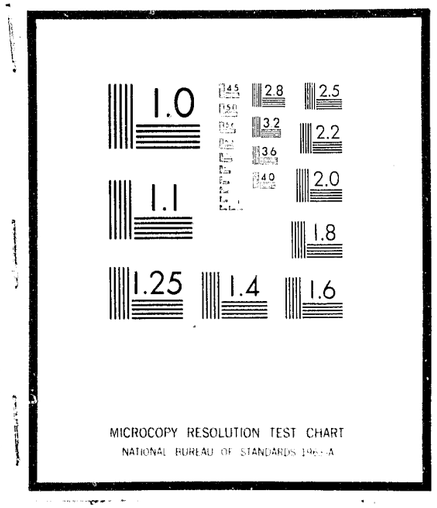

Reliability of Chart Interpretation and Sources of Errors in Polygraph Examinations

Report No. 76-3
Contract 75-NI-99-0001
National Institute of Law Enforcement and Criminal Justice
Law Enforcement Assistance Administration
U.S. Department of Justice

David C. Raskin, Ph.D.
Project Director

Department of Psychology
University of Utah
Salt Lake City, Utah 84112
June 7, 1976

Although field polygraph examiners have generally claimed that polygraph examinations yield less than 1% errors (e.g., Arther, 1965, 1968; Reid and Inbau, 1966), there are many who feel that error rates are considerably higher than 1%. Abrams (1973) and Barland and Raskin (1973) have reviewed a great deal of that literature, and only the most important studies will be discussed in this report. Since the control-question technique is the most widely accepted and utilized field technique, it is the only technique which will be considered in this report. The guilty-knowledge technique (Lykken, 1974) is seldom appropriate in field situations, and the relevant-irrelevant technique has been strongly criticized as lacking in adequate controls and scientific foundation (Podlesny & Raskin, 1976a).

The only controlled laboratory studies of the control-question technique have been conducted by Raskin and his co-workers (Barland & Raskin, 1975; Raskin, 1975; Podlesny & Raskin, 1976b). In those studies a mock crime situation was employed and the federal modification of the Backster zone comparison technique was employed. Excluding inconclusives, the accuracy of decisions was 81% in the Barland and Raskin study, 96% in the Raskin study, and 87% in the Podlesny and Raskin study.

Although the accuracy rate obtained in laboratory studies of the control-question test is quite high (88% for the three studies combined), a number of critics have questioned the accuracy of such tests in the field situation (Lykken, 1974; Orne, 1975; U.S. Congress, 1965). Various arguments are raised against the accuracy of such tests, and the types of errors described by the critics fall into the two general categories of false negatives and false positives.

In the field situation there are three possible outcomes for a polygraph examination, i.e., deceptive, truthful, and inconclusive. When a guilty person

is judged truthful on the basis of the polygraph examination, that is termed a false negative error. When an innocent person is judged deceptive on the basis of the polygraph examination, then the error is of the false positive type.

When adequate control-question technique is utilized, errors may be caused by two general factors. One type of error is the result of erroneous chart interpretation by the examiner. That is to say, the examiner has either made a decision contrary to the physiological information recorded on the polygraph charts; or he has rendered a decision when the charts yielded an inconclusive result. The second type of error occurs when the examiner has accurately interpreted the charts, but the subject had produced a pattern of physiological responses which led to an erroneous result on the basis of generally-accepted chart interpretation procedures. Each of those problems will be dealt with separately in this report.

A number of studies have examined the question of reliability of polygraph chart interpretation. Horvath and Reid (1971) studied the accuracy of decisions when seven experienced and three inexperienced field examiners evaluated 40 confirmed field cases which were selected to exclude charts "which were dramatically indicative of truth or deception (p. 277)." The overall accuracy of interpretation was 88%, and experienced examiners were 91% accurate. Among experienced examiners, 38% of the errors were false positives and 62% of the errors were false negatives.

In a later study by Hunter and Ash (1973) 20 verified criminal cases were evaluated by seven polygraph examiners, and the overall accuracy rate was 86%. Of the errors 55% were false positives, and 45% were false negatives. Subsequently, Slowik and Buckley (1975) reported 87% accuracy in a similar experiment.

The three studies described above used private polygraph examiners trained at John E. Reid Associates. A recent study by Horvath (1974) utilized 10 examiners employed by a law enforcement organization. They independently evaluated 56 verified and 56 unverified polygraph examinations from the files of a large state police department. Their overall rate of accuracy was 64% on the verified cases and 62% on unverified cases. The breakdown of errors on verified cases was 68% false positives and 32% false negatives.

All of the above reliability studies used chart evaluation procedures which are subjective in nature. However, three experiments in our laboratory have used a numerical scoring procedure developed by Backster and modified by the U.S. Army (see Barland & Raskin, 1975). With that procedure, numerical values are assigned to comparisons of responses to control and relevant questions according to rules, and the total score indicates whether the result is truthful, deceptive, or inconclusive.

In the Barland and Raskin (1975) study, 72 sets of charts obtained in a mock-crime experiment were sent out for independent evaluation by five military examiners experienced in using the numerical scoring system. When inconclusives were excluded, the overall agreement among examiners was 96%. The mean correlation between numerical scores was .86. In a subsequent study with 102 criminal suspects (Barland & Raskin, 1976), an independent evaluator agreed with the decisions of the original examiner 84% of the time when inconclusives were included and 100% of the time when inconclusives were excluded. The correlation between the scores of two examiners was .91. A third study just completed with a mock-crime situation (Podlesny & Raskin, 1976b) yielded 88% agreement between the original examiner and the independent examiner when inconclusives were included and 100% agreement when inconclusives were excluded.

Considering the high reliability of numerical scoring techniques, it seems reasonable to use such scores generated by experienced examiners trained in that technique as the basis for determining whether a particular set of polygraph charts is indicative of physiological patterns of truthfulness or deception. Thus, decisions based upon careful application of numerical scoring procedures would constitute a criterion for determining if a polygraph examiner has correctly interpreted the outcome of a polygraph examination. This would be useful in discriminating between errors which resulted from incorrect chart interpretation and errors due to inappropriate physiological responses produced by the subject. One of the purposes of the present study was to assess the accuracy of chart interpretation performed by polygraph examiners with a variety of training, experience, and familiarity with numerical scoring.

The problem of errors caused by inappropriate physiological responses has been raised recently by critics with quite different points of view (Lykken, 1974; Orne, 1975). Lykken has expressed the concern that control-question tests are likely to produce an unacceptable rate of false positives, whereas Orne is concerned that under certain circumstances an undesirable rate of false negatives may occur.

According to Orne's (1975) position, motivation to deceive and the threat of serious consequences are essential requirements for successful detection of deception. Such principles are well established and supported by others (e.g. Podlesny & Raskin, 1976a). On the basis of those principles, Orne has speculated that polygraph examinations conducted at the behest of defense attorneys fail to meet the requirements for successful detection of deception among guilty suspects. He reasons that a subject in such a situation "knows that the results of the test <u>if he is found deceptive</u> 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 from an "arm's length" examiner such as a law enforcement examiner or one working for an employer. Orne concluded that such a situation will make the guilty suspect less detectable.

A second purpose of the present study was an evaluation of the "friendly polygrapher" hypothesis which compared outcomes and numerical polygraph scores obtained by the same examiners on cases referred by defense attorneys and those referred by law enforcement sources and employers.

The paper by Lykken (1974) challenged the ability of the control-question technique to accurately identify innocent suspects. He argued that it is impossible to design control questions which will produce the same level of responsiveness in innocent suspects as the relevant questions evoke from guilty suspects. As a result, errors using control-question tests would be expected to be predominantly false positives resulting from larger responses to relevant questions by innocent suspects.

At the present time there are five scientifically-executed studies which have investigated the accuracy of control-question tests. Three of those are laboratory experiments conducted by Raskin and his associates (Barland & Raskin, 1975; Podlesny & Raskin, 1976b; Raskin, 1975). The accuracy rates excluding inconclusives were 81%, 87%, and 96%, respectively. Of the combined errors from the three studies 56% were false positives.

There are also two recent scientifically-conducted field experiments using the control question test with criminal suspects (Barland & Raskin, 1976; Bersh, 1969). Both of those experiments used the judgments of a panel of legal experts to establish the criterion for ground truth (guilt or innocence), and the Barland and Raskin study also used the additional criterion of judicial

outcome. When the criterion of majority panel decision was employed, the overall accuracy was quite high, being 87.5% in the Bersh study and 85% in the Barland and Raskin study. Of the observed errors 44% were false positives in the Bersh study, and almost all errors were false positives in the Barland and Raskin study. However, the latter study indicated that the panel criterion is not nearly perfect, and caution should be used in interpreting the results.

It appears that laboratory studies of the control-question technique yield almost equal proportions of false positive and false negative errors. However, there is some evidence which suggests a higher proportion of false positives in the field situation with criminal suspects. Therefore, a third purpose of the present study was to attempt to determine some of the factors which lead to false positives in the field situation.

Method

Three different substudies were conducted to investigate each of the questions raised above. They are described separately.

Accuracy of Chart Interpretation

Procedure. Sixteen sets of charts were selected from those obtained in a previous study (Barland & Raskin, 1976). All of the examinations had utilized the federal modification of the Backster zone-comparison technique (Barland & Raskin, 1973). Each examination consisted of three or more charts of 10 questions each recorded on a field type polygraph. Sometimes the examinations included a silent answer test, a yes test or a SKY series, and a number test was typically included. All of the examinations had been confirmed by the confession of the guilty person; 12 were from guilty subjects and 4 from innocent subjects.

Examiners. The charts were independently evaluated by field polygraph examiners from a variety of training backgrounds and experiences. Each examiner evaluated all 16 sets of charts. They were instructed concerning the type of question structure employed, and they were informed that all cases had been confirmed. They were told to compare each of the three relevant questions to the control question preceding it and to render a conclusion of truthful, deceptive, or inconclusive on each case. They were also told that if they had been trained in the numerical evaluation method, they could use that technique if they felt comfortable doing so.

The participating examiners had received their initial training at the following schools: American Institute of Polygraph Technology (2), Backster School of Lie Detection (4), Gormac School (6), National Training Center in Lie Detection (2), John E. Reid Associates (5), U.S. Army Military Police School (6), Unknown (1). Eighteen examiners had at least one year of experience. Thirteen of the examiners had received formal training in the numerical scoring method, but only seven of those explicitly scored the charts numerically.

The "Friendly Polygrapher"

Three different samples of examinations were obtained. First, the files of an experienced polygraph examiner who conducts control-question examinations for both law enforcement and private attorneys were sampled. For a 1-year period, all cases of criminal suspects examined for law enforcement or defense counsel were tabulated. They included the full range of criminal charges. A total of 204 cases was obtained including 98 defense cases and 106 law enforcement cases. A tabulation of the number of truthful, deceptive, and inconclusive decisions was made for each of the two referral sources.

The second sample consisted of cases sampled from two private polygraph

firms included in a previous study (Raskin & Barland, 1976). A total of 19 different examiners conducted the testing, and the cases were restricted to control-question tests of criminal suspects. If more than one suspect was examined in a particular case, only the first person examined was included in the sample. The cases were separated into those referred by defense counsel without knowledge of law enforcement or prosecution (N=54) and those which were conducted with explicit knowledge and/or agreement on the part of law enforcement authorities or by referral from the suspect's employer (N=57). Each of the polygraph examinations was subjected to a detailed numerical evaluation (Raskin & Barland, 1976) prior to obtaining information concerning the issue tested or the source of referral. Since at least two charts were run on each subject, the numerical score from those two charts comprised the data utilized in the analyses.

The third sample consisted of control-question examinations of criminal suspects referred by defense counsel without the knowledge of law enforcement officials and those examined with the explicit knowledge and/or agreement with law enforcement officials. The 27 examinations were conducted by the author, and 14 of them were conducted at the request of defense counsel. Since all examinations included a minimum of three charts, the total numerical scores for those charts were used as the raw data for the analyses.

Errors in Examinations

In order to attempt to discover factors which contribute to errors consisting of inappropriate physiological response patterns produced by the subject, efforts were made to obtain examples of such cases. A total of 12 cases were obtained from law enforcement and private examiners as well as cases referred to the author's laboratory. In some instances, contradictory results were obtained by two or more different examiners. In other cases,

reexamination by the same examiner yielded contradictory results. A third category consisted of tests shown to be in error by examination of a second person, confession by another person, or other compelling contradictory evidence.

With each case, additional case information was obtained if possible, and careful analyses of the polygraph charts, case information, and the subject's background were made.

Results

Accuracy of Chart Interpretation

Of the 400 judgments made by the 25 examiners on the set of 16 polygraph examination, 78.8% were correct decisions, 8.2% were wrong decisions, and 13.0% were inconclusive. Excluding inconclusives, 90.5% of the decisions were correct. The lowest correct for any examiner was 53%, and the highest was 100%. The use of the inconclusive category ranged from 0-44%. Of the 33 errors made by the examiners, 20 were false positives and 13 were false negatives. Thus, 60.6% of the errors were false positives, which was significantly (Z=4.52, p<.001) higher than the 25% expected on the basis of the proportions of truthful and deceptive subjects in the sample of charts evaluated.

There was no significant difference in mean percent correct decisions for examiners with at least one year of experience (M=91.7%) and examiners with less than one year of experience (M=89.3%), $t(23)$ = .48, p>.60.

The type of training received by the examiners had an important effect. The 10 examiners who attended schools which place a great deal of emphasis on numerical scoring (Backster, U.S. Army) had a significantly higher accuracy of decisions (M=97.1%) than the 15 examiners who attended other schools (M=86.9%), $t(23)$ = 2.54, p<.025.

The effect of numerical evaluation of charts was evaluated in two ways. First, the mean accuracy of the 7 examiners who numerically evaluated the charts (M=98.9%) was significantly higher than the 18 who did not numerically evaluate the charts (M=87.9%), $t(23)$ = 2.49, p<.025. Furthermore, among examiners who had received specific training in numerical evaluation, the seven examiners who explicitly used numerical evaluation on the charts had significantly higher accuracy of decisions (M=98.9%) than the six examiners who did not explicitly employ numerical evaluation (M=88.5%), $t(11)$ = 2.60, p<.05.

The "Friendly Polygrapher"

For the first sample of cases obtained on decisions made by a single examiner, the outcomes for defense cases were 77.6% truthful, 20.4% deceptive, and 2% inconclusive. The results for cases involving law enforcement participation were 75.5% truthful, 19.8% deceptive, and 4.7% inconclusive.

The numerical scores for defense and law enforcement/employer cases obtained from the two private polygraph firms were compared. The mean score on the first two charts was -4.7 for defense cases and -2.0 for law enforcement/employer cases. Although the mean scores were in the opposite direction from that predicted by Orne's "friendly polygrapher" hypothesis, the difference fell short of significance, $t(109)$ = 1.79, p<.10.

The comparison of numerical scores on the first three charts for the examinations conducted by the author showed that the mean score for defense cases was - 10.4 and the mean score for cases which involved law enforcement participation was -.7. Contrary to Orne's hypothesis, the defense cases produced scores which were significantly more in the deceptive direction than the cases which involved law enforcement, $t(25)$ = 2.19, p<.05.

Errors in Examinations

A total of 12 cases were obtained which showed clear evidence of inappropriate physiological reactions on the polygraph charts. In all cases more than one polygraph test was conducted on the subject. In four cases, at least one person in addition to the suspect had been examined on the polygraph. In all but one case at least one of the examinations of the suspect was conducted by Dr. Raskin or Dr. Barland. Nine of the subjects had been examined by one or more examiners other than Dr. Raskin and Dr. Barland.

Three criteria were used to justify the determination that an error had been made. In two cases subsequent evidence clearly proved that an error had been made. 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 was confirmed by a numerical evaluation of the charts by Dr. Raskin or Dr. Barland.

It appears that all but one of the errors obtained in the sample were false positives. That is not surprising since a false negative result would not be reported by a guilty subject. In eight cases there was confirmation in the form of compelling evidence or a polygraph test on another person which contradicted the results of the test on the subject. In four cases the only confirmation consisted of a subsequent truthful result on a polygraph test which usually followed a restructuring of some of the questions designed to separate out a related but irrelevant concern expressed by the subject. Such concerns typically involved some aspect of the subject's involvement in the circumstances of the crime which was not incriminating but caused the subject concern when the relevant questions were asked. Such concerns were typically incorporated into reworded control questions or simply

expressed as minor admissions by the subject following the deceptive outcome on the polygraph and prior to the subsequent polygraph test.

In attempting to discover the factors which contributed to the observed errors, a number of characteristics of the subject and the circumstances of the case were evaluated. Several of them appear to be related to the occurrence of the false positive results obtained.

In all but one of the false positive cases the subject 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 at being put in the situation of being charged with a violation of the law. This pattern of characteristics is not typical of criminal suspects who are examined on the polygraph.

In five of the cases the subject was examined twice by Dr. Raskin or Dr. Barland and was found deceptive on the first test. After being informed of the deceptive result, each of them described something which caused an emotional reaction to the relevant questions. In three instances minor, non-incriminating admissions were made, and a subsequent test produced truthful results on the same relevant questions. In the other two cases the subjects described a feature of the situation which they felt caused them to react emotionally to the relevant questions, e.g., use of a certain name or feelings of guilt or responsibility. When such material was incorporated into control questions on the subsequent test, truthful results were obtained.

In one case the subject was robbed while on duty at a store and was suspected of being an accomplice in the robbery. The robber had been a fellow employee, and it appears that the subject may have withheld his knowledge of the identity of the robber. In the remaining case, the subject was examined

twice by the same examiner who found him truthful on the second test. A careful inspection of the polygraph charts revealed clear evidence of countermeasures in the form of respiration manipulations. A subsequent examination at the University of Utah produced a deceptive result and also confirmed the use of respiration countermeasures. Thus, the original examiner had been misled by the countermeasures employed by the subject.

Discussion

In general the results of this study indicate that there is a relatively high accuracy in the decisions rendered solely on the basis of polygraph examiners' interpretation of the polygraph recordings. However, it is clear that training in numerical scoring of charts and the use of that procedure in the evaluation of all polygraph examinations results in a substantial increase in the accuracy of decisions. Even among those examiners who were trained in numerical scoring of charts, those who explicitly use the numerical evaluation procedure have a higher accuracy rate than those who do not.

The results obtained with regard to Orne's "friendly polygrapher" hypothesis do not support his assertion that there is a higher risk of false negatives in polygraph examinations conducted on a confidential basis at the request of defense attorneys. On the contrary, the findings generally indicated that suspects examined for defense counsel produced physiological responses significantly more indicative of deception than suspects examined with the knowledge and/or cooperation of law enforcement authorities. Thus, there appears to be no foundation for the opposition to defense-offered polygraph examinations which is based on the assertion of higher rates of false negatives in such circumstances.

The findings with regard to false positive errors point out some areas of caution in the acceptance of polygraph results in some cases. First, the risk

of erroneous evaluation of polygraph charts is higher when examiners do not employ numerical scoring procedures. In such instances, the errors are disproportionately of the false positive type. However, when numerical scoring was employed, there was only one error in 112 cases. Thus, it is clear that numerical scoring should be routinely employed in the evaluation of all control-question polygraph examinations.

The second problem regarding false positives involves the operation of psychological factors which occasionally leads to inappropriate reactions to relevant questions among innocent subjects. Although such false positive errors appear to be relatively infrequent, examiners should be on guard against their possible occurrence in certain types of subjects. Specifically, such errors seem to occur with individuals who have no prior criminal history, are relatively well-educated, and have middle-class values. In such cases the subjects may be overly concerned and anxious about the harm to their reputations in the community and the distress and embarrassment caused by their being suspected of a criminal act. The examiner must make great efforts to alleviate their anxiety, demonstrate the effectiveness of the polygraph by means of a number test, and create a focus of attention on carefully chosen control questions. In general, a psychologically sensitive and understanding approach is required with such subjects. An accusatory manner or interrogation of any sort could cause serious risks of error with such subjects.

Finally, there were several instances in which false positive errors included concern or anxiety over issues which the subject was unable to separate from the truthfulness of his answer to relevant questions. Thus, minor aspects of the situation, feelings of responsibility or nonspecific feelings of guilt caused reactions even though the subject was truthful in his answers to relevant questions. In some instances a discussion of the problems prior to a

second test was sufficient to eliminate the inappropriate reactions. In other cases, it was necessary to formulate new control questions which incorporated the concerns expressed by the subject. Using the same relevant questions, a second test with new controls produced a truthful outcome.

In summary, the data lead to the following conclusions and recommendations:

1. All examiners should be trained in numerical scoring techniques and should utilize numerical scoring on all control-question examinations. If the results are to be used as evidence or as the basis of an important decision, independent numerical evaluation should be obtained.
2. There is no scientific foundation for the "friendly polygrapher" hypothesis. If anything, examinations conducted on a confidential basis for defense counsel are more likely to produce deceptive results than those conducted with the involvement of law enforcement authorities.
3. There is a small, but demonstrable risk of false positives among certain types of subjects. Examiners should approach the situation with psychological sensitivity and understanding and make efforts to incorporate material related to the subject's expressed concerns into control questions on a retest in order to resolve the problem.

References

Abrams, S. Polygraph validity and reliability: A review. Journal of Forensic Sciences, 1973, 18, 313-326.

Arther, R. O. The scientific investigator. Springfield, Illinois: Thomas, 1965.

Arther, R. O. Re-examinations. Journal of Polygraph Studies, 1968, 3(2), 1-4.

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 validation study of polygraph examiner judgments. Journal of Applied Psychology, 1969, 53, 399-403.

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.

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. 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. Submitted for publication, 1976. (a)

Podlesny, J. A., & Raskin, D. C. The effectiveness of different question structures and physiological measures in detection of deception. In preparation, 1976. (b)

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., & 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. Congress, House Committee on Government Operations. Use of polygraphs as "lie detectors" by the federal government. 89th Congress, 1st Session. House Report No. 198. Washington, D.C.: U.S. Government Printing Office, 1965.

END