 = x_1^2$ and $x_4 = x_1x_2$. The right side of Equation 1 provides an estimate of the mean for y at given values of x.

If linear regression were used to provide probabilities of deception, the model would produce probabilities greater than one and less than zero. The range of the linear model on the right of Equation 1 and the range of the response on the left are mismatched. No valid probability interpretation exists. The odds (the ratio of the probability of nondeceptive to the probability of deceptive) range between 0 and ∞, so the log of the odds ranges between $-\infty$ and $+\infty$. Since a linear model has the same range as the log of the odds, it is reasonable to consider a model of the form

$$\log(\text{odds}) = a_0 + a_1x_1 + \ldots + a_nx_n .$$

The expected value of the log of the odds is more naturally modeled as linear than is the expected probability. This model, called the logit model, is well documented.[18,19] The assumptions for this model are simple: the data (y's) are statistically independent, and the log of the odds is a linear model of feature-vector elements.

Once the coefficients, a_i's, are estimated, the probability of deception is easily calculated by using

$$P = \frac{e^{AX}}{1 + e^{AX}} ,$$

where $AX = a_0 + a_1x_1 + \ldots + a_px_p$.

This model has been used successfully for many years, for a variety of applications. Comprehensive statistical software packages can find the maximum likelihood estimates of the coefficients. The SAS Institute procedure LOGISTIC will perform a "stepwise" search to find those feature-vector elements that best discriminate physiological responses of deceptive subjects from the responses of those who are nondeceptive. The coefficients, a_i's, provide information about how the algorithm works. For example, the coefficients show how the different measurements are weighted and which elements need to be computed.

THE ALGORITHM

The available mock-crime data were taken from a control question test with eleven questions per chart and three charts per subject. The first question for each chart was a neutral question such as, Are you in the state of Alabama? The second question was a sacrifice (not scored) relevant question such as, In regard to the theft, do you intend to answer the questions truthfully? The first two questions give the subject time to adjust to the test questions; they are not used by the examiner when scoring charts. Next, a neutral (N), a control (C), and a relevant (R) question are asked in that order. The control question is a probable lie, whereas the relevant question is directed at the subject's guilt or innocence. This series of three questions is repeated three times with some variations in wording. Thus, the question format on the chart is

N R N C R N C R N C R .

The data consisted of forty-two CAPS parameters computed for each of eleven questions on each of three charts.

In scoring charts with this format, examiners compare adjacent control and relevant responses. This approach, however, does not easily generalize to other test formats and does not allow for the use of the first two questions or the neutral questions, except as buffers. For this reason, the logit model is used to find the best way to combine the information, for a given parameter, on a chart. The data could be displayed as shown in Table 1. The columns in the table are feature vectors. A score is computed for each row in the table and could provide, for example, the probability of deception by using a single feature, such as skin conductivity amplitude, for the chart. For each element, x_{ij}, of the feature vector, the a_{ij}'s in the equation are estimated, using all charts, where

$$\text{score}_i = \log(\text{odds})_i$$

$$= a_{i,0} + a_{i,1}x_{i,1} + a_{i,2}x_{i,2} + \ldots + a_{i,11}x_{i,11} ,$$

and where the x_{ij}'s correspond to the ith feature or parameter for each of the eleven questions.

Since there are forty-two CAPS parameters, this step provides, for each chart, forty-two estimated probabilities of deception, some of which provide similar or correlated information. The log(odds) scores for each parameter are put into a second logit model to determine which of the parameters can best be used for discrimina-

Table 1. Format and feature-vector display for the control question test analyzed.

	Questions on Chart K										
	N	R	N	C	R	N	C	R	N	C	R
Logit computed score	1	2	3	4	5	6	7	8	9	10	11
$Score_1$, for Feature 1	$X_{1,1}$	$X_{1,2}$	.	.	.	.	.	.	.	$X_{1,10}$	$X_{1,11}$
$Score_2$, for Feature 2	$X_{2,1}$	$X_{2,2}$	.	.	.	.	.	.	.	$X_{2,10}$	$X_{2,11}$
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
$Score_{42}$, for Feature 42	$X_{2,1}$	$X_{42,2}$	.	.	.	.	.	.	.	$X_{42,10}$	$X_{42,11}$

tion and to find the best combination of the scores; that is,

$$\log(\text{odds}) = b_0 + b_1(\text{score}_1) + b_2(\text{score}_2) + \ldots + b_{42}(\text{score}_{42}) .$$

A stepwise logit procedure is used to find which $score_i$'s (or parameter combination) can make a statistically significant contribution to the discrimination. If several $score_i$'s provide similar information, only one will have a nonzero B_i coefficient. A parameter may at first appear to have little discriminating capability but may provide useful insights when used with other parameters. The stepwise procedure chooses a subset of statistically important parameters.

Finally, the log(odds)'s from each chart for a subject are averaged and converted to a probability of deception. For purposes of comparison with the examiner, subjects with probabilities between 0.45 and 0.55 are scored as inclusive.

Automatic Scoring Results

When all the test data from the sixty available subjects are used to estimate the coefficients (a_i's and b_i's) in the logit models, and the resulting algorithms are used to score these same subjects, fifty-nine of sixty are scored correctly. Training the algorithm (estimating the a_i's and b_i's) and testing using the same data are not valid procedures. The algorithm essentially memorizes the data. Instead, all subjects but one are used for training, and the remaining one is used for testing. Then, the algorithm is again trained by using fifty-nine subjects, but this time a different subject is held out for testing. The process of training and testing is repeated sixty times; each time the test uses a different subject. This process is known as cross-validation and is used to arrive at the results provided in Table 2.

Since population characteristics for the sampled Fort McClellan subjects may be different from those for other populations, it can be argued that an algorithm trained on this population should perform better for this population than other algorithms or standard procedures used by examiners for a more general population. Since the training data set is very small, however, the results are not expected to be as favorable as other methods. (The training can be influenced by a single subject.) The CAPS training is based on a larger data set containing both real and mock crimes. Another reason not to expect exceptional performance from this training stems from the difficulties with mock-crime data sets discussed earlier. Guilty subjects may be more threatened by control questions than relevant questions. Some subjects will even fall asleep during the test. An additional reason not to expect results better than those of other methodologies is that the parameters are limited to those provided by CAPS.

Table 2. A comparison of scoring methods.

Method of scoring	Percent inconclusive	Percent of remaining correct
Original examiner	23	85
Independent (blind) examiner	23	83
Computer Assisted Polygraph System	40	89
First APL algorithm (0.45 to 0.55 scored as inconclusive)	7	84

Table 2 provides the results for the mock-crime study used to train the algorithm and compares the cross-validated results with those of the CAPS system and standard scoring methods. The first APL algorithm was as accurate as the human examiners and, surprisingly, could remain so even while scoring the more difficult charts. It therefore produced a significantly smaller inconclusive rate. If, when using the new algorithm, probabilities between 0.30 and 0.70 are scored as inconclusive, the results are identical with those for CAPS.

The first attempt at developing an algorithm was limited by the small size of the training data set and

preselected parameters. New algorithms will examine many other parameters computed from recently provided and expected data. Some of the data will be from paper charts digitized with an APL-developed digitizer.

Digitizing Data on Paper Charts

As part of the work at APL, a polygraph digitizer is being developed that takes a paper polygraph chart as input and generates a computer file as output that can be analyzed with the algorithms described previously. Canon scanners were purchased, and the microprocessor control chip was modified so that very long charts (sometimes 9 ft) could be read. Software was developed to scan charts, convert them into raster image files, and store them on disks. These files can be processed with the digitizer software to convert them into meaningful time-series data to be scored by the APL system under development.

Considerations. One problem with digitizing data from paper polygraph charts involves the pens used to trace the data. They are mounted on pivots, so that if large responses are registered, the pen tracings can actually move backward (see Figs. 4 and 5) in time. The effect could be removed if the actual center of the curve (vertical location on the paper) could be found. Therefore, a software module was written that prompts the user to identify the center of each curve to digitize.

Another problem occurs when two traces cross each other on a chart (see Figs. 4 and 5). Any algorithm used to follow data from left to right will be confused when a trace can go in two or more directions. This problem was solved by using a variety of different image processing techniques. First, the traces are thinned by using a classic line-thinning algorithm described by Deutsch.[20] In this process, each pixel of the digitized chart is compared with its neighbors to decide whether or not it is outside of a boundary. If so, it is deleted. This process continues until the skeleton is only one pixel wide at all points. A line-following algorithm was written that traverses this skeleton and converts it into data curves for use with the analysis program.

Regression-like techniques are used to predict where a curve will appear in the future. The predicted direction lags the actual direction in which the curve moves and is less subject to variabilities resulting from the line thinning. When two traces cross, the thinned line branches off in two directions. The program needs only to choose which direction the curve is likely to take. Because the prediction is less subject to anomalies in a particular trace, the program makes a reasonable choice when selecting which of two lines to follow.

Digitization System. The digitization system requires some user intervention. In the future, this prototype will be developed into a system that will require less intervention and will digitize the large majority of polygraph charts, independently of their source. The user will be able to provide chart annotations, grid markings, adjustment and question markings, and a host of other polygraph-specific information.

The New Polygraph Analysis System

A significant effort and an important part of the APL project is the development of a system for entering, editing, displaying, and printing data. On paper charts, tracings cross each other and trend up or down. Pens often hit stops or are reset. The resulting discontinuities in the data make the charts difficult to score. Figure 7 provides an example of a chart with the normal line crossings and pen adjustments. With computer-generated charts, these difficulties can be eliminated. Long-term trends, which make pen adjustments necessary, are removed with a detrending algorithm. Data input definitions can be constructed to allow a range of values to accommodate any reasonable measurements. Figure 8 provides a view of the data shown in Figure 7 after the APL-developed system has detrended them and removed pen artifacts. (These data were digitized by using another system[21] but contained the expected chart artifacts.) The pen used to plot skin conductivity is made longer than the others to reduce the number of pen collisions. This difference creates a 7-s delay in the tracing and an added complication for the examiner trying to score charts. This delay has been removed from the computer image in Figure 8.

FUTURE WORK

Most of our future efforts in scoring the standard physiological measurements are apparent and have been described. The Laboratory will soon have digitized data from law enforcement agencies for use in developing new algorithms. These data will help in the selection of new features and the treatment of different test formats. In the longer term, we hope to investigate other physiological measurements. Evoked potentials from electroencephalograms, muscle movements from electromyographs, QRS waves (a diagnostic feature) from electrocardiograms, eye shifts and saccadic eye movement, and pupillometry have all been shown to have diagnostic value for lie detection in a laboratory setting.

CONCLUSION

In normal conversation, we continually evaluate people's convictions in what they tell us by voice intonations, eye movements, gestures, and general "body-English." Attempted deception by children is often obvious from observing these mannerisms. A jury cannot help but evaluate witnesses and the accused by watching their behavior. In some sense, jurors perform their own ad hoc lie-detection tests.

A polygraph test monitors physiological mannerisms in a controlled environment. Tests are conducted by a knowledgeable examiner asking carefully worded questions. Whether or not polygraphs are measuring lies or emotional responses, they provide a signal whose characteristics can be used to detect attempted deception. Researchers at APL are using scientific methods to determine which characteristics are most important and how they should be combined to provide the best possible estimate of the probability of deception. Since more than one thousand people each day take polygraph tests, our efforts to improve these tests could affect many lives by keeping criminals off the street, monitoring probation, preventing the conviction of innocent people, and improving our national security.

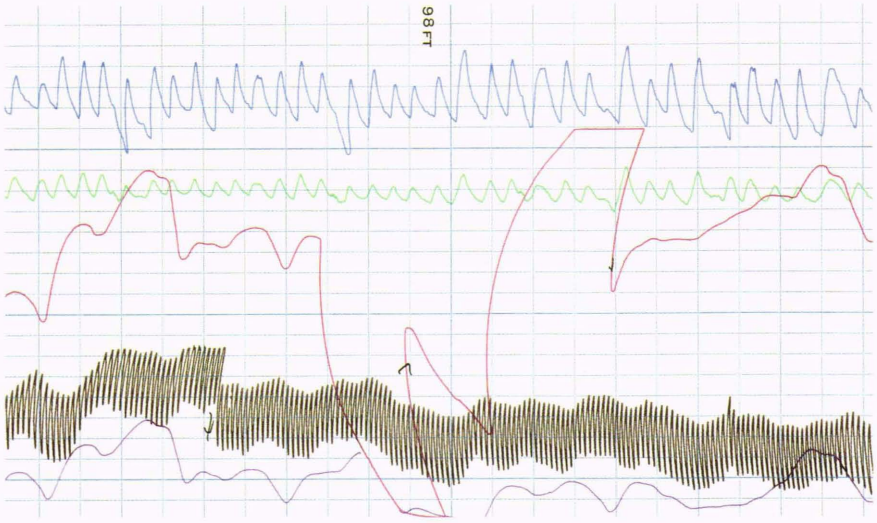


Figure 7. Sample polygraph data on a chart as they appear to the examiner.

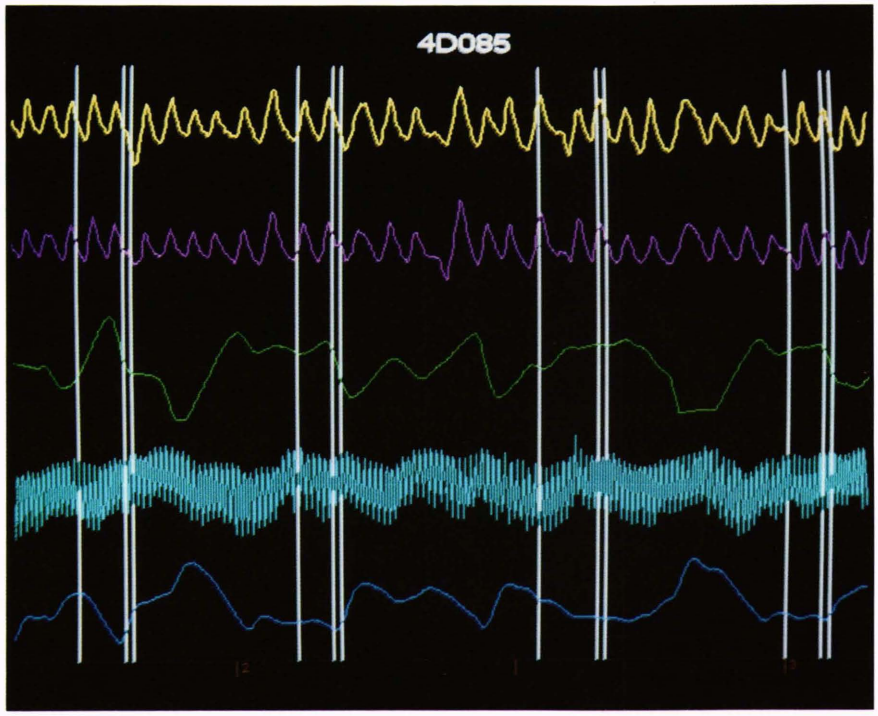


Figure 8. The same polygraph data shown in Figure 7 as they appear to the examiner on a video screen.

REFERENCES

[1] Wilkerson, O. M., "Peak of Tension Tests Utilized in the Ashmore Kidnapping," *Polygraph* **7**(1), 16–20 (1978).

[2] Geddes, L. A., and Newberg, D. C., "Cuff Pressure Oscillations in the Measurement of Relative Blood Pressure," *Psychophysiol.* **14**(2), 198–202 (1977). Reprinted in *Polygraph* **6**(2), 113–122 (1977).

[3] Jacobs, J. E., "The Feasibility of Alternate Physiological Sensors as Applicable to Polygraph Techniques," in *Legal Admissibility of the Polygraph*, Ansley, N. (ed.), Charles C. Thomas, Springfield, Ill., pp. 266–272 (1975).

[4] Summers, W. G., "Science Can Get the Confession," *Fordham Law Rev.* **5**, 334–354 (1939).

[5] Prokasy, W. F., and Raskin, D. C., *Electrodermal Activity in Psychological Research*, Academic Press, New York (1973).

[6] Lykken, D. T., *A Study in Anxiety in Sociopathic Personality,* Ph.D. thesis, Univ. of Minnesota (1955).

[7] Hare, R. D., *Psychopathy: Theory and Research*, John Wiley & Sons, Inc., New York, pp. 37–57 (1970).

[8] Raskin, D. C., and Hare, R. D., "Psychopathy and Detection of Deception in a Prison Population," *Psychophysiol.* **15**, 126–136 (1978).

[9] Patrick, C. J., and Iacono, W. G., "Psychopathy, Threat, and Polygraph Accuracy," *J. Appl. Psych.* **74**, 347–355 (1989).

[10] Hammond, D. L., *The Responding of Normals, Alcoholics and Psychopaths in a Laboratory Lie Detection Experiment*, Ph.D. thesis, California School of Professional Psychology (1980).

[11] Balloun, K. D., and Holmes, D. S., "Effects of Repeated Examinations and the Ability to Detect Guilt with a Polygraphic Examination: A Laboratory Experiment with Real Crime," *J. Appl. Psych.* **64**(3), 316–322 (1979).

[12] Barland, G. H., and Raskin, D. C., *Psychopathy and Detection of Deception in Criminal Suspects*, Society for Psychophysiological Research presentation, Salt Lake City, Utah (Oct 1974).

[13] Kubis, J. F., "Analysis of Polygraphic Data: Dependent and Independent Situations," *Polygraph* **2**(1), 42–58; **2**(2), 89–107 (1973).

[14] Raskin, D. C., Kircher, J. C., Honts, C. R., and Horowitz, S. W., *A Study of the Validity of Polygraph Examinations in Criminal Investigations, Final Report*, National Institute of Justice (May 1988).

[15] Timm, H. W., "Effect of Altered Outcome Expectancies from Placebo and Feedback Treatments on the Validity of the Guilty Knowledge Technique," *J. Appl. Psych.* **67**(4), 391–400 (1982).

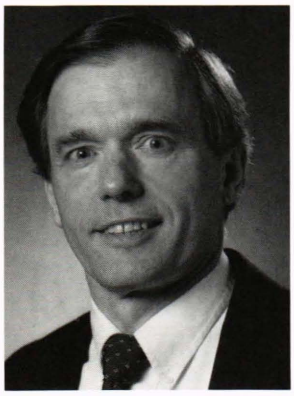
[16] Kircher, J. C., and Raskin, D. C., "Human Versus Computerized Evaluations of Polygraph Data in a Laboratory Setting," *J. Appl. Psych.* **73**(2), 291–302 (1988).

[17] Priebe, C. E., and Marchette, D. J., "Adaptive Mixture Density Estimation," *J. Pattern Recognition* (in press).

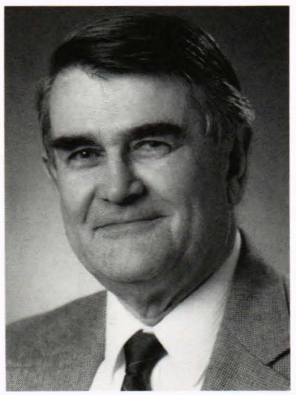
[18] Cox, D. R., and Snell, E. J., *Analysis of Binary Data*, Chapman & Hall, London (1989).
[19] Hosmer, D. W., and Lemeshow, S., *Applied Logistic Regression*, John Wiley & Sons, Inc., New York (1989).
[20] Deutsch, E. S., "Thinning Algorithms on Rectangular, Hexagonal, and Triangular Arrays," *Commun. ACM* **15**(9), 827–837 (Sep 1972).
[21] Yankee, W. J., Giles, F. G., and Grimsley, D. L., *A Comparison Between Control Question and Relevant/Irrelevant Polygraph Test Formats in a Screening Situation*, MDA904-86-2191, A. Madley Corporation, Charlotte, N.C. (Sep 1987).

ACKNOWLEDGMENTS: We gratefully acknowledge the research staff at the Department of Defense Polygraph Institute at Ft. McClellan, Alabama, for collecting the mock-crime data, for supplying the CAPS parameters, and for providing technical support.

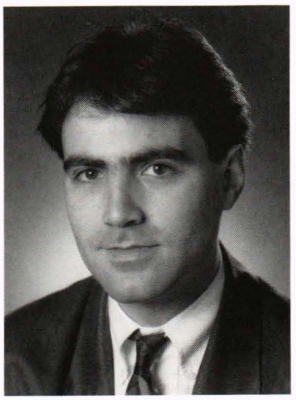
THE AUTHORS



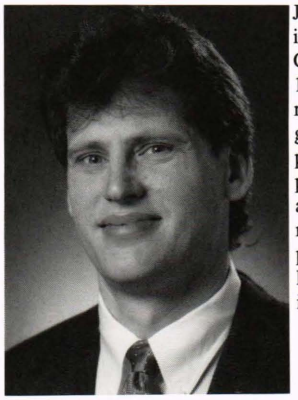
DALE E. OLSEN received his A.B. degree in mathematics from California State University, Chico, in 1965. After working as a research engineer for the Boeing Company for three years, he returned to school and earned an M.S. and Ph.D in statistics from Oregon State University. Dr. Olsen joined APL in 1973 and began developing and applying statistical methodology in the analysis of the reliability and accuracy of weapon systems. In 1988, he received Laboratory funding to support work in the analysis of electroencephalogram data. Government interest in this work resulted in funding for the polygraph scoring algorithm. Currently, Dr. Olsen is an Assistant Group Supervisor and a project manager at APL. He is also a member of the Principal Professional Staff.



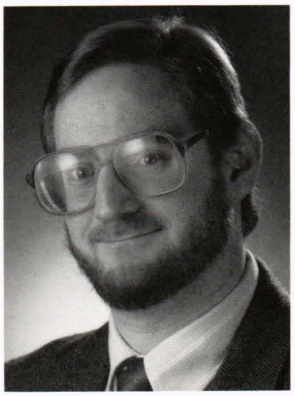
NORMAN ANSLEY received an A.B. degree from San Jose State College in 1950 and received his basic polygraph training at the Keeler Polygraph Institute in 1951. He also attended graduate school at Stanford University and The University of Maryland. Mr. Ansley is employed by the Department of Defense. He edits the quarterly journal *Polygraph* and is the author and editor of several books and more than fifty journal articles on the polygraph and other forms of lie detection. He is the recipient of the American Polygraph Association's John E. Reid Award for his contributions to polygraph research and teaching and the Leonarde Keeler Award for long and distinguished service to the polygraph profession.



IAN E. FELDBERG received a B.S. degree in electrical engineering and computer science from The Johns Hopkins University in 1984 and an M.S. in computer science from Johns Hopkins in 1987, specializing in computer graphics and computer vision. Since joining APL in 1984, he has contributed to a variety of software projects. In 1989, Mr. Feldberg joined the Computer Engineering Group of the Technical Services Department. Since then, he has worked primarily on a desktop system for designing application-specific integrated circuits and a system for digitizing paper polygraph charts for digital analysis.



JOHN C. HARRIS received a B.A. in applied statistics from the George Washington University in 1980. Since then, he has been a resident consultant with the Strategic Systems Department of APL, principally providing software support for the Trident I and Trident II accuracy analysis programs. In that role, he has developed several high-performance subroutine libraries for linear algebraic, database, and user interface functions.



JOHN A. CRISTION received a B.S. degree in electrical engineering from Drexel University in 1986 and is pursuing an M.S. degree in electrical engineering from The Johns Hopkins University G.W.C. Whiting School of Engineering. He is a member of the Associate Professional Staff at APL and works in the Signal Processing Group of the Strategic Systems Department. Mr. Cristion is currently working on a system for automatic polygraph scoring, an Independent Research and Development (IRAD) effort to develop an automated seizure detection algorithm, and an IRAD effort to study the use of neural networks in nonlinear adaptive controllers.