A Study of the Validity of Polygraph Examinations in Criminal Investigation

Final Report to the National Institute of Justice

Grant No. 85-IJ-CX-0040

David C. Raskin, Ph.D.

Principal Investigator

John C. Kircher, Ph.D.

Co-Principal Investigator

Charles R. Honts, Ph.D.

Research Associate

Steven W. Horowitz, M.S

Research Assistant

Department of Psychology

University of Utah

Salt Lake City, Utah 84112

May 1988

Abstract

This project was designed to answer many of major questions concerning the validity of the control question polygraph technique for assessing truth and deception in criminal investigations. Confirmed and unconfirmed polygraph charts from examinations conducted by the U. S. Secret Service in criminal investigations were sampled and blindly interpreted by six polygraph examiners from that agency and one psychophysiologist at the University of Utah, and they were also subjected to computer interpretation using algorithms developed at the University of Utah.

The accuracy of human and computer interpretations was very high. Decisions by the original examiners on individual relevant questions ranged from 91-96% correct on confirmed truthful answers and 85-95% correct on confirmed deceptive answers. Blind interpretation produced somewhat lower accuracies, ranging from 63-85% on truthful answers and 84-94% on deceptive answers. However, the accuracy of the computer interpretations was higher than the blind interpretations, and it ranged from 95-96% on confirmed truthful suspects and 83-96% on confirmed deceptive subjects. The results provide considerable support for the accuracy of decisions made by the original examiners and for the use of computer interpretations for quality control of decisions concerning the outcomes of polygraph tests.

The generalizability of laboratory research on control question polygraph tests was analyzed using computer-generated response profiles and double cross-validation of models developed from laboratory and criminal suspects. The results indicated that laboratory findings may provide considerable information about the underlying processes and accuracy of field polygraph examinations. They also indicated a need to improve the choice of relevant questions in multiple issue testing and a need for modifications to improve the accuracy of field numerical evaluation.

Introduction

Although the use of polygraph examinations in criminal investigations and security applications by the Federal Government more than tripled during a 10-year period, there appears to be a lack of adequate scientific research on the accuracy of such field applications (Office of Technology Assessment, 1983). The OTA study was mandated by the House Committee on Government Operations, and it provided an extensive review of the existing literature on polygraph research and applications. It concluded that although there is evidence that polygraph accuracy exceeds chance in field applications, there is a strong need for further research.

Every federal investigative agency, including those within the Department of Defense, uses polygraph examinations in criminal investigations (OTA, 1983). State and local law enforcement agencies, courts, and attorneys make extensive use of such techniques to screen suspects, to dispose of cases, to elicit confessions following deceptive results, to generate evidence for court proceedings, to provide information for pre-sentence investigations, and for various other applications within the criminal justice system. The extent to which these applications provide valid information and the weight that should be accorded to such results in various contexts are hotly debated issues (Lykken, 1981; OTA, 1983; Raskin, 1982, 1986). The OTA report highlighted the pressing need for additional research on this problem. In response to concerns expressed in their report, this project was designed to provide information that is crucial to enlightened decisions regarding the range of useful applications of polygraph techniques in the criminal justice system and ways to improve existing techniques.

Objectives of the Research Project

The first objective of this project was to provide a definitive study of the validity of control question polygraph examinations in criminal investigation and to provide reliable estimates of the accuracy of truthful and deceptive outcomes. The research was designed to generate important data that will be useful in guiding policy decisions in different settings, such as the extent to which polygraph tests should be used in different contexts and the amount of confidence that can be placed in the outcomes of such tests.

The second objective of this project was to assess the performance of polygraph examiners with different educational backgrounds and different types and amounts of experience with polygraph techniques. The analytic techniques that we applied to the data provided information about the qualitative and quantitative differences in the ability of different polygraph examiners to interpret polygraph recordings accurately.

The third objective of the project was to assess the efficacy of an automatic and objective computer method for interpreting the outcomes of polygraph examinations. At the present time, most federal investigative agencies have quality control procedures that require materials from all polygraph examinations conducted in the field to be sent to a central office for an independent evaluation before the results receive final approval. Independent evaluations are intended to minimize mistakes in interpretations caused by subjective influences or insufficient skill or experience, and they are also used to identify examiners in the field who are experiencing difficulties in their performance. However, the current procedures are slow and costly and may not solve all of the operational problems. Computer analysis might perform better than independent human

interpreters and be less costly in terms of time and resources.

Research has established that there is wide variation in the abilities of polygraph examiners to interpret correctly the physiological recordings obtained in such tests (Raskin, Barland, & Podlesny, 1978), and computer methods hatves been demonstrated to perform as well as the most experienced and sophisticated human interpreters (Kircher & Raskin, 1988). If a computer method could provide the same information as that obtained from human interpreters and at a significantly lower cost and within minutes instead of days or weeks, problems could be identified more readily and with greater speed. All subjectivity would be removed from the process, more accurate decisions would be available immediately, examiners in the field would receive immediate feedback that they could consider before the examination is terminated, polygraph examination results could be utilized in a more effective manner, and additional training could be provided on the basis of the computer identification of particular examiner deficiencies. The entire process could benefit from a powerful, rapid, and scientific approach to diagnosis of truth and deception.

The fourth objective of the project was to assess the extent to which laboratory mock-crime experiments provide information and results that have implications for field applications of polygraph examinations. Although a large amount of scientific laboratory research has investigated problems such as the influence of personality factors, the effectiveness of countermeasures and drugs, the usefulness of different physiological measures and examination techniques, and the accuracy of control question polygraph examinations (Raskin, 1986); the extent to which the results of such studies can be generalized to field applications of polygraph techniques is not entirely clear. The use of computer analystic techniques may provide information concerning the extent to which the findings of

laboratory research can be utilized in making decisions and formulating policy regarding applications of polygraph techniques in the criminal justice system.

Methodological Issues

In order to assess the accuracy of control question polygraph tests in criminal investigation, a reliable criterion of ground truth must be available against which the test results may be evaluated. Complete confidence in the criterion can be obtained using laboratory simulations that employ mock crimes and field polygraph techniques (Raskin, 1982). The results of such experiments have frequently produced accuracies in excess of 90% (Bradley & Ainsworth, 1984; Dawson, 1980; Gatchel, et al., 1983; Kircher & Raskin, 1988; Podlesny & Raskin, 1978; Raskin & Hare, 1978; Rovner, Raskin, & Kircher, 1979). However, critics of laboratory research have argued that the motivational structure of the field situation cannot be simulated in the laboratory and the greater consequences of the test outcomes in criminal investigations produce different physiological reactions and higher rates of error (Lykken, 1981).

If the possible problems of motivation and context inherent in the laboratory simulations are to be overcome, it is necessary to use examinations from actual criminal investigations. On the basis of 10 studies that met minimal criteria for methodological adequacy, OTA concluded that the average accuracy of polygraph tests in the field situation is 90% on guilty subjects and 80% on innocent subjects. However, these studies raise several additional problems.

A criterion for ground truth is more difficult to establish in the field situation than in the laboratory, and it is necessary to develop criteria with a high degree of reliability and accuracy. Three approaches

have been taken to that problem. One method is to submit all of the case information except the polygraph results to a panel of experts who are asked to make judgments of guilt or innocence using the available information and disregarding legal technicalities (Bersh, 1969; Raskin, Barland, & Podlesny, 1978). Accuracy of the polygraph tests is then determined by comparing the test outcomes to the composite judgments of the panel. The problem with this method is the fallibility of panel judgments that are based on a vague evaluation of evidence of unknown and variable quality and quantity. Therefore, the findings of panel studies of polygraph accuracy are open to serious question. Similar and more severe problems arise when polygraph accuracy is assessed against a criterion of judicial outcomes (Raskin, 1982, 1987).

It is generally agreed that the best criterion for assessing the accuracy of field polygraph tests is confirmation by means of confessions by guilty persons (Horvath, 1977; Lykken, 1979; Raskin, 1987). In such studies, polygraph charts are obtained from cases in which the guilty person subsequently confessed. Sets of such confirmed deceptive and confirmed truthful polygraph charts are assembled, and they are then submitted to other polygraph examiners for blind interpretation. The accuracy of their interpretations is assessed against the criterion of ground truth independently established by the confessions. The accuracies reported by such studies range from 64% in the Horvath study (1977) to 98% in the Raskin study (1976).

Only limited conclusions can be drawn from the available field studies that have used a confession criterion. Questions have been raised about the method of selecting the charts to be evaluated, their representativeness with regard to polygraph tests in general, and the training and qualifications of the polygraph interpreters (Raskin, 1987).

For example, the Horvath study included examinations of victims and witnesses as well as suspects (Barland, 1982), which complicates the interpretation of the results. In Barland's re-analysis of Horvath's data, he found that all but one of the false positive errors occurred on victims and witnesses, indicating that the Horvath study cannot be used to estimate the accuracy of polygraph tests on criminal suspects.

Another major problem with the Horvath study and those from the Reid organization (Horvath & Reid, 1971; Hunter & Ash, 1973; Kleinmuntz & Szucko, 1982; Slowik & Buckley, 1975; Wicklander & Hunter, 1975) is the failure to use control question techniques that are accepted by most federal agencies and supported by scientific research (Raskin, 1987). Furthermore, the interpreters in these studies were not adequately trained or experienced in the use of numerical evaluation of polygraph charts, and only the Horvath study even attempted to employ numerical methods. All of the examiners and interpreters in these studies were trained in a method that involves the observation and utilization of so-called "behavior symptoms" in the diagnosis of truth and deception. Such methods have been shown to be useless for diagnosing truth and deception, and they produce lower rates of accuracy than numerical interpretation (Raskin et al., 1978). The Reid studies also suffer from the additional weakness of having used cases where employers referred their employees for polygraph tests with no option to decline to take the test, and the Kleinmuntz and Szucko study did not even use qualified polygraph examiners or accepted methods of chart interpretation.

Finally, there is the problem of case selection and the generalizability of the results of validity studies based on confession criteria. In addition to the above problems, which indicate that many

studies used cases that are not representative of polygraph examinations on criminal suspects, the methods used to select the cases in the Reid studies have not been specified in a manner that permits a definite evaluation of the whether or not the cases were selected in an unbiased manner (Raskin, 1987). Even if these problems did not exist, there is a more fundamental problem with the use of cases confirmed by confessions. If tests are selected because someone (either the person who took the test or another suspect) confessed to the crime subsequent to the polygraph test, there arises the question of whether or not such tests are representative of the population of tests of suspects who agree to take polygraph tests in connection with a criminal investigation.

In all but one of the studies mentioned above, no data were presented concerning the proportion of cases resolved by confessions in the population of cases from which the data were drawn. Only the Raskin (1976) study provided that information, and it indicated a confession rate of only 17% in the set of tests from which the sample was drawn. It is possible that cases in which confessions are obtained are not representative of polygraph tests of criminal suspects in general, and the generalizability of the results of such studies is thereby limited.

It has been argued that only those subjects whose polygraph charts are most strongly indicative of deception are interrogated (Iacono, in press). Therefore, the resulting confessions may inflate reported accuracy by biasing the selection of charts selected in field validity studies. Subsequent blind interpretations of those charts are likely to produce correct deceptive decisions more frequently than would occur if subjects who produced weaker deceptive polygraph charts were also interrogated to attempt to obtain confessions. Therefore, we also performed analyses to determine if differences in the strength of physiological results

indicative of deception are obtained from suspects who were considered deceptive and subsequently confessed and from suspects who were considered deceptive and did not confess.

In order to overcome many of the methodological problems cited above, this project investigated the accuracy of the control question test in actual criminal investigations where standard field polygraph examination techniques were used and numerical evaluation was employed by adequately trained interpreters, including blind interpreters. The data also allowed us to assess the effectiveness of the computer methods to analyze polygraph charts for the automatic and objective diagnosis of truth and deception (Kircher & Raskin, 1988).

The use of computer algorithms and software and extensive multivariate statistical analyses made it possible to assess the relationships between polygraph recordings obtained in field examinations and the qualitative and quantitative nature of polygraph recordings obtained in mock crime laboratory experiments. These analyses provided the basis for estimating the extent to which laboratory research evokes emotional and physiological responses that are similar to those observed in the field situation. The obtained data allowed us to evaluate how far the results of laboratory research on polygraph techniques may be generalized to the application of polygraph examinations in the criminal investigation context.

Research Methods

Our initial objective was to obtain a sample consisting of the polygraph charts from 200 examinations conducted by a federal law enforcement agency. The U.S. Secret Service agreed to provide the materials from their files and the services of some of their examiners to participate in the study. The Secret Service was a logical choice for this study because they have a very high quality polygraph program with more than 20 experienced and well trained examiners (Raskin, 1984). They conduct in excess of 1,000 polygraph examinations per year in the context of criminal investigation, and they utilize standard control question procedures and numerical interpretations of the polygraph charts. Furthermore, OTA (1983) reported that they achieve very high rates of admissions and confessions that provide confirmation of more than 90% of their polygraph diagnoses. That high rate of confirmations would ensure that the results are not dependent on the selection of a small, non-representative sample of cases, a common problem in studies that rely on a confession criterion for establishing ground truth.

Cases were to be selected to provide 80 tests of suspects who were confirmed as deceptive at some time after their polygraph test and 80 tests of suspects subsequently confirmed as truthful. An additional sample of 20 unconfirmed deceptive results and 20 unconfirmed truthful results was also sought. A confirmed deceptive suspect is one who was examined on the polygraph and subsequently admitted having lied to one or more of the relevant questions that pertained to the crime under investigation. A confirmed truthful suspect is one who was examined on the polygraph and was later cleared of the allegation or suspicion by the admission or confession of another person. In this study, we required independent corroboration of the confession in the form of some type of physical evidence. Unconfirmed

results are those for which no admission or confession was obtained either to inculpate or exculpate the person who took the test. By including a sample of unconfirmed polygraph results, we were able to determine if there are qualitative or quantitative differences in the physiological reactions between cases in which deception was indicated by the polygraph charts and the polygraph subject confessed and those in which the charts indicated deception and no confession was obtained.

For each of the categories above, we planned to select tests so as to obtain half from cases where there was only one suspect and half where there was more than one suspect. That would have permitted us to determine if there are differences in outcomes when the examiner expects that at least one of the suspects will produce a truthful outcome (multiple-suspect cases) as compared to single-suspect cases where there would be a higher probability that the suspect is guilty. The resulting design of the sample was to be as follows:

	Confirmed Deceptive	Confirmed Truthful	Unconfirmed Deceptive	Unconfirmed Truthful
Single Suspect	40	40	10	10
Multiple Suspect	40	40	10	10

The polygraph charts were selected only on the basis of the type of case as described above. The decision regarding truth or deception made by the examiner and the quality or characteristics of the polygraph charts themselves played no role in the selection process, with two exceptions. First, the tests must have included at least three charts with a control question format. Tests that were incomplete in those respects were not used. Second, if there was an equipment malfunction or examiner error that rendered the charts technically unusable or incomplete, the examination was

not included in the sample. The cases were selected first, and the polygraph charts were then inspected to determine if they were to be retained or discarded for failure to meet the standards of completeness or technical adequacy.

Subject Selection

Three strategies were employed in selecting cases. Initially, the U.S. Secret Service case logs for all 1,757 polygraph examinations conducted during FY1983 and FY1984 were coded for type of case, examiner, pretest admissions, and posttest confessions and entered into a computer file. All of the examinations were then screened by a computer program that selected all cases with posttest admissions or confessions. The 241 cases selected by the computer program were then requested from the Washington, D. C. Headquarters of the Secret Service. The Secret Service personnel then requested the case files from the field offices where they were located. When they received the files, they removed all identifying information from the polygraph charts and recoded them with new identification numbers that we supplied. These recoded charts were taken from the case files and sent to the University of Utah. The case files without the polygraph charts were sent to the Secret Service Field Office in Salt Lake City, where they were evaluated by members of our University of Utah research team.

Confirmation of truthfulness or deception by the polygraph subject was based on a two-step criterion. The first step required an admission or confession by the subject who took the polygraph test or by another suspect in the case who either inculpated or exculpated the subject who was tested. The second step required that admissions and confessions be supported by independent evidence that corroborated the admission or confession, such as

recovering counterfeit notes or printing plates described in the confession, recovering the money stolen from a bank, or an analysis of the handwriting of a forged signature.

A very stringent criterion for confirmation was employed to increase the reliability and validity of the criterion so as to avoid errors in the subsequent analyses of the accuracy of the polygraph results and other types of analyses based on the confirmed polygraph results. The use of such a stringent criterion also made it more difficult to confirm cases that were otherwise confirmed by admissions or confessions. Therefore, it was difficult to fill all of the cells in the planned sample as described above. Although it appeared that the Secret Service had a lower rate of confirmation than that reported to OTA (1983), our stringent requirements for purposes of this research eliminated many cases that can reasonably be asssumed to have been confirmed for other purposes.

From this initial sampling, we obtained 127 sets of polygraph charts, 93 from multiple-suspect cases and 34 from single-suspect cases. Of the 93 multiple-suspect cases, 19 polygraph subjects were confirmed as having answered one or more relevant questions truthfully, 32 were confirmed as having answered one or more relevant questions deceptively, 7 were confirmed as having answered at least one relevant question truthfully and at least one relevant question deceptively in the same test, and 35 were not confirmed on any question. Of the 34 single-suspect cases, 14 polygraph subjects were confirmed as having answered one or more relevant questions deceptively, 4 were confirmed as having answered at least one relevant question truthfully and at least one relevant question deceptively in the same test, and the remaining 16 subjects were not confirmed on any question. We obtained no confirmed truthful single-suspect subjects from this initial sampling.

In order to increase the likelihood of obtaining confirmed truthful subjects, we used another approach of requesting all cases with pretest as well as posttest admissions and/or confessions. Cases from the first six months of FY1985 were coded as described above, and cases were then selected by a computer program. Of the 325 cases examined, 95 were selected by the computer program and requested from the Secret Service. Materials from those cases were sent to Salt Lake City in the same manner as previously described. Only charts from subjects needed to fill incomplete group categories were selected and coded.

From this second sampling we selected 32 multiple-suspect subjects and 5 single-suspect subjects. Of these 32 multiple-suspect subjects, 14 were confirmed as having answered one or more relevant questions truthfully, 11 were confirmed as having answered one or more relevant questions deceptively, 1 was confirmed as having answered at least one relevant question truthfully and at least one relevant question deceptively in the same test, and 6 were not confirmed on any question. Of the 5 single-suspect subjects, 4 were confirmed as having answered one or more relevant questions deceptively and 1 was confirmed as having answered at least one relevant question truthfully and at least one relevant question deceptively in the same test. Again, we obtained no confirmed truthful single-suspect subjects.

The third strategy obtained an exhaustive sample of multiple-suspect cases. By this time it was clear that it was not possible to fill the confirmed-truthful, single-suspect category, so we concentrated on trying to fill the multiple-suspect cells. We hoped that an exhaustive sample of all multiple-suspect cases would enable us to obtain additional confirmed truthful subjects. From the 440 cases from the first six months of FY1986,

we selected all of the 35 multiple-suspect cases and requested them from the Secret Service. Materials from those cases were sent to Salt Lake City in the same manner as previously described. From this third sample, we obtained 12 subjects, 6 of whom who were confirmed as having answered one or more relevant questions truthfully, 5 who were confirmed as having answered one or more relevant questions deceptively, and 1 who was confirmed as having answered at least one relevant question truthfully and at least one relevant question deceptively in the same test.

The polygraph charts obtained from the total of 176 cases from the three samples consisted of 39 subjects confirmed to have answered one or more relevant questions truthfully, 66 subjects confirmed to have answered one or more relevant questions deceptively, 14 subjects confirmed to have answered at least one relevant question truthfully and at least one relevant question deceptively on the same test, and 57 subjects who were not confirmed on any questions.

Blind Interpretations

Blind interpretations were conducted by seven interpreters, six of whom were U. S. Secret Service polygraph examiners who had been trained at the U. S. Army Military Police School. Of the Secret Service examiners, two were experienced examiners who performed quality control evaluations at their Washington, D. C. headquarters (quality control), two were stationed at field offices and had more than one year of experience as polygraph examiners (experienced examiners), and two were stationed in field offices and had less than one year of experience as examiners (inexperienced examiners). The other interpreter was a doctoral level psychophysiologist who had been licensed as a polygraph examiner for 10 years.

One hundred of the obtained cases were selected for scoring by the seven blind interpreters using random processes to fill three categories.

Forty deceptive subjects were selected from the total sample of subjects confirmed to have answered at least one relevant question deceptively, but not confirmed to have answered any relevant question truthfully. The 13 subjects confirmed to have answered at least one relevant question truthfully and at least one relevant question deceptively were coded as truthful subjects and were combined with the other subjects confirmed to have answered at least one question truthfully. Forty subjects were selected at random from this population of truthful subjects. The random procedure resulted in the selection of 13 of the 14 subjects who had been confirmed to have answered at least one relevant question truthfully and at least one relevant question deceptively. Twenty subjects were selected randomly from the sample of unconfirmed cases.

After the charts had been blindly interpreted, it was discovered that 1 confirmed truthful and 3 confirmed deceptive subjects did not meet the criteria for selection because their polygraph results were from a second test. Therefore, they were discarded from the sample and could not be replaced. That reduced the sample to 26 subjects confirmed to have answered one or more relevant questions truthfully, 37 subjects confirmed to have answered one or more relevant questions deceptively, 13 subjects confirmed to have answered at least one relevant question truthfully and at least one relevant question deceptively in the same test, and 20 unconfirmed subjects.

Division of Cases for Analysis

The cases appeared to belong to three natural categories of verification. Complete Verification occurred when responses to all relevant questions in an examination were confirmed as either truthful or deceptive. Partial Verification occurred when responses to some relevant

questions in an examination were confirmed as either truthful or deceptive, but there was also at least one response to a relevant question that remained unconfirmed. Mixed Verification occurred when suspects were confirmed to have answered at least one relevant question truthfully and at least one question deceptively within the same polygraph examination. Subjects were initially separated into these three categories of verification for purposes of data analysis.

Numerical Scoring

The original examiners and the Secret Service interpreters used the numerical scoring system developed and taught to federal polygraph examiners at the U. S. Army Military Police School. The psychophysiologist used the numerical scoring system developed and validated at the University of Utah. Although, the psychophysiologist used a different numerical scoring system than the other interpreters, differences in effectiveness of these systems are slight (Weaver, 1985). In general, both numerical scoring systems follow the scoring system described by Raskin and Hare (1978) and Podlesny and Raskin (1978). Differences in physiological reactions to relevant and control questions in electrodermal activity, respiration, peripheral vasomotor activity, and relative blood pressure were evaluated. The following characteristics were used to assess the strength of the responses: electrodermal response amplitude and duration; decrease in amplitude and rate of respiration, increases in respiration baseline; duration and amplitude of decreases in finger pulse amplitude, and amplitude and duration of baseline increase in relative blood pressure. Reactions were not scored if they began more than 5 seconds following the subject's answer. Minimum latencies of 0.5 second and 2.0 seconds were adopted for skin conductance and finger pulse amplitude responses, respectively, and reactions that began prior to the minimum latencies were

not scored.

For each physiological system, each pair of control and relevant questions was assigned a score from -3 to +3 (except by the two Secret Service quality control interpreters and one of the inexperienced Secret Service interpreters who elected to assign scores from -1 to +1) depending on the strength of the difference between the reactions to the two question types. Positive scores were assigned when reactions to control questions were stronger, negative scores were assigned when reactions to relevant questions were stronger, and scores of zero were assigned when the strength of reactions to relevant and control questions were approximately equal.

Computer Scoring

Data Entry.

The physiological data had been recorded at 2.5 mm per second on standard polygraph chart paper that was 20 cm in width. Physiological responses to each control and relevant question in the first three repetitions of the question sequence were manually traced on a digital tablet, the output of which was read by a laboratory microcomputer. The laboratory assistants who traced the response waveforms had no knowledge of the subjects' criterion status.

The computer was programmed to sample skin resistance (SR) and thoracic and abdominal respiration(R) channels at 10 Hz for 20 seconds following the onset of each test question. The program also read the times and levels of systolic and diastolic points of the blood pressure (BP) tracings. From the series of systolic and diastolic points for each question, average changes in BP were computed for 2 seconds immediately preceding the onset of question presentation and 20 seconds following question onset. The data for each chart were stored on a floppy disk in a

file identified by subject and chart numbers and date.

Data Editing.

A second program was written to read the data files from the floppy disks, display the physiological response waveforms on the computer screen, and edit movement artifacts. The editing program also rescaled the data when sensitivity adjustments had been made between charts. Artifacts of approximately 1-3 seconds in duration were replaced with interpolated values. A response containing multiple artifacts or artifacts greater than 3 seconds in duration was considered unusable and was not used.

Data Quantification.

The SR and BP response curves were divided into segments, and each segment was tested for positive slope. Approximate times of occurrence of low points in the waveform were identified by changes from zero or negative slope to positive slope. High points in the curve were isolated between successive pairs of low points. The exact times and levels of low points were then isolated between successive pairs of high points.

The procedures for locating high and low points in the SR and BP waveforms differed in two respects. Tests for positive slope were performed between successive samples (seconds) of the BP response curve and between every fifth sample (500 ms segments) of the SR response curve. In addition, a stepwise averaging procedure smoothed the SR response curve prior to testing the 500 ms intervals for positive slope. After the approximate times of low points in the response curve had been identified in those intervals, the exact times and levels of high and low points were isolated in the original sequence of 100 ms time samples.

The times and levels of high and low points in the response curves provided the information needed to quantify all of the physiological variables listed below, except respiration length that was quantified with

a separate algorithm (Timm, 1982). The first six of the following seven types of measurements were obtained from the SR and BP response waveforms:

Amplitude. Differences were computed between each low point and every succeeding high point identified in the response curve. Amplitude was defined as the greatest obtained difference.

Rise time. Time to the nearest 100 ms for SR and 1,000 ms for BP was measured between response onset and the occurrence of the maximum.

Half recovery time. Time of occurrence of the maximum was subtracted from the time at which the recovery limb reached a level that was half of the amplitude. When the response did not recover sufficiently to reach the criterion, the interval was measured to the end of the 20-second sampling period.

Rise rate. Amplitude was divided by rise time.

Half recovery rate. Half of amplitude was divided by half recovery time.

Latency to response onset. Time to the nearest 100 ms for SR and 1,000 ms for BP was measured from stimulus onset to response onset.

Respiration length. Linear distance was measured between successive pairs of 100-ms samples from question onset to the 10th poststimulus second. The 100 measurements were summed to yield a length measure in relative units for each respiration channel. After standardizing the measurements for the two respiration channels as described below, standard scores for the two channels were averaged to obtain a combined index of respiratory suppression (R Length) for each control and relevant question.

Variable Generation Procedures

For each subject and each response parameter, repeated measures were obtained across the control and relevant questions for the first three repetitions of the question sequence. The number of measurements depended on the number of control and relevant questions presented, and they ranged from four to eight per chart. The set of measurements for each response parameter was converted to standard scores. The transformation to standard scores within each subject established a common metric among the various types of response parameters. Since unit variance was partitioned among the repeated observations for each response parameter, it also controlled for the tendency of some individuals to react more strongly in one response system than in another.

The relative magnitudes of reactions to each relevant question were assessed separately for each response parameter. The mean standard score for repetitions of a given relevant question was subtracted from the mean standard score for reactions to all of the control questions on the test. The size of the Z-score difference indexed the magnitude of differential reactivity, and its sign indicated if the average response to the relevant question was greater or less than the average response to the control questions.

Variable Selection Procedures

Since it is difficult to obtain a stable prediction model from a large set of redundant measures (McNemar, 1969), three all-possible-subsets regression analyses (Pedhazur, 1982) were performed to identify a reliable subset of variables that was optimal for discriminating between truthful and deceptive responses to relevant questions. The first regression analysis was performed using only those cases in which all answers to relevant questions had been confirmed as either truthful or deceptive

(Complete Verification). The second regression analysis was conducted using only those cases in which some but not all answers to relevant questions were confirmed as either truthful or deceptive (Partial Verification). The Complete and Partial Verification samples were combined for the third analysis. Cases in which some answers to relevant questions had been confirmed as truthful and others had been confirmed as deceptive (Mixed Verification) were not included in these preliminary analyses.

The best subset of variables for discriminating between confirmed truthful and deceptive subjects in the Complete Verification sample consisted of four variables: SR Amplitude, SR Rise Rate, BP Amplitude, and R Length. The same set of four variables was the seventh best subset with four variables for the Partial Verification sample of subjects, but three of the four measures appeared as the best subset of three variables for that sample. When the Complete and Partial Verification samples were combined (Pure Sample), the four-variable model was again selected as optimal for discriminating between the groups. Therefore, the four-variable model was adopted for assessing the discriminant validity of the computer method.

Structure of the Probability Model

A probability-generating model was developed to calculate the probability of group membership for each subject. The probability of group membership was defined as the probability of truthfulness for a confirmed Truthful subject or the probability of deception for a confirmed Deceptive subject. Its complement, one minus the probability of group membership, was the probability that the subject was a member of the wrong criterion group.

The model consisted, in part, of a discriminant function that was used

to calculate a discriminant score for each subject. The discriminant score was a weighted combination of the subjects' scores on the four physiological variables. The weights for the variables were those that maximized the discrimination between confirmed truthful and deceptive individuals in the sample.

The model also incorporated two likelihood functions that were used to calculate the conditional probability of group membership given the obtained discriminant score. The two likelihood functions formed partially overlapping normal curves, the parameters of which were specified by the means and variances of the distributions of discriminant scores for confirmed truthful and deceptive subjects in the sample. To calculate the probability of group membership for a subject, two maximum likelihood estimates were computed using the subject's discriminant score and the equation for the normal probability density function (Winkler & Hays, 1975). The two likelihoods were then combined according to Bayes' Theorem to calculate the probability of group membership for each individual (Kircher & Raskin, 1988).

Results

Numerical Scores

Original Examiners

Differences in the numerical scores assigned by the original examiners for the three verification categories were tested by a 2-way ANOVA comprised of Confirmation (Truthful/Deceptive) and Verification (Complete/Partial/Mixed). The means for the 6 cells of the ANOVA are shown in Table 1.

Table 1

Original Examiners' Mean Scores for Confirmed Single Questions

	Verification		
	Complete	Partial	Mixed
Truthful	4.1	6.0	2.7
Deceptive	-5.6	-4.3	-2.8

The analysis indicated a main effect for Confirmation, F (1, 164) = 247.13, $p < 0.0001$. Positive numerical scores were associated with questions confirmed to have been answered truthfully, whereas negative numerical scores were associated with questions confirmed to have been answered deceptively. The analysis also indicated a significant Confirmation X Verification interaction, F (2, 164) = 5.35, $p = 0.006$. An examination of the means indicates that this effect was primarily due to a reduction in numerical scores for confirmed truthful responses in the Mixed Verification Group. A further ANOVA failed to find differences between the Complete and Partial Verification Groups, so the Complete and Partial Verification Groups were combined to form a Pure Verification Group that was then compared to the Mixed Verification Group. That ANOVA also

revealed a similar interaction of Confirmation and Verification, F (1, 166) = 8.90, p = 0.003.

The extent to which the original examiners' numerical scores predicted the truthful/deceptive criterion was assessed by correlating the numerical scores with the confirmation criterion for individual questions. For Pure Verification subjects, the correlation with the criterion was significant, $r(136) = 0.79$, $p < 0.001$. The correlation with the criterion was also significant for the Mixed Verification subjects, $r(33) = 0.61$, $p < 0.01$, but the correlation for the Mixed Verification subjects was significantly smaller than the correlation for the Pure Verification subjects, $z = 1.84$, $p = 0.03$ (one-tailed).

Blind Interpretations

Complete, Partial, and Mixed Verifications. Possible differences in numerical scores assigned by various blind interpreters for the three categories of Verification were assessed by a repeated measures ANOVA. An analysis of Interpreters by Confirmation (Truthful/Deceptive) by Verification (Complete/Partial/Mixed) indicated a significant main effect for Confirmation, F (1, 162) = 99.40, $p < 0.001$. The analysis failed to find a main effect for Verification, but there was a significant Confirmation X Verification interaction, F (2, 162) = 6.60, p = 0.002. Inspection of the means indicated that this interaction was primarily due to a reduction in the size of the numerical scores for confirmed truthful responses by subjects in the Mixed Verification Group (M = 0.41) as compared to confirmed truthful responses by subjects in the Complete (M = 2.20) and Partial (M = 2.68) Verification Groups. No interaction of Verification with Interpreters was found. A significant interaction of Interpreters and Confirmation was found, F (5, 830) = 3.26, p = 0.006, and it is discussed below in the section on Interpreter Characteristics.

An Interpreters by Confirmation by Verification (Complete/Partial) ANOVA was conducted to determine if there were differences between numerical scores for cases with Complete Verification and those with Partial Verification. This analysis indicated a significant main effect for Verification, F (1,122) = 4.04, p = 0.047. Inspection of the means indicated that for suspects with Complete Verification the numerical scores for individual questions tended to be more negative (M = -0.72) than the numerical scores to confirmed questions for suspects with only Partial Verification (M = -0.095). There was no significant interaction between Verification and Interpreters or Confirmation.

Since the difference in numerical scores for the Partial and Complete Verification was quite small, these groups were combined (Pure Verification) and compared to the Mixed Verification Group using an Interpreters by Confirmation (Truthful/Deceptive) by Verification (Pure/Mixed) ANOVA. This analysis indicated a strong main effect for Confirmation, F (1, 164) = 69.12, $p < 0.001$, and an interaction of Confirmation and Verification, F (1, 164) = p = 0.001. This effect was due to the reduction in the numerical scores for confirmed truthful responses in the Mixed Verification group (M = 0.41) as compared to the Pure Verification group (M = 2.33).

Reliability. All confirmed questions were used to assess interrater reliability in the assignments of scores, since ANOVA failed to indicate that Interpreters performed differently on the three Verification groups. A complete pairwise correlation matrix was calculated among the numerical scores assigned by the six Secret Service blind interpreters, and the interrater correlations were all significant, ranging from 0.80 to 0.88 (M = 0.84). The pairwise correlations between the scores of the

psychophysiologist and the Secret Service blind interpreters were also significant, ranging from 0.76 to 0.82 (M = 0.79).

Interpreter Characteristics. The numerical scores assigned by the six Secret Service blind interpreters were subjected to a Confirmation (Truthful/Deceptive) by Interpreter repeated measures ANOVA. This analysis indicated that the main effect of Interpreters was not significant, F (5, 830) = 1.59, but there was a significant interaction between Interpreters and Confirmation, F (5, 830) = 3.26, p = 0.006. The means for the six Secret Service blind interpreters shown in Table 2 indicate that the interaction of Interpreter and Confirmation was primarily due to lower scores assigned by the two quality control interpreters on confirmed truthful responses. This may have been a consequence of their use of scores of only +1, 0, and -1.

Table 2

Mean Numerical Scores on Individual Questions

and Correlations With The Criterion

for the Seven Blind Interpreters and The Original Examiners

	Confirmed Truthful	Confirmed Deceptive	Correlation With Criterion
Original Examiners	4.7	-4.8	0.79
Quality Control Examiner A	1.9	-3.1	0.62
Quality Control Examiner B	2.0	-3.4	0.64
Experienced Examiner A	3.0	-3.4	0.65
Experienced Examiner B	2.3	-3.3	0.57
Inexperienced Examiner A	2.2	-2.7	0.62
Inexperienced Examiner B	2.2	-3.6	0.62
Psychophysiologist	2.6	-4.8	0.66

The performance of the interpreters was further assessed by point-biserial correlations between the interpreters' numerical scores on individual questions and the Truthful/Deceptive criterion. These correlations are also shown in Table 2. The differences among interpreters appeared to be individual differences not associated with examiner experience. The best performance was shown by an experienced field examiner, r = 0.65, and the poorest performance was by the other experienced field examiner, r = 0.57. The difference between these two correlations was significant, t (190) = 5.01, p < 0.01. The inexperienced examiners performed at a level similar to that shown by the quality control evaluators, and the performance of the psychophysiologist was approximately midway between the best and poorest performance shown by the Secret Service examiners.

Accuracy of Outcomes

Decisions on individual questions using an inconclusive zone of +2 to -2 are shown in Table 3 for the original examiners and for the average of the six Secret Service blind interpreters. For Pure Verification subjects, the original examiners' were 77.6% correct, 3.6% incorrect, and 18.8% inconclusive, and the blind interpreters averaged 59.1% correct, 5.8% incorrect, and 35.1% inconclusive. The decision accuracy on individual questions for Mixed Verification subjects was poorer than for the Pure Verification subjects. For the original examiners, the overall accuracy was 95.5% for Pure Verification and only 87.5% for the Mixed Verification subjects. The overall accuracy of the blind interpreters averaged 90.5% for Pure Verification subjects and only 74.5% for Mixed Verification.

Table 3

Percent Accuracy on Individual Questions

for Original Examiners and Blind Interpreters

Pure Verification

	Truthful (N=26)					Deceptive (N=37)				
	(n)	C	W	?	Dec	(n)	C	W	?	Dec
Original Examiners	(62)	76	3	21	96	(76)	79	4	17	95
Blind Interpreters	(68)	52	9	39	85	(83)	65	4	31	94

Mixed Verification

	Truthful (N=13)					Deceptive (N=13)				
	(n)	C	W	?	Dec	(n)	C	W	?	Dec
Original Examiners	(15)	67	7	26	91	(20)	55	10	35	85
Blind Interpreters	(19)	29	17	54	63	(23)	47	9	43	84

It can be seen in Table 3 that the accuracy of decisions on confirmed truthful and deceptive answers differed as a function of type of verification, especially for the blind interpreters. For the original examiners, accuracy on questions answered deceptively was somewhat higher for Pure (95%) as compared to Mixed Verification (85%), and a similar pattern occurred on questions answered truthfully (Pure = 96% and Mixed = 91%). A stronger effect of verification type was observed for the blind interpreters. Again, accuracy of decisions on questions answered deceptively was somewhat higher for Pure (95%) as compared to Mixed Verification (84%). However, for questions answered truthfully there was a large drop in accuracy from 85% for Pure Verification to 63% for Mixed Verification subjects.

Comparison of Strength of Reactions by Confirmed and Unconfirmed Subjects

The magnitudes of numerical scores assigned to individual questions that yielded definite decisions (truthful or deceptive) were tested for possible differences between those decisions that were subsequently confirmed and those that were not confirmed. A 2-way ANOVA of Decision (Truthful/Deceptive) and Confirmation (Confirmed/Unconfirmed) was performed on the numerical scores that exceeded +2 or -2 assigned by the blind interpreters to the questions from the 100 cases, as described above. The mean numerical scores are shown in Table 4. ANOVA showed a significant main effect for Decisions, F (1, 212) = 1340.26, $p < 0.0001$. The main effect for Confirmation was not significant, F (1, 212) = 1.57, but the interaction of Decision and Confirmation approached significance, F (1, 212) = 3.84, $p = 0.051$. That was due to the slightly smaller scores for the Unconfirmed as compared to the Confirmed deceptive questions.

Table 4

Mean Numerical Scores for Blind Decisions on Confirmed and Unconfirmed Questions

	Confirmed	Unconfirmed
Truthful	5.9	5.7
Deceptive	-6.0	-4.9

Computer Analyses

Discriminant Validity

The discriminant validity of the computer method was initially assessed separately for the Complete, Mixed, and Partial Verification Groups. Subjects in the Mixed Verification Group had answered some of the

relevant questions truthfully and other relevant questions deceptively. For purposes of the analysis of cases with Mixed Verification, it was necessary to split the Mixed Group in half and assign confirmation group membership arbitrarily. When the subject was assigned to the Truthful group, only physiological responses to relevant questions confirmed to have been answered truthfully were included. Conversely, when the subject was assigned to the Deceptive group, only responses to relevant questions confirmed as having been answered deceptively were included in the analysis.

A discriminant function was computed for each verification group and was used to generate a discriminant score for each subject in that group. A subject was defined as correctly classified when the discriminant score yielded a probability of correct group membership that exceeded .50. If the probability was less than .50, the classification by the computer model was considered an error. Since it is known that a small subject-to-variable ratio causes discriminant analysis to capitalize on chance and produce inflated estimates of diagnostic validity (McNemar, 1969), standard statistical tests were also performed to assess the reliability of the findings. The results obtained for the three verification groups are presented in Table 5.

Table 5

Percent Correct Dichotomous Computer Classifications, Magnitude of Effect (R^2), and Tests of Statistical Significance (F) for Complete, Partial, Mixed, and Pure Verification Groups

	Percent Correct Classification				Statistics		
	(n)	Truthful	(n)	Deceptive	R^2	F	p
Complete	(17)	88.2	(13)	92.3	.79	24.09	<.0001
Mixed	(7)	85.7	(6)	83.3	.27	.73	ns
Partial	(9)	88.9	(24)	87.5	.56	9.01	<.0001
Pure	(26)	96.2	(37)	83.8	.62	23.91	<.0001

As shown in Table 5, the accuracy of the computer model was highest for cases with Complete Verification. In those cases, answers to all of the relevant questions had been confirmed as either Trutnful or Deceptive. A significant proportion of criterion variance was explained by the optimal linear combination of the four computer variables (R^2 = .79). The lowest accuracy was obtained for the Mixed Verification cases. Although the correct classifications in the Mixed Group exceeded 80%, it is clear that the result was unreliable since the F-ratio was not significant.

Complete versus Mixed Verification. A MANOVA with the four physiological parameters as dependent variables was performed to determine if the accuracies obtained for the Complete Verification Group differed significantly from those obtained for the Mixed Verification group. The MANOVA revealed that the Verification (Complete/Mixed) X Confirmation (Truth/Deception) interaction was significant, $F(4,36) = 2.69$, $p < .05$. The discrimination between truthful and deceptive answers was significantly better in Complete Verification cases than in Mixed Verification cases. This finding suggests that there are important differences between Complete

and Mixed Verification cases and that the two types of cases should be considered separately. A within-subjects MANOVA conducted using only Mixed Verification cases revealed that the physiological reactions associated with deceptive answers to relevant questions were not significantly stronger than those associated with truthful answers to relevant questions, $F(4,9) = 2.99$, $p = .08$.

Complete versus Partial Verification. MANOVA revealed no main effect for Complete versus Partial Verification Groups, $F(4,56) = 1.01$, and no evidence of a Verification X Confirmation interaction, $F(4,56) = .89$. Thus, cases in which answers to only some relevant questions were confirmed as either truthful or deceptive were indistinguishable from those in which answers to all relevant questions were confirmed as either truthful or deceptive. Since little would be gained from treating these two subgroups separately, they were pooled to form the Pure Verification sample for all subsequent analyses. The results obtained from the Pure sample are presented in the bottom row of Table 5.

Discriminant Validity in the Pure Verification Sample. Table 6 presents the percentage of correct truthful and deceptive decisions and inconclusives subjects in the Pure Verification sample as a function of various decision criteria. Withn the .50 cutoff, a correct decision was defined as a probability of correct group membership greater than .50, and an error occurred if the probability was less than .50. With the .90 cutoff, a correct decision was scored if the probability of correct group membership was .90 or greater; an error was scored if it was equal to or less than .10 ; and the result was inconclusive if the probability was between .90 and .10.

Table 6

Percent Correct Classifications and Inconclusives for Various Decision Criteria

	Probability Cutoffs for Decisions				
	.50	.60	.70	.80	.90
Truthful (n=26)	96	96	96	95	95
Deceptive (n=37)	84	83	93	93	96
Inconclusive	0	5	11	21	24

With the .50 cutoff, 96% of the Truthful and 84% of the Deceptive subjects were correctly classified, and there were no inconclusive outcomes since no probability was exactly .50. Predictably, there was a progressive increase in the percentage of inconclusive outcomes as the criterion for a definite truthful or deceptive diagnosis approached unity. Using the .90 criterion, 95% of the Truthful and 96% of the Deceptive subjects were correctly classified, and 15 of the 63 cases (24%) were inconclusive. Examination of the data in Table 6 suggests that an optimal cutoff to maximize the accuracy of decisions and minimize inconclusive outcomes is a probablity of approximately .70.

Relative Utility of Physiological Components. The univariate point-biserial correlations (r_{pb}) between each of the four physiological variables and the Truth/Deception criterion are presented in Table 7. This statistic provides a measure of the discriminant validity of each physiological parameter. Table 7 also presents the correlations between each of the physiological measures and the discriminant scores (structural coefficients). The structural coefficient for a variable indicates the extent to which the discriminant scores were dependent on changes in that

variable.

Table 7

Validity and Structural Coefficients for the Physiological Measures

	Validity Coefficient	Structural Coefficient
SR Amplitude	.73	.92
SR Rise Rate	.48	.61
BP Amplitude	.69	.87
R Length	-.39	-.49

It may be seen that SR Amplitude was clearly the most diagnostic measure, and it predicted over 53% of the criterion variance (r_{pb}^2). Not surprisingly, SR Amplitude was also correlated most highly with the discriminant scores. BP Amplitude was that next most diagnostic measure, followed by SR Rise Rate and R Length. The relative importance of the variables, as measured by the structural coefficients, followed a similar pattern.

Characteristics of Physiological Responses in Laboratory and Field Examinations

Profile analyses were performed to determine if there were reliable differences between physiological data obtained in laboratory simulations and data obtained from polygraph examinations conducted in the course of actual criminal investigations. The laboratory sample was composed of 26 Truthful and 37 Deceptive adult males randomly selected from a pool of 100 subjects who had participated in a previous mock crime experiment (Kircher & Raskin, 1988). The field subjects were the 26 confirmed Truthful and 37 confirmed Deceptive subjects in the Pure Verification sample. Field cases with Mixed Verification were excluded from the profile analyses because no

attempt had been made in the laboratory experiment to represent that condition.

The physiological measures for the profile analyses were obtained from subjects' electrodermal, cardiovascular, and respiration responses to control and relevant test questions. Although the procedures for recording blood pressure and respiration data in the laboratory and field settings were similar, different measures of electrodermal activity had been recorded. Specifically, skin conductance (SC) had been recorded in the laboratory, whereas skin resistance (SR) had been recorded in the field examinations. Although there is a well-defined, nonlinear relationship between SC and SR, the transformation from one to the other requires absolute measures of conductance and resistance that were not available for most of the field cases. Since the original units of measurement in the two data sets were not linearly related and it was not possible to transform the electrodermal measures to a common metric, any observed difference between laboratory and field measures of electrodermal activity was confounded with the method of measurement and should be viewed with caution.

Three physiological variables were selected for the profile analyses: SC or SR Amplitude, BP Amplitude, and R Length. These measures were selected because they comprised the largest subset of measures that had been independently, empirically, and consistently identified as diagnostic in the laboratory (Kircher & Raskin, 1988) and in the Pure Verification sample of field cases.

Parameter Standardization Procedures. In the above analyses, raw measurements of physiological reactions were transformed to Z-scores. However, for the profile analyses a Z-score transformation is inappropriate since the mean of a set of Z-scores is always zero. As a consequence, the

$\underline{Z}$-score for a reaction to one type of test question would necessarily be counterbalanced by a $\underline{Z}$-score of the same absolute magnitude but of opposite sign for the other type of question. The dependency introduced by use of a $\underline{Z}$-score transformation would preclude interpretation of differences in physiological response profiles associated with control and relevant test questions.

In order to establish a common metric among the three response variables, within-subject range-adjusted scores were computed separately for each physiological variable according to the following formula:

$$\underline{X}' = 100 * (\underline{X} - \underline{X}_{min}) / (\underline{X}_{max} - \underline{X}_{min})$$

where $\underline{X}$ was a raw score associated with one of the control or relevant questions in the first three repetitions of the question sequence; $\underline{X}_{max}$ was the greatest obtained score in the set of repeated measurements; $\underline{X}_{min}$ was the smallest obtained score in the same set; and $\underline{X}'$ was the range-adjusted value of $\underline{X}$. This transformation produced $\underline{X}' = 0$ for the smallest observed score in the original set of raw measurements for the subject ($\underline{X}_{min}$) and $\underline{X}' = 100$ for the greatest observed score for that subject ($\underline{X}_{max}$).

As noted by Nunnally (1978), the levels of response profiles are interpretable only when the variables are "pointed in the same direction" (p. 439). Since relatively strong physiological reactions yielded relatively high scores on the electrodermal and cardiovascular measures but low scores on the respiration measure, all measurements of R Length were reversed in sign prior to projecting the scores onto a standard scale of constant range.

For each subject, the mean of range-adjusted scores associated with each of the two types of test questions was calculated for each physiological measure. A single measure of R Length was obtained for each

question by averaging the means of the range-adjusted lengths of thoracic and abdominal respiration tracings. The mean reaction profiles for Truthful and Deceptive subjects in the laboratory and field samples are presented in Figure 1. The order of presentation of the three variables along the abscissa was arbitrary.

To examine possible differences among the response profiles exhibited by laboratory and field subjects, two independent sources of variance were assessed with MANOVA: differences in the levels of response profiles and differences in their shapes (Harris, 1975; Van Egeren, 1973). The level of a subject's response profile was the mean of the range-adjusted scores for the three physiological measures that comprised the profile. The level of a response profile may be viewed as a measure of the relative magnitude of generalized arousal associated with control or relevant questions. Observed differences between the shapes of response profiles would suggest qualitative differences in the patterns of physiological responses associated with particular questions (Control or Relevant), criterion status (Truthful or Deceptive), or context for the examination (Laboratory or Field).

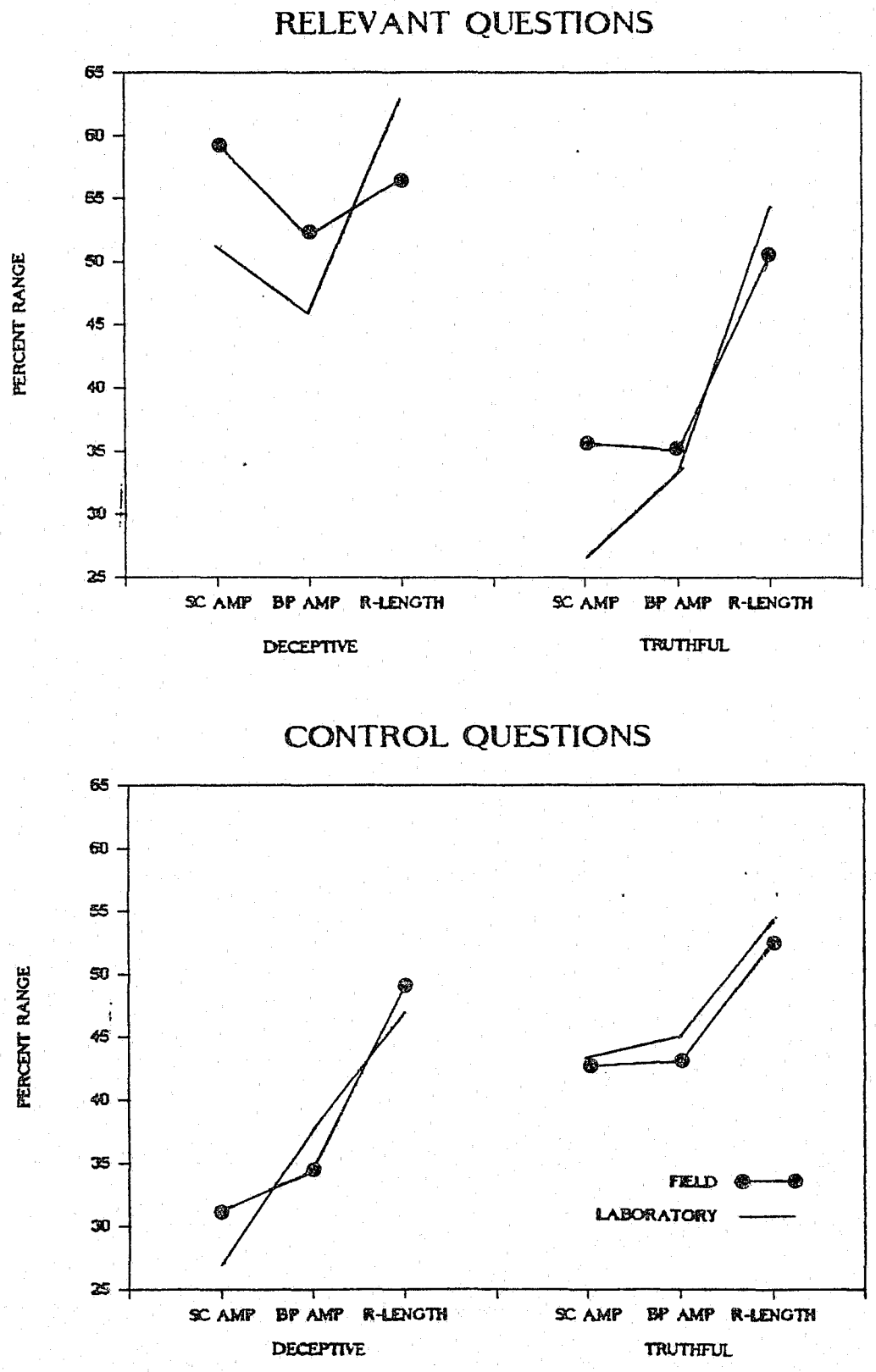

Figure 1. Mean profiles of physiological reactions of Truthful and Deceptive Laboratory and Field Subjects

Simple-effects MANOVAs were performed to compare the responses of laboratory and field subjects separately for Control and Relevant questions and for Truthful and Deceptive subjects. The results of the profile analyses are summarized in Table 8.

Table 8

Multivariate Comparisons of Response Profiles for Laboratory and Field Subjects

	Control Questions	Relevant Questions
Truthful (n=26)		
Profile level	$F(1,122) = .53$	$F(1,122) = 1.12$
Profile shape	$F(2,121) = .05$	$F(2,121) = 2.69$
Deceptive (n=37)		
Profile level	$F(1,122) = .36$	$F(1,122) = 1.73$
Profile shape	$F(2,121) = 2.17$	$F(2,121) = 5.71$

Among all comparisons of levels and shapes of response profiles produced by laboratory and field subjects, only one significant effect was observed. This was a significant difference between laboratory and field subjects in the shapes of their response patterns associated with deceptive answers to relevant questions, ($p < .01$). In order to assess the magnitude of this effect, a discriminant analysis was performed between the laboratory and field samples using the level-adjusted profiles for physiological responses to relevant questions answered deceptively. Level-adjusted scores were obtained for each subject and each response variable by subtracting the mean of the three scores that comprised a profile from each variable in that profile. The differences between laboratory and field subjects accounted for 9.8% of the variance in the shapes of these

profiles. By comparison, differences between Truthful and Deceptive subjects accounted for 56.9% of the variance in the physiological measures. In other words, the differences between Truthful and Deceptive subjects accounted for almost six times the amount of variance in physiological responses associated with the differences between the laboratory and field subjects.

The laboratory-field differences between the shapes of subjects' response profiles associated with deceptive answers to relevant questions were examined in greater detail by performing separate univariate tests using level-adjusted scores for the three physiological measures. Univariate tests revealed that the significant effect for profile shape was due to differences in the SR/SC Amplitude, $F(1,122) = 4.49$, $p < .04$, and R Length measures, $F(1,122) = 11.41$, $p < .001$. Level-adjusted scores on BP Amplitude did not distinguish between the groups, $F(1,122) = 1.97$.

Double Cross-Validation. Separate discriminant functions were developed from the 63 subjects in the Pure Verification sample (37 confirmed Deceptive and 26 confirmed Truthful) and from 50 Guilty and 50 Innocent subjects who had participated in a mock crime experiment (Kircher & Raskin, 1988). Each discriminant function was used to classify the subjects in the sample on which it was developed and also the subjects in the other sample. The discriminant functions developed from the laboratory and field samples incorporated the same variables, SC or SR Amplitude, BP Amplitude, and R Length. Generalizability from laboratory to field and vice-versa was first assessed by comparing the accuracy of classification made by each model when applied to the data from laboratory and field samples. Classification accuracies were calculated by comparing the actual status of each subject with the computer-generated probability of group

membership using a dichotomous decision rule that defined a correct decision as a probability of correct group membership that exceeded .50, and defined an error as a probability of correct group membership that was less than .50. The results are presented in Table 9.

Table 9

Accuracy of Classifications Based on Laboratory and Field Models

Laboratory Model	Classification		
Laboratory Sample	Deceptive	Truthful	% Correct
Deceptive	45	5	90
Truthful	6	44	88
Field Sample			
Deceptive	34	3	92
Truthful	6	20	77
Field Model			
Field Sample	Deceptive	Truthful	% Correct
Deceptive	31	6	84
Truthful	2	24	92
Laboratory Sample			
Deceptive	38	12	76
Truthful	1	49	98

The results indicated that each model performed similarly when applied to the two samples. Thus, the accuracy of the laboratory model was approximately the same when applied to the original sample of laboratory subjects and to the validation sample of field subjects. Similarly, the accuracy of the field model was approximately the same when applied to the

original sample of field subjects and to the validation sample of laboratory subjects. However, it should be noted that the laboratory model showed a drop in performance on Truthful subjects when applied to the field subjects (88% versus 77%), and the field model showed a drop in performance on Deceptive subjects when applied to the laboratory subjects (84% versus 76%).

The laboratory and field results were also compared by calculating univariate point-biserial correlations with the criterion (validity coefficients) and multivariate structural coefficients for the physiological variables used in the two models. The validity coefficients and structural coefficients for the laboratory and field samples are shown in Table 10.

Table 10

Validity and Structural Coefficients

for Laboratory and Field Samples

	Validity Coefficients		Structural Coefficients	
	Laboratory	Field	Laboratory	Field
SC/SR Amplitude	.77	.73	.94	.92
BP Amplitude	.61	.69	.74	.87
R Length	.55	.39	.67	.49

The validity and structural coefficients were similar for laboratory and field samples. These findings suggest that the relationships among the physiological variables obtained from polygraph tests of subjects in mock crime laboratory experiments are similar to those obtained from suspects in field polygraph tests. However, correlational analyses are not sensitive to differences in the means of the variables obtained from laboratory and

field subjects, and the analyses of classification accuracies presented in Table 9 suggest that mean differential physiological reactivity for Deceptive and Truthful subjects may not be symmetrical around zero in both samples. The findings that the laboratory model showed a drop in accuracy on Truthful field subjects and the field model showed a drop in accuracy on Deceptive laboratory subjects may indicate such asymmetry.

In order to examine the possibility of a lack of symmetry in the means of the differential physiological reactivity of laboratory and field subjects, the means of the computer-generated indices of differential physiological reactivity to relevant and control questions were calculated for Truthful and Deceptive laboratory and field subjects and are presented in Table 11.

Table 11

Computer Indices of Differential Reactivity

to Control and Relevant Questions for Laboratory and Field Subjects

	Laboratory		Field	
	Truthful	Deceptive	Truthful	Deceptive
SC/SR Amplitude	1.89	-2.41	.67	-2.95
BP Amplitude	1.53	-.93	.88	-2.02
R Length	.25	-1.64	.31	-1.07

Truthful laboratory and field subjects reacted more strongly to control than to relevant questions for all three physiological indices (positive means), and the Deceptive laboratory and field subjects responded more strongly to relevant than to control questions (negative means). However, the means for Truthful and Deceptive laboratory subjects were approximately equidistant from zero, whereas the means for the field sample

were generally shifted in the negative direction. Deceptive field subjects showed stronger differential reactivity to relevant questions than did Deceptive laboratory subjects, and Truthful field subjects showed weaker differential reactivity to control questions than did laboratory subjects.

Since the means for Truthful and Deceptive laboratory subjects are approximately symmetrical around zero, the model derived from those data "expects" that Truthful subjects will produce differential reactions to control questions as strong as those produced by Deceptive subjects to relevant questions. Since Truthful field subjects did not show that pattern to the same degree, there was a fairly high rate of false positive errors when the laboratory model was applied to the field subjects. On the other hand, the laboratory model "expects" only moderately strong differential reactions to relevant questions from Deceptive subjects. Since Deceptive field subjects showed much stronger differential reactions to relevant questions than to control questions, the laboratory model produced very few false negative errors when applied to field subjects. These results suggest that computer models developed on laboratory subjects are biased against Truthful field subjects, and they also suggest modifications of the decision cutoffs for numerical scoring based on the results of laboratory experiments. It appears that the cutoffs should be asymmetrical and shifted in the negative direction.

Human Versus Computer Scoring (Lens Model Analyses)

The subjects used in the lens model analyses were the Secret Service examiners who had conducted the polygraph examinations (Original Examiners) the six Secret Service examiners and one psychophysiologist who independently interpreted the polygraph charts. Only judgments made on examinees in the Pure Verification sample were included in the lens model

analyses. To facilitate comparisons among the polygraph interpreters, a forced-choice decision rule was adopted to produce an equal number of decisions for each interpreter. For confirmed relevant questions any positive total numerical score was considered a truthful outcome and any negative total score was considered a deceptive outcome. The physiological measures used to predict the criterion were the four parameters identified by the previous all-possible-subsets regression analyses as the subset that best discriminated between the Truthful and Deceptive subjects in the Pure Verification sample.

Brunswik's lens model (Slovic & Lichtenstein, 1971) was used to compare the performance of the blind numerical interpreters and the computer. The lens model was also used to examine possible differences among the polygraph examiners in their use of information from the polygraph charts to diagnose truth and deception. For the present problem, the lens model organized three sources of information and the relationships among them, as illustrated in Figure 2.

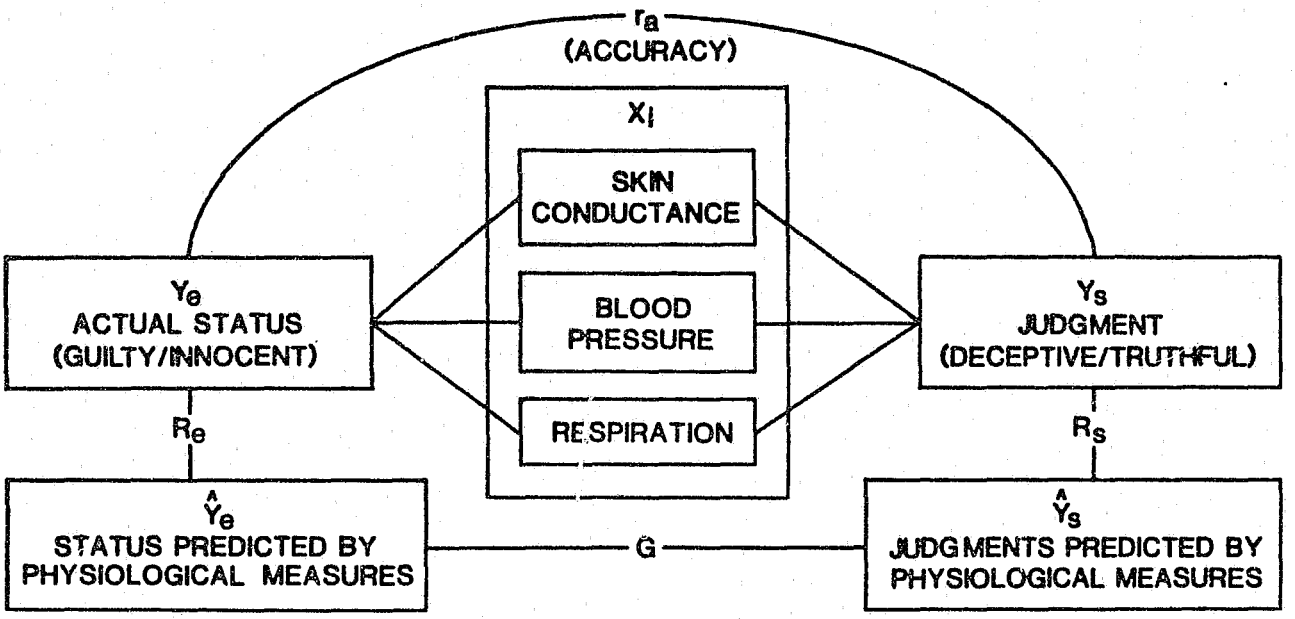

Figure 2. The Lens Model

As shown on the left side of Figure 2, the statistically optimal classification strategy is operationally defined in terms of a multiple regression equation that predicts the actual deceptive status of an individual ($\underline{Y}_e$) by means of a linear combination of weighted physiological measures or cues ($\underline{X}_i$). The subscript $\underline{e}$ in the lens model stands for the environment, which is the criterion of truth or deception. The obtained multiple correlation $\underline{R}_e$ provides a measure of the validity of the combination of physiological measures for predicting group membership.

The decision policy of the polygraph interpreter is represented on the right side of Figure 2 by the regression of diagnoses of truth and deception ($\underline{Y}_s$) on the multiple physiological measures ($\underline{X}_i$). The subscript $\underline{s}$ refers to the polygraph interpreter who served as the subject of the lens model analysis. The obtained multiple correlation $\underline{R}_s$ measures the extent to which the interpreter used information that was contained in the computer-generated physiological variables in making his decisions.

The correlation between the interpreter's decisions ($\underline{Y}_s$) and the criterion ($\underline{Y}_e$) provides a measure of achievement ($\underline{r}_a$). This correlation is the most important component of the lens model since the magnitude of $\underline{r}_a$ indicates how well the interpreter discriminated between guilty and innocent subjects on the basis of his blind evaluations of the polygraph charts. According to Tucker (1964), the relationship between achievement ($\underline{r}_a$) and other components of the lens model can be represented in terms of the following equation:

$$\underline{r}_a = \underline{G}\,\underline{R}_e\,\underline{R}_s + \underline{C}\sqrt{(1 - \underline{R}_e^2)}\sqrt{(1 - \underline{R}_s^2)}$$

where $\underline{G}$ is the correlation between the predicted criterion scores ($\hat{\underline{Y}}_e$) and the predicted decisions by the interpreter ($\hat{\underline{Y}}_s$), and $\underline{C}$ is the correlation between the residuals ($\underline{Y}_e - \hat{\underline{Y}}_e$) and ($\underline{Y}_s - \hat{\underline{Y}}_s$). Since both sets of predictions were made from the same physiological measures, the magnitude

of G specifies the degree of similarity between the model used to predict group membership and the model used to predict decisions. Conceptually, G specifies how closely the interpreter's use of information contained in the physiological measures generated by the computer matched the optimal linear combination of these variables. The C component represents the degree to which errors in predicting the criterion from the physiological measures were correlated with errors in predicting examiner judgments. The magnitude of C may be taken as a measure of the amount of diagnostic information available in the physiological recordings that was used by the blind interpreter to make valid diagnoses but was not contained in the four features of response waveforms that were quantified by the computer. Therefore, C provides an index of the extent to which the computer failed to use diagnostic information available in the physiological recordings that was effectively used by the human interpreters.

The results of the lens model analysis are presented in Table 12. The interpreters are listed in order of their achievement coefficients (r_a), which ranged between .53 and .87, with a mean of .76. On the average, human judgments based on numerical evaluations of the polygraph charts accounted for approximately 58% of the criterion variance. The multiple correlation between the physiological variables and the criterion (R_e) provided an overall estimate of the validity of the combination of the physiological measures for diagnosing truth and deception. The optimal linear combination of physiological measures produced a multiple correlation of .79 and accounted for 63% of the criterion variance. The average level of discrimination between Truthful and Deceptive subjects achieved by the human interpreters was slightly less than that achieved by the computer model (.76 vs. .79), but the difference was not significant.

Table 12

Lens Model Components

for the Original Examiners and Seven Blind Interpreters

	r_a	R_e	R_s	G	C
Original Examiners	.87	.79	.74	.99	.70
Experienced Examiner	.87	.79	.81	.99	.64
Quality Control	.84	.79	.76	.99	.62
Psychophysiologist	.77	.79	.77	.96	.47
Inexperienced Examiner	.77	.79	.75	.99	.45
Quality Control	.71	.79	.69	.99	.38
Inexperienced Examiner	.67	.79	.75	.99	.20
Experienced Examiner	.53	.79	.60	.93	.17
Mean (r-to-z-to-r)	.76	.79	.74	.99	.43

The G component is also important for summarizing the performance of a human interpreter (Slovic & Lichtenstein, 1971; Tucker, 1964). The G component, or matching index, exceeded .93 for each of the human interpreters. These findings indicate that most of the human interpreters made optimal use of the information contained in the four computer-generated physiological measures.

Variability in performance was observed among the blind numerical interpreters. Judgments made by the original examiners were highly accurate and were slightly more accurate than those made by the blind interpreters, all of whom used numerical scoring procedures. Since the original examiners interacted with the subjects and had detailed knowledge of the case facts, it is possible that their decisions were influenced by the case facts and the verbal and nonverbal behavior of the subjects during

the examinations.

Although the performance of the human interpreters was not clearly related to level of experience, it was directly related to $\underline{C}$. This finding may indicate that the major factor that distinguished among the blind numerical interpreters was their ability to extract more diagnostic information from the physiological recordings than was represented by the four response parameters quantified by the computer. The large value for $\underline{C}$ for the original examiners is another indication that they may have adjusted their numerical scoring of the physiological data by using nonphysiological, auxiliary sources of information that were available only to them.

The mean $\underline{C}$ component of the lens model indicated that on the average the blind evaluators were able to predict 18% ($\underline{C}^2$) of the criterion variance that was not predicted by the four computer-generated variables. This finding suggests that significantly more diagnostic information was available in the physiological recordings than was represented in the four parameters quantified by the computer. Some of that variance may be attributed to the human interpreter's ability to make reasonable approximations of the amplitudes of physiological reactions even when the recording pens exceeded the limit of travel because the examiner had set the amplifier sensitivity too high, a common occurrence in the polygraph charts used in the present study. The computer merely quantified the amplitude of the response as it appeared on the chart, and no attempt was made to estimate the true amplitude of the response when the limit of pen travel was exceeded.

Discussion

This study evaluated the accuracy of control question polygraph examinations in criminal investigations conducted by U. S. Secret Service personnel during FY1983 through FY1985. The cases were obtained from their files and were confirmed using a very stringent criterion of admissions and confessions that were independently corroborated by physical evidence. The results of this study clearly indicate that control question polygraph examinations used for purposes of criminal investigation can be highly accurate when conducted by qualified examiners and numerically evaluated by experienced interpreters or assessed using computer methods developed at the University of Utah.

Accuracy

Human Interpreters

The overall accuracy of decisions made by the Secret Service examiners on individual relevant questions was 96% for confirmed truthful answers and 95% for confirmed deceptive answers in those cases where suspects were either truthful to all confirmed relevant questions or deceptive to all confirmed relevant questions (pure verification). When suspects were confirmed as deceptive to at least one relevant question and also truthful to at least one relevant question in the same test (mixed verification), the accuracy of the decisions made by the original examiners dropped to 91% on confirmed truthful answers and 85% on confirmed deceptive answers. It should be noted that this high level of accuracy was achieved even though the level of analysis at individual questions would be expected to produce lower reliability and accuracy than analyses of all relevant questions combined.

The results also indicated that the accuracy of decisions made by examiners who made blind interpretions of the polygraph charts was also

high, but not quite as high as the original examiners. The accuracy of blind interpreters on pure verification suspects was 85% on truthful answers and 94% on deceptive answers. However, when there was mixed verification, their accuracy dropped to 63% on truthful answers and 84% on deceptive answers. From these results, it appears that control question polygraph tests perform best when the relevant questions deal with issues that elicit either all truthful or all deceptive answers from the subject. It should also be noted that the blind interpreters made more false positive than false negative errors, a result that consistently appears in the data from laboratory and field studies (Raskin, 1986). However, the original examiners did not show that pattern.

The effects of context of the interpretion (original or blind) and interpreter experience or type of training on the accuracy of chart interpretations were assessed by comparisons of the performance of the original examiners, highly experienced quality control interpreters, experienced and inexperienced field examiners, and an experienced field examiner-psychophysiologist. Analyses of the numerical scores and lens model analyses were used for these purposes, and the results produced two somewhat unexpected findings.

There was no demonstrable effect on accuracy as a function of experience or type of training among all of the blind interpreters. However, the original examiners clearly outperformed all of the blind interpreters and the computer model. The lens model analyses indicated that level of performance of the human interpreters was directly related to the extent to which they either extracted more diagnostic information from the polygraph charts than did the computer model or used nonphysiological information to adjust their numerical scoring to increase their accuracy.

The original examiners, one quality control, and one experienced blind interpreter outperformed the computer, but the computer outperformed the remaining five blind interpreters. The superior performance of the original examiners suggests that they used their knowledge of the case facts and their interactions with the subjects to achieve more effective use of the physiological information contained in the polygraph charts.

Computer Interpretations

The computer interpretations of the polygraph recordings also produced a high degree of accuracy. Using the discriminant function generated from these data and various probabilities to defire truthful and deceptive decisions, the accuracies ranged between 95% and 96% on confirmed truthful suspects and between 83% and 96% on confirmed deceptive suspects. As the probability required for a decision was increased, the accuracies and the rate of inconclusive outcomes increased. The optimal cutoffs of .70 probability of truthfulness for truthful decisions and .30 probability of truthfulness for deceptive decisions yielded accuracies of 96% on Truthful suspects and 93% on Deceptive suspects, with only 11% inconclusive outcomes. These analyses seem to indicate that the use of cutoffs of approximately .70 and .30 for probabilities of truthfulness yield the best results in field applications.

Comparisons of the computer-generated decisions and those produced by the human interpreters indicated that the computer was generally more accurate than the blind interpreters, but not as accurate as the original examiners. These findings are consistent with a recent review of the literature concerning clinical versus statistical prediction (Wiggins, 1981), indicating that statistical methods are frequently, but not always, superior to clinical judgments. If the computer could take advantage of the case information and observations of the suspect's behavior that were

available to the original examiners, computer models might equal or exceed the performance of the original examiners. Achievement of that goal would require additional research to determine the factors that account for the increment in performance of the original examiners and how to incorporate that information in the computer decision models. Toward that end, research that explores relationships between individual differences in expressive behavior, case information, and truthfulness seems feasible and desirable.

Research Issues

Validity of the Confession Criterion

Questions have been raised with respect to the validity of results obtained in field studies that select polygraph examinations for analysis using a criterion of ground truth based on confessions (Iacono, in press; Raskin, 1987). Iacono argued that such studies overestimate accuracy because they do not include the polygraph charts of innocent suspects who failed tests and did not confess and guilty suspects who passed tests and were not interrogated or failed to confess. Iacono also argued that guilty suspects selected for confession studies were only those who produced charts that were strong enough to cause the examiner to elicit a confession. The latter argument seems specious since it implicitly recognizes the accuracy of polygraph charts that are strongly indicative of deception. It also implies that the test results of suspects who failed the test and did not confess are weaker than those who failed the test and did confess. These arguments were addressed by the methods and results of this study.

The manner of selecting cases prevented the problem of not selecting innocent suspects who failed tests (false positive errors) because all of

the confirmed truthful suspects were obtained from multiple-suspect cases. Since the truthfulness of these suspects was established by corroborated confessions of other suspects, all truthful suspects who might have failed the tests were included in the sample and would have contributed to the observed error rate. Similarly, the large majority of confirmed deceptive suspects were obtained from multiple-suspect cases in which there was usually more than one deceptive person who could, and often did, confess and incriminate one or more of the other suspects who were tested. Thus, the potential problems of false positives and false negatives proposed by Iacono were reduced or eliminated by the methodology of this study.

This study also evaluated the suggestion that suspects who failed the tests and confess produced stronger deceptive charts than those who failed the tests and did not confess. In order to answer that question, we compared the strengths of the deceptive results produced by suspects who confessed to the original examiners and deceptive results produced by suspects who were scored as deceptive by the original examiners but did not confess. The analyses indicated a difference of approximately 20% between the magnitude of negative scores assigned to confirmed and unconfirmed deceptive results. However, the mean scores for unconfirmed deceptive results were 63% higher than the minimum score required for a conclusive deceptive decision. Therefore, it appears that the success or failure in eliciting a confession was unrelated to the strength of the physiological reactions to relevant questions. These results provide little support for Iacono's argument concerning the lack of validity of confession-based field polygraph studies.

Generalizability of Laboratory Results

Two types of analyses were conducted to assess the extent to which the results of laboratory experiments can be used to make inferences about the

accuracy and processes that underly control question polygraph examinations of criminal suspects. The first compared profiles of physiological responses of confirmed truthful and deceptive laboratory subjects and criminal suspects. The results indicated that although there was a small but significant difference in the shape of the profiles of deceptive laboratory and field subjects, the size of the effect was very small in comparison to the differences between the physiological responses to control and relevant questions produced by truthful and deceptive laboratory and field subjects. Since the latter is the basis for rendering decisions in the field as well in realistic simulations of the field situation (Kircher, Horowitz, & Raskin, 1987), the findings lend support to the generalizability of the results of such laboratory studies to applications of polygraph examinations in criminal investigation.

The second type of analysis used a double cross-validation procedure to determine the accuracy of computer classifications of criminal suspects based on a discriminant function derived from laboratory data and the accuracy of computer classifications of laboratory subjects based on a discriminant function developed on criminal suspects. The results indicated that the accuracies of each model were similar when applied to laboratory and field data. However, the laboratory model produced an increase in false positive errors when applied to field suspects and the field model showed an increase in false negative errors when applied to laboratory subjects. The structural coefficients and univariate validity coefficients also were consistent with the principle of generalizability.

The suggestion of asymmetry in false positive and false negative errors produced by the laboratory and field models was further assessed by a comparison of the means of the computer-generated indices of differential

reactivity to control and relevant questions by laboratory subjects and criminal suspects. The differential reactivity indices for laboratory subjects were symmetrical around zero, but the means for the field suspects were shifted in the negative direction. These results reinforce an interpretation that compared to deceptive laboratory subjects, deceptive field suspects show stronger differential reactions to relevant questions than to control questions; and compared to truthful laboratory subjects, truthful field suspects showed much weaker differential reactions to control than to relevant questions. Although it appears that the underlying structure of physiological responses in laboratory subjects is similar to that obtained in polygraph examinations of criminal suspects, the obtained differences suggest using somewhat different numerical cutoffs for decision-making in the two situations.

Implications of the Results for Investigative Applications

Three major conclusions for applications and procedures for control question polygraph examinations of criminal suspects are suggested by the results of this study. They concern the accuracy of such tests, the optimal composition of relevant questions to be used in such tests, and the optimal methods for interpreting the outcomes of such tests. The overall pattern of results indicates that properly conducted and interpreted examinations have a high degree of accuracy and can be of considerable benefit in evaluations of the credibility of criminal suspects. However, certain changes in current practices should be considered.

The results suggest that blind numerical scoring procedures using cutoffs that are symmetrical around zero may be biased against truthful criminal suspects. Although the scores assigned by the original examiners did not show this effect, the blind interpreters made relatively more errors on confirmed truthful responses. Apparently, the original examiners

used other information to compensate for the inherent bias of the test against truthful suspects. Even though the six U. S. Secret Service blind interpreters scored the charts using the federal system that compares the reactions to relevant questions to the control questions that evoke stronger physiological responses (Weaver, 1980, 1985), they still made more false positive errors than did the original examiners and the computer model. Thus, it appears that blind numerical interpretation would be more accurate if stronger negative scores were required for deceptive decisions and somewhat weaker positive scores were required for truthful decisions. The present data seem to suggest cutoffs of -3 and +2 for individual questions and -7 and +4 for overall decisions. However, additional analyses are required in order to establish definitive cutoffs for decisions based on blind numerical evaluations.

A related problem is raised by the finding of higher false positive rates for questions answered truthfully by suspects who were also deceptive to at least one relevant question in the same test. It appears that answering deceptively to at least one relevant question in the test tends to weaken the reactions to the control questions, thereby making it difficult for them to produce reactions that are larger than those to relevant questions that are answered truthfully. Therefore, field polygraph examiners should attempt to devise sets of relevant questions that the suspect can be expected to answer all truthfully or all deceptively. The case information and the importance of each relevant question should be carefully considered in formulating the set of relevant questions to be asked, and separate question series should be used whenever it seems likely that the suspect might answer some of the relevant questions truthfully and some of them deceptively.

Finally, the results of this research clearly support the utility of computer models for the analysis and interpretation of polygraph test outcomes. The results obtained with computer models derived from the data on criminal suspects demonstrated higher accuracy than blind numerical interpretations. Computer evaluations have the additional virtues of being objective and providing a rapid and readily available form of quality control for field examiners. Computer analyses would be especially useful when performing examinations in important cases and another examiner is not available for independent interpretation when decisions must be made on the spot. In most cases, decisions must be made in order to determine if the the suspect is to be excused, interrogated, or administered additional examinations. Under such circumstances, an independent computer analysis may be increase confidence in the decisions and guide the course of further testing.

References

Barland, G. H. (1982). On the accuracy of the polygraph: An evaluative review of Lykken's Tremor in the Blood. Polygraph, 11, 258-272.

Bersh, P. J. (1969). A validation study of polygraph examiner judgments. Journal of Applied Psychology, 53, 399-403.

Bradley, M. T., & Ainsworth, D. (1984). Alcohol and the psychophysiological detection of deception. Psychophysiology, 21, 63-71.

Dawson, M. E. (1980). Physiological detection of deception: Measurement of responses to questions and answers during countermeasure maneuvers. Psychophysiology, 17, 8-17.

Gatchel, R. J., Smith, J. E., & Kaplan, N. M. (1983). The effect of propranolol on polygraphic detection of deception. Unpublished manuscript, University of Texas.

Harris, R. J. (1975). A primer of multivariate statistics. New York: Academic Press.

Horvath, F. S. (1977). The effect of selected variables on interpretation of polygraph records. Journal of Applied Psychology, 62, 127-136.

Horvath, F. S., & Reid, J. E. (1971). The reliability of polygraph examiner diagnosis of truth and deception. Journal of Criminal Law, Criminology and Police Science, 62, 276-281.

Hunter, F. L., & Ash, P. (1973). The accuracy and consistency of polygraph examiners' diagnoses. Journal of Police Science and Administration, 1, 370-375.

Iacono, W. G. (in press). Can we determine the accuracy of polygraph tests? In P. K. Ackles, J. R. Jennings, & M. G. H. Coles (Eds.), Advances in psychophysiology (Vol. 4). Greenwich, CT: JAI Press.

Kircher, J. C., Horowitz, S. W., & Raskin, D. C. (1987). 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.

Kleinmuntz, B. J., & Szucko, J. J. (1982). On the fallibility of lie detection. Law & Society Review, 17(1), 85-104.

Lykken, D. T. (1979). The detection of deception. Psychological Bulletin, 86, 47-53.

Lykken, D. T. (1981). A tremor in the blood. New York: McGraw-Hill.

McNemar, Q. (1969). Psychological statistics (4th ed.). New York: Wiley.

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

Office of Technology Assessment (1983). Scientific validity of polygraph testing: A research review and evaluation. Washington, D. C.: U. S. Government Printing Office.

Pedhazur, E. (1982). Multiple regression in behavioral research: Explanation and prediction. (2nd ed.). New York: Holt.

Podlesny, J. A., & Raskin, D. C. (1978). Effectiveness of techniques and physiological measures in the detection of deception. Psychophysiology, 15, 344-358.

Raskin, D. C. (1976). Reliability of chart interpretation and sources of errors in polygraph examination. (Report 76-3, Contract 75-NI-99-0001, U.S. Department of Justice.) Salt Lake City, Utah: Department of Psychology, University of Utah.

Raskin, D. C. (1982). The scientific basis of polygraph techniques and their uses in the judicial process. In A. Trankell (Ed.), Reconstructing the past: The role of psychologists in criminal trials. Stockholm: Norstedt and Soners.

Raskin, D. C. (1984). An evaluation of the polygraph policies and programs of U. S. Department of the Treasury. Unpublished report to the U. S. Department of the Treasury.

Raskin, D. C. (1986). The polygraph in 1986: Sceintific, professional, and legal issues sorrounding applications and acceptance of polygraph evidence. Utah Law Review, 1986, 29-74.

Raskin, D. C. (1987). Methodological issues in estimating polygraph accuracy in field applications. Canadian Journal of Behavioural Science, 19, 389-404.

Raskin, D. C., Barland, G. H., & Podlesny, J. A. (1978). Validity and reliability of detection of deception. (Contract 75-NI-99-0001, U. S. Department of Justice). Washington, D. C.: U. S. Government Printing Office.

Raskin, D. C., & Hare, R. D. (1978). Psychopathy and detection of deception in a prison population. Psychophysiology, 15, 126-136.

Rovner, L. I., Raskin, D. C., & Kircher, J. C. (1979). Effects of information and practice on detection of deception. Psychophysiology, 16, 197-198. (Abstract)

Slovik, P., & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. Organizational Behavior and Human Performance, 3, 305-309.

Slowik, S. M., & Buckley, J. P. (1975). Relative accuracy of polygraph examiner diagnosis of respiration, blood pressure, and GSR recordings. Journal of Police Science and Administration, 3, 305-309.

Timm, H. W. (1982). Effect of altered outcome expectancies stemming from placebo and feedback treatments on the validity of the guilty knowledge technique. Journal of Applied Psychology, 67, 391-400.

Tucker, L. R. (1964). A suggested alternative formulation in the developments by Hursch, Hammond, and Hursch and by Hammond, Hursch, and Todd. Psychological Review, 71, 528-530.

Van Egeren, L. F. (1973). Multivariate statistical analysis. Psychophysiology, 10, 517-532.

Weaver, R. S. (1980). The numerical evaluation of polygraph charts: Evolution and comparison of three major systems. Polygraph, 9, 94-108.

Weaver, R. S. (1985). Effects of differing numerical chart evaluation systems on polygraph examination results. Polygraph, 14, 34-41.

Wicklander, D. E., & Hunter, F. L. (1975). The influence of auxiliary sources of information in polygraph diagnoses. Journal of Police Science and Administration, 3, 405-409.

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

Winkler, R. L., & Hays, W. L. (1975). Statistics: Probability, inference, and decisions. New York: Holt.