(2)

Bootstrap Decision Making for Polygraph Examinations

Final Report of DOD/PERSEREC

Grant No. N00014-92-J-1794

Charles R. Honts, Ph. D.
Principle Investigator
and
Mary K. Devitt
Research Assistant

24 August 1992

University of North Dakota
Psychology Department
7187 University Station
Grand Forks, ND 58202

Bootstrap Decision Making for Polygraph Examinations

Final Report of DOD/PERSEREC

Grant No. N00014-92-J-1794

Charles R. Honts, Ph. D.
Principle Investigator
and
Mary K. Devitt
Research Assistant

24 August 1992

University of North Dakota
Psychology Department
7187 University Station
Grand Forks, ND 58202

Table of Contents

Executive Summary **4**

Background **5**

Polygraph Tests in the National Security Systems 5

Weakness in Current Polygraph Practices 5

Statistical Decision Making, A Possible Solution 6

Possible Weaknesses of the Discriminant Analysis Approach 7

Bootstrapping: A New Approach to Statistical Decision Making **10**

A Conceptual Introduction to Bootstrapping 10

Our Approach To Bootstrapping a Polygraph Decision 11

Method: Phase I **13**

Software For The Project 13

Hardware for The Project 13

The Project Data Base 13

Procedure 13

Results & Discussion: Phase I 16

Phase II: A Second Bootstrap **19**

Method: Phase II 19

Results and Discussion: Phase II 19

UND
UNIVERSITY
OF
NORTH
DAKOTA

Phase III: An Exploration of Inconclusive Zones 21

Method 21

Results and Discussion of Phase III 22

General Discussion 28

Follow-Up Research Projects 32

References 34

Executive Summary

We examined human numerical evaluation, discriminant analysis, and a bootstrap approach to decision making in the psychophysiological detection of deception with the control question test. The data for these analyses were obtained from the Utah Cooperative Working Group Database and consisted of 100 innocent and 100 guilty subjects of mock crime experiments. We found statistically equivalent performance for the three approaches. However, it should be noted that the human evaluators used in this study were not representative of the average polygraph examiner, and the human evaluation data reported in this study are likely to have substantially overestimated the accuracy of human numerical evaluation in the field. Taken in that context, the performance of the statistical classifiers should be viewed very favorably. In absolute terms, the bootstrap approach outperformed the other two approaches. As compared to discriminant analysis, the bootstrap has much to recommend it. It avoids the restrictive mathematical assumptions of discriminant analysis, and since it is not tied to any empirical standardization sample, the bootstrap is likely to be widely generalizable. It was concluded that statistical decision making has come of age in the detection of deception and should see universal application in the field in the near future.

UND
UNIVERSITY
OF
NORTH
DAKOTA

Background

Polygraph Tests in the National Security Systems

Polygraph examinations play an important part in the Personnel Security Programs of the Department of Defense and other United States government agencies (Department of Defense [DOD], 1991; Honts, 1991a). Virtually all federal agencies concerned with national security and law enforcement employ a staff of polygraph examiners. Many, if not most, of those agencies use polygraph tests for personnel security screening, both before hiring and during employment. In the Counterintelligence Scope Polygraph Program alone, the Department of Defense ran 17,443 such tests in fiscal year 1990. With so many individuals being affected, and with such important national security decisions being made, it is critical that such polygraph tests be as accurate as possible. Unfortunately, recent research (Barland, Honts, & Barger, *1989*; Honts, *1989*, *1991a*, *1991b*, 1992) suggests that the accuracy of security screening polygraph tests is well below that of federal law enforcement applications (for example see Raskin, Kircher, Honts, & Horowitz, 1988).

Weakness in Current Polygraph Practices

One area where most federal polygraph examinations appear to be weak is in their diagnostic methods. That is, they appear weak in the analysis of polygraph data (charts), and in the decision making that follows data analysis. When a control question test is used, the current state of practice in federal polygraph data analysis has the examiner evaluate the data by applying a semi-objective system of 29 criteria to each relevant and control question pair. The examiner derives a total numerical score for each question and applies a relatively complex decision rule for deciding the outcome of the examination (Department of Defense Polygraph Institute [DODPI], 1990).

However, this approach to diagnostic decision making in the detection of deception has been criticized. In particular, the 29 semi-objective criteria taught at the Department of Defense Polygraph Institute have been criticized because they lack empirical support and because

some of them violate basic psychophysiological principles (Honts & Perry, 1992; Raskin, 1986). The decision rules used by federal examiners can also be criticized because, on their face they appear to be arbitrary, and they are without empirical validation. Further, it is generally known that the validity of any diagnostic technique that relies on human interpretation of complex data can be adversely affected by bias, drift, inexperience and incompetence (Nunnally, 1978). In support of this notion one study found that federal examiners produced inter-rater reliabilities as low as 0.36 in scoring respiration tracings (Honts, 1989). It may well be that the poor accuracy found in security screening applications of polygraph testing is due to the unavoidable biases of human decision making under the extreme base rate conditions of the national security setting (see Honts, 1991a for an elaboration of the base rate problem in national security screening).

Statistical Decision Making, A Possible Solution

One possible solution to the problems associated with human-based data analysis and decision making in the detection of deception may lie in the adoption of a computer based statistical decision making approach. There is an extensive scientific literature concerning clinical versus statistical decision making that suggests that statistical decision models usually outperform human evaluators (Wiggins, 1981). At least one such statistical decision making system is currently available commercially, and is known as the Computer Assisted Polygraph System ([CAPS] Kircher & Raskin, 1991). That system uses a multivariate classification technique know as discriminant analysis followed by Bayesian probability modeling to provide a posterior probability of truthfulness. The data analysis used by the CAPS and the data upon which it was originally based have been described in detail by Kircher and Raskin (1988).

Briefly, Kircher and Raskin (1988) conducted a laboratory experiment with 100 subjects, half of whom enacted a mock theft. The physiological data were collected on-line and were converted to digital form by a computer. Features were extracted from the physiological waveforms associated with relevant and control questions. These feature values were converted to a common metric through standardization. An average value was calculated for each of the features for control and relevant questions. Then the average value for relevant questions was subtracted from the average value for control questions. The feature difference scores were then subjected to discriminant analysis. The number of variables retained for analysis were reduced by empirical and *a priori* methods (ultimately to three in Raskin et al., 1988). Discriminant analysis then provided optimal weights for the variables to be used in a

classification equation. The discriminant classification equation was then applied to each subject's data to generate a single discriminant score for that subject. The mean and standard deviation of the discriminant scores for innocent and guilty subjects were then calculated and used as an estimate of the population values for those distributions. Cross validation trials were conducted by applying the discriminant equation to subjects from other experiments. Discriminant scores were then calculated, and probability densities for those discriminant score were computed with respect to the empirical truthful and deceptive populations. Those probabilities, along with the base rate, were entered into Bayes Theorem to determine an *a posteriori* probability of truthfulness for each subject. This soft information was then converted into a hard decision by the application of either an arbitrary or empirical decision rule (see, Honts, Raskin, Kircher, & Horowitz, 1988; Kircher & Raskin, 1988; Raskin, et al., 1988).

Possible Weaknesses of the Discriminant Analysis Approach

The original Kircher and Raskin (1988) five variable model, and the simpler three variable model sold in the commercial application, have shown considerable general validity on cross validation, performing as well as, or better than, human evaluators in five laboratory (Gatchel, Smith, & Kaplan, 1984; Honts, 1986, Honts & Carlton, 1990; Horowitz, 1989, Rovner, 1986) and two field (Raskin, Kircher, Honts, & Horowitz, 1988; Raskin, Horowitz & Kircher, 1989) studies. However, discriminant analysis is limited in its application by three potential problems.

Problem 1: Small Sample Sizes. Discriminant analysis operates on a vector of *k* variables, and retains a subset of weighted variables that provides for the maximum linear discrimination of the groups (Cooley & Lohnes, 1971). With two criterion groups, only one discriminant equation results. Unfortunately, the stability of the weights, and to some extent the stability of the variables chosen for the equation, depend upon the number of subjects used in the analysis. Estimates of the minimum number of subjects needed to obtain stable variable subsets and variable weights are usually given as a subject to variable ratio. Estimates vary widely, with recommended ratios as low as 5/1 or as high as 200/1. Clearly, discriminant analysis modeling in the detection of deception has been operating toward the lower end of this suggested range. Therefore, the stability of the variable subsets and the obtained weights in the current studies is suspect.

Problem 2: Restrictive Assumptions. Discriminant Analysis makes the following three assumptions: (1) the relationship between the variables

is linear, (2) the continuous variables come from multivariate normal populations, and (3) the covariance matrices for the groups are equal. These assumptions are often not met in actual data sets. For example, the covariance matrices in the Honts and Carlton (1990) study were not equal. To the extent that these basic assumptions are not met, the predictive accuracy of discriminant analysis will be decreased (Betz, 1987).

Problem 3: Issues Of Generalizability. The generalizability of a given predictive discriminant equation is dependent upon meeting the assumptions of the technique, and upon the origin of the data on which the equation was developed. The Kircher and Raskin (1988) approach was largely based and cross validated on laboratory data. The extent that laboratory data generalize to field situations is a matter of considerable heated debate. Many of the critics of lie detection state opinions that laboratory data are essentially worthless for generalization to the field (Iacono & Patrick, 1978; Lykken, 1981).

One response to potential problems of generalizability of laboratory data is to collect field data. Raskin et al. (1988) collected field data from the United States Secret Service and tested the Kircher and Raskin laboratory statistical model. Some modifications of the discriminant equation were found to be necessary. However, the cross validation and/or model building on field data is only a partial solution to the generalizability problem. Currently, extant confirmed field data sets are small in size. They are also subject to a number of sampling biases and to at least some uncertainty in the quality of their ground truth confirmations. Some authors have stated that the sampling problems associated with conducting field studies in the detection of deception are so severe that they may not be solvable (Patrick & Iacono, 1991). These sampling biases may severely limit the generalizability of data analysis conducted on field data sets. Even if solutions to the problems of sampling bias can be found, field studies are notoriously difficult and expensive to conduct. Therefore, the prospects for the development of large *N*, high quality, highly generalizable, field data sets for model building in the near future seems dim.

The generalizability problem may be particularly acute when attempting to apply statistical decision making to security screening. Laboratory research on security screening polygraph tests has failed to find accuracies as high as those found in laboratory studies of law enforcement-type testing (Honts, 1991a). This suggests that statistical models built in the context of law enforcement testing may be not be appropriately applied to screening situations. Statistical models could be built on data obtained in laboratory screening studies, but then those results would be subject to the same criticisms and concerns as those associated with laboratory data in law enforcement simulations. Moreover,

field data for national security screening studies will be extremely difficult to develop. It seems likely that large numbers of confirmed deceptive subjects will be almost impossible to obtain in the context of national security screening since the base rate of espionage is so low in the field, and because current research suggests that the false negative rate for detecting security violations in the field is very high (Honts, 1991a).

Bootstrapping: A New Approach to Statistical Decision Making

In the present study we examined the potential utility of a relatively new, computationally intensive method of data analysis known as bootstrapping (Diaconis & Efron, 1983; Efron, 1979; Wasserman & Brockenholt, 1989) for polygraph decision making. Bootstrapping was an attractive choice for use as a polygraph decision maker for two reasons. First, it made no restrictive mathematical or distributional assumptions of the data, and second, it potentially avoided problems of generalizability since each decision was based only on the data available from the subject in question.

UND
UNIVERSITY
OF
NORTH
DAKOTA

A Conceptual Introduction to Bootstrapping

The bootstrap technique has been described in detail in a number of places, and accessible descriptions have been provided by Diaconis and Efron (1983) and Wasserman and Brockenholt (1989). Conceptually, the bootstrap procedure is as follows. Consider the following simple example adapted from Simon and Bruce (1991). Two samples size $N = 10$ are obtained. The mean of sample A = 29.5 and the mean of sample B = 28.2. You are interested to know if those means are really different. Formally, we have a null hypothesis that the two means are not different and an alternative hypothesis that the means are different. The statistical question is: How likely are we to have drawn two samples of size $N = 10$ from the null hypothesis population that are as different as, or more different, than 1.3 units. This statistical question might be answered traditionally through the use of a *t*-test. However, use of the *t* test requires certain assumptions, such as homogeneity of variance, normal population distributions, and interval scale measurement, that may not be met by the data at hand. Bootstrapping answers this question without reference to such restrictive assumptions.

A bootstrap solution to the above problem is as follows. First, since our null hypothesis is that there is no difference between A and B, we can therefore consider all 20 of the subjects to be representative of the null

hypothesis population. To bootstrap a solution to this problem, each subject is assigned a number and those twenty numbers are thrown into a hat. A random sample, A_{sam}, of size $N = 10$ is drawn from the hat with replacement and is used to represent A. Then a second random sample, B_{sam}, of size $N = 10$ is drawn from the hat with replacement and is used to represent B. The difference between the mean of A_{sam} and B_{sam} is calculated and is stored. This process is repeated many times. The more times the sampling process is repeated, the more exact the approximation of the true sampling distribution of the difference between the means. Generally, at least 400 resamplings are necessary for a useful approximation of the sampling distribution, and 1500 resamplings generally gives a close approximation to the true sampling distribution (Searls, 1991). Once the resampling process is completed, the sampling distribution created by the retained difference scores can be used to evaluate the likelihood of obtaining a difference score as large as, or larger than, the one obtained between the two original samples. This is accomplished by calculating the probability associated with the area under the curve as extreme as, or more extreme than, the obtained value. Simulation studies have been conducted on bootstrap approaches to hypothesis testing and those studies have found the bootstrap to perform comparably to traditional parametric statistics when the assumptions of those statistics are met, and to outperform them when the parametric assumptions are violated (Thompson, 1991). Furthermore, the bootstrap has been recommended for consideration in complex psychological problems where no appropriate parametric statistics are available (Thompson, 1991).

Our Approach To Bootstrapping a Polygraph Decison

Our approach to polygraph decision making via bootstrapping was as follows. As is described in detail below, the physiological responses of subjects to relevant and control questions were quantified by extracting 8 physiological features from each of the 9 presentations of control and relevant questions in a control question test polygraph examination. Thus, there were 72 data points for the relevant questions and 72 data points for the control questions. These 144 data points were then used to create 2 vectors of data representing the responses to relevant and control questions. Next, the relevant question vector was subtracted from the control question vector to create a difference score vector. Then the difference score vector was summed to produce a single score. This sum of the difference scores was the target statistic for the bootstrapping process.

The bootstrapping process was as follows: The control and relevant question vectors were concatenated into a single data vector. The order of

the data in this combined vector was then randomized. Two new vectors of length 72, were then created by random sampling with replacement from the randomized combined vector. One of the new data vectors was arbitrarily called a control question vector and the other a relevant question vector. The relevant question vector was subtracted from the control question vector and a difference score vector was created. The data within the difference score vector were summed. The sum of the difference scores was stored. This process was repeated 2000 times.

After the 2000th repetition, the mean and standard deviation of the bootstrapped sampling distribution of the difference scores were calculated. The obtained difference score was then evaluated as follows: Using the bootstrapped parameters, the obtained difference score was converted into a *z*-score. It was expected that innocent subjects would produced positive *z*-score values, while guilty subjects were expected to produce negative *z*-score values. The probability of obtaining a difference score equal to, or less positive than, the obtained difference score was then calculated using standard tables for the *z* statistic. Using this approach it was predicted that innocent subjects would produce high probability values. That is, we predicted that for innocent subjects most of the bootstrapped difference scores would be smaller than the obtained difference scores. It was also expected that guilty subjects would produce small probabilities. That is, we predicted that for guilty subjects most of the bootstrapped difference scores would be larger than the obtained difference scores.

Performance of the bootstrapping approach was tested in relation to the performance of independent human evaluators who used the semi-objective scoring system developed at the University of Utah (Kircher & Raskin, 1988). Bootstrap performance was also examined in relation to the results of a discriminant analysis of the same data that had been adjusted for overfitting with a jackknife procedure.

Method: Phase I

Software For The Project.

Bootstrapping was conducted with programs written in the *Resampling Stats* software language (Simon, Puig, & Bruce, 1991). Other statistical analyses were conducted with SPSS/PC+ (Norusis, 1988).

Hardware for The Project

The data analysis was conducted on a 80386 25 MHz PC clone, equipped with an 80387 math coprocessor and 4 Megabytes of RAM. The system used DOS 5.0 and Windows 3.1 for maximum memory availability to *Resampling Stats*.

The Project Data Base.

Data were obtained from the Utah Cooperative Working Group Data Base. This data base contained the digitized data in CAPS format from the following studies: Honts (1986), Honts and Carlton (1990), Horowitz, (1989), and Kircher and Raskin (1988). All of these were mock crime studies of the version of the control question test developed at the University of Utah. This control question test contained three relevant and three control questions that were presented on each chart. Within the data base, there were complete data for 110 Innocent and 114 Guilty subjects from the control groups in those studies. From that data base, 100 Innocent and 100 Guilty subjects were randomly selected for analysis using the TAKE function of *Resampling Stats*. Data from the first three charts of each examination were used in this study.

Procedure

Feature Extraction / Data Reduction Phase. Eight physiological features were extracted for each of the relevant and control question presentations, using the *ARCHIVE* software developed by Kircher (1990). Those features were: electrodermal response amplitude, electrodermal response duration, the number of peaks in the electrodermal response,

UND
UNIVERSITY
OF
NORTH
DAKOTA

cardiovascular response amplitude, cardiovascular response duration, the number of peaks in the cardiovascular response, thoracic respiration length and abdominal respiration length. These variables were chosen on an *a priori* basis because they appeared to capture the major reactive dimensions of the three physiological systems most often measured during polygraph testing. Feature extraction was done for three repetitions of the questions (charts) and for three relevant-control pairs on each repetition. This resulted in a 72 variable relevant question vector and a 72 variable control question vector. Those data were then placed on a common metric by conversion to *z*-scores. Our *a priori* assumption was that in all of our physiological features, except the two respiration measures, larger feature values reflected larger psychophysiological response. However, in the respiratory system larger psychophysiological response is indexed by a decrease in physiological activity. This is indexed by smaller feature values. To facilitate interpretation and subsequent calculations, the respiration length *z*-scores were reflected by multiplying all respiratory feature values by (-1). A 72 variable difference score vector was then created by subtracting the relevant question vector from the control question vector. Thus, from our *a priori* perspective, all of the difference scores were expected to have a positive predictive correlation with the guilt criterion. That is, guilty subjects were expected to produce negative difference scores, while innocent subjects were expected to produce positive difference scores. The difference score vector was then summed to yield a single difference score value. The obtained sum of the difference scores served as the target statistic for bootstrapping.

The Bootstrapping Process. Bootstrapping was performed as follows: The 72 variable control question vector was concatenated to the relevant question vector. The order of the 144 variables in the concatenated vector was then randomized. Two 72 variable vectors were then created by random sampling with replacement from the 144 variable concatenated vector. The first of the two 72 variable vectors was labeled as a mock control question vector and the second was labeled as a mock relevant question vector. The mock relevant question vector was then subtracted from the mock control question vector resulting in a 72 variable mock difference score vector. The mock difference score vector was summed to give a single value. This sum of the mock difference scores was then stored and the bootstrapping process repeated. Two thousand repetitions of the bootstrapping processes were conducted for each subject.

Decision Making. The mean and standard deviation of the bootstrapped sampling distribution of the sum of the difference scores were calculated. Using those bootstrapped parameters, the obtained sum of the difference scores was converted to a *z*-score. Then the probability of

obtaining a *z*-score as positive as, or less positive than, the obtained *z*-score was determined from a standard table of areas under the normal curve (Pagano, 1990). Initially, for purposes of classification, all probabilities greater than 0.50 were considered as truthful outcomes and all probabilities less than 0.50 were considered as deceptive outcomes. No probabilities equal to 0.50 were obtained.

Discriminant Analysis. The data base for discriminant analysis was developed as follows: The same eight physiological features were extracted for each relevant and control question presentation. Those values were then averaged across relevant and control questions and across charts. Difference scores were calculated by subtracting the average values for relevant questions from average values for control questions. The data base for discriminant analysis was the resultant 8 variable difference score vector.

Our approach to discriminant analysis was to force simultaneous entry of all eight variables into the discriminant equation. In order to control for discriminant analysis' tendency to overfit the standardization data set, a jackknife procedure was employed. In the jackknife, 200 discriminant analyses were conducted using a remove one subject, build the discriminant model, classify the one subject, jackknife procedure similar to the one previously used by Honts and Devitt (1991). This analysis produced a residual discriminant score and a dichotomous classification as either truthful or deceptive for each subject. Those residual discriminant scores and classifications were retained for comparison with the results of the bootstrap analysis.

Human Evaluation. All of the data in the present study had been evaluated by an independent human evaluator who was blind to the subject's guilt status at the time the evaluation was conducted. The independent evaluators in these studies were trained in and used the numerical scoring developed at the University of Utah (Kircher & Raskin, 1988). All of the independent evaluators held Ph. D. degrees in Psychology and they had all received graduate level training in psychometrics and psychophysiology. Total numerical scores for the first three charts were recorded for each subject included in this analysis. Initially, subjects with positive scores were considered truthful outcomes, and subjects with negative scores were considered deceptive outcomes. There were five subjects with total numerical scores for three charts of zero. Those five subjects were dropped from the initial comparison.

Results & Discussion: Phase I

The results of the three evaluations from Phase I are shown in Table 1. The human evaluators correctly classified 82.0% of the subjects. The human evaluators' false positive rate was 12.2%, while their false negative rate was 23.7%. The classification efficiency (Kircher, Horowitz, & Raskin, 1988) of the human evaluators was calculated to be $r = 0.64$. The jackknifed discriminant analysis correctly classified 84.1% of the subjects. The false positive rate with the discriminant analysis was 13.3%, and the false negative rate was 18.6%. Classification efficiency for the discriminant analysis was $r = 0.68$. The bootstrap classifications were 77.9% correct. The bootstrap classification false positive rate was 18.4%, while the false negative rate was 25.8%. Classification efficiency for the bootstrap classifications was $r = 0.56$. The differences between the classification efficiencies of the bootstrap classifier and the other two classifiers were tested using the procedures described by Klugh (1970). The bootstrap classifiers efficiency was not statistically different from the efficiency of the human evaluators, but was significantly less than the efficiency of the discriminant analysis $t(192) = 2.82$, $p < 0.01$. The efficiencies of the discriminant analysis and the human evaluators were not statistically different.

Table 1. Outcomes by the three classifiers in Phase I. ($N = 195$)

Decision ⇒ Condition ⇓	Truthful	Deceptive
Human Evaluation		
Innocent	86	12
Guilty	23	74
Discriminant Analysis		
Innocent	85	13
Guilty	18	79
Bootstrap Analysis		
Innocent	80	18
Guilty	25	72

The discriminative power of these three techniques was also assessed by correlating the numerical scores, the discriminant scores, and bootstrap-obtained *z*-scores with the guilt criterion. All of the 200 subjects were used for this analysis. The resulting point biserial correlations are shown in Table 2. The differences between these correlations were tested. The bootstrap-obtained *z*-scores were significantly less discriminative than the discriminant scores, $t(197) = 2.29$, $p < 0.05$. However, the bootstrap obtained *z*-scores and the numerical scores were not statistically different

Table 2. Correlations of the Phase I Classifier Scores With the Guilty Criterion.

Evaluation ⇒	Human Evaluation	Discriminant Analysis	Bootstrap Analysis
Point Biserial r	0.62	0.66	0.59

in their discriminative power nor were the numerical scores and the discriminant scores.

The initial results with the bootstrap classifier were somewhat disappointing. The bootstrap classifier performed more poorly than the discriminant analysis and although the differences between the bootstrap classifier and the human evaluators were not statistically significant, in absolute terms the bootstrap classifier performed more poorly than the human evaluators.

In order to explore this outcome, we examined the data from the subjects that the bootstrap procedure had misclassified. Our subjective impression was that, although the electrodermal and cardiovascular response amplitudes were often in the correct direction on these misclassified subjects, the number of peaks variables often seemed to be varying in the opposite to the predicted direction. In order to explore this possibility, we correlated the averaged feature difference scores from the discriminant analysis with the guilt/innocence criterion. The results of those correlations are presented in Table 3. Both number of peaks variables produced significant negative correlations with the guilty criterion.

Despite the fact that most numerical scoring systems treat a response with larger number of peaks as a stronger physiological response than a response with fewer peaks, these results strongly suggest that the opposite is true. These results indicate that responses with less complexity should be considered as the stronger responses. It should also be noted that the overall discriminant analysis and the individual jackknife analyses assigned negative weights to the number of peaks variables. Given these results, we felt that the initial bootstrap analysis was run at a disadvantage because of our *a priori* assumption that the number of peaks variables were positive predictors rather than negative predictors. We, therefore, decided to conduct a second wave of bootstrap analyses with the values for the number of peaks variables transformed to indicate their correct predictive orientation.

Table 3. Inter-Correlations of Physiological Feature Difference Scores and the Guilt Criterion.

[N = 200 for all correlations, $p < 0.05$ for all values > 0.13.]

	EDRD	EDRN	CDRA	CDRD	CDRN	TRL	ARL	GUILT
Electrodermal Response Amplitude (EDRA)	.79	-.30	.49	.41	-.22	.19	.30	.65
Electrodermal Response Duration (EDRD)		-.15	.43	.41	-.15	.18	.23	.55
Number of Electrodermal Peaks (EDRN)			-.13	-.10	.14	-.20	-.20	-.31
Cardiovascular Response Amplitude (CDRA)				.76	-.34	.09	.15	.39
Cardiovascular Response Duration (CDRD)					-.15	.13	.18	.34
Number of Cardiovascular Peaks (CDRN)						-.17	-.21	-.15
Thoracic Respiration Length (TRL)							.72	.30
Abdominal Respiration Length (ARL)								.35

Phase II: A Second Bootstrap

Method: Phase II

The method for the Phase II analyses was the same as for Phase I, except that all of the variables in the initial 72 variable control and relevant question vectors were multiplied by a 72 variable vector of 1s and -1s, arranged to result in a transformation of the signs of all of the number of peaks variables before the calculation of difference scores. This process effectively changed the number of peaks variables from negative predictors to positive predictors. The bootstrap analyses were then conducted as before. No changes were made in the analyses of the numerical scores or the discriminant analyses.

Results and Discussion: Phase II

The results of the analyses from Phase II are illustrated in Table 4. The bootstrap analyses now correctly classified 83% of the subjects, compared to 82% and 84% for the human evaluators and the jackknifed discriminant analysis, respectively. The false positive rate with the bootstrap classifications was 18.4% as compared to 12.2% and 13.3% for the human evaluators and the jackknifed discriminant analysis, respectively. The bootstrap false negative classification rate was 15.5% as compared to 23.7% and 18.6% for the human evaluators and the jackknifed discriminant analysis, respectively. Classification efficiencies for the three classifications were: $r = 0.66$ for the bootstrap classifier, $r = 0.64$ for the human evaluators, and $r = 0.68$ for the jackknifed discriminant analysis. Differences between the detection efficiencies were tested and none were found to be significant. The improvement of the classification efficiency of the bootstrap classifier in Phase II over the performance of the bootstrap classifier in Phase I was tested and was found to be significant, $t(192) = 2.53$, $p < 0.02$.

The numerical, discriminant, and bootstrap *z*-scores, from all 200 subjects were correlated with the guilt criterion. The respective correlations were: 0.62, 0.66, and 0.64. These correlations were not found to be statistically different from each other.

Table 4. Outcomes By The Three Classifiers In Phase II. (N = 195)

Decision ⇒	Truthful	Deceptive	Detection Efficiency *r*
Human Evaluation			.64
Innocent	86	12	
Guilty	23	74	
Discriminant Analysis			.68
Innocent	85	13	
Guilty	18	79	
Bootstrap Analysis			.66
Innocent	80	18	
Guilty	15	82	

Phase III: An Exploration of Inconclusive Zones

Traditionally, decision making in polygraph examinations has included an inconclusive zone. By accepting some examinations as indeterminate, accuracy rates can generally be improved. We conducted a parametric analysis to examine the effects of the use of a number of inconclusive zones on the three classifiers from Phase II.

Method

Human Evaluation. For the numerical scores, nine different inconclusive zones were examined. Initially, the inconclusive zone was set so that only total numerical scores of zero were considered inconclusive. Then the inconclusive zone was expanded by 1 numerical score point in the positive direction and by 1 numerical score point in the negative direction. Thus, the second inconclusive zone was from -1 to +1, inclusive. The inconclusive zone was expanded by 1 point in either direction for 9 steps. Thus, the largest inconclusive zone examined was from -8 to +8, inclusive.

Jackknife Discriminant Analysis. Nine inconclusive zones were also examined for the jackknife discriminant analysis. These inconclusive zones were based on the *a posteriori* probability of truthfulness generated for the residual subject in each of the jackknife steps. These probabilities were such that *p*-values greater than 0.50 would indicate that the control questions were stronger than the relevant questions and *p*-values less than 0.50 would indicate that the relevant questions produced the stronger response. The initial inconclusive zone was set up so that *p*-values greater than 0.45 and less than 0.55, exclusive, were considered inconclusive outcomes. The inconclusive zone was expanded by 0.05 in both directions for each step for 9 steps. Thus, the largest inconclusive zone ranged from 0.05 to 0.95, exclusive.

Bootstrap Classification Analysis. Nine inconclusive zones were also tested on the results of the bootstrap classification analysis. Here the inconclusive zones were based on the probability of obtaining a *z*-score

value as extreme as, or less extreme than, the one obtained under the assumption that the subject was truthful. These *p*-values behave in a manner similar to the *a posteriori p*-values generated by a discriminant classification analysis. Thus, bootstrap *p*-values greater than 0.50 indicated that the control questions were stronger than the relevant questions, and *p*-values less than 0.50 indicated that the relevant questions produced the stronger response. The initial inconclusive zone was set up so that *p*-values greater than 0.45 and less than 0.55, exclusive, were considered inconclusive outcomes. The inconclusive zone was expanded by 0.05 in both directions for each step for 9 steps. Thus, the largest inconclusive zone ranged from 0.05 to 0.95, exclusive.

Results and Discussion of Phase III

The results of the various inconclusive zones imposed on the numerical scores are given in Table 5. The true positive decision, true negative decision, and the inconclusive rates are summarized in Figure 1. The overall performance of the numerical scoring peaked when the inconclusive zone was set between -4 and +4, exclusive. With that inconclusive zone, the numerical scores correctly classified 79.5% (159) of the subjects and achieved a detection efficiency coefficient of $\underline{r} = 0.67$. The true positive rate was 81.1% (60 of 74 decisions), and the true negative rate was 91.9% (79 of 86 decisions) At that inconclusive zone, 20% (40) of the subjects produced inconclusive outcomes. There were approximately twice as many inconclusive outcomes with guilty subjects. Twenty-six inconclusive outcomes occurred with guilty subjects as compared to 14 with the innocent.

The results of the various inconclusive zones imposed on the discriminant score probabilities are given in Table 6. The true positive decision, true negative decision, and the inconclusive rates are summarized in Figure 1. The overall performance of the discriminant analysis peaked when the inconclusive zone was set between *p* values of 0.40 and 0.60, exclusive. With that inconclusive zone, the discriminant analysis correctly classified 75.0% (150) of the subjects and achieved a detection efficiency coefficient of $r = 0.68$. The true positive rate was 84.1% (74 of 88 decisions), and the true negative rate was 89.4 (76 of 85 decisions) At that inconclusive zone, 13.5% (27) of the subjects produced inconclusive outcomes. The inconclusive outcomes were approximately equal for guilty (12) and innocent (15) subjects.

The results of the various inconclusive zones imposed on the bootstrap $\underline{z}$-score probabilities are given in Table 7. The true positive decision, true negative decision, and the inconclusive rates are

Table 5. Outcomes and Detection Efficiencies Across Inconclusive Zones for Numerical Scores. (*N* = 200)

Inconclusive Zone (Values Exclusive)	Condition	Outcome			Detection Efficiency *r*
		Truthful	Inc.	Deceptive	
-1/+1	Innocent	86	2	12	.64
	Guilty	23	3	74	
-2/+2	Innocent	82	8	10	.64
	Guilty	21	9	70	
-3/+3	Innocent	81	11	8	.66
	Guilty	16	20	64	
-4/+4	Innocent	79	14	7	.67
	Guilty	14	26	60	
-5/+5	Innocent	72	21	7	.64
	Guilty	11	34	55	
-6/+6	Innocent	70	24	6	.65
	Guilty	9	38	53	
-7/+7	Innocent	64	30	6	.61
	Guilty	9	45	46	
-8/+8	Innocent	60	37	3	.64
	Guilty	5	52	43	
-9/+9	Innocent	57	41	2	.64
	Guilty	3	59	38	

Table 6. Outcomes and Detection Efficiencies Across Inconclusive Zones for Discriminant Score Probabilities. (*N* = 200)

Inconclusive Zone (Values Exclusive)	Condition	Outcome			Detection Efficiency *r*
		Truthful	Inc.	Deceptive	
.45/.55	Innocent	82	7	11	.68
	Guilty	17	6	77	
.40/.60	Innocent	76	15	9	.68
	Guilty	14	12	74	
.35/.65	Innocent	70	21	9	.65
	Guilty	14	15	71	
.30/.70	Innocent	68	25	7	.67
	Guilty	12	19	69	
.25/.75	Innocent	59	37	4	.67
	Guilty	9	26	65	
.20/.80	Innocent	52	45	3	.64
	Guilty	8	32	60	
.15/.85	Innocent	46	52	2	.62
	Guilty	5	46	49	
.10/.90	Innocent	40	59	1	.62
	Guilty	3	50	47	
.05/.95	Innocent	32	67	1	.54
	Guilty	3	60	37	

Table 7. Outcomes and Detection Efficiencies Across Inconclusive Zones for Bootstrap *z*-score *p*-values. (*N* = 200)

Inconclusive Zone (Values Exclusive)	Condition	Outcome			Detection Efficiency *r*
		Truthful	Inc.	Deceptive	
.45/.55	Innocent	77	6	17	.65
	Guilty	15	3	82	
.40/.60	Innocent	76	9	15	.68
	Guilty	12	7	81	
.35/.65	Innocent	74	13	13	.69
	Guilty	16	20	64	
.30/.70	Innocent	71	17	12	.71
	Guilty	8	13	79	
.25/.75	Innocent	65	23	12	.67
	Guilty	8	17	75	
.20/.80	Innocent	63	25	12	.64
	Guilty	8	25	67	
.15/.85	Innocent	55	33	12	.60
	Guilty	7	29	64	
.10/.90	Innocent	50	40	10	.60
	Guilty	6	32	62	
.05/.95	Innocent	41	56	3	.60
	Guilty	6	39	55	

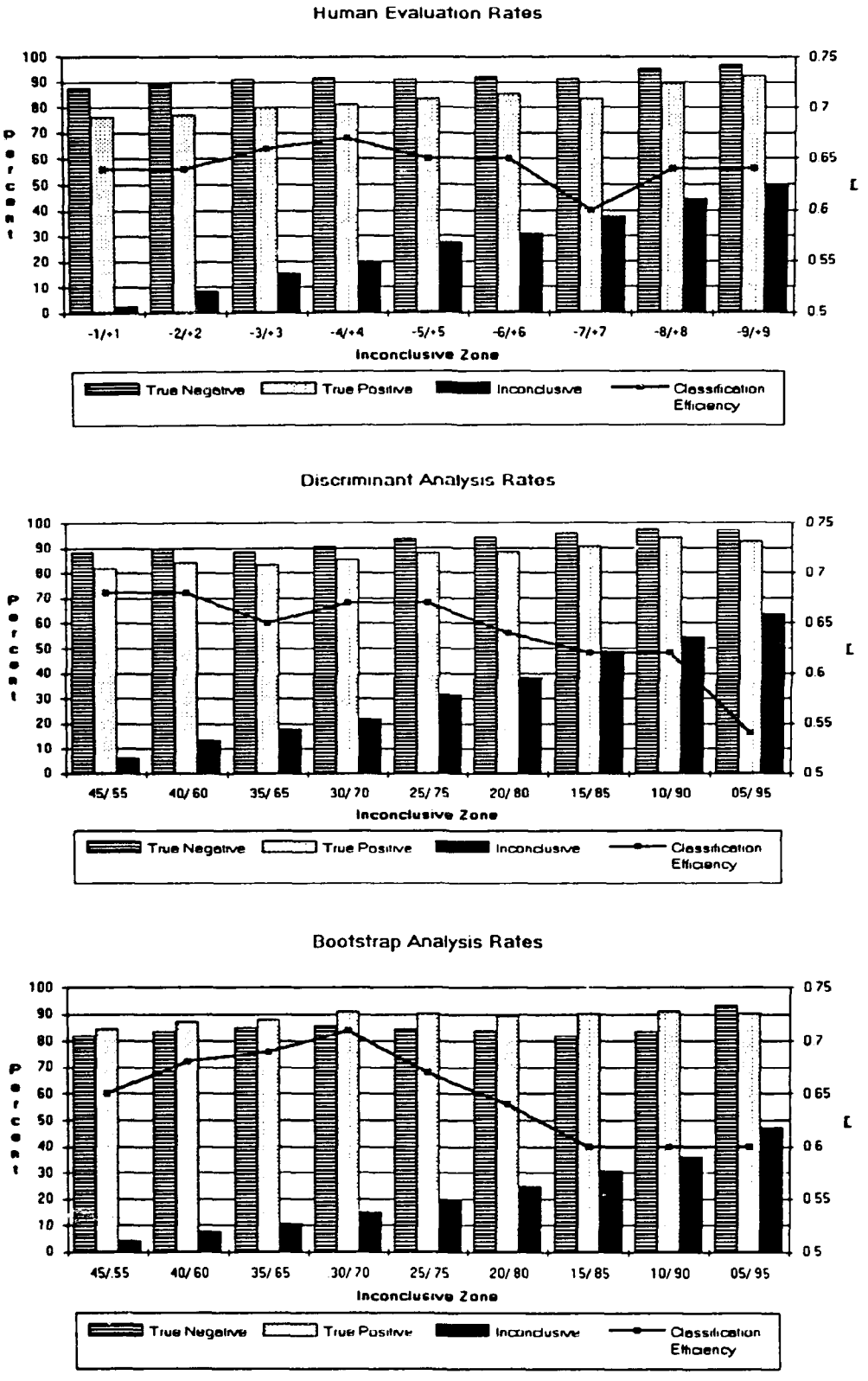

Figure 1. Rates of correct decisions and inconclusives plus the detection efficiency coefficient for the classifiers across inconclusive zones.

summarized in Figure 1. The overall performance of the bootstrap analysis peaked when the inconclusive zone was set between p values of 0.30 and 0.70, exclusive. With that inconclusive zone, the discriminant analysis correctly classified 75.0% (150) of the subjects and achieved a detection efficiency coefficient r = 0.71. The true positive rate was 90.8% (79 of 87 decisions), and the true negative rate was 85.5% (71 of 83 decisions) At that inconclusive zone, 15% (30) of the subjects produced inconclusive outcomes. The inconclusive outcomes were approximately equal for guilty (13) and innocent (17) subjects.

A visual inspection of Figure 1 leads to several observations. Across the range of inconclusive zones, the bootstrap analysis seems to have a slightly better performance as indexed by the detection efficiency coefficient. The bootstrap analysis generally produced better performance with guilty subjects, and gave many fewer inconclusive outcomes. However, the performance of the bootstrap analysis with innocent subjects lagged slightly behind the other two classifiers.

General Discussion

In the present study we compared the performance of a bootstrap approach to polygraph decision making to independent numerical scoring and jackknife-adjusted discriminant analysis applied to the same data. The bootstrap analysis performed well, producing results that were statistically not different from the other two classifiers. However, in absolute terms, the bootstrap classifier produced the largest detection efficiency coefficient of any of the classifiers. Furthermore, across the range of inconclusive zones we examined, the bootstrap produced the highest rate of correct decisions with guilty subjects and by far the smallest inconclusive rate.

The relative performance of the statistical classifiers in this study is enhanced by the fact that the performance of the human evaluators used in this study must be considered as near the maximum attainable by the numerical scoring technique. The human evaluators who provided the numerical scoring data in these studies were all scientists who held advanced degrees. They all had advanced training in psychometrics and psychophysiology. Moreover, they used the scientifically-based numerical scoring system developed at the University of Utah. That system has consistently been shown to be highly reliable. This favorable set of factors may not represent the case for numerical scoring in the field. Field examiners in most federal agencies use the numerical scoring system taught at the Department of Defense Polygraph Institute. That numerical scoring system has been widely criticized (Honts, 1991a; Honts & Perry, 1992; Raskin, 1986, 1989) in the scientific literature for the reasons noted earlier in this report. In application, the Department of Defense Polygraph Institute numerical scoring system has, on at least one occasion, produced reliability estimates that were no, or only slightly, better than, chance in the scoring of some components (Honts, 1989). Considering that the Department of Defense may have serious reliability problems with their numerical scoring system, then performance across the organization cannot be nearly as high as the performance of the numerical evaluators in this study.

Given that the estimates for numerical evaluation in this study must be nearly asymptotic, the performance of the two statistical classifiers is all the more impressive. In considering the two statistical classifiers, the

bootstrap has a number of things to recommend it. First, bootstrapping either solves, or avoids, the technical problems associated with discriminant analysis we discussed earlier in this report. The bootstrap analysis makes no measurement or distributional assumptions of the data. Sample size of the database is not of importance to the bootstrap, since bootstrap decision making is based on the single subject without reference to a standardization group. Furthermore, since the bootstrap model is based on assumptions derived from psychophysiological knowledge and the rationale of the control question test, and not on an empirical standardization group, issues of generalizability are moot. The bootstrap is generalizable, as long as the psychophysiological and theoretical assumptions hold. The generalizability of these theoretical assumptions is, of course, the subject matter of considerable controversy, and needs to be the subject of considerable basic science research. However, such issues are beyond the scope of this project. If the rationale of the control question test holds, the bootstrap analysis should work.

It is also interesting to note that the bootstrap analysis, which gives equal weight to all of the physiological variables, performs as well as discriminant analysis which optimally weights the variables. This result suggests that the process of weighting variables in the psychophysiological detection of deception may not be very useful. It may be that the process of weighting variables results in too much dependence on the standardization sample upon which the classification model was built, whereas flat models based on assumptions about the psychophysiology of the situation are more robust. This area deserves additional attention and research in the future.

One finding from this study has implications for immediate application in the field, and that concerns the evaluation of the complexity of electrodermal and cardiovascular responses. The data base for this study is, to our knowledge, the largest single sample of control question test data ever examined. In this sample, contrary to expectations, we found that the number of peaks of electrodermal and cardiovascular responses were negatively associated with strength of reaction. That is, simple responses rather than complex responses were associated with a stronger psychophysiological response. Most numerical scoring systems currently evaluate responses with more complexity (more peaks) as stronger responses. These results suggest that this is incorrect and should be changed. Changing the orientation of our bootstrap decision maker toward those variables resulted in a significant increase in performance. Such a change should be strongly considered for human numerical scoring systems.

Finally, some general comments regarding statistical versus clinical decision making in the psychophysiological detection of deception. It is interesting that after more than a decade of effort toward applying statistical techniques to decision making in the psychophysiological detection of deception, no major breakthroughs have occurred. In general, statistical classifiers have been shown to perform as well as, or perhaps slightly better than, the **best** human evaluations. Even with equivalent performance to the best human evaluators, there is much to recommend the immediate universal application of statistical decision making in the field. This is especially true considering the sorry state of much of the practice in the polygraph profession at this time (Honts, 1991a, 1992; Honts & Perry, 1992; Raskin, 1986). Statistical decision making is worthwhile for the advantage it gives in reliability alone. Furthermore, even if statistical decision making is only a few percentage points more accurate than human evaluation, that is a very significant improvement in actuarial terms. Given all of these demonstrated advantages, it is truly surprising that statistical decision making is not already widespread in the polygraph profession, especially when one considers that statistical decision makers have been commercially available since 1986. However, to date, only the United States Secret Service has made the move to widespread statistical decision making. The results of the present study add to an already considerable body of research that clearly suggests that statistical decision making has come of age in the detection of deception. Universal implementation of statistical decision systems should be pursued with vigor by the policy makers in agencies that use polygraph tests.

However, despite our strong endorsement of the application of statistical decision making in the field, it should be acknowledged that statistical decision making is not a panacea. As we noted earlier, a considerable amount of research has failed to yield a major breakthrough in detection accuracy. Furthermore, it appears unlikely that such a breakthrough is possible within the context of the current physiological measures. The research in this study has been paralleled by research being conducted for a Master's thesis in the Psychology Department at the University of North Dakota. That study (Devitt, 1992), examined these same data with logistic regression and a back propagation neural network. Again, only small differences were found between the accuracy rates produced by the classifiers. Across the five classification techniques tested, numerical analysis, neural network analysis, discriminant analysis, logistic regression analysis, and bootstrap classification analysis, there was a group of subjects who were misclassified by all of the techniques. Examination of the data from those subjects suggests that the failure is not due to a problem with the classification techniques. Rather, the problem is that the control question test failed to work. That is, the problem appears to be psychophysiological, rather than one of classification technique. If

this is true, then no classifier will ever achieve a major breakthrough in accuracy. Rather, we need to concentrate on basic science research that will allow for the scientific development of improvements in testing techniques. Progress in such research is likely to be slow, breakthroughs may require a great deal of research before there is a direct payoff, but such research is necessary. Unfortunately, such research will not be conducted until the priorities of the funding agencies change and foster a basic science research agenda rather than one focused almost exclusively on application.

Follow-Up Research Projects

There seem to be three natural follow-up projects to the present research:

First, these results need to be replicated with field data. Although we believe that these results will be generalizable to any application where our basic assumptions about psychophysiological response and control question test hold, the effectiveness of bootstrapping needs to be demonstrated in the field. We have access to the data to do this demonstration in the form of the dataset obtained from confirmed field polygraph examination conducted by the United States Secret Service (Raskin et al., 1988, 1989). It is our intention to submit a proposal to DoD/PERSEREC/DODPI in the near future to conduct such a study

UND
UNIVERSITY
OF
NORTH
DAKOTA

Second, these results need to be examined in the context of threats posed by physical and mental countermeasures. The applicability of bootstrapping to decision making with countermeasure subjects is unknown. It seems likely that the success of countermeasures in studies such as Honts (1986) was due to a psychophysiological failure of the control question test, rather than to classification problems, however, this is not known with certainty. Again, we have the data on hand (from Honts, 1986) to test the effectiveness of bootstrapping with subjects using mental and physical countermeasures. Our intention is to include such an analysis in a proposal to DoD/PERSEREC/DODPI in the near future.

Third, the possibility of nesting several statistical classifiers in a hierarchical multilayer system should be explored. As we have noted earlier in this report there is a core of subjects who were misclassified by five different approaches to polygraph decision making, and unfortunately, there seems to be little hope of developing statistical techniques that will correctly classify those subjects. However, there were a number of subjects who were correctly classified by one of the classifiers, but were misclassified by the others. It may be possible to improve the overall accuracy of classification by using one, two or even three of the statistical approaches in a nested fashion. That is, multiple analyses would be conducted one after the other. One possible form for this hierarchical approach would be to initially conduct a bootstrap analysis, then for those subjects classified as truthful by the bootstrap, a logistic regression

analysis would be conducted in an effort to discriminate the true negatives from the false negatives. A similar analysis would be conducted with those subjects who failed the bootstrap in an effort to discriminate the true positives from the false positives. Thus, by maximizing the relative strengths of the various classifiers, we might be able to improve overall accuracy by as much as perhaps 10%, although that may be optimistic. This study should probably be done on the same data used in this study, so that results would be comparable to the analyses already conducted. Again, we hope to submit a proposal in the near future to DoD/PERSEREC/DODPI to do this study.

References

Barland, G. H., Honts, C. R., & Barger, S. D. (1989). *Studies of the Accuracy of Security Screening Polygraph Examinations*. Department of Defense Polygraph Institute, Fort McClellan, Alabama.

Betz, N. E. (1987). Use of discriminant analysis in counseling psychology research. *Journal of Counseling Psychology, 34*, 393-403.

Cooley, W. W., & Lohnes, P. R. (1971). *Multivariate data analysis*. New York: Wiley.

Department of Defense (1991). *Fiscal Year 1990 Report to Congress on the DoD Polygraph Program*. Washington, D.C.: Department of Defense.

Department of Defense Polygraph Institute (1990). Materials for the basic polygraph examiner's training course. Fort McClellan, AL: Department of Defense Polygraph Institute.

Devitt, M. K. (1992). *A study of the relative accuracy of discriminant analysis, logistic regression, and back propagation neural network classifiers in a psychophysiological detection of deception problem*. Unpublished master's thesis, University of North Dakota, Grand Forks.

Diaconis, P., & Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American, 248*, 116-130.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics, 7*, 1-26.

Gatchel, R. J., Smith, J. E., & Kaplan, N. M. (1984). *The effect of propranolol on polygraphic detection of deception*. Unpublished manuscript.

Honts, C. R. (1986). Countermeasures and the physiological detection of deception: A psychophysiological analysis. *Dissertation Abstracts International, 47*, 1761B. (Order No. DA8616081)

Honts, C. R. (1989). *The relative validity of two CSP question series*. Department of Defense Polygraph Institute, Fort McClellan, Alabama.

Honts, C. R., (1991a). The emperor's new clothes: Application of polygraph tests in the American workplace. *Forensic Reports, 4*, 91-116.

Honts, C. R. , (1991b). Converging evidence indicates invalidity for national security screening polygraph tests. *Psychophysiology, 28*, S30. (Abstract).

UNIVERSITY OF NORTH DAKOTA

Honts, C. R., (1992). Counterintelligence scope polygraph (CSP) test found to be a poor discriminator. *Forensic Reports*, *5*, 215-218.

Honts, C. R., & Carlton, B. (1990). The effects of incentives on the detection of deception. *Psychophysiology*, *27*, S39. (Abstract)

Honts, C. R., & Devitt, M. K. (1991, October). *Jackknife analyses of discriminant, logistic regression and back propagation neural network classifiers in a psychophysiological detection of deception problem*. Paper presented at the annual meeting of the Society for Psychophysiological Research, Chicago, Illinois.

Honts, C. R., & Perry, M. V. (1992). Polygraph admissibility: Changes and challenges. *Law and Human Behavior*, *16*, 357-379.

Honts, C. R., Raskin, D. C., Kircher, J. C., & Horowitz, S. W. (1988, March). *A field validity study of the control question test*. Paper presented at the American Psychology and Law Society / Division 41 Midyear Conference, Miami, Florida.

Horowitz, S. W. (1989). The role of control questions in physiological detection of deception. *Dissertation Abstracts International*, *50*, 1138. (Order No. AAC8911518)

Iacono, W. G., & Patrick, C. J. (1988). Assessing deception: Polygraph techniques. In R. Rogers (Ed.), *Clinical assessment of malingering and deception*. New York: Guilford. (205-233).

Kircher, J. C. (1990). *Archive, version 1.1*. Salt Lake City, UT: Scientific Assessment Technologies.

Kircher, J. C., Horowitz, S. W., & Raskin, D. C. (1988). Meta-Analysis of mock crime studies of the control question polygraph technique. *Law and Human Behavior*, *12*, 79-90.

Kircher, J. C., & Raskin, D. C. (1988). Human versus computerized evaluations of polygraph data in a laboratory setting. *Journal of Applied Psychology*, *73*, 291-302.

Kircher, J. C., & Raskin, D. C. (1991). *Computer assisted polygraph system, version 7.01*. Salt Lake City, UT: Scientific Assessment Technologies.

Klugh, H. E. (1970). *Statistics: The essentials for research*. New York: Wiley.

Lykken, D. T. (1981). *Tremor in the blood: Uses and abuses of the lie detector*. New York: McGraw-Hill.

Noursis, M. J. (1988). *SPSS/PC+ advanced statistics V2.0*. Chicago: SPSS, Inc.

Nunnally, J. C. (1978). *Psychometric theory (2nd ed.)*. New York: McGraw-Hill.

Pagano, R. R. (1990). *Understanding statistics in the behavioral sciences*. St. Paul: West.

Patrick, C. J., & Iacono, W. G. (1991). Validity of the control question polygraph test: The problem of sampling bias. *Journal of Applied Psychology*, *76*, 229-238.

Raskin, D. C. (1986). The polygraph in 1986: Scientific, professional and legal issues surrounding application and acceptance of polygraph evidence. *Utah Law Review*, *1986*, 29-74.

Raskin, D. C., Horowitz, S. W., & Kircher, J. C. (1989). *Computerized analysis of polygraph outcomes in criminal investigations*. Report of research and results of Phase II of Contract No. TSS 86-18 to the United States Secret Service, Salt Lake City: University of Utah, Department of Psychology.

Raskin, D. C., Kircher, J. C., Honts, C. R., & Horowitz, S. W. (1988). *A Study of the Validity of Polygraph Examinations in Criminal Investigations*. Final Report to the National Institute of Justice, Grant Number 85-IJ-CX-0400, Salt Lake City: University of Utah, Department of Psychology.

Rovner, L. I. (1986). The accuracy of physiological detection of deception for subjects with prior knowledge. *Polygraph, 15*, 1-39.

Searls, D. (1991, March). Data Analysis Workshop. Sponsored by the Neuroscience/Medical Education and Research Services of the University of North Dakota, Grand Forks.

Simon, J. L., & Bruce, P. C. (1991). *Resampling stats, probability and statistics a radically different way: User guide*. Arlington, VA: Resampling Stats.

Simon, J. L., Puig, C., & Bruce, P. C. (1991). *Resampling stats, version 2.O*. Arlington, VA: Resampling Stats.

Thompson, P. A. (1991). Resampling approaches to complex psychological experiments. *Multivariate Behavioral Research, 26*, 737-763.

Wasserman, S., & Brockenholt, U. (1989). Bootstrapping: Applications to psychophysiology. *Psychophysiology. 26*, 208-221.

Wiggins, J. S. (1981). Clinical and statistical prediction: Where do we go from here? *Clinical Psychological Review, 1*, 3-18.

Acknowledgments

The authors would like to thank David Raskin, John Kircher and Steve Horowitz for allowing the use of their data in this project. John Kircher deserves special thanks for his work in converting the data collected at the University of Utah from DEC System format to CAPS format so that it could be used in this study. We would also like to acknowledge and thank John Podlesny for first calling our attention to Bradley Efron's original work in developing the bootstrapping technique. Finally, we would like to thank Hank Slotnick for bringing Don Searls to the University of North Dakota. It was Dr. Searls' presentation at the University of North Dakota in the spring of 1991 that started us down the path toward this project.

The research depicted in this report was sponsored by the Department of the Navy, Office of the Chief of Naval Research. The contents of this report do not necessarily reflect the position or the policy of the United States government, and no official endorsement should be inferred.

UNIVERSITY OF NORTH DAKOTA