Dale E. Olsen,[1] Ph.D.; John C. Harris,[1] B.S.; Michael H. Capps,[2] M.S.; and Norman Ansley,[3] B.A.

Computerized Polygraph Scoring System

REFERENCE: Olsen DE, Harris JC, Capps MH, Ansley N. Computerized polygraph scoring system. J Forensic Sci 1997;42(1): 61–71.

ABSTRACT: Digitized polygraph data were collected during criminal investigations to develop a computerized algorithm for evaluating zone comparison polygraph examinations. The algorithm was incorporated in a software system and provides consistent and objective examination interpretation. The software system evaluates the data using methods that are fundamentally different than those used by examiners. It is currently in use in more than 40 of the United States and in other countries around the world.

KEYWORDS: forensic science, polygraph, algorithm, deception, zone comparison, discrimination, decision rule, logit, jackknife

Several different kinds of polygraph examinations are in common use today. Perhaps the most often used and the most controversial is the control question examination. This test procedure is used for criminal investigations and has many variations. A weakness of all polygraph procedures, including the control question examination, is that interpretations of the physiological data (chart evaluation) may vary. Although substantial differences in interpretation are uncommon, practitioners can and do differ in their conclusions. Evaluation of the physiological recordings has not seen a history of unified, rigorous scientific inquiry, in part because most of the current concepts appear to have come from the codified observations of experienced polygraph examiners.

Computerized statistical analyses of physiological data from polygraph examinations have been suggested by several investigators (1,2). In 1983, Kircher (3) reported the first computerized scoring system but relied on laboratory-based cases. Recently, an effort has been made to collect digitized polygraph data from law enforcement cases so that statistically optimum scoring criteria could be developed, implemented in software, and evaluated. The data were collected using a common variant of the control question examination known as the zone comparison method. The resulting computerized scoring system removes nearly all of the variation in chart interpretation. This paper describes the research done to develop that algorithm and presents the findings.

[1]Program manager and senior staff, respectively, The Johns Hopkins University, Applied Physics Laboratory, Johns Hopkins Rd., Laurel, MD 20723-6099.

[2]Director, US Department of Defense, Polygraph Institute, Fort McClellan, AL.

[3]Consultant, Forensic Research, Inc., Severna Park, MD 21146.

Received 24 July 1995; and in revised form 2 Feb. 1996 and 30 May 1996; accepted 3 June 1996.

Polygraph Background

Instrumentation

For the traditional polygraph instrument, three physiological measurements (4) are normally recorded: 1. Volumetric measures of the cardiovascular activity in the upper arm. A "cardio" cuff is placed on the arm over the brachial artery and inflated between the systolic and diastolic pressure for a measure of blood volume in the arm, together with the strength and rate of pulsation from the heart (5). 2. Respiratory measures of the expansion and contraction of the thoracic and abdominal areas. These measurements are most often taken using rubber tubes placed around the subject in an effort to obtain data closely related to the actual gaseous exchange involved in breathing (6). 3. Skin conductivity (or resistance) measures of electrodermal activity. Electrodermal activity is largely influenced by eccrine sweat gland activity and is recorded by measuring the conductance or resistance to an electrical current (6,7).

Question Sequences

A polygraph question sequence or format is an ordered combination of (1) relevant questions about the issue, (2) control questions that provide physiological responses for comparison, (3) neutral questions that provide a baseline of responsitivity to questions that are not relevant to the issue under investigation, and (4) other questions that are essential elements of some polygraph test formats. The collection of recorded responses to the set of questions is called a chart. Normally, two to five charts are collected so that questions are repeated several times.

All questions asked during a polygraph examination are reviewed and discussed with the examinee before the examination. When necessary, the questions are reworded to assure understanding, accommodate partial admissions, and present a dichotomy answerable with a definite "yes" or "no." Polygraph examiners may choose from several standard test formats; selection is based on test objectives, experience, and training.

Control Question Examinations

Most criminal investigation examinations are conducted using one of the possible formats of the control question method (3). Of these, the two most widely used formats are the zone comparison (9,10) and the modified general question test (MGQT) (11,12). Control question examinations contain several types of questions, which are asked in a fixed sequence. Each question series is repeated two to five times, and each series produces a separate chart. Examiners who want to cover only one issue generally choose a zone comparison sequence. This sequence permits questions on only one relevant issue, with variations in the wording

of relevant questions during the repetitions. The scoring system described in this paper is for the zone comparison examination.

Zone Comparison Examination

An example of a relevant question is "Did you steal any of the missing money from the safe?" A control question, in this case, might be about stealing in general, but not the theft at issue. An example is "Before you were employed at this bank, did you ever steal from an employer?" When evaluating charts using traditional methods, the physiological responses that occur as a result of the control and relevant questions are compared. There are also irrelevant questions that will normally 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. Peculiar to the zone comparison format of the control question method is the "sacrifice relevant" question, such as "In regard to the theft, do you intend to answer the questions truthfully?" It is called a sacrifice relevant because it is not evaluated and it serves as a buffer, being the first apparently relevant question asked in the series. The sacrifice relevant question introduces the relevant issue on each chart (13,14). Finally, this format contains symptomatic questions, such as "Are you completely convinced that I will not ask you a question on this test that has not already been reviewed?" (15).

Zone Comparison Chart Evaluation

The zone comparison technique is a primary technique taught in polygraph training. This format is designed to be scored numerically for the purpose of making a diagnosis of truth or deception. Relevant question reactions are compared to nearby control question reactions, and a numerical score is given to each physiological measure for each relevant question. If the relevant response is significantly greater than the nearby control question response, a negative score is assigned to that relevant response; if the nearby control response is significantly greater than the relevant response, a positive score is assigned to that relevant question. Normally, in comparing an electrodermal control reaction to relevant reaction, a minus one point is assigned if the relevant reaction is somewhat greater in amplitude than that of the control. If the relevant electrodermal reaction is two to three times greater in amplitude than the control reaction, the relevant reactions may be assigned minus three points (see Fig. 1). If the reactions are about the same, no points are assigned and positive scores are assigned when the control question responses are greater in amplitude. Examiners will differ on how they award points. Some examiners will never assign more than one point to a component reaction whereas others will assign as many as three points. Similar methods are used to assign points to respiration and cardiovascular components. The scores from all relevant reactions are added together and compared with a threshold to determine whether the results are inconclusive, indicate deception, or indicate no deception (16).

A General Description of the Algorithm Development Approach

The purpose of our research was to learn how to effectively distinguish the reactions of those who are attempting deception to the relevant question from the reactions of those who are not. The first requirement was to collect data from a target population suspected of criminal activity. To do this, we narrowed our research to the zone comparison examination, but used a variety of examiners and case types. Because our interest was in real criminal cases, mock crime data were not used.

We used the computer first to calibrate and condition the data and then to calculate thousands of different characterizations of the reactions. Figure 1 provides a sample of computerized polygraph tracings showing responses to two questions. The first vertical line on the left of Fig. 1 indicates the time of the start of the question; the second indicates the end of the question; and the third indicates the time of the verbal response. The first reaction (labeled C4) is a response to a control question; the second reaction (labeled R5) is that of a deceptive subject to a relevant question. The respiration tracings are the top two tracings and show the expansion and contraction of the upper chest and the abdominal area as the subject inhales and exhales. The diminished cycles during the relevant question are called suppression and are believed to occur during attempts at deception. The third tracing from the top shows the electrodermal response; the large and rapidly increasing response (large range) in the tracing after the relevant question (R5) is believed to indicate deception. The cardio channel at the bottom of Fig. 1 shows both the pulse and the change in blood volume in the arm. The rapid increase in blood volume after the second question (R5) and the narrowing of the pulse amplitude are believed to be associated with deception. These possible indicators of deception, called features, were characterized numerically.

Digital data can be processed in many different ways to facilitate interpretation. There are many ways to characterize the reactions, assign a numeric value to those characterizations, and combine assigned values to reach a conclusion. Our aim was to determine which processing steps and methods of characterizing reactions could be used to determine most effectively which subjects were deceptive and which were not. We sought a set of features that would allow us to separate deceptive and nondeceptive subjects as clearly as possible.

Figure 2 is a two-feature scatter plot of measured electrodermal and cardiovascular responses. Circles represent reactions of subjects identified as deceptive (Deception Indicated or DI) and squares represent reactions identified as nondeceptive (No Deception Indicated or NDI). This figure shows separation of DI and NDI subjects by the decision line. Ideally, the DI and the NDI sets would not overlap, and a clear distinction could be made between the two. However, the two data sets do overlap, indicating that we cannot cleanly separate the two sets using just these two features or characterizations alone. By using other features that help to detect deception, we can better separate the two populations. For example, a subject incorrectly classified as NDI using just the two plotted features may be clearly in the DI portion of a plot when a respiration feature is used.

Algorithm Development

To learn how to separate best the DI and NDI data, we began by collecting data that are representative of the polygraph applications of interest.

Digitized Data Collection

In 1989, when data collection for this project began, the only computerized data collection systems available were the CODAS system (DATAQ Instruments, Inc., Akron, OH), the Computerized Polygraph System (CPS; University of Utah), and the Axciton System (Houston, TX). The CODAS system is a general-purpose

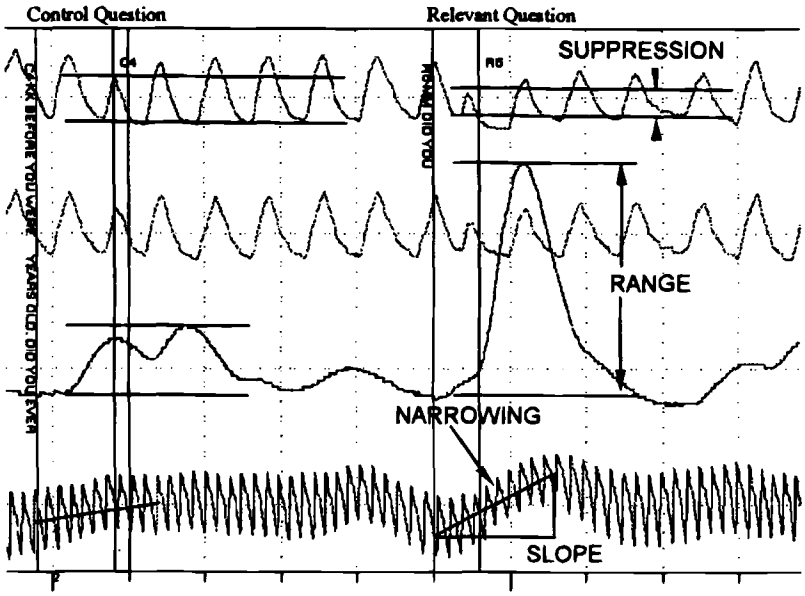

FIG. 1—*Control and relevant reactions for a deceptive subject.*

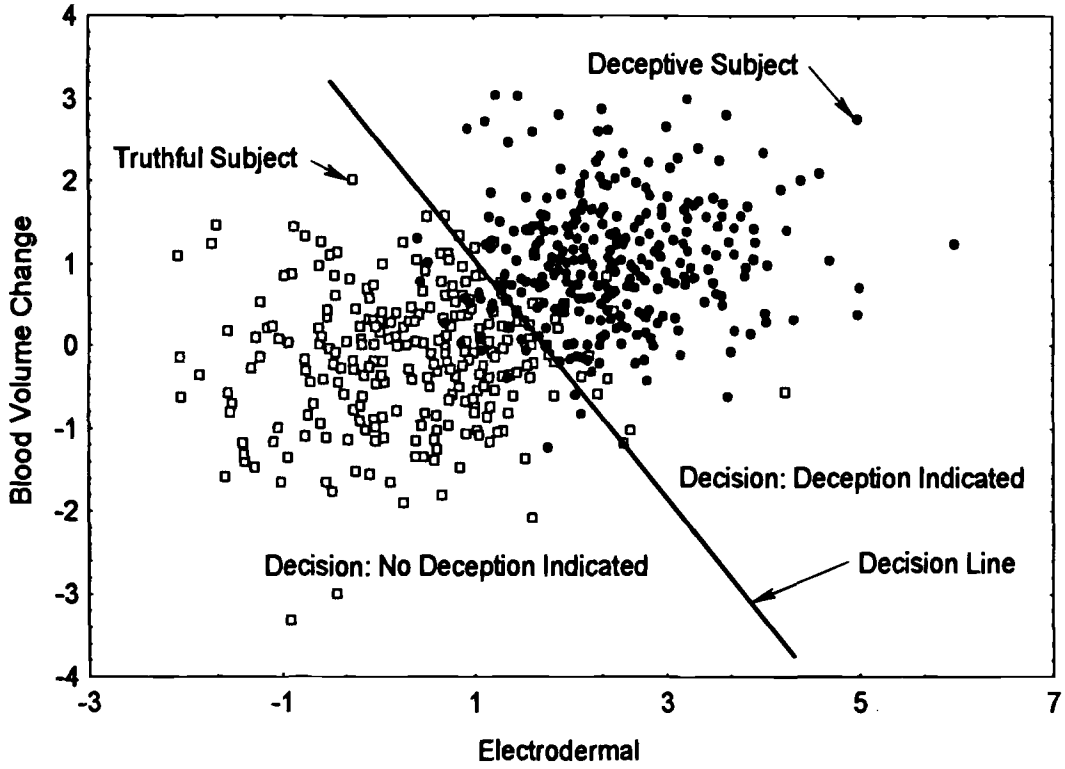

FIG. 2—*Scatter plot of electrodermal and blood volume change features.*

data collection system and is not as well-suited for standard polygraph examinations. The early CPS stored only 25 s of data for each question and did not store pulse information. This system included the first true automated scoring system (2,16).

The only computerized data collection system available at the time that met our requirements, that is, being user-friendly and providing all of the necessary data, was the Axciton System. It contained a system for the rank-order scoring of the responses, but the scoring system did not produce a probability of deception or suggest that the subject is either deceptive or not deceptive.

Field Examinations

Axciton Systems were purchased and placed with two federal, two state, and two county law enforcement agencies and one city agency. Ten examiners provided most of the cases, but others also contributed. The examiners were located in the eastern United States from Vermont to Florida. All data collected were saved and represented a reasonable mix of crimes including larceny, murder, witness statement verification, forgery, arson, assault, bribery, child molestation, incest, patient abuse, kidnapping, and drug violations.

A few preemployment screening cases were also used. (The zone comparison question sequence is not used as a primary test for preemployment screening, but is sometimes used to investigate further a single issue identified through other testing.) All examiners had experience with the zone comparison question sequence and were willing to take an additional week of training in the use of the computerized instruments at the Department of Defense Polygraph Institute.

The Data

Because the purpose of our effort was not to determine the accuracy of polygraph examinations but rather to determine how polygraph data can best be used to separate deceptive and nondeceptive sets of reactions, we did not have to use every case received to develop the algorithm. A random sample of cases is desirable but not necessary for learning how to discriminate reactions. The only requirement was that the subject's status (deceptive or nondeceptive) either be known with certainty or that the polygraph results be clear to experienced examiners. From these data, we could then identify traits that could be used to discriminate the subjects. By 31 March 1994, we had received data from 852 cases. The deceptive or nondeceptive status of the subject was determined in one of two ways: (1) a confession or guilty plea by the subject or someone else was obtained, which thereby cleared the subject; or (2) an agreement was obtained among the original and two blind examiners on the interpretation of the charts.

Of the 852 cases, 228 (27%) were discarded. The most common reason for not using the data (37% of the 228 cases) was that two or more examiners called the charts inconclusive (INC) (see Table 1). In 34% of the unused cases, a consensus could not be reached. For example, one examiner may have scored the charts as INC, whereas the others scored them as DI. In a few of these cases, all three examiners made different calls—NDI, DI, and INC. We used cases that one or more examiners called INC but that were eventually confirmed.

Initial work indicated that respiration reactions could last longer than 18 s, so examiners were required to wait at least 25 s between questions. However, data with as few as 18 s between questions were used to develop the algorithm.

Data collection began in April 1991. Just under half of the 228 discarded cases had been collected in the first four months when examiners were using computerized polygraph. During that period, examiners were becoming familiar with the computerized collection systems, and the Axciton hardware and software were being improved. The current version of the algorithm was implemented in software (PolyScore® version 3.0) and was developed with 624 cases; of those, 301 were nondeceptive cases and 323 were deceptive cases.

TABLE 1—*Disposition of cases not used in algorithm development.*

Reason for Discarding Cases	Number of Cases	Unused Cases (%)	All Cases (%)
Inconclusive for two or more examiners	85	37	10
No consensus	77	34	9
Excessive movements	19	8	2
Instrumentation malfunction or no electrodermal responses	17	8	2
Less than 18 s between questions	7	3	0.8
Improper question sequence	7	3	0.8
Countermeasures	6	3	0.7
Only one or two poor-quality charts	5	2	0.6
Examination discontinued before complete	3	1	0.3
No respiration channel	1	0.5	0.1
Unfit subject	1	0.5	0.1
Total	228	100	26.4%

The number of charts for each subject varied from two to five, with three charts for most subjects. Typically, there were three relevant question responses on each of three charts producing nine relevant question responses for each subject, in addition to nine control question responses.

Extreme control or relevant responses are classified as outliers and were therefore not used to develop the algorithm. (To be classified as extreme, the one reaction must have accounted for more than 89% of the variability among the 18 responses.) Reactions are sometimes distorted by movements of the subject or other events. These distorted reactions are called artifacts. We therefore built an algorithm to detect artifacts so that components of the reactions classified as artifacts could be eliminated. Other than dropping these artifacts and outliers, no data editing was done.

Processing Steps

The algorithm was developed using seven basic steps: (Similar procedures have also been used to develop a seizure detection algorithm and can be used to develop other detection, discrimination, and image matching systems.) 1. *Condition the data*—The digitized polygraph data or signals are transformed into other useful signals. 2. *Standardize the data*—The signals are scaled so that the various tracings have similar amplitude scales. 3. *Develop data features*—Many different characterizations of each question response are calculated. These quantified descriptions of the responses are called features. 4. Standardize features—The features for all relevant questions are standardized by subtracting the average of control question responses and dividing by a standard deviation developed using the features from both control and relevant questions. 5. Evaluate features—The features are evaluated to determine which combination can best be used to detect deception. 6. Develop the decision rule—A logistic regression decision rule (18,19) converts these features into a probability of deception. 7. Evaluate the algorithm—The algorithm-produced probabilities are compared with verified results and examiner decisions. These seven steps will now be discussed in more detail.

Condition the Data

As part of the first step, the algorithm transforms the signals to separate those portions that contain the information to be characterized. For example, the cardio signal (bottom tracing in Fig. 3) contains both a high-frequency component corresponding to pulse and a low-frequency component corresponding to overall blood volume. To make feature extraction easier, these signals are split into their frequency parts using a digital filter. The signal is passed through a finite impulse response filter to divide the signal at 0.25 Hz, which produces the two signals shown in Fig. 4. Although the scales are somewhat different, one can see that the blood volume signal overlies the middle of the original cardio signal and that the pulse signal is the movement about this middle. Thus, pulse and blood volume can now be characterized independently. Other transformations produce a series of other new channels of data.

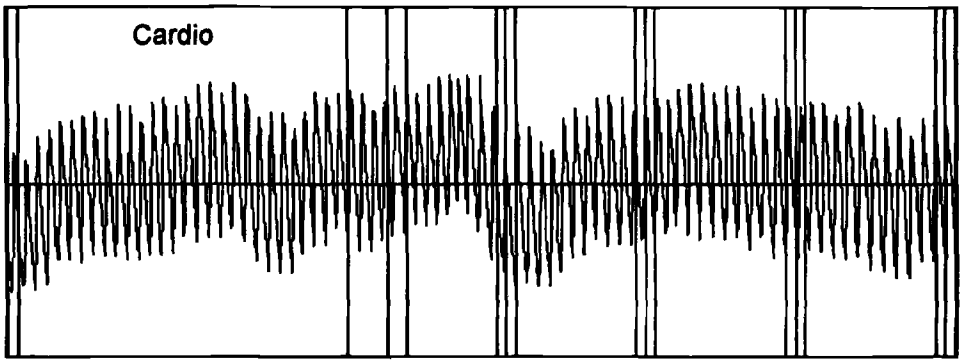

FIG. 3—*A typical cardio signal.*

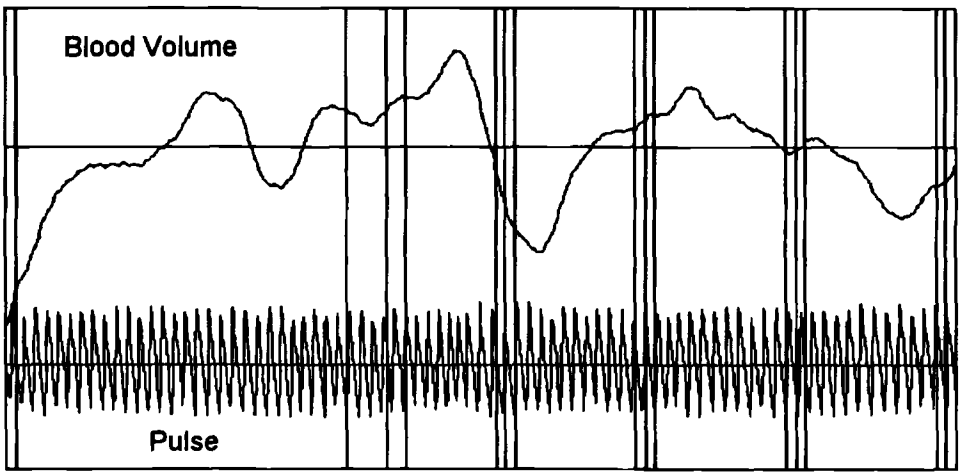

FIG. 4—*A cardio signal split into high-frequency and low-frequency components.*

Detrending is a technique used for removing gradual signal changes or trends unrelated to a particular question. For example, a downward trend in the electrodermal signal may be caused by changes in the subject's physiology that are not related to a particular questions. Trends often require the centering adjustments (see Fig. 5) that examiners make during the examination. Detrending is thus important for developing the best signals for processing by the algorithm and also makes it easier to accurately evaluate charts using traditional scoring. Figure 5 shows a sample of standard polygraph data and Fig. 6 shows the same data after detrending. The electrodermal signal (shown as the center tracing) is much easier to evaluate visually after detrending. (This segment of data contains distortions and should not be used for scoring.)

Standardize the Data

Signal standardization allows the amplitude measurements from different subjects or different charts from one subject to be evaluated using a common algorithm. The idea is to calibrate all instruments and people as nearly as possible to the same level. Typically, the mean of the signal is subtracted from each signal data point, and the difference is divided by the standard deviation. However, that method produces poor results when the data are not symmetric about some point. Because polygraph data are not symmetric, the extreme values (large or small) are not used when scaling the data.

Standardization affects the appearance of a channel. When data are collected using conventional analog equipment, reactions are often missed. Even with the gain turned up, the electrodermal tracing sometimes appears to contain only a few small, seemingly unimportant responses, similar to the bottom tracing in Fig. 7. Compare the two electrodermal tracings shown in Fig. 7. The tracings are plots of the same data, except one is placed on a different scale. Many examiners would identify significant reactions in the top tracing but would not attribute any significance to the bottom one. This problem commonly occurs when analog equipment produces relatively flat electrodermal signals. It cannot be rectified after the charts are created. When using digital equipment, the standardization step scales the signals so that the reactions in the lower tracing will be obvious, and these two identical signals will appear to be identical.

Develop Features

Once the scaled signals have been stored, they can be characterized numerically in many different ways. These characterizations (e.g., slope, area under the curve, and length of the line) are called features. We developed hundreds of features and evaluated them using different windows of data to determine the most effective intervals for characterizing a response. For example, to find the best response interval or window to characterize the electrodermal range (maximum-minimum) data (see Fig. 1), an initial guess of the beginning of the interval was set to 1 s, and various interval endings from 8 to 15 s were evaluated. Then various times for the start of the interval were evaluated. The resulting response intervals

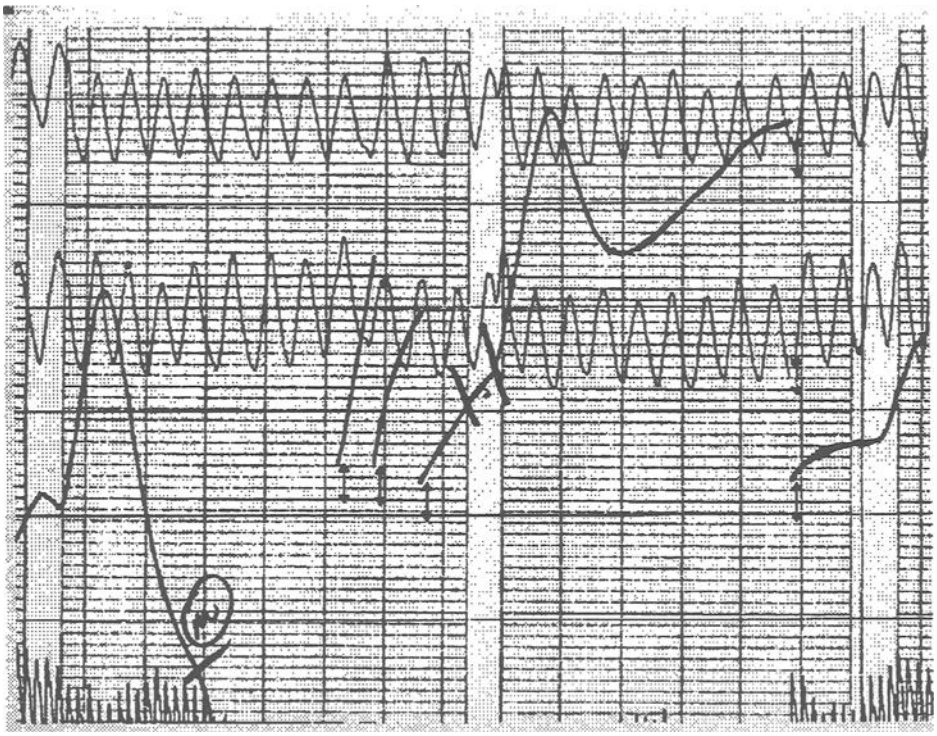

FIG. 5—*Sample of detrended polygraph signals, which are difficult to evaluate.*

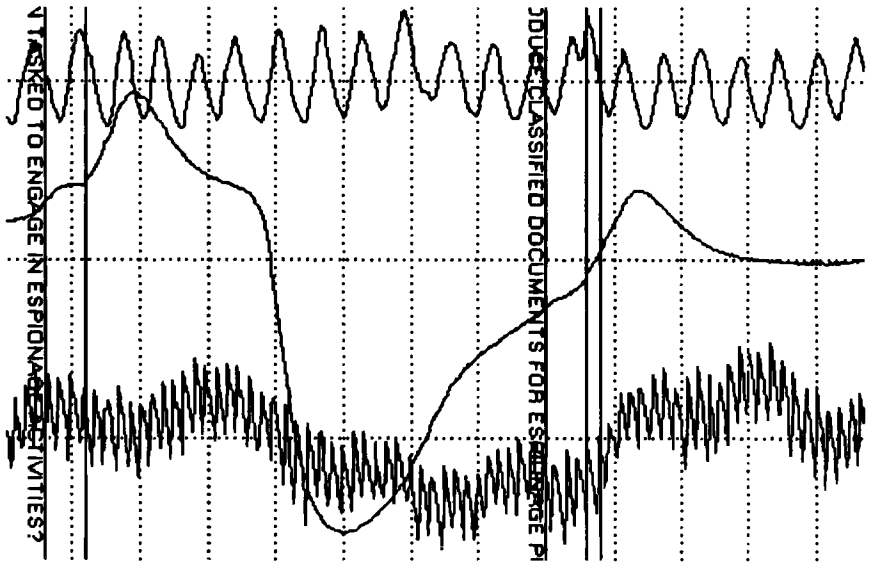

FIG. 6—*Sample of the same signals after detrending.*

FIG. 7—*Identical electrodermal reactions on different scales. The vertical lines indicate the beginning of the question.*

produced thousands of features to be evaluated. The optimum windows used by the algorithm depend on the feature and the processing.

Standardized Features

The next step in selecting the best group of features for the discrimination rule is feature standardization. Like signal standardization, feature standardization is intended to calibrate different subjects to the same scale. In addition, this step uses responses to the control question to determine a response standard of comparison.

Both control and relevant questions were used to determine the natural variability of the subject's responses. Standardization was achieved by first calculating the mean of the control responses for the feature and the pooled control and relevant standard deviation. The pooled standard deviation was calculated from the variability of the control questions about their mean and relevant questions about their mean. The standardization was accomplished using the equation:

$$R_i' = \frac{R_i - \mu_C}{S_{CR}}$$

where

R_i' is the ith standardized relevant question feature,
R_i is the ith relevant question feature,
μ_C is the mean of the control features,

$$S_{CR}^2 = \frac{\sum(R_i - \mu_R)^2 + \sum(C_i - \mu_C)^2}{\text{(number of questions-2)}}$$ is the pooled variance,

μ_R is the mean of the relevant features, and
C_i is the ith control question feature.

There is another important difference between the traditional scoring and the algorithm evaluation: Although as traditional scoring compares the relevant question reaction to nearby control question reactions, we compare each relevant reaction to a standard created using both control and relevant question reactions. This comparison was made possible by the standardization of the data from different charts and the use of the pooled variance to learn how a subject would react. If these preprocessing steps are not taken, information from control questions from different charts cannot be used for evaluating the relevant reactions.

Evaluate Features

Features were evaluated by testing them in a logistic regression-produced decision rule (18,19). The use of this process ensures that the features are selected in a manner consistent with the properties of the decision rule and enables the development of an effective algorithm based on a minimum number of features.

Ten features are used in version 3.0 of the PolyScore® model. Because of the way the algorithm was developed, this combination of 10 features is particularly useful in separating DI and NDI data. Many features not used by this algorithm, including those relating to traditional scoring criteria, are valuable for discrimination by themselves but do not improve the algorithm. For example, during the respiration baseline rises shown in Fig. 8, the subject is keeping more air in the lungs. Examiners would treat this change in baseline as an important reaction. The algorithm does not use this change in respiration directly as a scoring feature. The essential information in the respiration signal is contained in other features, none of which is associated directly with a change in respiration baseline. Just the opposite is true for a few of the 10 features used by PolyScore®. They are not particularly effective by themselves in detecting deception, but do, in combination with other features, help to separate the two groups of subjects.

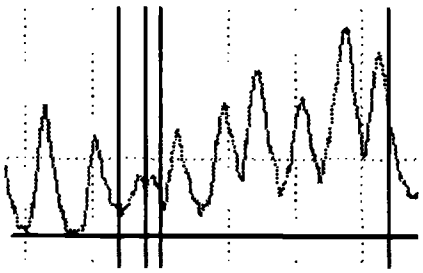

FIG. 8—*Sample of respiration change in baseline.*

Traditional chart interpretation assigns a value for each identified scoring criterion and adds those values together, often without regard to what other criteria appear. The approach of evaluating features in combination is fundamentally different than that of traditional chart interpretation.

The features used by the PolyScore® algorithm require a digitized signal and a computer to calculate. These features are complex, and the data must be extensively processed; they can, however, be related to traditional scoring criteria. Three features are used to characterize electrodermal reactions and typically contribute the most to a decision. Relatively large and long reactions, as shown in Fig. 9, indicate significant responses. A rapid increase in blood volume in the arm, as shown in Fig. 10, is also an indication of a significant reaction and is characterized using three features. Pulse and respiration are each characterized using two features. When the amplitude of the subject's pulse decreases (stroke length), as shown in Fig. 10, a significant reaction has occurred. Deceptive subjects are also more likely to suppress their breathing when responding to a relevant question (as shown in Fig. 11), whereas innocent subjects are more likely to suppress breathing when responding to a control question.

Develop the Decision Rule

We used a logistic regression model to create a decision rule having two parts. The first step is to linearly weight combinations

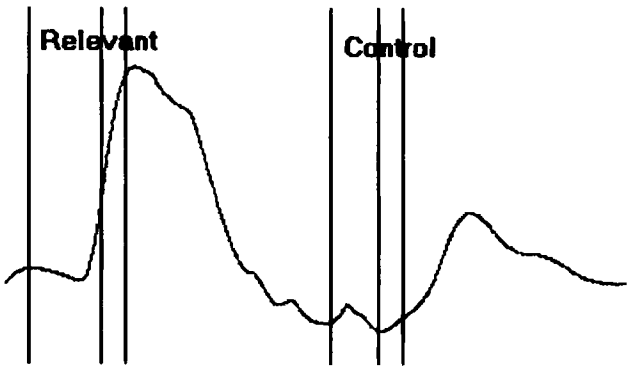

FIG. 9—*Electrodermal responses.*

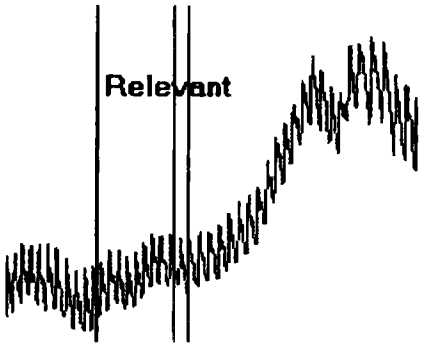

FIG. 10—*Cardio response.*

of the features to produce a "score." A statistical technique called maximum likelihood was used to obtain the optimal weights (β_is) of the features to form a score given by:

$$\text{Score} = \text{Intercept} + \beta_1 \cdot \text{feature}(1) + \ldots + \beta_{10} \cdot \text{feature}(10)$$

The score was developed using methods that allowed it to be converted to a probability of deception. The second part of the rule uses the "logit" conversion function to calculate a probability:

$$\text{Probability of deception} = \frac{e^{\text{Score}}}{1 + e^{\text{Score}}}$$

The probability of deception depends on the measured features of the relevant responses standardized using the control question responses. An example of the relationship between an electrodermal feature and the probability of deception is provided in Fig. 12.

This logit conversion function models many naturally occurring phenomena and is widely used. It produces a number between 0 and 1 that reflects the probability that the given set of reactions comes from a subject attempting deception. Unlike some methods, when the model is valid, the logistic regression rule produces numbers that are valid probabilities rather than just a number between 0 and 1. If a measure from the examination is the number 0.95, the correct interpretation is that, based on the data in the database, 95% of the time when similar features are present, deception has been attempted.

Algorithm Evaluation

Our intent in developing the computerized scoring algorithm was to produce a system that could objectively and consistently evaluate polygraph examination data. However, the system would be worthless if it did not produce results that by some measure were accurate. Because we were working with criminal cases, sampling problems and our inability to establish the subject's status (deceptive, nondeceptive) with certainty made it impossible to determine quantitatively the accuracy of the system. Even if it were possible to address these two problems, accuracy would still depend on examiner skill. We were, nevertheless, able to determine how well PolyScore® performs relative to our database.

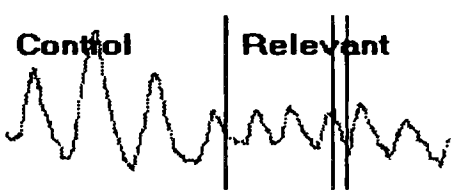

FIG. 11—*Respiration suppression.*

First, the algorithm was used to score the cases from which it was built. Because it used only 10 features, it could not memorize the 624 cases; however, it is not expected to perform as well on an independent data set. Because the algorithm produces a probability of deception, it will, on occasions, produce probabilities near 0.5. Any value between 0.10 and 0.90 is identified as INC. Scores at or above 0.90 are interpreted as DI; scores at or below 0.10 are interpreted as NDI. The algorithm scored 8% (50) of the cases as INC and agreed with "known-truth" or examiner decision in 99.8% (573) of the remaining cases. It disagreed with the examiners in one case. As shown in Fig. 13, most of the cases were scored with probabilities less than 0.01 or greater than 0.99.

A jackknife procedure was used to drop one subject from the database, to refit a model with the same features, and then to classify the subject that was dropped from the data set using the modified algorithm. This process was repeated for all 624 subjects. Table 2 compares the jackknife results and those obtained from scoring the data using version 3.0 of the algorithm.

Conclusion 1—Because subject status, for many of the cases in the database, was determined by examiner decision, these results demonstrate that the algorithm is consistent with experienced examiner decisions.

Of the 624 cases, just over one third or 218 were confirmed by confessions. Most of these subjects were deceptive: only 21% (45) of the 218 subjects were not attempting deception. For these confirmed cases, 25 were scored as INC, producing a rate of 11%. All of the other cases (193 subjects) were scored correctly.

The confirmed cases did not represent a random sample of criminal cases. Because many of these cases were confirmed by a confession given to the examiner shortly after the examination, confirmation of the result depended on the examiner's decision. In many cases, the examiner decision was made based on a confession obtained within a day or two of the examination. This selection of confirmed cases introduces a sampling bias in favor of polygraph. It is also true that for the 45 truthful subjects, no errors were made, and for these subjects, confirmation was made through a confession of another person combined with supporting evidence. Although none of these subjects was incorrectly scored, we cannot use these results to estimate the accuracy of the computerized system when applied to other data. However, there were 218 chances for the algorithm to make an error, and it made none.

Conclusion 2—The results indicate that an algorithm has been built that can separate the confirmed truthful and deceptive subjects in this database.

The computerized evaluation system differs fundamentally from systems used by examiners. The two methods of chart evaluation use different scoring criteria and weight the channels differently. To determine if the computerized algorithm mimics examiners, we identified eight cases that were scored as INC by the original examiner and were later confirmed. Six of those subjects were deceptive and two were not. To make this evaluation independent of the training data, these eight examinations were dropped from the database and the model was refit using the same features. For these difficult cases, the refit algorithm probabilities and the PolyScore probabilities are shown in Table 3. A clear decision was made in seven of the eight cases, and all of those decisions were correct. For one case, the result was inconclusive. Thus, the rules used by examiners and the rules used by the algorithm can

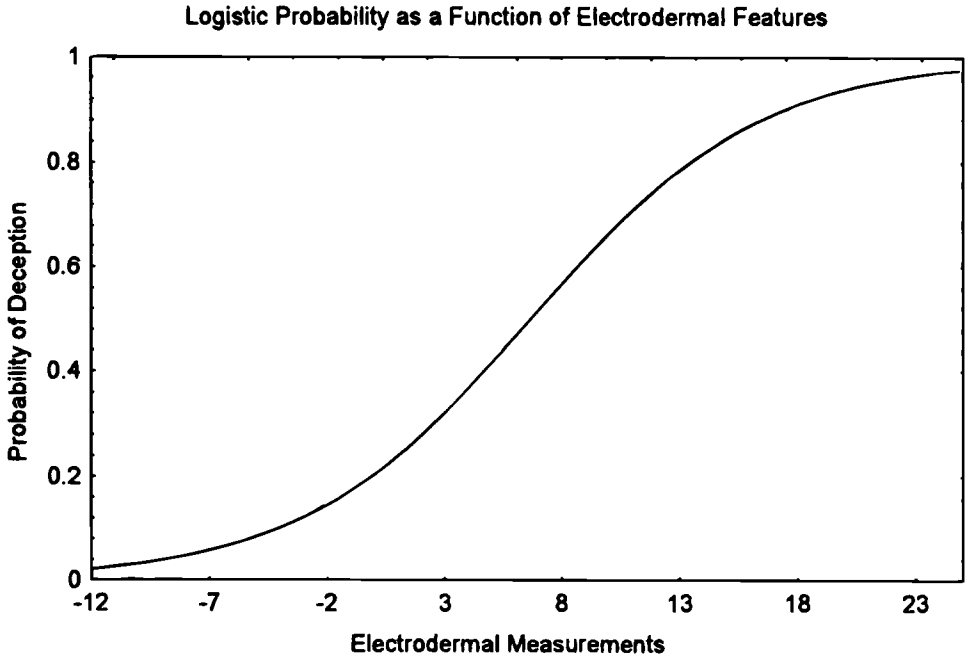

FIG. 12—*The curved line is the plot of the probability of deception as a function of an electrodermal feature.*

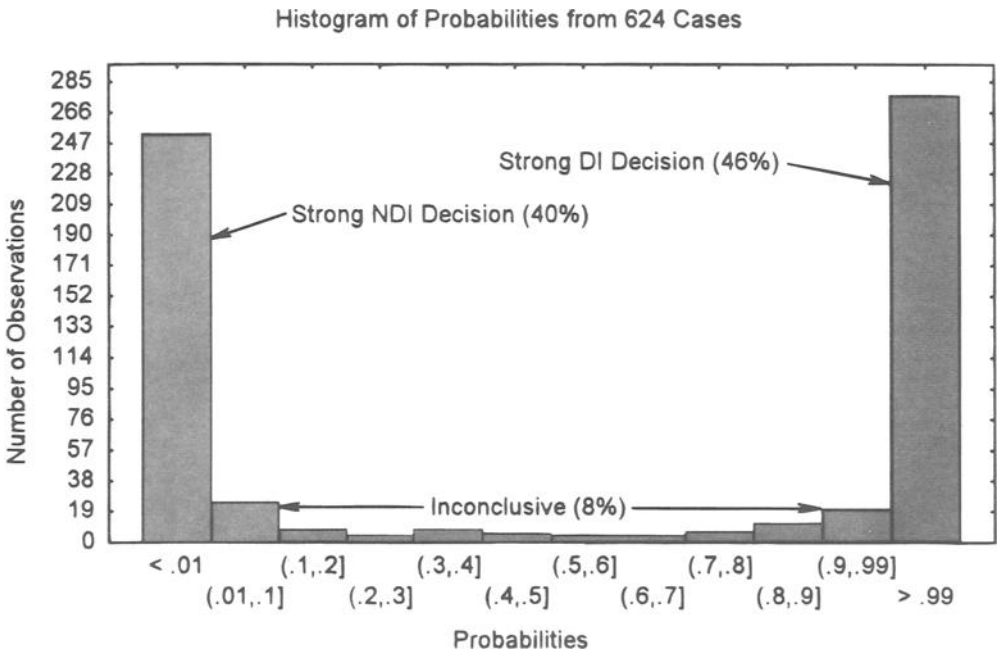

FIG. 13—*Histogram of probabilities generated using the zone comparison algorithm.*

TABLE 2—*A comparison of jackknife and version 3.0 scoring.*

	Jackknife	Version 3.0
Correct	570	573
Inconclusive	52	50
Incorrect	2	1

lead to a different set of inconclusive cases implying that the algorithm is evaluating the charts differently.

Conclusion 3—The algorithm does not mimic traditional chart interpretation.

Because the algorithm and traditional methods process data in fundamentally different ways, we were able to reach the third conclusion without empirical support, but it is reassuring that the algorithm did properly evaluate these difficult cases. This result also suggests that the algorithm could be effective in reducing the number of inconclusive decisions by law enforcement examiners.

The South Carolina Law Enforcement Division (SLED) uses polygraph examinations carefully. Every chart is reviewed by an

TABLE 3—*Polyscore and refit results for subjects scored inconclusive by the examiner and later confirmed.*

Actual Subject Status	Refit Algorithm	PolyScore®	Decision
Innocent	0.009	0.009	NDI
Innocent	0.025	0.026	NDI
Guilty	0.273	0.326	INC
Guilty	0.957	0.961	DI
Guilty	0.999	0.999	DI
Guilty	1.000	1.000	DI
Guilty	1.000	1.000	DI
Guilty	1.000	1.000	DI

TABLE 4—*South Carolina Law Enforcement Division Examination results.*

	1992 Traditional		1993 Version 2.1		1993 Version 2.3	
	Number	%	Number	%	Number	%
DI	181	37	132	52	100	48
NDI	214	43	96	38	93	44
INC	100	20	26	10	16	8
Total	495	100	254	100	209	100

independent examiner and all scoring must be carefully justified. Subjects scored as NDI or INC are not interrogated. Careful records are kept allowing us to compare results both with and without PolyScore.

During 1992, the SLED used only conventional equipment. Version 2.1 of PolyScore® was used from Feb. 1, 1993 until June 30, 1993. This software contained an early version of the algorithm and was used to acceptance-test the system. Results from using version 2.3 of the algorithm were collected and tabulated from July 1, 1993 until Dec. 31, 1993. All the polygraph examinations were conducted by the same two examiners throughout the various periods. The results (Table 4) show a dramatic drop in the inconclusive rate.

Conclusion 4—The use of the algorithm can resolve some inconclusive cases and reduce the overall inconclusive rate for some law enforcement agencies.

Remarks

We emphasize that PolyScore® can only effectively evaluate properly collected data. If a subject is not a suitable candidate for a polygraph examination, if the issue is not clearly defined or of sufficient intensity, or if proper zone comparison procedures are not used, accuracy will suffer.

Version 3.0 of PolyScore® is now available from two vendors of computerized polygraph equipment. The software is being used in more than 40 of the United States and in countries around the world. In addition to the zone comparison algorithm, the PolyScore® software includes the capability to score other control question examinations, the ability to rank-order question responses, and the ability to re-scale and display data in a way more suitable for traditional scoring. Work is continuing on developing algorithms for other question sequences.

References

1. Kubis JF. Studies on lie detection: Computer feasibility considerations. Tech. Rep. 62-205, prepared for Air Force Systems Command, Contract No. AF30 (602)-2270, Project 5534. Bronx, NY: Fordham University, 1962.
2. Kubis JF. Analysis of polygraph data: Dependent and independent situations. Polygraph 1973;2:42–58.
3. Kircher JC. Computerized decision-making and patterns of activation in the detection of deception [dissertation]. Salt Lake City, The University of Utah, 1983.
4. Abrams S. The Complete Polygraph Handbook. Lexington Books, Lexington, MA, 1989.
5. Geddes LA, Newberg DC. Cuff pressure oscillations in the measurement of relative blood pressure. Psychophysiology 1977;14: 198–202.
6. Jacobs JE. The feasibility of alternate physiological sensors as applicable to polygraph techniques. In: Ansley N, editor Legal admissibility of the polygraph. Springfield, IL: Charles C Thomas, 1975;266–72.
7. Summers WG. Science can get the confession. Fordham Law Rev, 1939;5:334–54.
8. Prokasy WF, Raskin DC. Electrodermal activity in psychological research, Academic Press, New York, 1973.
9. Backster C. Technique, tips and polygraph chart interpretation. Newsletter of the Academy for Scientific Interrogation 1962;4–6.
10. Backster C. New standards in polygraph chart interpretation—Do the charts speak for themselves? Law and Order 1963;11:67–71.
11. Weaver RS. The numerical evaluation of polygraph charts: Evolution and comparison of three major systems. Polygraph 1980;9:94–108.
12. Reid JE. A revised questioning technique in lie detection tests. J Crim. Law Criminol. 1947;37:542–7.
13. Capps MH. Predictive value of the sacrifice relevant. Polygraph 1991;20:1–6.
14. Horvath F. The value and effectiveness of the sacrifice relevant question: An empirical assessment. Polygraph 1994;23(4):261–79.
15. Capps MH, Knill BL, Evans RK. Effectiveness of the symptomatic question. Polygraph 1993;22:285–98.
16. Capps MH, Ansley N. Numerical scoring of polygraph charts: What examiners really do. Polygraph 1992;21:264–320.
17. Kircher JC, Raskin DC. Human versus computerized evaluations of polygraph data in a laboratory setting. J. Appli. Psychol. 1988;73:291–302.
18. Cox DR, Snell EJ. Analysis of binary data. Chapman & Hall, London, 1989.
19. Hosmer DW, Lemeshow S. Applied logistic regression Wiley, New York, 1989.

Additional information and reprint requests:
Dale E. Olsen, Ph.D.
Johns Hopkins University, Applied Physics Laboratory
Johns Hopkins Rd.
Laurel, MD 20723-6099