Validity and Reliability of Detection of Deception

National Institute of
Law Enforcement and Criminal Justice
Law Enforcement Assistance Administration
U. S. Department of Justice

Validity and Reliability of Detection of Deception

by
David C. Raskin, Ph.D.
Gordon H. Barland, Ph.D.
John A. Podlesny, M.A.

June 1978

National Institute of
Law Enforcement and Criminal Justice
Law Enforcement Assistance Administration
U. S. Department of Justice

National Institute of Law Enforcement
and Criminal Justice
Blair G. Ewing, Acting Director

Law Enforcement Assistance Administration
James M. H. Gregg, Acting Administrator

This project was supported by Grant Number 75-NI-99-0001, awarded to the University of Utah by the National Institute of Law Enforcement and Criminal Justice, Law Enforcement Assistance Administration, U.S. Department of Justice, under the Omnibus Crime Control and Safe Streets Act of 1968, as amended. Points of view or opinions stated in this document are those of the authors and do not necessarily represent the official position or policies of the U. S. Department of Justice.

For sale by the Superintendent of Documents, U.S. Government Printing Office
Washington, D.C. 20402
Stock Number 027-000-00692-2

CONTENTS

ABSTRACT

This project was designed to provide information on the usefulness of polygraph techniques in detecting truth and deception in criminal investigations. The report describes the methodology of each of the eight experiments and studies conducted. In addition to studies involving criminal suspects in a field situation, other studies involved laboratory experiments with a mock-crime paradigm. The lab experiments investigated aspects of the general problems of accuracy and reliability not easily studied in the field situation. They also assessed the usefulness of a number of physiological measures that had received little attention in previous scientific research. Studies also were undertaken to evaluate the commonly-held belief that psychopaths can "beat the polygraph," the adequacy of current practices used by field polygraphists, the usefulness of different question structures in polygraph examinations, and the risks of different types of errors in field applications. The project results indicate that polygraph examinations using control-question or guilty-knowledge tests are highly accurate.

I. METHODS

The eight separate phases of this project can be divided into two categories. One category consists of two laboratory experiments in which the subjects were offered monetary incentives for participating in a mock-crime and attempting to produce truthful outcomes on the polygraph examination. Such studies have certain advantages. First, the laboratory setting allows complete and certain determination of ground (factual) truth. Therefore, the accuracy of outcomes and other results can be assessed against the certain knowledge of truthfulness or deception on the part of the subject. Second, in a laboratory situation it is possible to compare and evaluate different question techniques (test structures) and various physiological measures which may or may not have been extensively employed in previous research and application. Finally, it is possible to investigate the influence of subject characteristics such as psychopathy (sociopathy) in a controlled situation.

The results of laboratory experiments are very useful in making generalizations to the field situation with criminal suspects. However, such inferences should be made cautiously and tested by research in the field setting. Furthermore, there are many questions concerning field practices which can be answered only by studies of field applications. Therefore, the remaining six studies involved investigations which utilized data obtained from polygraph examinations conducted on criminal suspects for real-life purposes. Some of those examinations were conducted at the University of Utah, and others were provided by a variety of law enforcement and private polygraph examiners.

A. Laboratory Experiments

The two laboratory experiments were similar in design and procedures employed. Both utilized a mock-crime situation in which the subjects were informed about the nature of the crime. Half of the subjects in each experiment were instructed to commit the crime (guilty subjects), and the other half were merely informed about the nature of the crime (innocent subjects). Each subject was subsequently administered a polygraph examination by an examiner who had no knowledge concerning the guilt or innocence of the subject.

All subjects had been instructed to deny having committed the theft, and they were offered a cash bonus if they could produce truthful results on the polygraph test. The polygraph examiner (who was trained and experienced in field polygraph techniques) conducted a standard pretest interview with each subject, administered the polygraph test, and made his decision on the basis of numerical evaluation of the polygraph charts (Raskin, 1975). The first three charts for each subject were later subjected to detailed quantitative analyses utilizing computer techniques.

1. *Experiment I.* This research (Raskin, 1975) was conducted at a provincial prison in British Columbia, Canada with 48 male volunteers from the prison population. All of the subjects were convicted felons, and half of them had been clinically diagnosed as psychopathic (sociopathic). The crime consisted of stealing $20 from a drawer in a room which was off-limits to inmates, and all subjects (guilty and innocent) were instructed to deny the theft and attempt to produce truthful results on the polygraph test. All subjects who produced truthful polygraph charts received a $20 bonus.

The polygraph test was a federal zone-comparison control-question test (Barland & Raskin, 1975) consisting of a number test followed by a minimum of three charts. A typical question sequence was as follows:

- (neutral) Were you born in Canda? Yes.
- (sacrifice relevant) Regarding that $20, do you intend to answer truthfully each question about that? Yes.
- (outside issue) Are you completely convinced I will not ask a question that hasn't been reviewed? Yes.
- (control) Other than what you told me, before you were 18 did you ever steal any money? No.
- (relevant) Did you take that $20? No.
- (control) Did you ever steal anything else from someone who trusted you? No.
- (relevant) Did you take that $20 from the drawer? No.
- (neutral) Is your last name _____? Yes.

- (control) Have you ever taken anything of value from an employer? No.
- (relevant) Do you have that $20 now? No.

According to the theory of control-question tests (Backster, 1962; Podlesny & Raskin, in press; Reid & Inbau, 1966) the subject will respond most to the questions which are the greatest threat at that time. For a guilty person the relevant questions are the most threatening, and he will show larger reactions to the relevant as compared to the control questions. However, the innocent subject knows he is being truthful to the relevant questions, and he should be more concerned about the control questions. Therefore, the innocent subject should produce larger reactions to the control than to the relevant questions.

Skin conductance (SCR), thoracic and abdominal respiration, plethysmographic measures of finger blood volume (FBV) and finger pulse amplitude (FPA), heart rate (HR), and skin potential (SPR) were measured on a Beckman Dynograph. The time between the beginning of consectuve questions ranged between 25 and 35 seconds. Following each chart, the examiner asked the subject if any question bothered him and if he would like to change the wording of any questions. The subject's attention was directed toward the control questions, which were frequently modified following admissions or expression of concern by the subject. If the results did not seem to be conclusive after three charts, the fourth chart was a silent answer test (Horvath & Reid, 1972). Additional charts were sometimes obtained up to a maximum of seven charts.

The charts were then scored numerically using the three components of respiration, SCR, and cardiovascular (plethysmograph). If the total score was +6 or larger, the decision was truthful; if it was −6 or smaller, the decision was deceptive; and scores between ±6 were called inconclusive. All of the physiological measures were then subjected to a detailed quantitative analysis by a person who had no knowledge of the outcome or experimental group of any subject.

2. *Experiment II.* This experiment (Podlesny et al., 1976) was conducted at the University of Utah using 60 male subjects recruited from the community by newspaper advertisements. They were paid $5 for participation and were offered a $10 bonus if they produced truthful results. The crime consisted of stealing a gold wedding ring from a drawer in a secretary's desk on another floor of the building. All subjects were informed that they were also suspected of having stolen a watch. That was introduced to allow the evaluation of the effectiveness of a guilt-complex question as a type of control question (Barland & Raskin, 1973; Lykken, 1974; Reid & Inbau, 1966).

The testing procedures were the same as in Experiment I with a few exceptions. For 20 subjects the control questions were of the type which clearly excludes the crime being investigated (Backster, 1969), and for 20 subjects the control questions overlapped the relevent issue (Reid & Inbau, 1966). A typical question sequence with Reid type control questions was as follows:

- (neutral) Is your name _____? Yes.
- (sacrifice relevant) Regarding the ring and watch, do you intend to answer the questions about them truthfully? Yes.
- (outside issue) Are you convinced I will only ask questions on this test that you've already okayed? Yes.
- (Reid control) Have you ever stolen any money? No.
- (relevant) Did you take that ring? No.
- (Reid control) Besides what you told me about, have you ever taken anything of value? No.
- (relevant) Did you take that ring from that desk? No.
- (guilt complex) Did you take that watch from Room 702? No.
- (Reid control) Have you ever taken anything from someone who trusted you? No.
- (relevant) Do you have that ring with you now? No.

For subjects who received Backster type control questions, the words "Have you ever" were replaced with wording which clearly excluded the crime being investigated. That was done by specifiying a certain time period such as "While you were in high school." A typical Backster control question was "Between the ages of 15 and 20 did you take something of value?"

An additional 20 subjects were examined using the guilty-knowledge test (Lykken, 1959). The rationale for the guilty-knowledge test is that the absence or presence of differential responsivity to items of information known only to a guilty person provides the basis for conclusions about truth or deception concerning the crime. The subject was administered a series of five charts each having six alternatives to a different question concerning information related to the crime. The items consisted of a set of equally plausible alternatives, one of which was the correct

(critical) alternative. The question sequence was as follows:

Chart 1

Regarding the type of ring that may have been taken,
(1) Do you know if it was a sapphire class ring?
(2) Do you know if it was a pearl engagement ring?
(3) Do you know if it was a silver and turquoise ring?
*(4) Do you know if it was a gold wedding ring?
(5) Do you know if it was a ruby class ring?
(6) Do you know if it was a diamond engagement ring?

Chart 2

Regarding the floor of this building that the ring was hidden on,
(1) Do you know if it was the 1st floor?
(2) Do you know if it was the 12th floor?
(3) Do you know if it was the 6th floor?
(4) Do you know if it was the 4th floor?
*(5) Do you know if it was the 8th floor?
(6) Do you know if it was the 10th floor?

Chart 3

Regarding the number of the room that the ring was hidden in,
(1) Do you know if it was Room 800?
*(2) Do you know if it was Room 820?
(3) Do you know if it was Room 810?
(4) Do you know if it was Room 816?
(5) Do you know if it was Room 814?
(6) Do you know if it was Room 803?

Chart 4

Regarding the type of envelope that the ring was hidden in,
(1) Do you know if it was an inter-campus mail envelope?
(2) Do you know if it was a medium-sized manila envelope?
*(3) Do you know if it was a business-sized white envelope?
(4) Do you know if it was a small-sized manila envelope?
(5) Do you know if it was a small-sized white envelope?
(6) Do you know if it was a large-sized manila envelope?

Chart 5

Regarding the name of the doctor that the guilty person was instructed to ask for,
(1) Do you know if it was Dr. Trumbull?
(2) Do you know if it was Dr. Tolman?
(3) Do you know if it was Dr. Heisse?
(4) Do you know if it was Dr. Jordan?
(5) Do you know if it was Dr. Calvin?
*(6) Do you know if it was Dr. Mitchell?

The first alternative was included to buffer initial responding and was not scored. The remaining five alternatives consisted of a critical item and four noncritical items. The critical item was the correct alternative, and noncritical items were all incorrect. The critical items were positioned among the noncritical items in a pseudo-random order across charts. In the above list, critical items are identified with an asterisk. The rationale of the guilty-knowledge technique was explained to each subject in that group, and prior to each chart the question was reviewed, but no alternatives were stated until the test was administered except with Chart 4. Prior to that chart, an example of each type of envelope was shown to the subject and specifically named. Subjects were instructed to answer "no" to each alternative on all of the charts.

The probability that an innocent subject without knowledge of the details of the crime would show his largest reaction to a critical item is 1/5 for each chart. Thus, only 6 of 100 innocent subjects would produce their largest reactions to the critical item on three or more charts. However, the probability of a series of such large reactions from a subject with guilty knowledge is very high. This guilty knowledge procedure is designed to protect against false positives (an innocent person producing deceptive results) which Lykken (1974) claims are frequent occurrences with control-question tests.

With both control-question and guilty-knowledge tests, a minimum of 15 seconds elapsed between the verbal response to the question or alternative item and the beginning of the next question or item. Control-question tests consisted of a number test and a minimum of three charts. Responses were recorded on a Beckman Dynograph and included thoracic respiration, skin conductance (SCR), plethysmographic measures of finger pulse amplitude (FPA) and finger blood volume (FBV), relative blood pressure measured by a low-pressure (50–50 mmHg) cardio cuff, skin potential (SPR), heart rate (HR) measured on a second-by-second basis, and a dry cardio activity monitor (CAM) placed on the palmar tip of the second finger.

Following the last chart, the results were immediately evaluated to determine whether the subject was truthful or deceptive. Control-question tests were numerically evaluated in the same manner as Experiment I using the SCR, respiration, relative blood pressure (cardio), and plethysmograph measures. For guilty-knowledge tests, the size of the SCR to each critical item was measured to determine if it was the largest of the responses to the five alternative items for that question. If the response to the critical item was largest on at least three of the five charts, the subject was called deceptive. If he showed less than three such responses, he was called truthful.

All of the charts were subsequently scored independently by an examiner who had no contact with the subjects and no knowledge of their guilt or innocence. All of the analyses of numerical scores reported here are based on the results of the inde-

pendent evaluations. The control-question tests were evaluated in the same manner as was originally performed. For the guilty-knowledge tests, the ranking system used by Lykken (1959) was employed. If the critical item produced the largest SCR, it was assigned a value of 2; if it was the second largest, it received a 1; and ranks lower than second largest was assigned a 0. The ranks for the critical items were summed over the five charts. If the total was 6 or higher, the subject was called deceptive. If the total was less than 6, the subject was called truthful. All of the physiological measures were then subjected to a detailed, quantitative analysis by persons who had no knowledge of the outcome or experimental group for any subject.

B. Field Studies

A total of six different studies were completed using polygraph examinations on criminal suspects. Except for one source in the study described below under the heading of "current field practice," all of the examinations used a control-question technique. Some of them were Backster zone-comparison tests (Bailey & Rothblatt, 1970), some were federal zone-comparison tests (Barland & Raskin, 1975), and some were Reid control-question tests (Reid & Inbau, 1966).

1. *Reliability and validity with criminal suspects.* In spite of careful attempts to simulate field situations, there are a number of important differences between most laboratory experiments and application of detection of deception in the field situation with criminal suspects. Those include profound differences in the consequences of the outcome and the resulting differences in subject motivation, differences in subject populations, the availability of information about the case which may influence the examiner and the subject, and frequent differences in the type of techniques utilized and the training and experience of field examiners as compared to the typical laboratory researchers. In the two experiments described above, many of those differences were eliminated. However, it was not possible to eliminate some of them.

This study (Barland & Raskin, 1976) was designed to extend the results obtained by Bersh (1969) and to overcome some of the limitations of that study by using nonmilitary criminal suspects obtained by referral from both law enforcement and defense sources. It also studied the relationship between various personality, behavioral, socioeconomic, and crime categories and the results obtained on the polygraph tests.

A total of 102 criminal suspects were examined at the request of police, defense attorneys, or prosecuting attorneys; and 92 independent cases were selected from those. All but one subject was tested on field model polygraphs with a federal zone-comparison control-question technique, and the results were evaluated using the numerical scoring procedures described above. The charts were subsequently evaluated by an independent examiner who had no knowledge of the case or the original outcome, and the results reported here are based on those evaluations. All tests included a minimum of three charts, and numerical evaluations were based on the standard field measures of respiration, skin resistance, and cardiovascular activity.

Background information was obtained from each subject, and he or she responded to several scales from the *Minnesota Multiphasic Personality Inventory* (MMPI) including the Lie, K, Psychopathic Deviancy, Hypochondriasis, and Depression scales. The examiner also observed both spontaneous behavior cues and those elicited by specific questions reported to be helpful in differentiating truthful and deceptive persons (Horvath, 1973). On the basis of that behavioral information, the examiner made two covert predictions of the outcome of the polygraph test, the first immediately following the advisement of rights and the second just prior to the administration of the polygraph test.

Three criteria were developed for assessing ground truth in order to evaluate the accuracy of the polygraph results. The first consisted of the independent judgments of a 5-member panel of experts composed of two criminal defense attorneys, two criminal prosecuting attorneys, and a judge. In cases where at least 3 of the 5 panel members agreed on a decision of guilty or innocent, the judgment of the panel was used as the criterion for ground truth. The second criterion consisted of judicial outcomes in which the polygraph results played no role and the case was not dismissed for insufficient evidence. The third consisted of a full confession or plea of guilty to the original charge. The latter criterion was used only for analyses to assess the effectiveness of the three physiological components with guilty subjects.

2. *Effectiveness of physiological measures with criminal suspects.* In order to test the findings from laboratory Experiment II on a sample of subjects tested on real-life criminal issues, cases were obtained by referral from defense and prosecution sources. A total of 26 subjects was tested on the criminal charge involved in the case using the federal zone-comparison control-question test.

All subjects were examined at the University of

Utah laboratory using the same instrumentation employed in Experiment II (above). From that group of subjects, 14 were selected to analyze the effectiveness of the non-standard physiological measures evaluated in Experiment II. Since the standard measures of respiration and skin conductance had repeatedly been demonstrated to identify guilty and innocent subjects in laboratory and field situations, the subjects were separated into guilty and innocent groups on the basis of the total numerical scores for respiration and skin resistance responses. Only subjects who produced decisive outcomes were utilized, and 7 subjects were eliminated for analyses reducing the innocent group to the same size (N=7) as the guilty group. The mean total score on those measures was −8.7 for the guilty group and +6.3 for the innocent group. On the assumption that there was a high degree of accuracy in assigning the subjects to the proper group, it was then possible to evaluate with a sample of criminal cases, all of the nonstandard measures found to produce significant results in Experiment II.

3. *Evaluation of current practices by law enforcement and private polygraph examiners*. Although there are several recent reports concerning the accuracy and reliability of decisions made by field polygraph examiners (Barland & Raskin, 1975; Horvath, 1974; Horvath & Reid, 1971; Hunter & Ash, 1973; Slowik & Buckley, 1975), the sampling of cases was highly selective in all but the Barland and Raskin study. In the other four studies, only cases in which the original examiner had made a definite decision were utilized, and most of those were selected from cases which had been verified by confession of the guilty person. Those facts, plus the laboratory nature of the Barland and Raskin experiment, limit the representatives and generalizability of the results to the typical field situation.

This study (Raskin & Barland, 1976) was designed to overcome the above objections and to extend the investigation to answer the questions of adequacy of techniques employed by law enforcement and private examiners, relative effectiveness with different crime categories, and relative effectiveness of the standard measures of respiration, skin resistance (galvanic skin response), and relative blood pressure (cardio).

The procedure involved sampling approximately 60 recent cases from each of seven different locations, three law enforcement agencies and four well-known polygraph firms. Cases were selected to include examples from the crime categories of crimes against people (homicide, assault), economic crimes (robbery, burglary, theft), and sex and drug offenses. Only tests performed on suspects were included in the sample, and only one examination from any specific case was included. Using those criteria, a total of 419 examinations was obtained for analysis.

Each examination was independently evaluated by Dr. Barland and Dr. Raskin. All of the standard control-question charts were subjected to our standard numerical scoring procedure, and decisions were made using the ±5 inclusive region for inconclusives. The relevant-irrelevant tests obtained from one location were evaluated subjectively since there is no system for numerical scoring of such charts. Characteristics such as chart quality, chart markings, case information, question structure, source of referral, length of pretest interview, and name of examiner were noted for each examination. A total of 43 polygraph examiners was represented in the sample.

4. *Accuracy of chart interpretation*. Although there has been a number of studies of reliability of chart interpretation, the only reported studies which have utilized numerical scoring procedures have been conducted at the University of Utah (Barland & Raskin, 1975, 1976; Podlesny, Raskin, & Barland, 1976). Those experiments have produced very high rates of agreement on decisions (96–100%) based on numerical scoring of charts. Those results together with the high accuracy rates obtained with that technique (Barland & Raskin, 1975; Podlesny, Raskin, & Barland, 1976) seem to indicate that numerical evaluation of polygraph charts might increase the accuracy of decisions made on the basis of such numerical scores. One purpose of this study (Raskin, 1976) was to investigate that possibility. Since polygraph examiners differ widely in their training and experience with chart interpretation, a second purpose of this study was to study the accuracy of chart interpretation performed by polygraph examiners with a variety of training, experience, and familiarity with numerical scoring techniques.

Polygraph charts from 16 independent criminal cases were selected from those obtained in a previous study (Barland & Raskin, 1976). Each examination consisted of three charts recorded on a field model polygraph using the federal zone-comparison control-question technique. All of the examinations had been confirmed by confession of the guilty person; 12 were from guilty subjects and four from innocent subjects. The 16 sets of charts were independently evaluated by 25 field polygraph examiners from a variety of training backgrounds and experiences. They were not informed about any aspect of the case or the outcomes and were asked to render a conclusion of truthful, deceptive, or inconclusive

for each case. They were told to employ numerical scoring if they had been trained in it and wished to do so. Of the 25 examiners, 18 had at least one year of experience, 13 had received formal training in numerical scoring, but only 7 of those explicitly scored the charts numerically.

5. *The "friendly polygrapher."* In a recent paper, Orne (1975) stated that under certain conditions the motivation of a guilty subject would be reduced to the point that false negative errors would be greatly increased. Since it is well known that motivation to deceive and the threat of serious consequences of detection are essential to successful detection of deception, Orne speculated that polygraph examinations conducted at the behest of defense attorneys fail to meet those motivational requirements for guilty subjects. He reasoned that a subject in such a situation "knows that the results of the test *if he is found deceptive* will not be used against him... As a consequence, the client's fears about being detected are greatly reduced (p. 114)." He also speculated that the so-called "friendly polygrapher" employed by the defense attorney will treat the subject differently than an "arms length" examiner such as a law enforcement examiner or one working for the subject's employer. Orne concluded that such a situation will make the guilty subject less detectable.

This study (Raskin, 1976) was designed to test Orne's hypothesis using the results of polygraph examinations conducted on a confidential basis for defense attorneys and those conducted with explicit knowledge and/or agreement with law enforcement authorities or the subject's employer. The "friendly polygrapher" hypothesis predicts that examinations conducted confidentially for defense attorneys would produce more truthful-appearing polygraph charts and more truthful decisions than those performed with the knowledge on the part of the subject that the results would be reported to the law enforcement authorities and/or his/her employer. However, the theory and experience with control-question tests would predict that such an effect would not occur and the most to be expected would be an increase of inconclusive results.

Three different samples of control-question examinations were obtained. The first sample consisted of all examinations of criminal suspects conducted during a 1-year period by an experienced examiner for law enforcement authorities or private attorneys. The cases consisted of 106 examinations for law enforcement authorities and 98 examinations for defense attorneys. The outcomes of those examinations were obtained in the form of the number of truthful, deceptive, and inconclusive determinations.

The second sample consisted of control-question examinations of criminal suspects obtained from two private polygraph firms included in a previous study (Raskin & Barland, 1976). A total of 19 different examiners conducted the testing which consisted of 54 confidential examinations referred by defense attorneys and 57 examinations performed with subjects' knowledge that the results would be reported to law enforcement authorities and/or his/her employer. Each set of polygraph charts was numerically evaluated by Dr. Barland or Dr. Raskin prior to their obtaining any information concerning the issue tested, the source of referral, or the decision by the original examiner. Since at least two charts were obtained for each subject, the numerical score from the first two charts comprised the data utilized in the data analysis.

The third sample consisted of 27 control-question examinations of criminal suspects conducted at the University of Utah. Fourteen of the examinations were performed on a confidential basis for defense attorneys, and 13 were performed with the subject's knowledge that the results would be reported to law enforcement authorities. All examinations contained a minimum of three charts, and the total numerical scores for the first three charts were utilized in the data analysis.

6. *Errors in examinations.* In contrast to the concern about false negative results expressed by Orne (1975), another critic (Lykken, 1974) has focused a great deal of attention on the problem of false positive errors (a deceptive result obtained from a truthful person). Lykken asserted that the control-question technique cannot accurately identify innocent suspects since he believes that it is impossible to design control questions which will produce the same level of responsiveness in innocent subjects as is produced by the relevant questions with guilty subjects. A number of scientifically conducted studies have investigated the accuracy of control-question tests. Three of those are laboratory studies conducted by Raskin and his associates (Barland & Raskin, 1975; Podlesny, Raskin, & Barland, 1976; Raskin, 1975). Of the errors obtained in those studies, 56% were false positives. Two recent field experiments (Bersh, 1969; Barland & Raskin, 1976) used the control-question technique with criminal suspects and defined ground truth by means of the judgments of a panel of experts. In the Bersh study, 44% of the errors were false positives, and all but one of the errors in the Barland and Raskin study were of the false positive type. Although the error rate is relatively low (approximately 10%), there is evidence that false positive errors may comprise a substantial

proportion of the errors obtained with criminal suspects. Therefore, the major purpose of this study (Raskin, 1976) was to attempt to determine some of the factors which are associated with the occurrence of such errors.

The type of error investigated was restricted to the occurrence of inappropriate physiological responses which occurred in the polygraph examination. Specifically, instances were sought in which other evidence indicated innocence but the subject produced a deceptive pattern on the polygraph charts. In all cases more than one polygraph test had been conducted on the suspect, and in some cases one or more additional persons had received polygraph examinations. In all but one case at least one of the examinations of the suspect had been conducted at the University of Utah.

In the 12 cases obtained for the study, three criteria were used to justify the determination that an error had been made. In two cases evidence obtained subsequent to the polygraph test clearly proved that an error had occurred, in nine cases opposing results were obtained from two or more examinations on the same subject, and in four cases conflicting results were obtained from another person examined on the same issue. In all cases the original deceptive result with the suspects was confirmed by a numerical evaluation by Dr. Barland or Dr. Raskin.

II. RESULTS

A. Accuracy of Decisions

1. *Laboratory experiments*. The decisions made by the polygraph examiner in Experiment I and Experiment II are shown in Table 1. They are based on the total numerical scores using the criterion of +6 or larger for truthful outcomes and −6 or less for deceptive outcomes. It can be seen that the accuracy rates were quite high with a combined accuracy of decisions which exceeded 90%. Approximately 10% of the subjects yielded inconclusive results, and the errors were almost equally distributed between false positives and false negatives.

TABLE 1 *Accuracy of Control-Question Decisions and Types of Errors in Two Laboratory Experiments*

	% Correct	% False Positive	% False Negative	% Inconclusive	% Correct Decisions
Experiment I	88	4	0	8	95
Experiment II	80	2	8	10	89

In order to compare the effectiveness of control-question tests in identifying quilty and innocent subjects, the total numerical scores for the first three charts were obtained for both types of subjects. Those mean scores for innocent and guilty subjects in both experiments are shown in Table 2. Statistical analyses indicated that the technique was equally effective in identifying innocent and guilty subjects.

TABLE 2 *Total Numerical Scores for 3 Charts Obtained in Two Laboratory Experiments*

	Experiment I	Experiment II
Guilty	−11.1	−9.0
Innocent	+9.4	+14.8

Since the criterion of requiring a score of at least ±6 in order to render a decision was developed from the experiences of field examiners, the data from these two experiments were combined to evaluate the effectiveness of those cutoff points. Figure 1 presents the rates of correct decisions for guilty and innocent subjects and the rates of inconclusives using cutoffs for decisions which ranged from nonzero scores up to scores exceeding ±12 for all charts. It can be seen that the accuracy of decisions reaches an optimal level in the region of ±4, and inconclusives are relatively low (9%) up to cutoffs of ±5. Thus, it appears that the field practice of using scores which fall outside an inconclusive region of ±5 provides a good balance between accuracy and rate of inconclusives. Furthermore, there seems to be no compelling reason to alter that inconclusive region.

2. *Field study*. The accuracy of polygraph examinations with criminal suspects was evaluated using the decisions based on the numerical scores obtained from the independent chart interpretations. The standard ±5 boundaries were used for the inconclusive region. Those decisions were compared to the combined judgments of the panel and also to the judicial outcomes.

The results of the comparisons between the outcome of the polygraph examinations and the decisions based on agreement among at least a majority of the panel are shown in Table 3. When both the panel and the polygraph scores yielded a decision, the polygraph outcome agreed with the majority panel in 86% of the cases. More than half of the suspects found truthful with the polygraph produced inconclusive outcomes from the panel, and six of the seven disagreements were false positives (deceptive polygraph results on subjects considered innocent by the panel).

TABLE 3 *Comparison of Outcomes Based on Independent Examiner's Numerical Score and Panel Majority Decisions*

Independent Numerical Evaluation	Panel Majority Decision			
	Guilty	Innocent	Inconclusive	Total
Deceptive	39	6	13	58
Truthful	1	5	7	13
Inconclusive	7	6	8	21
Total	47	17	28	92

The polygraph results were also compared to the judicial outcomes which were considered conclusive and were not influenced by the polygraph results. Those results are presented in Table 4 and they

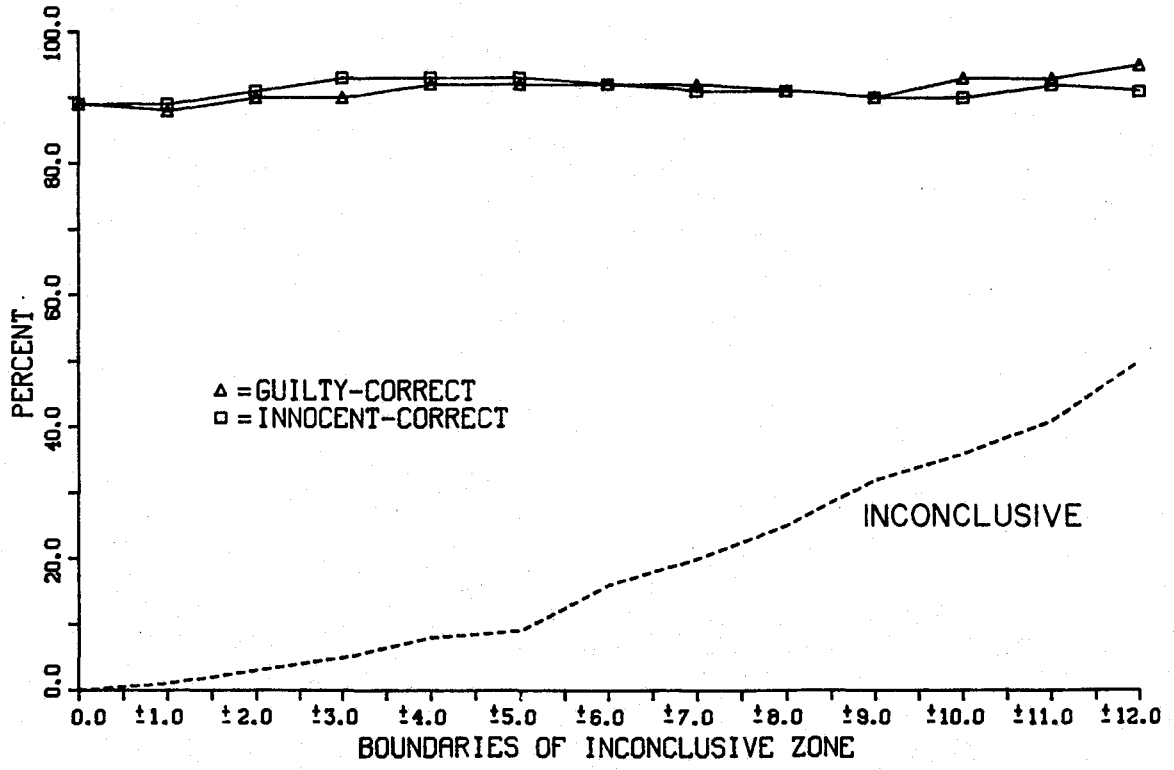

Figure 1. Rates of accuracy and inconclusives with different boundaries of the inconclusive zone.

indicated that there was 88% agreement between the polygraph decisions and the judicial outcomes. All of the disagreements (4) occurred on suspects who produced deceptive polygraph charts and who were acquitted in court.

TABLE 4 *Comparison of Decisions Based on Independent Examiner's Numerical Scores and Independent Judicial Outcomes*

Independent Numerical Evaluation	Judicial Outcome		
	Guilty	Innocent	Total
Deceptive	27	4	31
Truthful	0	3	3
Inconclusive	6	1	7
Total	33	8	41

It is of interest to note that there was less than complete agreement between the panel majority decisions and the judicial outcomes. In the 35 cases where a definite decision was made using both criteria, there was 89% agreement between the two criteria. Three of the four disagreements occurred with persons judged guilty by the panel and acquitted in court, and only three of the eight acquitted by the judicial process were judged innocent by the panel. Therefore, the extent of disagreement between polygraph outcomes and the two criteria for guilt and innocence should be interpreted in light of the fact that the two criteria were in less than perfect agreement.

3. *Psychopaths*. Two types of data were obtained to assess the effectiveness of examinations performed on persons diagnosed psychopathic (sociopathic). The first consisted of the accuracy of decisions obtained in Experiment I. The outcomes based on numerical evaluation of those charts are presented in Table 5. Among the 24 subjects who had been diagnosed as psychopaths, decisions were 96% correct. The single error was a false positive, and not a single guilty psychopath was able to produce a truthful polygraph outcome. Although the polygraph tests appeared to be slightly more effective with the psychopaths than with the nonpsychopaths, there was not a statistically significant difference between the accuracy rates for the two groups.

TABLE 5 *Examiner Decisions Based on Total Numerical Scores in Experiment I*

	Correct	Wrong	Inconclusive
Psychopaths	23	1	0
Nonpsychopaths	19	1	4

The second set of data with regard to psychopaths and polygraph outcomes was obtained from the field study with criminal suspects. On the bases of their MMPI scores, two groups were obtained from the 36 suspects judged to be guilty by the panel majority. The 14 guilty suspects with the highest psychopath scores were compared to the 12 guilty suspects with the lowest psychopath scores. Using the total polygraph scores for the first three charts, the guilty psychopaths had a mean score of −7.8, and the guilty nonpsychopaths had a mean score of −7.1. There was no significant difference between those groups which indicated that the guilty psychopaths were detected by the polygraph as readily as were the guilty nonpsychopaths.

4. *Other subject characteristics*. The subjects from the study of criminal suspects were also compared on a number of biographical and personality variables. Those comparisons were made for sex, education, number of previous arrests, religiousness, previous polygraph tests, age, and the MMPI scores for the lie scale, K-scale, hypochondriasis scale, and depression scale. There were no indications that any of those variables were related to the polygraph results.

5. *Type of crime*. The strength of polygraph reactions exhibited by deceptive suspects was compared for different crime categories. When the total scores for the first three charts were compared for suspects accused of sex crimes, drug crimes, crimes of violence, and crimes of financial gain, there were no discernible differences among the groups. A similar analysis was performed to compare the categories of sex crimes, drug crimes, and crimes involving confrontation between criminal and victim, and crimes without confrontation between the criminal and victim. Again, there were no discernible differences among polygraph scores for deceptive suspects divided into those categories. Thus, there was no evidence that the type of crime affected the strength of polygraph reactions among suspects found deceptive on the polygraph test.

6. *Behavior symptoms*. The predictions based upon the observation of behavior during the pretest phase of the polygraph examinations of criminal suspects were compared to the judgments of guilt or innocence made by a majority of the panel. The initial predictions agreed with the panel in 56% of the cases, and the later predictions agreed with the panel in 69% of the cases. Neither of those results was significantly above chance, indicating that systematic observation of behavior during the pretest phase of the polygraph examination was of no value in determining truth or deception.

B. Reliability of Chart Interpretation

This project included four different assessments of the reliability of chart interpretation. Two of those involved comparisons between the decisions based on numerical scoring by the Project Director and Co-Director, the third compared decisions based on numerical scoring by the Project Director and Co-Director with the original decisions made by 43 examiners employed by law enforcement and private firms, and the fourth involved a study of accuracy of chart interpretation by 25 field examiners who evaluated the same set of 16 confirmed polygraph examinations of criminal suspects. The results of each is described below.

1. *Experiment II.* The 40 sets of polygraph charts obtained with the control question technique were scored numerically by the original examiner (Dr. Barland) and were independently scored by Dr. Raskin. Both examiners made a definite decision on 36 of the 40 subjects, and they were in agreement on 100% of them.

2. *Criminal suspects study.* The outcomes based on numerical scores by the original examiner (Dr. Barland) and those based on the blind evaluaton of the charts by Dr. Raskin are shown in Table 6. Both examiners obtained the same categorization in 86 of the 102 cases (84.3%) when inconclusives were included. On cases in which both examiners made a decision, they were in agreement 100% of the time. The correlation between the numerical scores assigned by the two examiners was very high, $r = .91$.

TABLE 6 *Comparison of the Original Examiner's Chart Evaluation and the Blind Evaluation of the Charts By an Independent Examiner*

	Original Examiner's Scores			
Independent Evaluation	Truthful	Deceptive	Inconclusive	Total
Truthful	10	0	5	15
Deceptive	0	61	1	62
Inconclusive	1	9	15	25
Total	11	70	21	102

3. *Current practices by law enforcement and private examiners.* The extent of agreement between polygraph decisions by law enforcement and private examiners and those based on independent numerical evaluation by the Project Director and Co-Director is presented in Table 7. Location G utilized the relevant-irrelevant technique, and our evaluation of those charts was subjective since numerical scoring could not be utilized. The overall rate of agreement was 57% when inconclusives were included and 85% when inconclusives were excluded. There was a significant difference in agreement across the different locations sampled, and our evaluations showed significantly higher rate of agreement with decisions made by law enforcement (92%) as compared to private polygraph firms (79%). The results were separated by type of crime, and there were no significant differences in rate of agreement for crimes against people, economic crimes, or sex and drug crimes.

TABLE 7 *Agreement Between Independent Evaluations and Original Decisions at Each Location*

	% Agreement	
	Including Inconclusives	Excluding Inconclusives
Police		
A	88.3%	98.0%
B	42.6%	95.8%
C	58.6%	82.1%
Combined	64.0%	92.1%
Private		
D	60.0%	89.7%
E	43.3%	76.5%
F	55.0%	80.6%
G	50.8%	75.6%
Combined	52.2%	79.9%

4. *Accuracy of chart interpretation.* Of the 400 judgments made by the 25 polygraph examiners on the set of 16 polygraph examinations, 79% were correct decisions, 8% were errors, and 13% were inconclusive. Excluding inconclusives, 90% of the decisions were correct. Accuracy ranged from 53% correct decisions for one examiner to 100% correct decisions for nine examiners. The proportion of the errors that were false positives (60.6%) was more than twice as high as would be expected by chance. There was no significant difference in accuracy of decisions for examiners with at least one year of experience (92%) as compared to those with less than one year of experience (89%). However, the seven examiners who employed numerical scoring of the charts were significantly more accurate in their decisions (99%) than the 18 examiners who did not use numerical scoring (88%). Furthermore, even among examiners who had received formal training in numerical scoring, the seven examiners who explicitly employed numerical evaluation achieved significantly higher accuracy of decisions (99%) than the six examiners who knew how to numerically score charts but did not explicitly employ the technique (88%).

C. Relative Effectiveness of Standard Measures

In four studies the standard field measures were compared in terms of their relative effectiveness of

identifying truthful and deceptive subjects using numerical scoring of responses. That was accomplished in both laboratory experiments and two of the field studies.

1. *Laboratory experiments.* In Experiment I the finger plethysmograph was used instead of a pressurized cardio cuff, and the numerical scores for the first three charts are shown in Table 8 for Experiments I and II. In both experiments the skin conductance measure provided the best discrimination between guilty and innocent subjects. In Experiment I all three measures showed significant capacity to identify both guilty and innocent subjects, and the respiration component identified innocent subjects better than it identified guilty subjects. In Experiment II the plethysmograph measures produced better discrimination between guilty and innocent subjects than did the cardio and respiration measures. The cardio and respiration measures showed significant identification of innocent but not guilty subjects. It should be noted that respiration in Experiment II was measured with a device different from that typically employed in the laboratory or field. Also, the cardio was measured using a low-pressure cuff at an inflation pressure between 50 and 60 mmHg.

TABLE 8 *Mean Numerical Scores for Each Standard Component for the First 3 Charts in Two Experiments*

	Experiment I		Experiment II	
	Guilty	Innocent	Guilty	Innocent
Respiration	−2.3	+4.3	+0.3	+1.8
Skin Conductance	−5.8	+3.8	−5.0	+5.6
Cardio	—	—	0.0	+3.6
Plethysmograph	−3.0	+1.3	−4.2	+3.1

2. *Field studies.* The three standard components used in the study of accuracy with criminal suspects were evaluated using the numerical scores on the first three charts of 32 suspects who subsequently made full confessions or pleaded guilty to the original charge. Those numerical scores were assigned by the independent evaluator who had no knowledge of the confessions or pleas. All three of the components yielded significant deceptive scores for those confirmed guilty subjects, and the skin resistance (galvanic skin response) measures produced significantly better results than respiration or cardio measures.

Finally the three standard components were evaluated using the numerical evaluations of the charts obtained from the six law enforcement and private locations which used control-question tests. Two types of data were utilized. First, our numerical scores were tabulated for the 179 cases in which our decisions agreed with the decisions made by the original examiner. The mean scores for each component for the 147 deceptive and 32 truthful decisions are shown in Table 9. All measures contributed significantly to both deceptive and truthful outcomes, and the skin resistance measure was significantly larger than respiration and cardio for deceptive outcomes. The other analysis was performed on 56 deceptive results which were confirmed by confessions or admissions. All three measures showed significant identification of guilty subjects, and the mean skin resistance score (−5.8) was significantly larger than that obtained for respiration (−2.9) or cardio (−2.8).

TABLE 9 *Mean Numerical Scores for Each Measure in Cases Where the Original Decision and the Independent Numerical Evaluation Agreed*

	Respiration	Skin Resistance	Cardio-vascular	Total
Deceptive (N = 147)	−3.8	−7.4	−4.7	−15.9
Truthful (N = 32)	+4.5	+4.8	+2.7	+12.0

D. Quantitative Analyses of Physiological Responses

In order to identify the characteristics of the various physiological measures which appear to be useful in detecting truth and deception, detailed quantitative analyses were performed on the polygraph recordings obtained in Experiments I and II and on the 14 sets of charts using those measures obtained from criminal suspects. Using control-question tests, significant effects were demonstrated when guilty subjects showed larger reactions to relevant as compared to control questions and innocent subjects showed larger reactions to control as compared to relevant questions. Only those measures which showed some effectiveness in the laboratory were tested with the sample of criminal suspects. Since measures of respiration and skin conductance amplitude were used to categorize the criminal suspects as truthful or deceptive, they were not analyzed quantitatively.

1. *Respiration amplitude.* Experiment I measured both thoracic and abdominal respiration. Both measures of respiration produced clear indications of greater suppression in respiration amplitude following relevant questions for guilty subjects and control questions for innocent subjects. Furthermore, thoracic respiration showed an increase in amplitude following relevant questions for innocent subjects. A similar effect did not occur in abdominal respiration.

No significant results were obtained with respiration amplitude in Experiment II. However, the transducer used in Experiment II appears to have been inadequate. Based on the results of Experiment I, it appears that suppression of respiration is a clearly effective indicator, and increase in respiration amplitude should not be used as in index of deceptive reaction but as a lack of reaction.

2. *Respiration cycle time.* There were significant effects in respiration cycle time demonstrated in Experiment I. The effects were accounted for by the reactions of the innocent subjects who showed a slowing in respiration following control questions and a speeding of respiration following relevant questions. Respiration rate showed no significant effects in Experiment II, but the poor transducer may have prevented any significant findings. Based on the results obtained, it appears that slowing of respiration may be considered as a reaction, but speeding of respiration should be viewed as a lack of reaction.

3. *Skin conductance response amplitude.* In both Experiment I and Experiment II skin conductance response amplitude was larger following relevant questions for guilty subjects and following control questions for innocent subjects. With guilty-knowledge tests, guilty subjects produced larger responses to critical items. Thus, skin conductance response amplitude was found to be very effective.

4. *Skin conductance response rise time.* In Experiment II the time required for skin conductance to reach its maximum level from the beginning of a response was found to be shorter to relevant questions for both guilty and innocent subjects. Therefore, it was not useful in identifying deceptive reactions.

5. *Skin conductance response recovery.* In Experiment II the amount of time required for skin conductance responses to return half of the distance back toward their base level before the response began was shown to be effective for two different measurement methods. Basically, the results showed that skin conductance responses which began after the onset of a question recovered more slowly following relevant questions for guilty subjects and following control questions for innocent subjects. However, those effects occurred only when Backster type control questions were used and not with tests employing Reid control questions. With guilty-knowledge tests, there was some indication of slower recovery of responses following critical items for guilty subjects. Measurement of skin conductance response recovery also produced significant effects with the sample of criminal suspects. Deceptive suspects showed slower recovery of responses to relevant questions. Thus, the slowness with which a skin conductance response returns to its pre-response level may be given some consideration as an index of response.

6. *Cardio responses.* The cardio responses were measured in Experiment II with the low pressure cuff inflated to a pressure between 50 and 60 mmHg. The changes in diastolic pressure were similar and somewhat better than those obtained with changes in systolic pressure. Diastolic pressure was measured on a second-by-second basis, and the changes in diastolic pressure are shown in Figure 2 for the 14 seconds following the beginning of the questions. The guilty subjects failed to show any differential response to control and relevant questions. However, the innocent subjects showed an increase in diastolic pressure following control questions and a decrease in diastolic pressure following relevant questions. The subsequent tests conducted with criminal suspects utilized inflation pressures of approximately 70 mmHg which appear to yield better reactions and more stable baselines. Using that pressure with criminal suspects yielded significantly greater increases in diastolic pressure following relevant questions for deceptive suspects and following control questions for truthful suspects. In addition, significant decreases in diastolic pressure showed greater decreases in diastolic pressure following control questions, and truthful suspects showed greater decreases following relevant questions. Since no reliable changes in cardio pulse amplitude were found in Experiment II, pulse amplitude changes were not measured with the criminal suspects. However, the improved quality of recordings obtained with 70 mmHg pressure showed some indications of appropriate decreases in pulse amplitude. Based on the obtained results with laboratory subjects and criminal suspects, it seems safe to conclude that increases in diastolic pressure are good indicators of reactions and decreases in diastolic pressure should be considered as lack of reaction.

7. *Finger blood volume.* Using a photoelectric plethysmograph, significant changes in finger blood volume were obtained in Experiments I and II. In both experiments guilty subjects showed greater decreases in finger blood volume following relevant questions, and innocent subjects showed larger reactions following control questions. With the guilty-knowledge test, guilty subjects showed significantly larger reactions to critical items. Significant results were obtained with the sample of criminal suspects, but only the deceptive suspects showed the effects. In

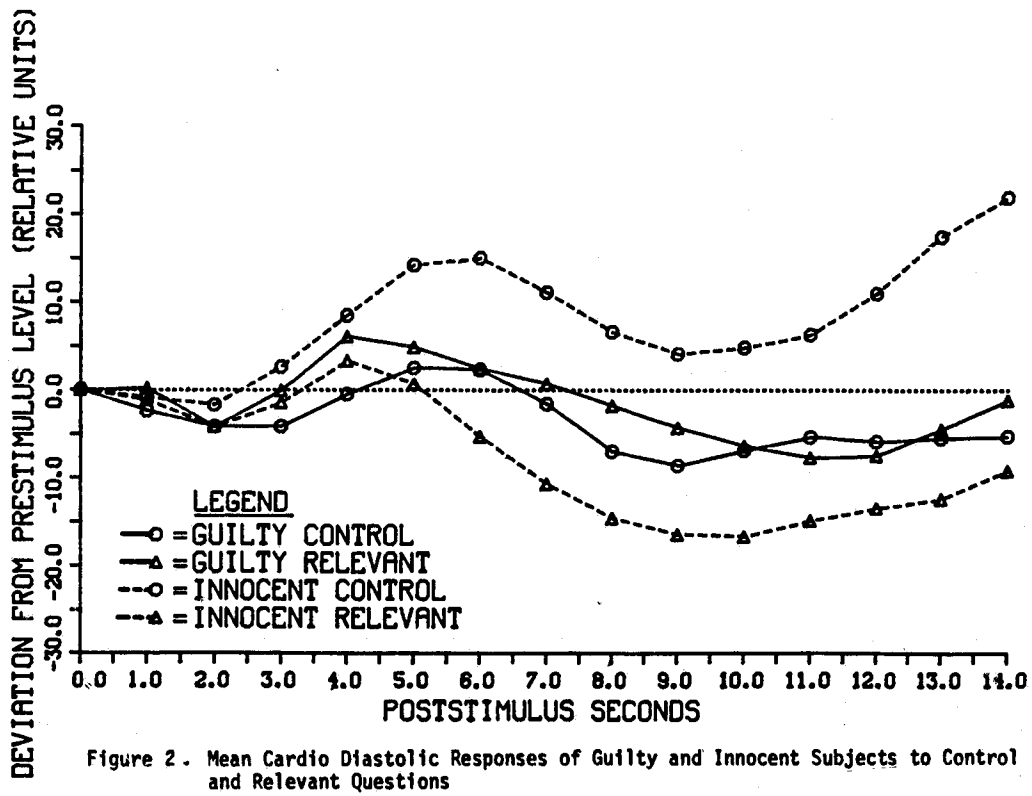

Figure 2. Mean Cardio Diastolic Responses of Guilty and Innocent Subjects to Control and Relevant Questions

Experiment II the duration of responses was longer to relevant questions for guilty subjects and to control questions for innocent subjects. Duration effects were not obtained with the criminal suspects. On the basis of the obtained results, it is clear that decrease in finger blood volume is a good indication of reaction, and duration of reaction may also be of some use.

8. *Finger pulse amplitude.* The amplitude of pulses was obtained from the same plethysmograph used for blood volume by recording with a short time constant coupling. In Experiment I guilty subjects showed significantly larger decreases in pulse amplitude following relevant questions, but for innocent subjects there was no difference in magnitude of pulse amplitude changes following control and relevant questions. In Experiment II the changes in pulse amplitude were measured on a second-by-second basis for the 14 seconds following the beginning of the questions, and the results are shown in Figure 3. The guilty subjects showed significantly greater decreases in pulse amplitude following relevant questions, and the innocent subjects showed longer-lasting decreases following control questions. With guilty-knowledge tests, significantly greater decreases occurred to critical items for guilty subjects. The measurements of decrease in pulse amplitude in the criminal suspects showed significantly larger decreases in pulse amplitude following relevant questions for deceptive suspects, but no differences between reactions to control and relevant questions were obtained for truthful suspects. Based on the obtained results, decrease in finger pulse amplitude can be considered a useful index of reaction.

9. *Heart rate changes.* In Experiment I heart rate in beats per minute was measured on a second-by-second basis from prior to the beginning of the questions through 20 seconds following question onset. The changes in heart rate are shown in Figure 4. Guilty and innocent subjects showed an initial increase in heart rate to control and relevant questions. Following that initial increase, heart rate returned to previous levels except for guilty subjects following their answer to relevant questions. They showed a clear slowing of heart rate which was more pronounced for the psychopathic group. In Experiment II virtually identical results were obtained with guilty subjects producing slowing of heart rate following their answers to relevant questions and innocent subjects showing no differences in heart rate responses to control and relevant questions. However, analysis of heart rate responses by the criminal suspects failed to produce any differentiation between control and relevant questions. Although the laboratory results appear to be encouraging with regard to slowing of heart rate as an indicator of reaction in guilty subjects, the results with criminal suspects failed to produce a similar result. At this time, the use of heart rate slowing as an index of reaction should be viewed with caution.

10. *Cardio activity monitor.* In Experiment II the systolic and diastolic levels obtained, from the cardio activity monitor (CAM) were measured on a second-by-second basis from just prior to the beginning of the question through 14 seconds following question onset. Only the changes in systolic level were significant, and those results are presented in Figure 5. There was an early increase in systolic levels which did not differentiate between control and relevant questions for guilty or innocent subjects. However, there was a subsequent decrease in systolic levels which appeared to be greater for guilty subjects in response to relevant questions. A similar but substantially stronger result was obtained with the guilty knowledge test, and those data are presented in Figure 6. There was an early, non-differential increase in systolic level followed by a very pronounced decrease in guilty subjects responding to critical items. Those results are very similar to the results obtained with finger blood volume and finger pulse amplitude as measured by a photoelectric plethysmograph. However, CAM measures of changes in systolic levels failed to produce any significant results with the sample of criminal suspects. At this time, it appears that the CAM measures do not provide very much useful information, and what is provided by the CAM can be better obtained from a photoelectric plethysmograph.

11. *Skin potential responses.* The negative and positive components of the skin potential response were measured in Experiments I and II. In Experiment I the significant results indicated that guilty subjects gave larger negative and positive skin potential responses to relevant questions than to control questions, but there was no differentiation for innocent subjects. There was also a significant effect for psychopathy which consisted of psychopaths producing positive skin potential responses which were disproportionately larger to relevant questions than those produced by nonpsychopaths. In Experiment II significant effects were obtained with the control-question technique only for negative skin potential, and the results with mean negative skin potential amplitude are presented in Table 10. They show that with Backster control questions the guilty subjects produced larger responses to relevant questions and

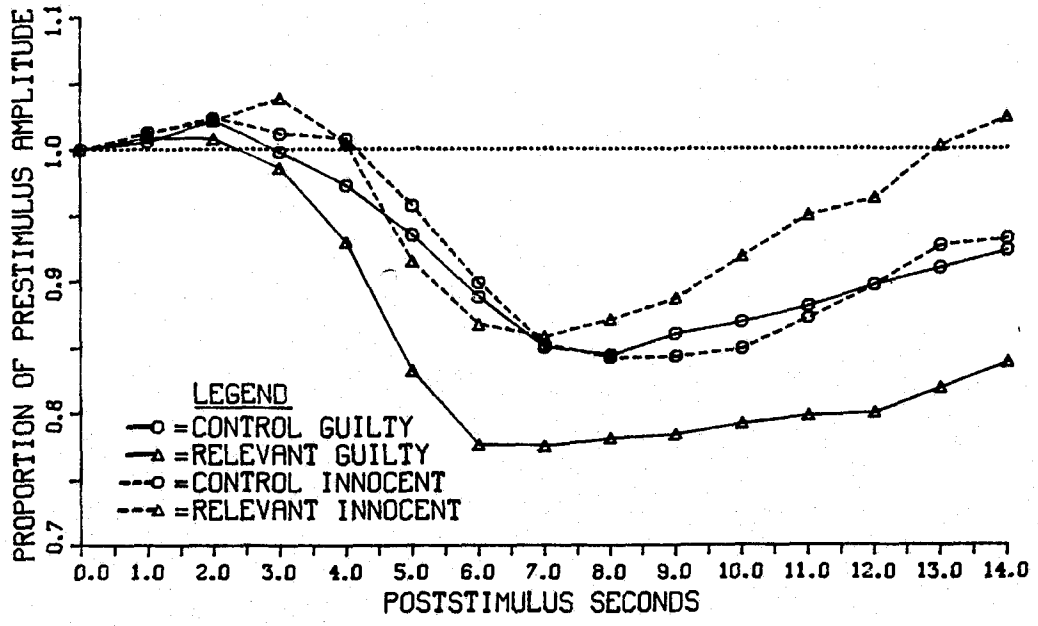

Figure 3. Mean Finger Pulse Amplitude Responses of Guilty and Innocent Subjects to Control and Relevant Questions

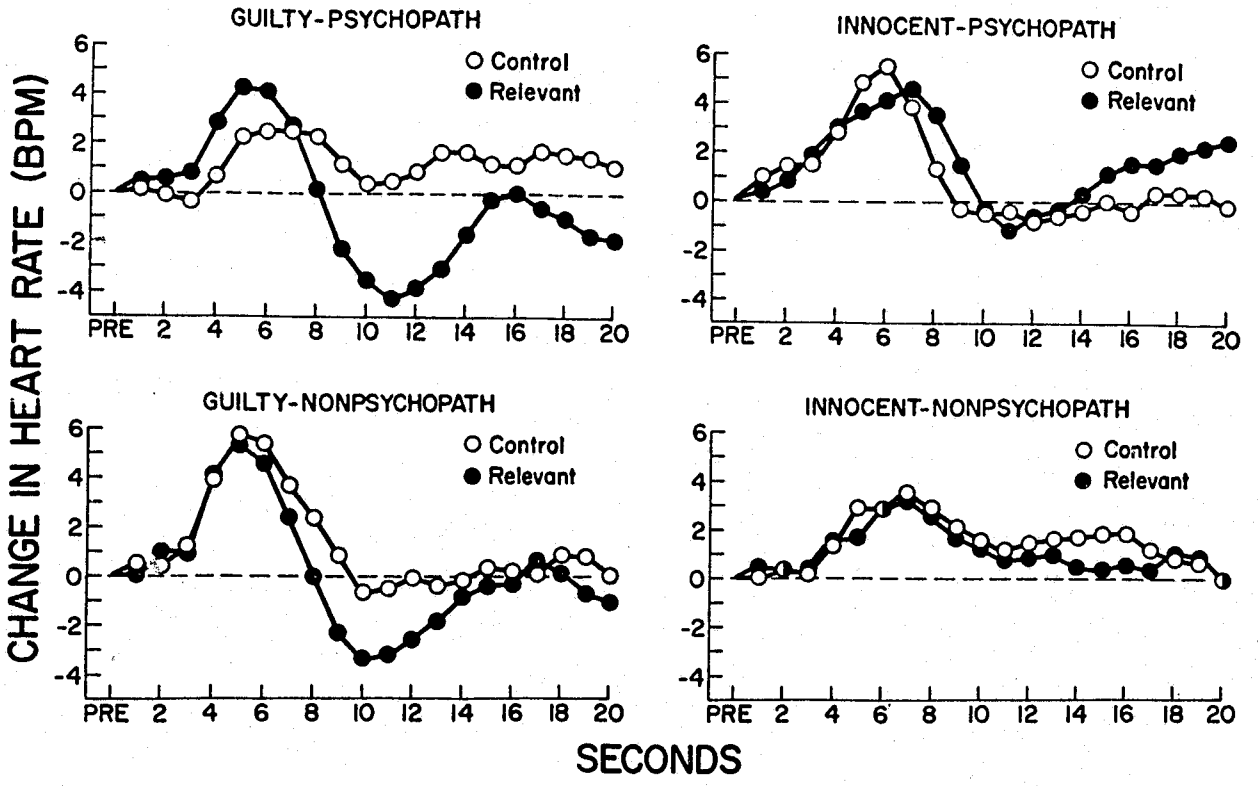

Figure 4. Mean Heart Rate Responses to Control and Relevant Questions for Guilty and Innocent Psychopaths and Nonpsychopaths.

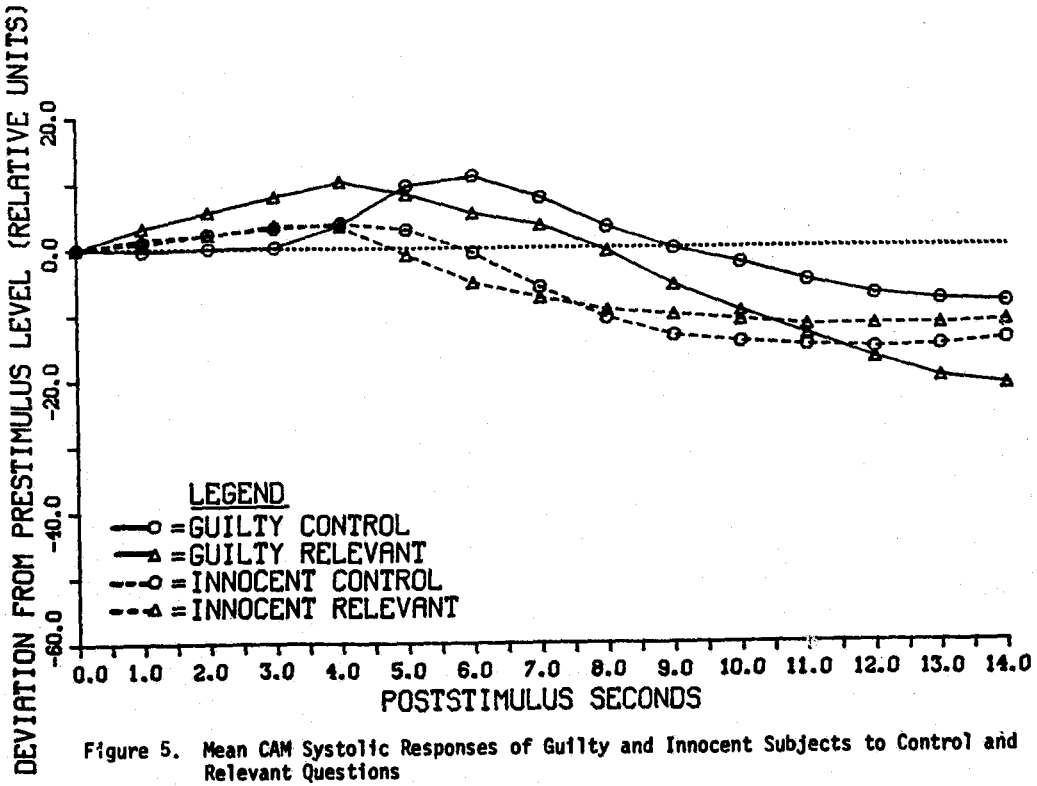

Figure 5. Mean CAM Systolic Responses of Guilty and Innocent Subjects to Control and Relevant Questions

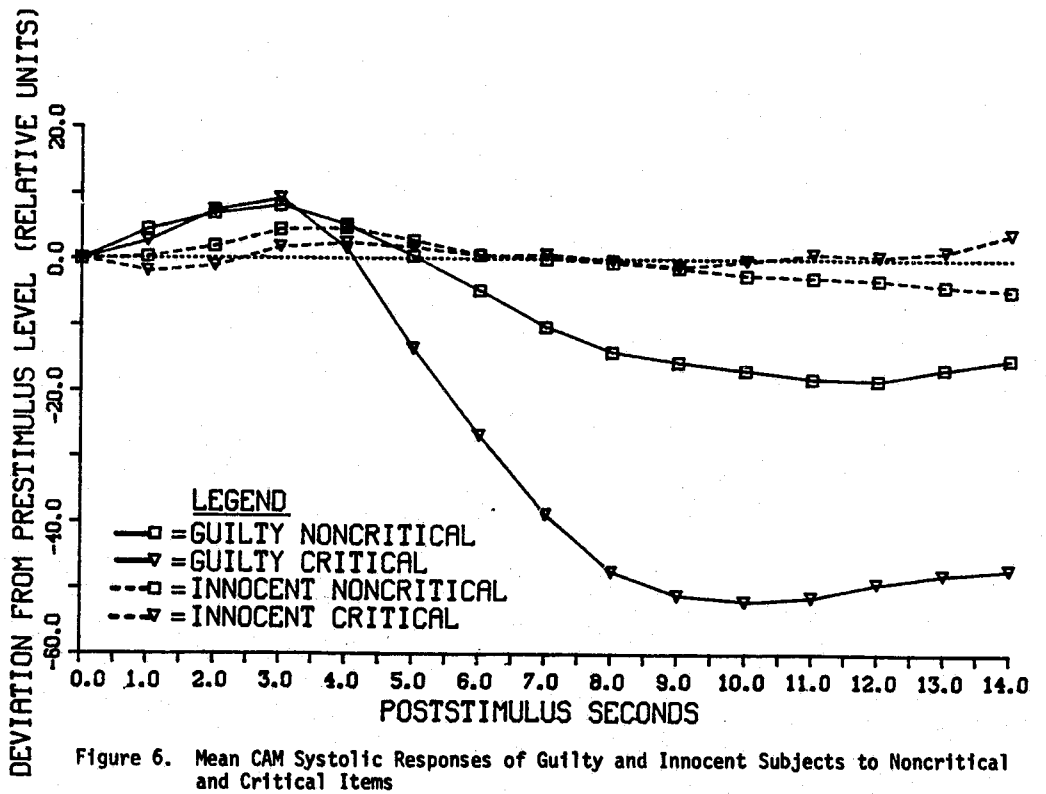

Figure 6. Mean CAM Systolic Responses of Guilty and Innocent Subjects to Noncritical and Critical Items

the innocent subjects produced larger responses to control questions. No differentiation was obtained with Reid control questions. The guilty-knowledge tests also produced significantly larger negative and positive skin potential responses by guilty subjects to critical items. The measurement of skin potential responses with criminal suspects showed larger amplitude of positive responses by deceptive suspects to relevant questions, but no significant results were obtained with negative skin potential. Thus, it appears that measures of skin potential may be of some use as an index of deception, but they appear to be inferior to other measures of electrodermal activity such as skin conductance amplitude and recovery time.

TABLE 10 *Mean Negative Skin Potential Response Amplitude (mV) with the Control-Question Technique and Guilty-Knowledge Technique in Experiment II*

		Control Question	
		Control	Relevant
Backster	Guilty	1.0	1.7
	Innocent	1.4	1.1
Reid	Guilty	1.1	1.2
	Innocent	.9	.8
		Guilty Knowledge	
		Noncritical	Critical
GKT	Guilty	1.3	2.5
	Innocent	1.4	1.2

E. Comparisons of Question Techniques

Laboratory Experiment II was designed to allow evaluations of several different approaches to question structure employed in polygraph examinations. Those features included comparisons of the effectiveness control-question tests with that of guilty-knowledge tests, evaluations of the relative effectiveness of Backster and Reid types of control questions, and the relative usefulness of a guilt-complex question utilized as a control question. The results obtained with each of those questions are described below.

1. *Relative accuracy of control-question and guilty-knowledge tests.* The outcomes based on numerical scoring of control-question and guilty-knowledge tests are presented in Table 11. The rate of accuracy of decisions using control-question tests was 94% with Backster type control questions and 83% with Reid type control questions. Since the difference in accuracy rate for those two types was not significant, their combined accuracy rate was compared to that obtained with the guilty-knowledge test. The accuracy of guilty-knowledge tests was 90% as compared to the combined decision accuracy of 89% with control-question tests. The types of errors which occurred consisted of false negative errors in all but one subject who was tested with a control-question technique. Thus, the overall accuracy of decisions was virtually identical for control-question and guilty-knowledge tests. In addition, quantitative analyses of physiological responses produced a large number of significant results using both control-question and guilty-knowledge tests as previously described.

TABLE 11 *Accuracy of Decisions and Types of Errors Using Backster and Reid Control Questions and Guilty-Knowledge Tests in Experiment II*

	% Correct	% False Positive	% False Negative	% Inconclusive	% Correct Decisions
Backster	85	0	5	10	94
Reid	75	5	10	10	83
Guilty-Knowledge	90	0	10	0	90

2. *Comparison of Backster and Reid control questions.* As indicated above, the tests which utilized Backster control questions produced a slightly higher accuracy rate than those which utilized Reid control questions, but the difference was not statistically significant. However, when the total numerical scores for the first three charts were compared, some differences were observed. The tests using Backster control questions produced significant identification of innocent (mean score = +13.6) and guilty (mean score = −11.7) subjects, but the results with Reid control questions were significant for innocent (mean score = +14.2) but not guilty (mean score = −6.3) subjects. Quantitative analyses of physiological responses also produced some results which indicated a superiority for tests utilizing Backster control questions. As previously described, measures of skin conductance response recovery times and amplitude of negative skin potential responses showed stronger reactions to relevant questions by guilty subjects and to control questions by innocent subjects only with Backster control questions. The test which utilized Reid control questions showed no discrimination for either of those measures. Thus, it appears that control questions which are separated from the relevant issue by age or time of occurrence have some advantages over control questions which do not have those exclusionary characteristics.

3. *Effectiveness of a guilt-complex question.* In order to evaluate the usefulness of a guilt-complex question as a control question, Experiment II included a guilt-complex question as the eighth question on each chart. The reactions to that question

were evaluated by using it as a control question for the relevant question at position 7 and performing a standard numerical scoring for that question pair on the first three charts. The scores obtained in that manner were then compared to those obtained by a numerical evaluation using the control question at position 9 and the same relevant question. The results of those evaluations are presented in Table 12. When the reactions to the standard control question were compared to those produced by the relevant question, the mean scores for guilty and innocent subjects were of approximately the same magnitude but in opposite directions, showing significant discrimination for guilty and innocent subjects. The guilt-complex question produced a negative mean score for guilty subjects but failed to produce any difference for the innocent subjects. In addition, the frequencies of scores in the wrong direction (positive scores for guilty subjects and negative scores for innocent subjects) were tabulated. With the guilt-complex question, scores of "0" with guilty subjects were considered to be in the wrong direction since the theory of guilt-complex questions predicts that only innocent subjects would produce scores of "0." Using standard control questions, there were three scores in the wrong direction with guilty subjects and three scores in the wrong direction with innocent subjects. With the guilt-complex question, there were three such errors with guilty subjects and nine with innocent subjects. Thus, it appears that the standard control questions were clearly more effective than guilt-complex questions in identifying innocent subjects.

TABLE 12. *Mean Scores of Guilty and Innocent Subjects when a Control Question and the Guilt-Complex Question were Compared with a Relevant Question*

	Comparison Question	
	Control	Guilt Complex
Guilty	−4.2	−3.7
Innocent	4.0	0.0

F. The "Friendly Polygrapher"

Three sets of data were obtained in order to evaluate Orne's "friendly polygrapher" hypothesis. The first sample showed that defense cases produced 78% truthful, 20% deceptive, and 2% inconclusive outcomes. The law enforcement cases produced 76% truthful, 20% deceptive, and 5% inconclusive outcomes. Contrary to the "friendly polygrapher" hypothesis, there was no difference in frequency of truthful outcomes for defense and law enforcement examinations conducted by the same examiner. The second sample produced mean numerical scores of −4.7 for defense cases and −2.0 for law enforcement/employer cases. Although the difference between those means was not significant, it was in the opposite direction from that predicted by the "friendly polygrapher" hypothesis. The third sample produced mean numerical scores of −10.4 for defense cases and −0.7 for law enforcement cases. The difference between those means was statistically significant and in the opposite direction from that predicted by the "friendly polygrapher" hypothesis. Thus, the three samples of data obtained to test the predictions from the "friendly polygrapher" hypothesis not only failed to produce any evidence to support that hypothesis, but some of the results indicated effects which were totally contrary to Orne's speculations.

G. Errors in Examinations

This project provided four different sources of possible data concerning errors. The first was the two laboratory experiments utilizing mock crimes. The second was the field study of criminal suspects using criteria of ground truth developed with a panel of experts. The other two sources consisted of independent evaluations by field examiners of confirmed cases provided by us and cases from our own laboratory and those referred to us by other examiners.

In the two laboratory experiments, there was a total of eight errors in 108 subjects. They consisted of three false positives and three false negatives with control-question tests and two false negatives with the guilty-knowledge test. The laboratory results seem to indicate a low rate of errors equally divided among false positives and false negatives using control question tests and only false negatives using the guilty knowledge test.

The results obtained from examinations of criminal suspects were somewhat different. Using the panel criterion, there were six false positives and one false negative with 92 suspects. When polygraph examiners made 400 independent evaluations of polygraph charts of confirmed cases from our laboratory, 20 of the 33 errors were false positives. Only eight false positives would be expected if examiners were equally likely to make either type of error when interpreting charts.

The last sample concerning errors yielded 12 cases in which there was clear evidence of inappropriate physiological responses on one of the polygraph examinations. It appears that all but one of the cases

yielded false positive results on the first examination, and one suspect produced a false negative result. The preponderance of false positives is not surprising since a guilty subject would be unlikely to insist that a truthful outcome was erroneous. The one case of a false negative appeared to be the result of deliberate countermeasures which produced substantial respiration reactions to control questions. A subsequent examination at our laboratory confirmed the deception and the obvious use of countermeasures. Among the false positive results, four were resolved by a truthful result on a subsequent polygraph test which followed a restructuring of some of the questions in order to separate a related but irrelevant concern expressed by the suspect. Such concerns were typically incorporated into new control questions or simply expressed by the subject as minor admissions prior to the second test.

There were a number of characteristics which seemed to be prevalent among the false positive cases. In all but one case, the suspect had no previous experience of being in serious difficulty with regard to criminal activity. They were generally well-educated (six had college degrees), had middle-class values, and expressed strong concerns about their reputations and their personal distress over being charged with a criminal act. That pattern of characteristics is not typical of criminal suspects who are given polygraph examinations. In five of the cases, the suspect was examined twice by Dr. Raskin or Dr. Barland and produced deceptive results on the first test. After being informed of the outcome, each of them described something which caused an emotional reaction to the relevant questions. In three instances minor, non-incriminating admissions were made, and the second test produced truthful results with the same relevant questions. In the other two cases, the suspects described a feature of the situation which caused them to experience an emotional reaction to the relevant questions, e.g., the use of a certain name or feelings of guilt or responsibility. When such material was incorporated into control questions, truthful results were obtained on the second test.

III. Conclusions and Recommendations

The results of this project clearly indicate that polygraph examinations utilizing control-question or guilty-knowledge tests are highly accurate. In light of the available evidence from the laboratory and field, it seems reasonable to conclude that the accuracy of such tests is approximately 90% when they are properly conducted and evaluated.

With regard to specific techniques, it appears that the control-question test utilizing Backster type control questions is the most accurate test which is suitable for a wide variety of criminal investigations. Although the guilty-knowledge test also produced high levels of accuracy, it is seldom possible to utilize that type of test due to the nature of most crimes or the lack of significant items of information which would be known only to the guilty person and the investigators. The relevant-irrelevant test possesses many weaknesses (Podlesny & Raskin, in press) and should not be used as a substitute for control-question tests.

The results of this project clearly indicate that numerical scoring of polygraph charts produces higher rates of accuracy and reliability of chart interpretation than other methods of chart interpretation. The basic scoring system taught by the U.S. Army Military Police School seems to produce good results using scores of ±5 inclusive to define the inconclusive zone. However, the criteria for evaluating reactions may require some slight modifications on the basis of the results obtained from the studies of physiological responses performed on this project.

The results obtained with detailed analyses of physiological measures support the continued use of respiration, skin conductance (galvanic skin response), and cardiovascular (relative blood pressure) measures. Among all of the variables measured, the skin conductance response (galvanic skin response) was clearly superior in laboratory experiments and with criminal suspects in field situations. In addition, the results clearly indicated that a properly-designed photoelectric plethysmograph would make a useful addition to field polygraph instruments. Unfortunately, those which have been marketed in the past have not met the performance standards which are easily attainable. The substitution of an electronic, low-pressure blood pressure (cardio) device for the mechanical, high-pressure system seems clearly beneficial and desirable. The results obtained with recordings made at pressures of 70 mmHg were quite good, and the use of inflation pressures below diastolic blood pressure have the clear advantages of reducing discomfort to the subject and allowing a slower rate (25–30 seconds) of question presentation. Although some positive findings were obtained with the cardio activity monitor (CAM) and measures of skin potential and heart rate, the results do not seem to warrant adding those measures to field polygraph instruments at this time.

There were a number of findings which have implications concerning the criteria for defining reactions and lack of reactions in different physiological measures. With regard to respiration, the results support the use of suppression of breathing, slowing of rate, and apnea as indicators of reaction. Speeding of breathing and increases in respiration amplitude were found to be associated with truthfulness and should be viewed as indicative of lack of reaction. Although no quantitative analyses were made on respiration baseline changes, their use in obtaining accurate results with numerical scoring was consistent with an interpretation of baseline arousal as a reaction. However, detailed analyses should be done to assess the usefulness of baseline arousal as an indicator of reaction.

Measures of the recovery times of skin conductance responses clearly demonstrated that longer-lasting responses are associated with reactions. Therefore, additional weight may be given to scoring skin conductance responses which show slower recovery toward baseline levels. Since the short time constant measurement technique which is employed in the automatic mode of recording such responses has the effect of eliminating information concerning recovery time, skin conductance (galvanic skin response) should always be recorded in the manual (long time constant) mode. Also, the use of the automatic mode can greatly distort the relative size and shape of those responses, and the use of automatic mode should be eliminated.

The findings with regard to cardio (relative blood

pressure) responses indicate that increases in diastolic level represent good indications of reaction. Furthermore, decreases in diastolic level were found to be associated with truthfulness and should not be used as indicative of reactions. There was little support for the use of decreased pulse amplitude as an indicator, but the measurements with the low pressure system may have greatly diminished the possibility of finding such changes. Therefore, at this time we do not recommend eliminating the use of decrease in pulse amplitude as an indicator of reaction.

The measures of finger pulse amplitude and finger blood volume provided excellent results. It seems clear that decreases in finger pulse amplitude and finger blood volume are very useful indicators of reaction. Since those measures are obtained by recording with greatly different time constants, manufacturers of field equipment should consider providing a photoelectric plethysmograph which offers a choice between a very short time constant (.1 second) and a very long time constant (28 seconds). With that option, the examiner could select the setting which provides the more useful recording with each subject.

The changes in physiological activity obtained on this project were measured from the beginning of the questions, and reactions were frequently observed beginning shortly after the question started and prior to the subject's answer. Although many field examiners are of the opinion that reactions which begin prior to the subject's answer are "listening reactions" and should not be utilized in arriving at decisions of truth or deception, the scientific evidence supports the use of reactions which begin soon after the start of the question. The use of such reactions would be a problem only if the questions had not been reviewed with the subject such as in the current practice with relevant-irrelevant tests, but such tests do not meet scientifically acceptable standards.

The investigation of problems associated with personality and psychopathy yielded a good deal of useful information. The results were strongly contradictory to the common belief that psychopaths (sociopaths) can "beat the lie detector" (Barland & Raskin, 1973). With convicted felons who were diagnosed psychopathic, not a single guilty subject was able to produce a truthful result. In fact, there were some indications that psychopaths may be somewhat easier to detect using polygraph examinations. The results with criminal suspects supported the position that deceptive psychopaths are as physiologically reactive and as readily detected as non-psychopaths. Thus, the fears that psychopathic criminals are able to be successful in deception during polygraph examinations can be dispelled. Perhaps the greatest danger is that a clever and convincing psychopath can talk a polygraph examiner into believing him, even though the polygraph charts indicate deception. Adequate training in chart interpretation and numerical scoring should prevent that from occurring.

With regard to a variety of personality, biographical, and circumstantial factors, the results failed to show any relationship between those variables and the polygraph outcomes. There were no differences attributable to aspects of personality as measured by the MMPI, age, sex, previous arrests or polygraph examinations, educational attainment, or the type of crime involved. Thus, in the absence of very low intelligence or any incapacitating psychological or physical illness, it seems reasonable to conclude that polygraph examinations are effective with a wide variety of individuals with respect to the broad range of crimes typically investigated.

With regard to the risk of errors, the findings provided information on important questions which have been raised. The concept of the "friendly polygrapher" (Orne, 1975) has been used as an argument against the validity of polygraph examinations conducted confidentially at the request of defense attorneys (Dogin, 1974). The findings obtained with three different samples of criminal cases are contrary to the "friendly polygrapher" notion. There appears to be no increased risk of false negatives under such circumstances. If anything, the results indicated a higher likelihood of deceptive results on defense-conducted examinations. Therefore, the results of defense-offered polygraph examinations should be accorded careful consideration if the examinations have utilized adequate techniques employed by properly trained and competent examiners.

The findings showed that in the criminal suspect situation there may be a somewhat higher risk of false positives rather than false negatives. In the study with criminal suspects using the panel criterion, all but one of the obtained errors were false positives. In addition, the chart interpretation errors by polygraph examiners showed a disproportionate number of false positives decisions, which is consistent with the report of Horvath (1974). Finally, among the criminal cases in which we were able to obtain fairly strong indications of an error, all but one appeared to be false positives. In the latter study,

the increased risk of false positives occurred with suspects who were generally well-educated, had no prior history of criminal activity, and were very concerned about their reputation or standing in the community. When such subjects proclaim their innocence after having produced deceptive reactions on polygraph examinations, the examiner should provide opportunities for the subject to explain why he responded to the relevant question if he was actually being truthful. If the subject provides an explanation which could be incorporated into a restructured set of questions, the examiner should give serious consideration to administering a second test. That procedure should be followed in all cases where the deceptive results are strongly disputed by the subject, but particular attention should be paid in those cases where the subject fits the pattern described above.

The usefulness of behavioral cues (Horvath, 1973; Reid & Inbau, 1966) was investigated in the study of accuracy with criminal suspects. The results were not supportive of the claims that behavioral observations are effective in assessing truth and deception. Similar results were obtained in Experiment II, and they showed that decisions based on behavioral cues produced more that 50% incorrect designations of innocent subjects as deceptive. Unfortunately, many examiners are taught to place great emphasis on gestures, verbal behavior, and mannerisms in arriving at a decision. At this time the evidence does not support such procedures, and examiners should restrict their basis for decisions to the physiological recordings on the polygraph charts.

On the basis of the existing scientific literature and the findings of the research performed on this project, the following conclusions and recommendations are made:

- Polygraph examinations can be highly accurate in determining truth and deception regarding specific issues in criminal investigations.
- In order to obtain maximum accuracy, polygraph examinations should employ control-question tests or guilty-knowledge tests, when applicable. Control questions should employ time or age exclusions of the type utilized in the Backster control questions.
- Polygraph recordings should include respiration, skin conductance (galvanic skin response), and relative blood pressure (cardio cuff). Manufacturers of polygraph instruments are urged to develop and market an adequate photoelectric finger plethysmograph, and examiners are encouraged to use measures of finger pulse amplitude and finger blood volume. The use of an electronic low-pressure cardio cuff is encouraged, using inflation pressures of 70 mmHg. Lower inflation pressures appear to produce less satisfactory results.
- Examiners should be given formal training in numerical scoring of polygraph charts. The numerical system should be of the type utilized by the U.S. Army and modified on the basis of the findings described above. The use of a total score of +6 or higher as a criterion for truthfulness and −6 or lower as a criterion for deception is supported by this research.
- The results of control-question examinations should always be determined by numerical evaluation of the polygraph charts. When important investigative or judicial decisions may be influenced by the results of such tests, an additional numerical evaluation of the polygraph charts by an independent examiner is recommended.
- If the above conditions are met, the results of this research indicate an accuracy rate of approximately 90% with criminal suspects. In light of that high rate of accuracy, the results of polygraph examinations should be given careful consideration in criminal investigations and judicial proceedings.

IV. REFERENCES

Backster, C. *Technique Fundamentals of the Tri-Zone Polygraph Test*. New York: Backster Research Foundation, 1969.

Bailey, F. L., & Rothblatt, H. B. *Investigation and Preparation of Criminal Cases, Federal and State*. Rochester, New York: The Lawyers Cooperative Publishing Company, 1970, 296–299.

Barland, G. H., & Raskin, D. C. Detection of deception. In W. F. Prokasy & D. C. Raskin (Eds.), *Electrodermal Activity in Psychological Research*. New York: Academic Press, 1973.

Barland, G. H., & Raskin, D. C. An evaluation of field techniques in detection of deception. *Psychophysiology*, 1975, *12*, 321–330.

Barland, G. H., & Raskin, D. C. *Validity and Reliability of Polygraph Examinations of Criminal Suspects*. (Report No. 76–1, Contract 75–NI–99–0001, U.S. Department of Justice). Salt Lake City, Utah: University of Utah, Department of Psychology, March, 1976.

Bersh, P. J. A validiation study of polygraph examiner judgements. *Journal of Applied Psychology*, 1969, *53*, 399–403.

Dogin, H. S. Statement of Henry S. Dogin, Deputy Assistant Attorney General, Criminal Division, Department of Justice. In, *The Use of Polygraphs and Similar Devices by Federal Agencies* (Hearings before a Subcommittee of the Committee on Government Operations, House of Representatives, Ninety-third Congress, Second Session, June 4 and 5, 1974). Washington: U.S. Government Printing Office, 1974, 413–417.

Horvath, F. S. Verbal and nonverbal clues to truth and deception during polygraph examinations. *Journal of Police Science and Administration*, 1973, *1*, 138–152.

Horvath, F. S. *The Accuracy and Reliability of Police Polygraphic ("Lie Detector") Examiners' Judgments of Truth and Deception: The Effect of Selected Variables*. Unpublished doctoral dissertation, Michigan State University, 1974.

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

Horvath, F. S., & Reid, J. E. The polygraph silent answer test. *Journal of Criminal Law and Criminology*, 1972, *63*, 285–293.

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

Lykken, D. T. The GSR in the detection of guilt. *Journal of Applied Psychology*, 1959, *43*, 385–388.

Lykken, D. T. Psychology and the lie detector industry. *American Psychologist*, 1974, *29*, 724–739.

Orne, M. T. Implications of laboratory research for the detection of deception. In N. Ansley (Ed.), *Legal Admissibility of the Polygraph*. Springfield, Illinois: Thomas, 1975.

Podlesny, J. A., & Raskin, D. C. Physiological measures and the detection of deception. *Psychological Bulletin*, in press.

Podlesny, J. A., Raskin, D. C., & Barland, G. H. *Effectiveness of Techniques and Physiological Measures in the Detection of Deception*. (Report No. 76–5, Contract 75–NI–99–0001, U.S. Department of Justice). Salt Lake City, Utah: University of Utah, Department of Psychology, August, 1976.

Raskin, D. C. *Psychopathy and Detection of Deception in a Prison Population*. (Report No. 75–1, Contract 75–NI–99–0001, U.S. Department of Justice). Salt Lake City, Utah: University of Utah, Department of Psychology, June, 1975.

Raskin, D. C. *Reliability of Chart Interpretation and Sources of Errors in Polygraph Examinations*. (Report No. 76–3, Contract 75–NI–99–0001, U.S. Department of Justice). Salt Lake City, Utah: University of Utah Department of Psychology, June, 1976.

Raskin, D. C., & Barland, G. H. *An Evaluation of Polygraph Techniques Currently Practiced by Law Enforcement and Private Polygraph Examiners*. (Report No. 76–2, Contract 75–NI–99–0001, U.S. Department of Justice). Salt Lake City, Utah: University of Utah, Department of Psychology, April, 1976.

Reid, J. E., & Inbau, F. E. *Truth and Deception: The Polygraph ("Lie Detector") Technique*. Baltimore: Williams & Wilkins, 1966.

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

☆ U.S. GOVERNMENT PRINTING OFFICE: 1978 O—267-969

END